Classificação de subtipos de câncer de mama: Um estudo baseado em genes representativos

João Reis[1], **Rayol M. Neto**[1], **Fabíola G. Nakamura**[1], **Eduardo F. Nakamura**[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Manaus – AM – Brasil

{joao.reis,rayol,fabiola,nakamura}@icomp.ufam.edu.br

Abstract. *Breast cancer is the second most common cancer type and is the leading cause of cancer-related deaths worldwide. Since it is a heterogeneous disease, subtyping breast cancer plays an important role in performing a specific treatment. In this work, we propose an approach that uses different machine learning techniques for a broader analysis of the PAM50 list in the classification of breast cancer subtypes. The experiments show that the best method to be used in the classification of breast cancer subtypes is the SVM with linear kernel, which presented an F1 score of 0.97 for the Basal subtype and 0.83 for the Her 2 subtype, the two subtypes with worse prognosis, respectively.*

Resumo. *O câncer de mama é o segundo tipo de câncer mais comum e é a principal causa de mortes relacionadas ao câncer em todo o mundo. Por ser uma doença heterogênea, a subtipagem do câncer de mama desempenha um papel importante na realização de um tratamento específico. Neste trabalho, propomos uma abordagem que utiliza diferentes técnicas de aprendizado de máquina para uma análise mais ampla da lista PAM50 na classificação de subtipos de câncer de mama. Os experimentos mostram que o melhor método a ser utilizado na classificação dos subtipos de câncer de mama é o SVM com kernel linear, que apresentou valor F1 de 0,97 para o subtipo Basal e 0,83 para o subtipo Her 2, os dois subtipos de pior prognóstico, respectivamente.*

1. Introdução

O câncer de mama é o segundo tipo de câncer mais comum e é a principal causa de morte relacionada à doença em todo o mundo [Bray et al. 2018]. Como uma doença altamente heterogênea, o câncer de mama mostra distintas variações genéticas, resultados clínicos e estratégias de tratamento entre seus subtipos [Chen et al. 2016]. O câncer de mama tem quatro subtipos moleculares principais: Basal, Her 2, Luminal A e Luminal B. Basal e Her 2 são os subtipos com os piores prognósticos (apresentam maior taxa de letalidade), respectivamente, enquanto Luminal A e Luminal B estão ligados a um melhor prognóstico, pois existem terapias direcionadas eficazes para eles [Dwivedi et al. 2019].

Métodos de classificação são amplamente empregados na identificaçao dos subtipos de um câncer, pois ajudam a fornecer um diagnóstico eficiente, preciso e objetivo. Diagnosticar um tumor pelo seu subtipo biológico (ou "intrínseco") adiciona informações prognósticas e preditivas significativas para pacientes com câncer de mama [Parker et al. 2009], portanto, classificá-lo em seus subtipos corretamente é crucial para tratar pacientes de forma eficaz.

Recentes avanços na tecnologia de *microarray* de DNA permitiram monitorar os níveis de expressão de milhares de genes simultaneamente durante processos biológicos importantes [Jiang et al. 2004], resultando em dados de expressão gênica. Apresentando uma alternativa viável para empregar na classificação do câncer, esses dados de expressão gênica possuem um desafio quanto a sua análise, pois geralmente há milhares de genes para poucas centenas de amostras.

Apesar da riqueza desses dados, o mapeamento genético do câncer de mama e seus subtipos ainda está longe de ser completo. Atualmente, existe uma lista chamada PAM50 [Chia et al. 2012] que inclui cinquenta genes aceitos como representativos para a caracterização do câncer de mama, e é considerada o conjunto referencial de genes para diferenciar os subtipos. Entretanto, ainda há necessidade de maior investigação desses subtipos, pois não existe uma forma precisa de se diferenciá-los utilizando a lista de genes PAM50. Isso mostra que embora o estudo de expressão gênica já seja realidade, ainda não há uma compreensão definitiva de todos os genes relacionados com o câncer de mama e, principalmente, um entendimento definitivo das interações entre estes genes. Portanto, uma contribuição importante para o diagnóstico preciso é a identificação de um subconjunto de genes capazes de caracterizar os subtipos e diferenciá-los entre si.

Neste contexto, propomos uma abordagem que utiliza diferentes técnicas de aprendizagem de máquina na classificação dos subtipos de câncer de mama. Dadas as particularidades dos dados de expressão gênica, causados principalmente pela sensibilidade das diferentes tecnologias para sua aquisição, do ponto de vista computacional, não se trata de uma simples aplicação de métodos e pacotes de aprendizagem de máquina, há uma clara necessidade de proposição de novas técnicas e adaptação de outras existentes para tratar estes dados de entrada com características e confiabilidades distintas. Com base nisso, as principais contribuições desse trabalho são: (i) Estudo de diferentes métodos na tarefa de classificação de subtipos de câncer de mama; e (ii) análise da lista PAM50 na classificação dos subtipos de câncer de mama.

2. Trabalhos Relacionados

Graudenzi et al. [2017] propuseram um framework de classificação baseado em *Support Vector Machines* (*SVMs*) com uma estratégia de seleção de recursos baseada no conceito de *pathway activity*. Eles identificaram e analisaram uma lista de *enriched pathways* nos quatro subtipos diferentes de câncer de mama e usaram essas informações para realizar o método de seleção de características na implementação do classificador. Em termos de acurácia geral, o classificador proposto apresenta acurácia em torno de 85,00%, usando 400 genes do método de seleção de características.

Lee et al. [2020] usaram uma abordagem baseada em *pathways* para seleção de recursos e aplicou um modelo de aprendizado profundo com mecanismo de atenção e propagação de rede para classificação de câncer. Eles usaram cinco bases de dados de câncer do TCGA[1]. A *precisão* média de classificação de seu método foi de 66,91% para câncer de mama (BRCA). Eles selecionaram um total de 5.515 genes para a tarefa de classificação.

Mostavi et al. [2020] propuseram três redes neurais convolucionais distintas para a tarefa de classificação de câncer. Com relação à predição de subtipos de câncer de mama,

[1]The Cancer Genome Atlas Program

Tabela 1. Resumo dos trabalhos relacionados.

Autor	# de genes	Características	Classes	Classificador	Métricas de Avaliação
[Graudenzi et al. 2017]	400	Basedo em pathways	Multiclasse	SVM	Precisão, revocação e Acurácia
[Mostavi et al. 2020]	7,091	Desvio padrão e média	Multiclasse	1D-CNN	Precisão, revocação e F1
[Lee et al. 2020]	5,015	Baseado em Pathways	Multiclasse	GCN+MAE	Acurácia
[Li et al. 2017]	-	Algoritmo genético	Biclasse	KNN	Acurácia
[Lyu and Haque 2018]	-	Variância	Biclasse	KNN	Precisão, revocação, acurácia e F1

foi utilizado o modelo *1D-CNN*. Os autores usaram métodos de estatísticas fracos para a etapa de seleção de recursos, como desvio padrão e média. Depois de selecionar 7.091 genes, eles usaram seu modelo para a tarefa de classificação e alcançaram uma *precisão* média de 88,42% entre cinco subtipos.

No trabalho de Li et al. [2017], os autores dividiram o processo em duas etapas. Primeiro, um algoritmo genético é usado como mecanismo de seleção de genes e o algoritmo *KNN* (*k-nearest neighbors*) como método de classificação. O conjunto de dados contém 31 tipos de tumor. Para a tarefa de classificação usando *KNN*, k foi definido como 5 com uma regra de votação por maioria. Os resultados mostram que a acurácia da classificação foi superior a 90% para 28 dos 31 tipos de câncer.

Lyu and Haque [2018] incorporaram dados expressão gênica em imagens 2-D e usou uma Rede Neural Convolucional (CNN) para fazer a classificação de 33 tipos distintos de câncer. Os autores transformam a classificação do câncer com base no problema de expressão gênica em problema de imagem. O principal problema é que os dados de expressão gênica são altamente dimensionais, enquanto a maioria das arquiteturas de aprendizado profundo são para imagens 2-D. Como resultado, os autores alcançaram uma média de F1 entre os tipos de câncer de 95,43%.

A Tabela 1 apresenta os trabalhos relacionados. Em resumo, as pesquisas que exploram o problema de classificação de multiclasse (subtipos) encontram uma dificuldade em relação ao desempenho. Ao lidar com a classificação multiclasse, os trabalhos usam centenas de genes e não alcançam resultados tão expressivos, não chegando em 90% de *acurácia*. Como diferentes subtipos podem compartilhar genes importantes para sua identificação, classificar um tipo de câncer entre os subtipos torna uma tarefa muito mais complexa. Neste contexto, iremos trabalhar com a lista PAM50 pois é considerada representativa para os subtipos de câncer de mama e possui apenas cinquenta genes, com o intuito de gerar melhores resultados utilizando menos genes.

3. Abordagem proposta

A abordagem proposta neste trabalho consiste das seguintes etapas (ilustradas na Figura 1). (i) coleta de base de dados que possuam expressão gênica ; (ii) pré-processamento dos dados, a fim de selecionar somente os genes envolvidos no estudo; (iii) classificação

das amostras entre os subtipos de câncer mama ; e (iv) análise do desempenho dos classificadores com diferentes métricas de avaliação.

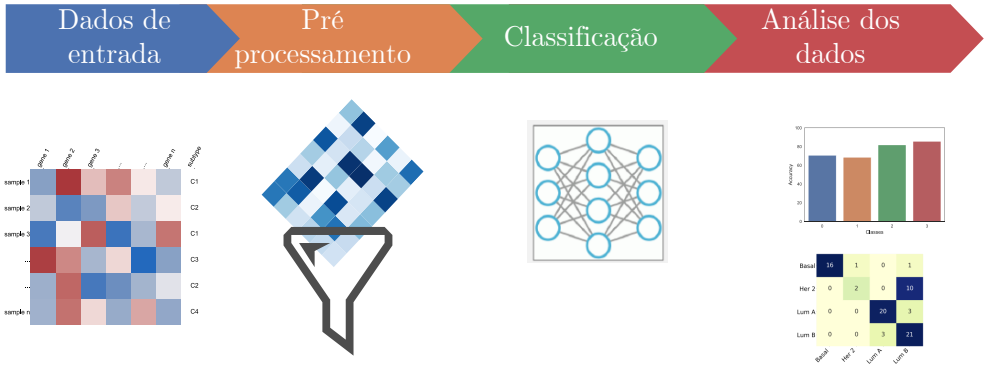


Figura 1. Abordagem proposta.

3.1. Base de dados e Pré-processamento

A abordagem proposta se inicia na fase da escolha da base de dados, onde os dados podem ser extraídos a partir de repositórios de dados genômicos e que possuam dados de expressão gênica. Após a fase de coleta de dados, o pré-processamento é necessário para identificar se a base de dados possui todos os 50 genes da lista PAM50. Com isso, dentre os todos os genes que existem na base de dados escolhida, são selecionados apenas os 50 genes da lista PAM50, caso não haja os 50 genes, outra base de dados necessita ser utilizada para validar o trabalho.

3.2. Classificação

Após selecionar somente os genes do PAM50 para a base de treino e teste, fazemos a classificação com diferentes classificadores. O objetivo dessa etapa é entender como diferentes métodos de classificação são capazes de distinguir os subtipos de câncer de mama utilizando dados de expressão gênica.

3.3. Análise de Dados

Para medir o desempenho dos métodos, aplicamos métricas tradicionais como *precisão*, *revocação*, *Medida F* e *acurácia*. Uma vez que os dados biológicos geralmente têm um conjunto de dados esparso [Chicco 2017], também medimos o desempenho dos métodos usando o coeficiente de correlação de Matthews (*MCC*) [Chicco 2017] e curva de *precisão x revocação* (*AUPRC*). Ambas as métricas foram selecionadas porque são adequadas para base de dados desequilibradas. Enquanto *MCC* é mais apropriada para classificação binária, a curva de *precisão x revocação* é um indicador mais confiável e informativo do desempenho estatístico em problemas multiclasse [Chicco 2017].

Também calculamos a *especificidade*, pois esta medida é usada para verificar o quão corretamente podemos classificar um indivíduo no subtipo de câncer correto. As métricas são definidas como:

$$Precisão = \frac{TP}{TP + FP}, \quad (1) \qquad Revocação = \frac{TP}{TP + FN}, \quad (2)$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}, \quad (3) \qquad Acurácia = \frac{TP + TN}{TP + FP + FN + TN}, \quad (4)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}}, \quad (5)$$

$$Especificidade = \frac{TN}{TN + FP}, \quad (6)$$

em que *TP* são verdadeiros positivos, *TN* são verdadeiros negativos, *FP* falsos positivos e *FN* são falsos negativos. Quanto maior o valor dessas métricas, melhor é o resultado.

4. Avaliação da abordagem proposta

Essa seção apresenta uma análise dos diferentes classificadores quando empregados na classificação de subtipos de câncer de mama utilizando o conjunto de genes PAM50. Adicionalmente, detalhamos a metodologia utilizada na aplicação da abordagem proposta.

4.1. Metodologia

Nesta subseção descrevemos a metodologia de avaliação utilizada neste trabalho. Detalhamos as características das base de dados utilizadas nos experimentos. Apresentamos os parâmetros escolhidos para os métodos de aprendizado de máquina e também apresentamos as métricas de avaliação empregadas.

Base dados

Para avaliar os métodos, usamos duas bases de dados do *Clinical Proteomic Tumor Analysis Consortium* (CPTAC). A base de dados Cptac 2C é usada para treinar os modelos, pois possui um maior número de amostras, enquanto a Cptac 2D é usada para os testes. As bases de dados apresentam a expressão gênica de pacientes com câncer de mama. Essas amostras são divididas em quatro subtipos intrínsecos de câncer de mama (consulte a Seção 1).

A Tabela 2 resume as características das bases de dados usadas nos experimentos:

Tabela 2. Descrição das bases de dados.

Dataset	# de genes	Subtipos	# de amostras	# total de amostras
Cptac 2C	23122	Basal	29	117
		Her 2	14	
		Luminal A	57	
		Luminal B	17	
Cptac 2D	16525	Basal	18	77
		Her 2	12	
		Luminal A	23	
		Luminal B	24	

Classificadores

Para avaliar o desempenho da lista PAM50 para a classificação de subtipos de câncer de mama, empregamos 5 métodos distintos. O *Grid search* [Bergstra and Bengio 2012]

foi usado para otimizar os parâmetros para cada classificador. A tabela 3 apresenta os classificadores e os parâmetros escolhidos. Os parâmetros restantes foram definidos para o padrão do scikit-learn[2].

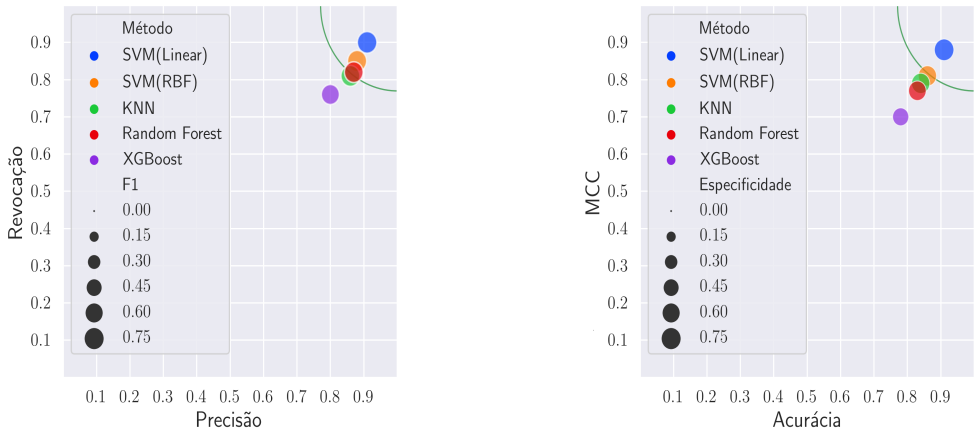
Tabela 3. Parâmetros dos classificadores.

Método	Parâmetros
SVM(Linear)	$C = 0.1$
SVM(RBF)	$C = 1.1$
KNN	$p = 1, n\ neighbors = 5, weights = uniform$
Random Forest	$bootstrap = False, min\ samples\ split = 6, n\ estimators = 28$
XGBoost	$gamma = 0.04\ , learning\ rate = 0.07$

Para os cáculos das métricas de avaliação utilizamos as bibliotecas scikit-learn e pandas-ml[3]. Comparamos os diferentes classificadores para verificar qual o método com melhor desempenho geral e para cada subtipo separadamente. Utilizamos as métricas de avaliação apresentadas na Seção 3.

4.2. Resultados

Os diferentes classificadores levam em consideração diferentes maneiras de classificar as amostras às diferentes classes. Alguns utilizam espaçamentos entre as classes para diferenciá-las (*SVM*), enquanto outros verificam qual classe predominante entre os elementos mais próximos da amostra analisada (*KNN*). Também tem os que utilizam árvores de decisão (*Random Forest*) para realizar a classificação e outros partem de uma hipótese básica e tentam melhorá-la para chegar a um melhor resultado (*XGboost*). Por isso, espera-se também que haja diferentes resultados para os classificados testados, mesmo que sejam submetidos às mesmas condições de teste.



(a) *Precisão, revocação e F1.* **(b)** *Acurácia, MCC e especificidade.*

Figura 2. Desempenho obtido pelos métodos utilizando macrométricas.

[2]https://scikit-learn.org/stable/
[3]https://pypi.org/project/pandas-ml/

No primeiro experimento, comparamos os resultados obtidos por todos os métodos na tarefa de classificação dos quatro subtipos. A Figura 2a ilustra o desempenho em termos de *precisão*, *revocação* e *F1* e a Figura 2b ilustra o desempenho em termos de *acurácia*, *MCC* e *especificidade* . Comparando os desempenhos obtidos pelos métodos, vemos que o *SVM(Linear)* superou todos os outros métodos nas seis macrométricas analisadas. Este desempenho pode ser explicado pelo fato de ser o classificador que melhor conseguiu separar as amostras do Luminal A com Luminal B.

Analisando a Tabela 4, contendo os resultados obtidos pelos cinco classificadores testados, podemos identificar que o subtipo Basal obteve *precisão* de 100% em 3 dos 5 classificadores. O subtipo Luminal A obteve a maior pontuação de *precisão* de 92% com o *SVM(Linear)*, além de um índice de *revocação* de 96% em 3 dos 5 métodos de classificação utilizados. Observando o subtipo Her 2, pode-se notar que este consegue obter uma pontuação de 100% de *precisão* com o classificador *Random Forest*, porém com um índice de *revocação* acima dos 80% apenas com o classificador *SVM(Linear)*. O subtipo Luminal B obteve uma precisão máxima de 93% quando utilizado o classificador *KNN* e *revocação* máxima de 88% utilizando-se o *SVM(Linear)*.

Tabela 4. Resultado micrométricas - *Precisão*, *Revocação* e *F1*.

Método	*Precisão*				*Revocação*				*F1*			
	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB
SVM(Linear)	**1,00**	0,83	**0,92**	0,88	0,94	**0,83**	**0,96**	**0,88**	**0,97**	**0,83**	**0,94**	**0,88**
SVM(RBF)	**1,00**	0,90	0,76	0,86	0,94	0,75	**0,96**	0,75	**0,97**	0,82	0,85	0,80
KNN	0,90	0,90	0,76	**0,89**	**1,00**	0,75	**0,96**	0,54	0,95	0,82	0,85	0,76
Random Forest	0,95	**1,00**	0,72	0,81	**1,00**	0,67	0,91	0,71	**0,97**	0,80	0,81	0,76
XGBoost	**1,00**	0,70	0,68	0,80	0,89	0,58	0,91	0,67	0,94	0,64	0,78	0,73

Analisando a pontuação *F1*, o classificador *SVM(Linear)* detém as maiores pontuações. Para o subtipo Basal, mais dois classificadores, além do *SVM(Linear)* obtiveram pontuação de 97%, o subtipo Her 2 obteve uma pontuação de 83%, enquanto o subtipo Luminal A obteve 94% e o subtipo Luminal B obteve 88%.

Tabela 5. Resultado micrométricas - *MCC*, *AUPRC* e *Especificidade*.

Método	*MCC*				*AUPRC*				*Especificidade*			
	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB
SVM(Linear)	0,96	**0,80**	**0,91**	**0,82**	**0,96**	**0,72**	**0,89**	**0,80**	**1,00**	0,97	**0,96**	0,94
SVM(RBF)	0,96	0,79	0,78	0,72	**0,96**	0,71	**0,74**	0,72	**1,00**	0,98	0,87	0,94
KNN	0,93	0,79	0,78	0,69	0,90	0,71	**0,74**	0,70	0,97	0,98	0,87	**0,96**
Random Forest	**0,97**	0,79	0,72	0,66	0,95	**0,72**	0,69	0,66	0,98	**1,00**	0,85	0,92
XGBoost	0,93	0,58	0,68	0,62	0,91	0,47	0,64	0,64	**1,00**	0,95	0,81	0,92

A Tabela 5 contém os dados das métricas *MCC*, *AUPRC* e *Especificidade*. Examinando os dados obtidos com a métrica *MCC*, notamos que o classificador *SVM(Linear)* possui as maiores pontuações para os subtipos Her 2 com 80%, Luminal A com 91% e Luminal B com 82%, enquanto o subtipo Basal obteve pontuação máxima de 97% com o classificador *Random Forest*.

Na métrica *AUPRC*, o classificador *SVM(Linear)* obteve pontuações de 96% para o subtipo Basal, 72% para o subtipo Her 2, 89% para o subtipo Luminal A e 80% para

o subtipo Luminal B. Analisando a *Especificidade*, notamos que o subtipo Basal obteve 100% com os classificadores *SVM(Linear)*, *SVM(RBF)* e *XGBoost*. O subtipo Her 2 obteve máxima de 100% ao utilizar-se o classificador *Random Forest*, o subtipo Luminal A obteve 96% com o classificador *SVM(Linear)* e o subtipo Luminal B também obteve 96% de *Especificidade*, porém utilizando o classificador *KNN*.

Tabela 6. Resultado macrométricas. *F1*, *Acurácia*, *MCC*, *AUPRC* e *Especificidade*.

Método	*F1 Macro*	*ACC*	*MCC*	*AVG AUPRC*	*AVG Especificidade*
SVM(Linear)	**0,90**	**0,91**	**0,88**	**0,84**	**0,97**
SVM(RBF)	0,86	0,86	0,81	0,78	0,95
KNN	0,84	0,84	0,79	0,76	0,95
Random Forest	0,83	0,83	0,77	0,75	0,94
XGBoost	0,77	0,78	0,70	0,67	0,92

Analisando os dados da Tabela 6, com os resultados das macrométricas, conclui-se então que o *SVM(Linear)* é o melhor classificador dentre os testados, tendo a melhor pontuação em todas métricas utilizadas para avaliação, com uma pontuação de 90% para a pontuação *F1*, 91% de *Acurácia*, 88% para o *MCC*, 84% para *AUPRC* e 97% de *especificidade*. Em linhas gerais, percebemos que o Basal, subtipo que possui o pior prognóstico é o mais característico entre todos, uma vez que os classificadores obtiveram melhores resultados nesse subtipo, com relação ao Her 2 (subtipo com segundo pior prognóstico), notamos que obteve o pior resultado entre os subtipos, sendo assim o mais difícil de ser classificado. Por fim, os resultados mostraram que a estratégia proposta combinada com a lista PAM50 foi capaz de obter bons resultados e se mostra promissora para classificação.

5. Conclusão

Este trabalho apresentou uma abordagem para classificação de subtipos de câncer de mama utilizando como base a lista de genes do PAM50. Utilizamos diferentes métodos de classificação, cada um com características distintas para analisar se há diferença entre eles na tarefa de classificação. Diversas métricas de avaliação foram empregadas para obter um panorama de como os métodos classificavam as amostras.

Como resultados, percebemos que o *SVM(Linear)* obteve resultados macro melhores que os demais. Verificamos também que o subtipo Basal (o de pior prognóstico e mais característico), os classificadores *SVM(RBF)* e *Random forest* alcançaram o mesmo valor *F1*, que foi de 97%. Além de que os demais cassificadores mantiveram-se com uma pontuação acima de 90% para este subtipo.

Percebe-se que o Her 2, o subtipo com o segundo pior prognósitco, possui os piores resultados na classificação, onde alcança pontuação *F1* de no máximo 83%, chegando a alcançar uma mínima de 64% com o classificador *XGBoost* onde as amostras se confundiam em geral com amostras Luminal B. Entre os subtipos Luminal A e Luminal B pode-se notar confusão entre as amostras, tendo em vista o fato de serem altamente correlacionados. Embora o PAM50 possua apenas 50 genes, este é um bom conjunto para classificação uma vez que obteve em quatro dos cinco classificadores uma pontuação *F1*

Macro acima dos 80%. A nível micro, o *SVM(Linear)* conseguiu manter também uma pontuação *F1* acima dos 80% para todos os subtipos.

Como trabalhos futuros, pretendemos estender a abordagem proposta para classificação multinível, onde iremos isolar os subtipos e investigar se existem classificadores que apresentam melhores desempenhos para o subtipo analisado, utilizando os genes da lista PAM50.

Referências

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305.

Bray, F., Ferlay, J., Soerjomataram, I., L. Siegel, R., Torre, L., and Jemal, A. (2018). Global cancer statistics 2018. *CA: A Cancer Journal for Clinicians*, 68:394–424.

Chen, X., Hu, H., He, L., Yu, X., Liu, X., Zhong, R., and Shu, M. (2016). A novel subtype classification and risk of breast cancer by histone modification profiling. *Breast cancer research and treatment*, 157(2):267–279.

Chia, S. K., Bramwell, V. H., Tu, D., et al. (2012). A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clinical cancer research*, 18(16):4465–4472.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1):1–17.

Dwivedi, S., Purohit, P., Misra, R., Lingeswaran, M., et al. (2019). Application of single-cell omics in breast cancer. In *Single-Cell Omics*, volume 2, pages 69–103.

Graudenzi, A., Cava, C., Bertoli, G., Fromm, B., et al. (2017). Pathway-based classification of breast cancer subtypes. *Front Biosci*, 22(10):1697–1712.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge & Data Engineering*, (11):1370–1386.

Lee, S., Lim, S., Lee, T., Sung, I., and Kim, S. (2020). Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics*, 36(12):3818–3824.

Li, Y., Kang, K., Krahn, J. M., Croutwater, N., et al. (2017). A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC genomics*, 18(1):508.

Lyu, B. and Haque, A. (2018). Deep learning based tumor type classification using gene expression data. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 89–96. ACM.

Mostavi, M., Chiu, Y.-C., et al. (2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*, 13(44):1–13.

Parker, J. S., Mullins, M., Cheang, M. C., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.