Redução dimensional de dados de expressão gênica para classificação de subtipos de câncer de mama

Arthur Bindá Alves[1], Rayol Mendonca-Neto[1], Eduardo Nakamura[1], Fabíola Nakamura[1], David Fenyö[2]

[1]Instituto de Computação – Universidade Federal Amazonas (UFAM)

[2]Institute for Systems Genetics - NYU School of Medicine (NYU)

{aba,rayol,fabiola,nakamura}@icomp.ufam.edu.br, david@fenyolab.org

Abstract. *Breast cancer is the most common cancer in women and the leading cause of death. Breast cancer subtyping is critical for effective treatment, but the high dimensionality of the data is a challenge to obtain the diagnosis. To mitigate this problem, dimensionality reduction techniques can be applied to extract more relevant attributes and reduce dimensionality. This work compares deep learning frameworks and proposes the use of the Siamese networks model as a new approach to improve classification in breast cancer subtypes. The results show that the new approach improved by 0.05 the F1 of the most difficult subtypes.*

Resumo. *O câncer de mama é mais comum em mulheres e o que mais causa mortes. A subtipagem do câncer de mama é fundamental para um tratamento eficaz, mas a alta dimensionalidade dos dados é um desafio para obtenção do diagnóstico. Para mitigar este problema, técnicas de redução dimensional podem ser aplicadas para extrair atributos mais relevantes e reduzir a dimensionalidade. Este trabalho compara estruturas de aprendizagem profunda e propõe o uso do modelo de redes siamesas como uma nova abordagem para aprimorar a classificação em subtipos de câncer de mama. Os resultados mostram que a nova abordagem melhorou em 0,05 o F1 dos subtipos mais difíceis.*

1. Introdução

O câncer de mama é uma doença heterogênea com subtipos que apresentam características biológicas distintas. Este é o câncer mais comum e com o maior número de vítimas por ano entre mulheres [Bray et al. 2018]. Por se tratar de uma doença complexa, cada subtipo possui um prognóstico distinto que leva a diferentes resultados clínicos [Yersal and Barutca 2014]. Os quatro subtipos moleculares do câncer de mama são: Basal, Her2, Luminal A e Luminal B. A identificação do subtipo durante o diagnóstico é de extrema importância para o sucesso do tratamento, visto que esses subgrupos tumorais apresentam diferenças tanto em ritmo de crescimento, como também em vias de sinalização, composição celular e sensibilidade terapêutica [Yersal and Barutca 2014].

O diagnóstico precoce e preciso aumenta em pelo menos 25% a chance de recuperação da paciente com câncer de mama [Rivera-Franco and Leon-Rodriguez 2018]. A classificação de amostras de câncer por meio de algoritmos de aprendizagem de máquina é uma das técnicas mais utilizadas para obter esse diagnóstico [Ang et al. 2015].

Entretanto, bases de dados de expressão gênica são afetadas pela "maldição da dimensionalidade", apresentando um grande número de genes em relação ao baixo número de amostras, o que pode comprometer a generalização do classificador e levar ao *overfit*. Essa alta dimensionalidade também aumenta a complexidade da tarefa de classificação e o custo computacional, podendo inviabilizar a abordagem.

Para viabilizar a classificação de amostras de expressão gênica, é possível aplicar técnicas de transformação dimensional, diminuindo o número de atributos através de representações reduzidas. Essas técnicas abrangem desde métodos estatísticos à abordagens mais complexas. Embora o uso de técnicas lineares alcance resultados promissores, ainda existem desafios e limitações a serem superados. Essas limitações ocorrem principalmente devido à suposição de que a relação entre os genes é linear. Nesta direção, modelos de aprendizagem profunda tem ganhado destaque por aplicarem transformações dimensionais mais complexas. Este trabalho busca descobrir qual o melhor método de redução dimensional para a classificação de subtipos de câncer de mama utilizando algoritmos de aprendizagem profunda.

A abordagem proposta neste estudo busca criar representações em dimensão reduzida mais relevantes, utilizando camadas sequenciais de compressão para extrair características complexas e não-lineares. A inovação deste trabalho está na aplicação da rede siamesa para a redução de dados gênicos, uma abordagem até então não explorada. Os resultados obtidos apontam que o modelo de redes siamesas foi capaz de melhorar os resultados dos subtipos mais difíceis enquanto reduziu em 47 vezes a quantidade de atributos utilizados.

As principais contribuições deste trabalho são: (i) estudo de diferentes métodos profundos para redução dimensional de dados de expressão gênica, (ii) apresentação de uma nova abordagem capaz de criar representações mais significativas. O restante deste trabalho está organizado da seguinte maneira: na Seção 2 apresentamos os trabalhos relacionados. Na Seção 3 apresentamos a abordagem proposta desta pesquisa. A Seção 4 descreve a metodologia utilizada e explica as técnicas de transformação dimensional utilizadas. Na Seção 5 apresentamos e discutimos os resultados dos experimentos. Por fim, na Seção 6 expomos nossas conclusões e apontamos direções para pesquisas futuras.

2. Trabalhos Relacionados

Existem diversos estudos aplicando técnicas de redução dimensional em dados de expressão gênica para melhorar a classificação do câncer de mama. O PCA (*Principal Component Analysis*) é uma das abordagens lineares mais utilizadas. Sahu et al. [2019] aplicaram o PCA para comprimir os genes e utilizaram uma rede neural para classificar o câncer de mama em maligno ou benigno. Os autores alcançaram 0,95 de precisão e 97% de acurácia com esta metodologia e atestaram a capacidade do PCA em reduzir o número de atributos sem perder características importantes.

Adem [2020] implementou um SAE (*Stacked Autoencoder*) e comprimiu os genes de um estudo binário de câncer de mama em oito vezes. Utilizando um classificador especializado em agrupar amostras em subespaços, o autor apresentou ganhos de 23% na acurácia quando comparado com nenhuma abordagem de seleção de atributos e 10% em relação ao PCA. O melhor resultado apresentado foi de 91% de acurácia e 0,89 de precisão com o classificador *subspace* KNN.

Danaee et al. [2017] aplicaram um SDAE (*Stacked Denoising Autoencoder*) na base de dados TCGA para melhorar a classificação biclasse do câncer de mama. Assim como em outros trabalhos, o desempenho do modelo foi comparado com o de um PCA e novamente o *autoencoder* obteve o melhor resultado. Os autores evidenciaram um ganho de quase 10% de acurácia e de 0,08 no F1 em comparação ao desempenho do PCA. O melhor resultado apresentado foi utilizando o classificador SVM-RBF, alcançando 98% de acurácia e 0,98 de F1.

No estudo conduzido por Xiao et al. [2018], um SSAE (*Stacked Sparse Autoencoder*) foi utilizado para classificar amostras de três tipos diferentes de câncer: Pulmão, estômago e mama, oriundos da base de dados TCGA. Nessa abordagem, os autores utilizaram um coeficiente de esparsidade como uma técnica para criar representações mais relevantes dos genes. O objetivo do estudo era melhorar a classificação biclasse dos diferentes tipos de câncer. O resultado mais promissor foi alcançado com uma acurácia de 96% e um valor F1 de 0,98, utilizando o classificador SVM.

Entre os trabalhos relacionados citados nesta seção, destacamos Danaee et al. [2017] que obtiveram os melhores resultados com as métricas acurácia e F1. Por este motivo, o trabalho serviu de base para esta pesquisa. Optamos por aplicar o PCA como abordagem linear para fins de comparação com os trabalhos relacionados. É importante ressaltar que os trabalhos expostos nesta seção aplicam seus métodos para melhorar a classificação binária entre câncer e não-câncer, tarefa mais simples que a classificação em subtipos do câncer de mama e, por este motivo, os resultados obtidos são próximos de 100%.

3. Abordagem Proposta

A abordagem proposta deste estudo consiste nas seguintes etapas: (i) coleta da base de dados de expressão gênica; (ii) aplicação de técnicas de redução dimensional; (iii) divisão dos dados reduzidos em treino e teste; (iv) classificação dos dados reduzidos; (v) análise dos resultados da classificação com diferentes métricas. A Figura 1 apresenta as etapas da abordagem proposta neste trabalho.

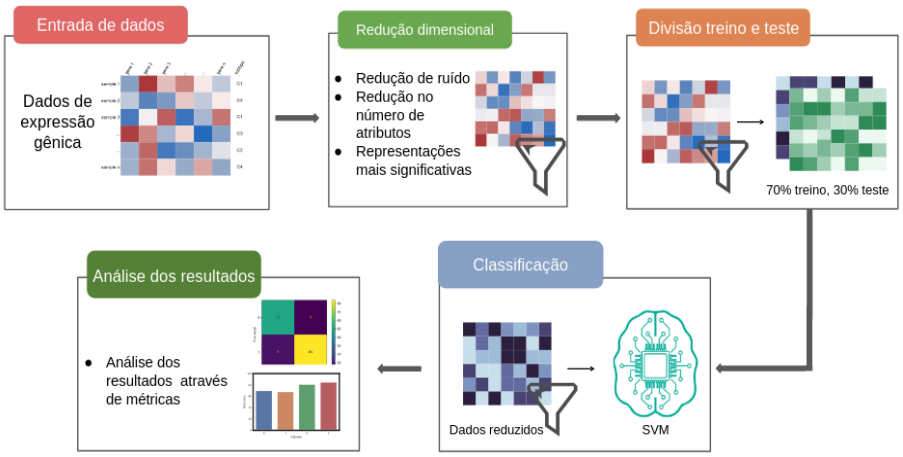

Figura 1. Representação das etapas da abordagem proposta.

O processo inicia selecionando uma base de dados contendo informações de expressão gênica onde cada coluna corresponde a um gene e cada linha representa uma

amostra. A dimensão original da base de dados é utilizada para definir o espaço de busca dos hiperparâmetros. Em seguida, técnicas de redução dimensional são aplicadas para reduzir a quantidade de atributos que serão inseridos no classificador. O objetivo desta etapa é extrair dados mais significativos que melhorem a diferenciação dos subtipos enquanto mitiga o problema da dimensionalidade presente em dados gênicos.

Com menos atributos, a complexidade computacional do modelo é reduzida, permitindo que a classificação seja realizada com maior confiabilidade e precisão. Após a redução dimensional, os dados são divididos randomicamente em 70% para treino do classificador e 30% para testagem. Um classificador é então utilizado para identificar nas amostras os subtipos de câncer de mama. A última etapa é composta pela análise dos resultados. Nesta, validamos as técnicas de redução dimensional através do seu desempenho no classificador, avaliando métricas gerais e específicas para cada subtipo.

4. Metodologia

Nesta Seção, descrevemos a metodologia deste trabalho, incluindo informações sobre a base de dados, métodos de redução dimensional, ferramentas utilizadas para a implementação desses métodos, etapa de classificação e análise dos resultados.

4.1. Base de Dados

Partindo do referencial de Danaee et al. [2017] e buscando aumentar a confiabilidade dos resultados, unimos os estudos TCGA e ACES [Staiger et al. 2013]. Ao combinarmos as bases, a quantidade de genes comuns entre elas foi reduzida para 9403. Ao total, utilizamos 2221 amostras gênicas de câncer de mama rotuladas em seus quatro subtipos. A Tabela 1 apresenta a quantidade total de amostras, o número de genes e amostras por subtipo. Encontramos efeitos de *batch* entre essas bases de dados, uma vez que as amostras são de estudos diferentes. Especificamente na base TCGA percebemos uma normalização diferente da ACES. Para mitigar esse problema, aplicamos uma das ferramentas mais utilizadas para corrigir efeitos de *batch* em dados de expressão gênica *pycombat*[1].

Tabela 1. Descrição do dataset utilizado.

Dataset	# de genes	Subtipos	# de amostras	# total de amostras
TCGA	17814	Basal Her 2 Luminal A Luminal B	143 84 312 170	709
ACES	12750	Basal Her 2 Luminal A Luminal B	297 191 584 440	1512
TCGA + ACES	**9403**			**2221**

4.2. Métodos de Redução Dimensional

A utilização de técnicas de redução dimensional é essencial para adaptar bases de dados de alta dimensionalidade a um número adequado de atributos [Mendonca-Neto et al. 2022]. Especificamente quando trabalhamos com dados de expressão gênica lidamos também

[1]https://github.com/epigenelabs/pycombat

com a complexidade das relações biológicas e grandes quantidades de ruído nos dados [Danaee et al. 2017]. Nesta pesquisa abordamos técnicas de redução dimensional através da transformação dos genes para representações em dimensão reduzida. Utilizando modelos de aprendizagem profunda podemos expandir a capacidade de extrair relações mais complexas entre os genes e mais significativas para o classificador.

Autoencoders são modelos de aprendizagem profunda treinados para reconstruir sua própria entrada a partir de representações das características mais importantes dos dados. No processo de treinamento, a entrada é comprimida pelo codificador para uma representação reduzida em um espaço de atributos. A partir da camada de representação reduzida o decodificador reconstrói o dado para sua dimensão original.

A partir do conceito do *autoencoder* é possível derivar modelos mais complexos. O SDAE é treinado adicionando ruído estatístico nos dados de entrada para forçar o modelo a aprender representações mais robustas. Isso acontece porque o ruído adicionado torna a tarefa de reconstrução mais difícil, aumentando a necessidade do modelo em aprender a ignorar o ruído e gerar representações mais significativas [Danaee et al. 2017].

O SSAE é uma abordagem que incorpora um coeficiente de esparsidade na camada de representação reduzida. A esparsidade força os neurônios da camada latente ficarem com a maioria dos valores próximos de zero, enquanto apenas alguns são significativos. Isso permite que o modelo se concentre nos genes relevantes, reduzindo o risco de *overfitting* [Xiao et al. 2018]. Este modelo busca criar representações mais compactas enquanto poupa recursos computacionais por utilizar menos neurônios que os outros modelos citados.

Com estrutura de compressão semelhante ao do *autoencoder*, as redes siamesas se destacam por utilizar a classe dos dados como parâmetro no seu treinamento. Esta técnica ainda não foi aplicada em dados de expressão gênica. A ideia por trás do modelo é aprender a representar dados em dimensão reduzida visando minimizar a distância entre exemplos da mesma classe e maximizar a distância entre exemplos de classes diferentes [Taigman et al. 2014]. Esta abordagem gera um modelo que aprende a agrupar os subtipos de câncer de mama com base nas suas semelhanças.

4.3. Implementação dos Métodos de Redução Dimensional

Nós utilizamos a biblioteca Keras para implementar os modelos de aprendizagem profunda escolhidos para esta pesquisa. A vantagem de utilizarmos estes algoritmos é a adaptação da estrutura aos dados inseridos. Porém, a escolha errada dos hiperparâmetros pode comprometer os resultados. Para otimizarmos os modelos fizemos um processo de *tuning* que retorna o melhor conjunto de hiperparâmetros.

Nesta pesquisa ajustamos sete hiperparâmetros na estrutura dos modelos. Na estrutura de *autoencoder* configuramos a quantidade de camadas de compressão, número de neurônios por camada, função de ativação das camadas e da última camada, tamanho da camada latente, taxa de aprendizagem e esparsidade. Durante o treinamento foi utilizado *K-fold* e os dados utilizados para teste foram os mesmos para todas as abordagens.

4.4. Classificação

Para avaliar o desempenho dos métodos de redução dimensional, comparamos com o SVM, um classificador amplamente utilizado em pesquisas de câncer de mama [Daoud

and Mayo 2019]. Para obter os melhores resultados utilizamos *tuning* nos seguintes hiperparâmetros do classificador: *kernel*, C e *gamma*. Ao final desta etapa de ajuste, obtemos o algoritmo mais adequado para a classificação dos dados gênicos reduzidos. Portanto, o SVM servirá para avaliar a eficácia dos métodos de redução dimensional enquanto está otimizado para obter os melhores resultados.

4.5. Análise dos Resultados

Para validar o desempenho da nossa abordagem escolhemos as métricas precisão, revocação, F1 e acurácia. A precisão é uma métrica que indica quantas classificações positivas foram acertadas. Já a revocação avalia quantas amostras positivas existentes foram classificadas corretamente. Buscamos modelos com boa precisão, mas a revocação é mais importante. Isso se deve ao fato de que resultados falsos negativos são mais prejudiciais que os falsos positivos. O F1 é calculado a partir da média harmônica entre a precisão e a revocação. Essencialmente, o F1 permite observarmos, em uma única métrica, as variações da precisão e revocação. Essa é a principal medida de comparação utilizada neste trabalho. Usaremos também a acurácia que é calculada através da divisão do número de acertos sobre o total de amostras avaliadas.

5. Resultados

Os resultados foram coletados da seguinte maneira: (i) cada abordagem de redução dimensional foi executada múltiplas vezes para encontrar a melhor combinação de hiperparâmetros,(ii) cada modelo otimizado gerou uma representação dos conjuntos de treino e teste, (iii) o classificador foi treinado e otimizado com a base de treinamento e os resultados da execução com a base de teste estão expostos nesta Seção.

A Tabela 2 expõe as métricas F1 macro, acurácia e F1 por subtipo. A Figura 2 mostra a matriz de confusão das técnicas avaliadas. A matriz de confusão apresenta o desempenho do classificador comparando as previsões corretas e incorretas para cada classe. Por questão de espaço e desempenho equivalente ao PCA, a matriz de confusão do SSAE não foi inserida.

Com o objetivo de comparar métodos mais complexos com uma abordagem linear, implementamos a redução dimensional utilizando o PCA. Observando a Tabela 2, notamos que o PCA não deve ser descartado para este problema, pois alcançou resultados próximos dos *autoencoders*, sendo mais simples e computacionalmente mais barato. Entretanto, a Figura 2 aponta a dificuldade dessa abordagem em melhorar a distinção das classes Luminal A e Luminal B.

Tabela 2. Desempenho da melhor configuração de hiperparâmetros por técnica

Desempenho melhor configuração por técnica.							
Técnica de redução dimensional	# final de atributos	F1 Macro	ACC	F1			
				Basal	Her2	Lum A	Lum B
PCA	700	0,87	0,88	0,96	0,83	0,91	0,80
SAE	100	0,88	0,89	**0,97**	**0,88**	0,90	0,80
SDAE	100	0,89	0,89	0,95	0,86	**0,92**	0,84
SSAE	100	0,87	0,88	0,94	0,85	0,91	0,81
Rede Siamesa	200	**0,90**	**0,90**	0,96	0,87	**0,92**	**0,85**

Entre os modelos de *autoencoder* é possível notar como o número de atributos reduzidos foram iguais, mas seus resultados divergiram. O SAE obteve o pior resultado

para a classe Luminal B, mas foi capaz de gerar a melhor representação para as classes Basal e Her2. O SDAE alcançou os melhores valores para as classes Luminal A e Luminal B entre os *autoencoders*, um indicativo de que o ruído presente na base pode estar penalizando estes subtipos nos outros modelos. Com resultados próximos do PCA, o SSAE criou representações muito complexas dos dados e não se destacou em relação às outras abordagens. A rede siamesa obteve o melhor valor nas métricas F1 macro e acurácia. Esta abordagem melhorou em 0,05 o F1 das classes Her2 e Luminal B em relação à abordagem linear.

Comparando as matrizes de confusão do PCA com a rede siamesa, fica evidente a qualidade da representação reduzida do modelo não-linear em relação às classes mais difíceis. Apesar de obter resultados próximos dos outros modelos profundos, a rede siamesa se destaca por criar representações que aprimoram a distinção entre os subtipos Luminal dos restantes. Este resultado é relevante do ponto de vista clínico em função das diferenças no tratamento destes subtipos. Portanto, os mecanismos da arquitetura de rede siamesa podem ser o próximo passo para melhorar o desempenho dos modelos profundos na tarefa de redução dimensional.

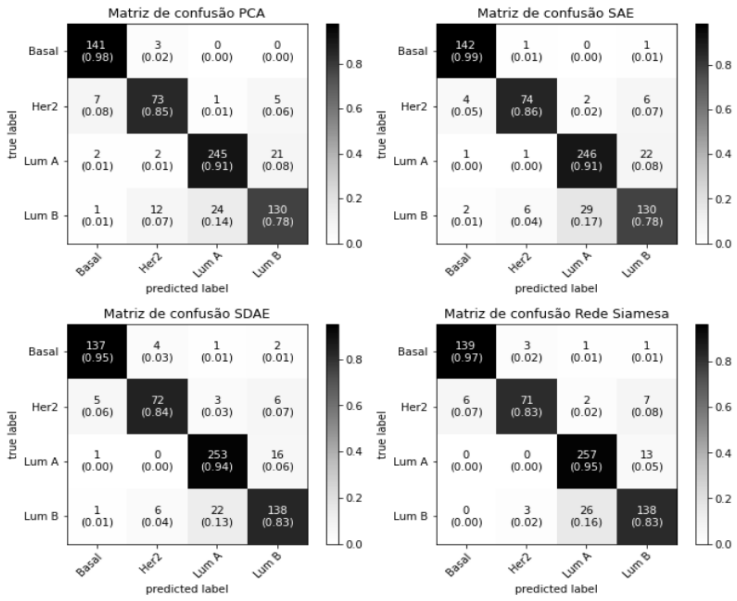

Figura 2. Matriz de confusão por técnica de redução dimensional.

6. Conclusão

Nesta pesquisa comparamos técnicas de redução dimensional em amostras de expressão gênica para melhorar a classificação de câncer de mama em subtipos. O método proposto considerou três estruturas de *autoencoder* apontadas pela literatura como alternativa superior para métodos lineares como PCA e aplicou uma nova abordagem para a solução deste problema. A alta adaptabilidade dos modelos de aprendizagem profunda possibilitou o ajuste preciso das técnicas aos dados, gerando representações mais significativas.

A nova abordagem proposta neste estudo, a rede siamesa, obteve os melhores resultados entre as técnicas testadas. Sua arquitetura foca em encontrar diferenças entre as classes, sendo uma vantagem sobre as outras abordagens profundas aplicadas

para redução dimensional. A redução com a rede siamesa resultou em melhorias na classificação do câncer de mama em todos os subtipos, sendo mais beneficiadas as classes Her2 e Luminal B, ambas de difícil separação.

Os resultados obtidos neste trabalho descrevem a complexidade desta tarefa. Enquanto o subtipo Basal possui características bem claras, os subtipos Luminal A e B são parecidos entre si e a classe minoritária Her2 muitas vezes é confundida com as classes Luminal. Este fato reforça os avanços obtidos nesta pesquisa. Como próximos passos pretendemos expandir o contexto da pesquisa para outros tipos de câncer, integrar técnicas de Auto-ML na etapa de ajuste dos modelos e avaliar o desempenho dos modelos em problemas binários, pois a redução no número de classes pode inviabilizar estas abordagens.

7. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas por meio do projeto POSGRAD 22-23.

Referências

Adem, K. (2020). Diagnosis of breast cancer with stacked autoencoder and subspace knn. *Physica A: Statistical Mechanics and its Applications*, 551:124591.

Ang, J. C., Mirzal, A., Haron, H., and Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5):971–989.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*.

Danaee, P., Ghaeini, R., and Hendrix, D. A. (2017). A deep learning approach for cancer detection and relevant gene identification. In *Pacific symposium on biocomputing*.

Daoud, M. and Mayo, M. (2019). A survey of neural network-based cancer prediction models from microarray data. *Artificial intelligence in medicine*, 97:204–214.

Mendonca-Neto, R., Reis, J., Okimoto, L., Fenyö, D., Silva, C., Nakamura, F., and Nakamura, E. (2022). Classification of breast cancer subtypes: A study based on representative genes. *Journal of the Brazilian Computer Society*, 28(1):59–68.

Rivera-Franco, M. M. and Leon-Rodriguez, E. (2018). Delays in breast cancer detection and treatment in developing countries. *Breast cancer: basic and clinical research*.

Sahu, B., Mohanty, S., and Rout, S. (2019). A hybrid approach for breast cancer classification and diagnosis. *EAI Endorsed Transactions on Scalable Information Systems*.

Staiger, C., Cadot, S., Györffy, B., Wessels, L. F., and Klau, G. W. (2013). Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Frontiers in genetics*, 4:289.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiao, Y., Wu, J., Lin, Z., and Zhao, X. (2018). A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using rna-seq data. *Computer methods and programs in biomedicine*, 166:99–105.

Yersal, O. and Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World journal of clinical oncology*, 5(3):412.