Universidade de São Paulo
Instituto de Matemática e Estatística
Bacharelado em Ciência da Computação

Correlações entre dados radiômicos e dados clínicos em tumores de boca e garganta de célula escamosa

Lourenço Henrique Moinheiro Martins Sborz Bogo

Monografia Final

mac 499 — Trabalho de Formatura Supervisionado

Supervisor: André Fujita
Cossupervisor: Gilberto de Castro Junior
Cossupervisor: Vinícius Jardim Carvalho
Cossupervisor: Mateus Cunha

São Paulo
2023

Resumo

Lourenço Henrique Moinheiro Martins Sborz Bogo. **Correlações entre dados radiômicos e dados clínicos em tumores de boca e garganta de célula escamosa**. Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

Neste projeto estuda-se a forma de prever fatores clínicos de tumores de boca e garganta de célula escamosa, como a sobrevida de pacientes e o grau do tumor, a partir de dados radiômicos, que são dados retirados de imagens médicas. Essa possibilidade traria grandes benefícios tanto para os pacientes, que passariam por riscos menores para serem diagnosticados, quanto para os médicos que teriam acesso a informações mais completas sobre cada caso.

Palavras-chave: Radiômica. Ciência de Dados. Aprendizado não supervisionado. Estatística. Regressão. Tumor. Sobrevida. Diagnóstico.

Abstract

Lourenço Henrique Moinheiro Martins Sborz Bogo. **Correlations between radiomic and clinical data in head and neck squamous cell carcinomas**. Capstone Project Report (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

This project aims to study how to predict clinical factors of head and neck squamous cell carcinoma, such as the pacients' survival and the tumors' grades, using radiomics, a type of data which is extracted from clinical images. If possible, this would bring great benefits to patients, which would be at a lesser risk to receive their diagnoses, and to the physicians, which would be able to obtain more complete information about each specific case.

Keywords: Radiomics. Data science. Non-supervised learning. Statistics. Regression. Tumor. Survival. Diagnosis.

Sumário

Introdução

Na oncologia, os diagnósticos se realizam com base em dados clínicos (conjunto de informações sobre a saúde dos pacientes contendo o seu histórico de doenças, os procedimentos e os tratamentos a que foram submetidos) e genéticos (informações conseguidas, ou passíveis de se conseguir, a partir do DNA humano).

Esses dados são obtidos por meio de questionários, exames e biópsias. No caso concreto das biópsias, existem algumas limitações. Por serem processos invasivos e perigosos, apresentam riscos para os pacientes. Também não abrangem a heterogeneidade dos tumores, porque são uma extração pontual do tumor.

Com o rápido avanço da tecnologia impactando em todas as áreas da vida humana, tornou-se possível agregar novas categorias de informação que diminuam o risco para os pacientes e ajudem a desenvolver modelos preditivos que detectem precocemente as doenças. Isto é de vital importância para pacientes com câncer, uma vez que existe uma correlação entre a precocidade do diagnóstico e a probabilidade de cura.

É neste contexto que uma nova área da medicina nuclear – a radiômica – pode contribuir para a obtenção de diagnósticos mais precisos e seguros. Os dados radiômicos, extraídos de imagens clínicas, podem emergir como uma inovadora categoria de informação.

O que se pretende, neste projeto, é estudar as relações entre os dados radiômicos e os dados clínicos (como a sobrevida, o grau do tumor e o tabagismo) em tumores de boca e garganta, com o objetivo de descobrir se é possível prever alguns fatores clínicos a partir de variáveis radiômicas. Isto é relevante para a medicina, porque poderá habilitar diagnósticos a partir de dados radiômicos sem a necessidade de biópsias ou outros processos invasivos, normalmente utilizados para diagnosticar casos de tumores.

Aqui, foram usados os dados de tumores de boca e garganta do *The Cancer Genome Atlas* (TCGA), um banco de dados público com informações sobre tumores. Exploraram-se dados de 111 pacientes e, a partir de cada um, extraíram-se mais de 2000 *features* radiômicas.

Para encontrar as relações entre os dados, aproximações diferentes são feitas no decorrer do projeto e, para cada etapa dos experimentos, vários algoritmos distintos foram utilizados na tentativa de identificar o mais performático.

Nos próximos capítulos será explicado, com profundidade, quais os motivos subjacentes à escolha da linguagem, qual o interesse da pesquisa, o que é a radiômica, quais os tipos de dados usados no projeto, quais experimentos realizados e qual o objetivo de cada um deles, seus resultados e a conclusão.

Capítulo 1

Escolha da Linguagem

Um dos fatores iniciais, e relevantes, em um projeto de pesquisa, a seguir à seleção e demarcação do tema, é a definição da linguagem, por impactar diretamente no desenvolvimento do trabalho e na qualidade dos resultados.

Com milhares de linguagens disponíveis atualmente, e em constante aperfeiçoamento, é difícil navegar nesse universo. No entanto, existem algumas que se destacam por sua maturidade e número de utilizadores. E é neste grupo que emergem normalmente as escolhas.

A opção por uma linguagem passa, em geral, por duas dimensões: uma técnica, associada às características e performance da própria linguagem, e outra pessoal, relacionada aos *hard skills* do pesquisador.

Na técnica, alguns critérios devem ser observados: disponibilidade de bibliotecas; existência de um ambiente, que ofereça apoio aos pesquisadores; alinhamento da linguagem aos objetivos do projeto.

Já na dimensão pessoal, pesa o domínio que o pesquisador tem da linguagem, porque isso facilita a utilização, emprestando velocidade à pesquisa.

Neste projeto, uma seleção inicial reduziu a variedade de linguagens disponíveis para duas: Python e R. Após análise das vantagens de cada uma delas, a opção recaiu no Python.

Os critérios que mais pesaram na escolha foi a experiência prévia com a linguagem e o imenso ambiente de bibliotecas que ela proporciona. Além disso, outra vantagem é que a implementação do UMAP (*Uniform Manifold Approximation and Projection*) original, utilizada aqui, é em Python.

A linguagem R tem, por padrão, a maioria das funções utilizadas neste projeto, o que seria a sua principal vantagem caso tivesse sido escolhida. Apesar disso, não foi difícil encontrar as bibliotecas necessárias para realizá-lo em Python.

Estes breves comentários, ilustram como a escolha da linguagem não é fácil. Isso significa também que, em casos como o deste projeto, só em retrospecto se percebe a totalidade das vantagens de uma linguagem.

Capítulo 2

Radiômica

Dados radiômicos são uma grande quantidade de dados numéricos, extraídos de imagens médicas como radiografias, tomografias e ressonâncias magnéticas.

Essas extrações, feitas por radiologistas por meio de softwares e bibliotecas especializadas (como a pyradiomics Van Griethuysen *et al.*, 2017), não são completamente automáticas e os parâmetros devem ser alterados manualmente.

A variedade desses parâmetros permite a extração de centenas ou até milhares de *features* de uma única imagem clínica (mais de 2000 no dataset usado neste projeto).

Existem diversos tipos de *features* radiômicas Mayerhoefer *et al.*, 2020. As utilizadas aqui foram extraídas de pacientes do TCGA e são separadas em 4 grupos:

- *Features* de intensidade são estatísticas baseadas nas intensidades dos *pixels* das imagens. Alguns exemplos são a soma desses valores, o máximo, a média, a mediana, desvio padrão, máximo e mínimo.
- *Features* de formato e tamanho. Cada uma representa alguma característica do formato do tumor em três dimensões (neste caso), como área, volume, diâmetros e eixos. Esse é o tipo de *features* mais simples e interpretável.
- *Features* de textura levam em conta a textura do tumor, ou seja, a variação dos *pixels* nas imagens. Isso é feito, por exemplo, usando o valor do gradiente local em alguns conjuntos de *pixels* da imagem.
- *Features* de filtro são calculadas a partir de transformações feitas na imagem original, por exemplo, usando o Filtro de Gabor.

Além desses quatro tipos, existem outros, como as *features* de modelo, que não foram explorados neste projeto, por não fazerem parte do *dataset* inicial.

Normalmente, dados radiômicos são utilizados para criar modelos de aprendizado supervisionado, como classificadores ou regressões. Aqui, exploraram-se aproximações alternativas para o problema, com maior enfoque em aprendizado não supervisionado e em estatística.

Capítulo 3

Motivação

Este projeto explora as relações entre as variáveis radiômicas e duas variáveis clínicas específicas, o grau do tumor e a sobrevida do paciente.

Usar dados radiômicos para diagnosticar tumores pode trazer grandes vantagens, com as principais sendo:

1. As biópsias retiram material de um ponto do tumor, sendo incapazes de capturar a sua heterogeneidade. Já os dados de imagem abrangem toda a região de importância, capturando, também, a sua heterogeneidade. Isso pode contribuir para a obtenção de diagnósticos mais precisos.
2. As biópsias são processos extremamente invasivos que podem ser perigosos para os pacientes. Conseguir prever alguns fatores sem sua necessidade evitará expor o paciente a riscos contornáveis.

Ou seja, encontrar uma relação entre dados clínicos e dados radiômicos, é útil tanto para os pacientes, que poderão ter seus diagnósticos por meio de processos mais seguros, quanto para os médicos, que poderão conseguir interpretações melhores e mais completas das regiões tumorais de interesse.

Capítulo 4

Dataset

Como já mencionado, os dados usados ao longo deste projeto foram adquiridos do TCGA (*The Cancer Genome Atlas*). São dados de tumores de boca e garganta de célula escamosa, divididos em duas tabelas, uma com dados clínicos e outra com radiômicos. A qualidade dos dados é essencial para garantir os resultados científicos. Não há como solucionar um problema científico

O primeiro passo necessário para trabalhar os dados foi a consolidação das duas tabelas. Como a ordem dos pacientes é diferente em cada uma delas, usou-se uma chave estrangeira para combiná-las. Sendo o índice do paciente um identificador único, presente em ambas as tabelas, selecionou-se essa coluna como chave estrangeira.

Percebe-se na imagem 4.1, que o formato dos índices era diferente nas duas tabelas, portanto desenvolveu-se um *script* para deixá-los iguais.

PAT_IND	Patients number
'Pat1'	
'Pat2'	
'Pat3'	Pat113
'Pat4'	Pat112
'Pat5'	Pat111
'Pat6'	Pat110
'Pat7'	Pat109
'Pat8'	Pat108
'Pat9'	-
'Pat10'	Pat107
'Pat11'	-
'Pat12'	-
'Pat13'	Pat106
'Pat14'	Pat105
'Pat15'	Pat104
'Pat16'	Pat103
'Pat17'	Pat102
'Pat18'	-
'Pat19'	Pat101
'Pat20'	Pat100
'Pat21'	-
'Pat23'	Pat99

Figura 4.1: *Figura com as colunas dos índices dos pacientes em ambas as tabelas (radiômica à esquerda e clínica à direita)*
Fonte: Autor

Durante o processo de consolidação notou-se que alguns pacientes estavam presentes somente em uma das tabelas. Esses dados foram descartados, pois muitas etapas demandavam informações de ambas as tabelas.

Outro pré-processamento foi realizado para remover as redundâncias das tabelas (por exemplo, a tabela de dados clínicos tem duas linhas com os *labels* das colunas).

Figura 4.2: *Figura evidenciando que a tabela de dados clínicos contém duas linhas com os labels das colunas.*
Fonte: Autor

Capítulo 5

Experimentos

No decorrer do projeto realizaram-se vários experimentos usando diferentes técnicas, para avaliar qual a mais performática.

Nesta seção, serão apresentados e analisados os experimentos, seus objetivos e resultados.

5.1 Experimento inicial

O objetivo inicial era fazer uma clusterização dos dados radiômicos usando algum dos fatores clínicos como *label* dos *clusters*, mostrando a relação entre os dois tipos de dados. Para que o resultado se tornasse visualizável, seria necessário diminuir a dimensionalidade dos dados, já que foram extraídas aproximadamente 2000 *features* radiômicas de cada paciente.

Porém, esse processo só seria possível se houvesse um tratamento anterior dos dados.

O primeiro passo foi normalizar os dados e, para isso, testaram-se dois processos diferentes. Começou-se por usar a normalização min-max. Nesse caso, cada *feature* x é substituída por $\frac{x-x_{min}}{x_{max}-x_{min}}$, onde x_{min} é o valor mínimo da *feature* no *dataset* e x_{max} é o valor máximo. Esse processo é interessante por colocar todas as *features* na mesma escala, entre zero e um.

Essa normalização foi abandonada quando se introduziu uma nova etapa no experimento, o PCA (*Principal Component Analysis*), algoritmo que exige que todos os valores tenham variância um, coisa que a normalização min-max não faz. Portanto, optou-se por usar a normalização Z, onde cada *feature* x é substituída por $\frac{x-\mu}{\sigma}$, tal que μ é a média da *feature* e σ é o desvio padrão. Esse processo faz com que todos os valores fiquem entre zero e um, como a normalização anterior, mas também faz todos os valores ficarem com média zero e variância um.

Em seguida, aplicaram-se algoritmos de diminuição de dimensionalidade, começando pelo PCA. Esse algoritmo combina diversas variáveis em uma, de modo que uma quantidade desejada da variância seja conservada. Isso se obtém combinando variáveis com correlação

positiva, ou seja, a partir da nova variável é possível reconstruir ou aproximar as variáveis iniciais (Bro e Smilde, 2014). Dois testes foram feitos para saber quanto da variância deveria ser mantida, um com 90% e outro com 95%. O desempenho com 95% não foi suficientemente melhor para justificar a quantidade de *features* necessária para manter os 5% a mais de variância. Por isso, decidiu-se conservar 90% da variância. Na figura 5.1, o gráfico revela quanto cada componente do PCA representa da variância inicial dos dados radiômicos.

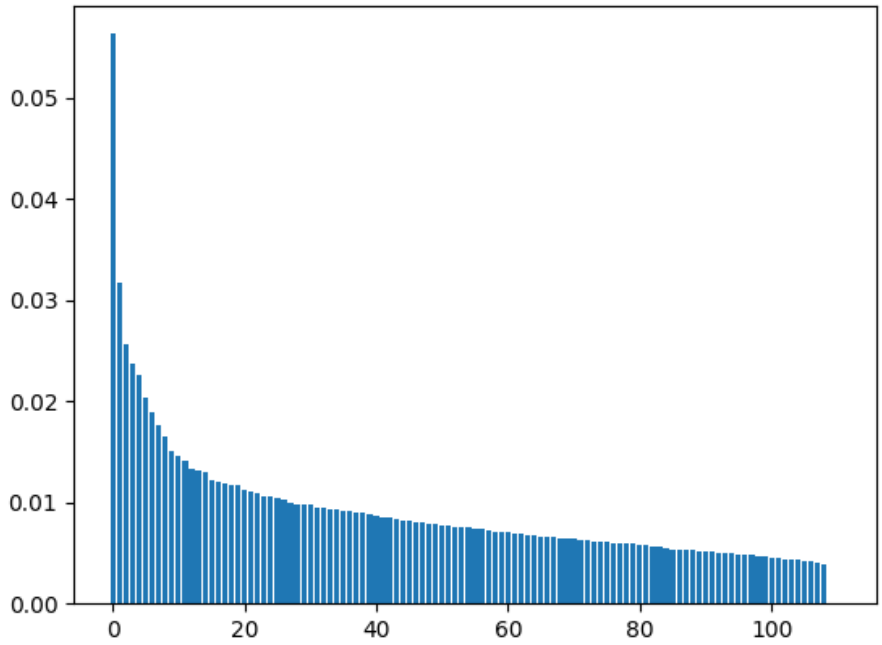

Figura 5.1: *Gráfico com quanto cada componente do PCA representa da variância do dataset. Interessante notar que a componente mais importante corresponde a aproximadamente 5% da variância. Fonte: Autor*

Esse passo trouxe resultados abaixo dos esperados. Idealmente, as primeiras componentes do PCA corresponderiam à grande parte da variância do *dataset*. Como isso não ocorreu, buscou-se outros algoritmos com propósitos parecidos, como o KernelPCA, uma variação do algoritmo PCA que normalmente se comporta melhor em dados não lineares, e o SVD (*Singular Value Decomposition*). Esses algoritmos tiveram performances praticamente iguais à do PCA, portanto o PCA foi mantido.

Com as novas variáveis, o passo seguinte foi diminuir, uma vez mais, a dimensionalidade, mas dessa vez diretamente para duas dimensões. Testaram-se novamente dois algoritmos. De início, aplicou-se o t-SNE (*t-Distributed Stochastic Neighbor Embedding*) nos dados resultantes do PCA. Esse algoritmo transforma as distâncias euclidianas entre dois pontos no espaço de dimensão alta em uma probabilidade de que esses dois pontos serão vizinhos no novo espaço. A probabilidade é calculada com base na similaridade entre os dois pontos.

Ao aplicar esse algoritmo, gerou-se um espaço bidimensional com os pontos do *dataset*. Vale lembrar que os eixos dos gráficos gerados a partir desse processo não são interpretáveis, ou seja, seus valores por si não têm significado. Abaixo, estão algumas figuras com exemplos dessas execuções.

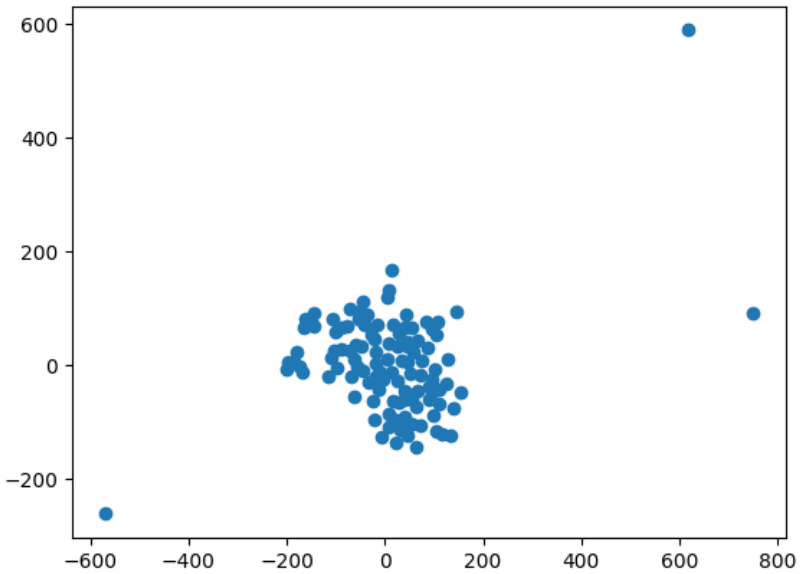

Figura 5.2: *Gráfico com execução do processo usando o PCA e o t-SNE.*
Fonte: Autor

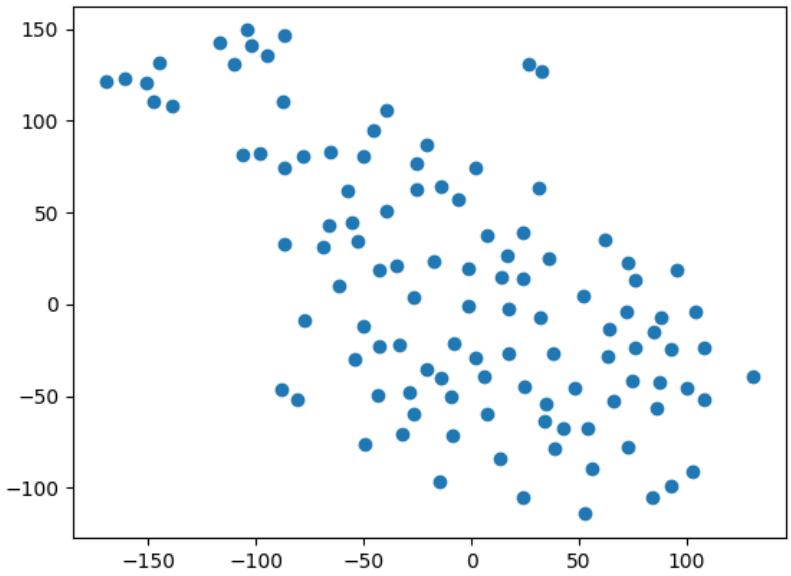

Figura 5.3: *Gráfico com execução do processo usando o PCA e o t-SNE. Os resultados são diferentes do gráfico anterior, pois a inicialização do algoritmo é aleatória, ou seja, sem uma semente, os resultados serão diferentes em cada execução.*
Fonte: Autor

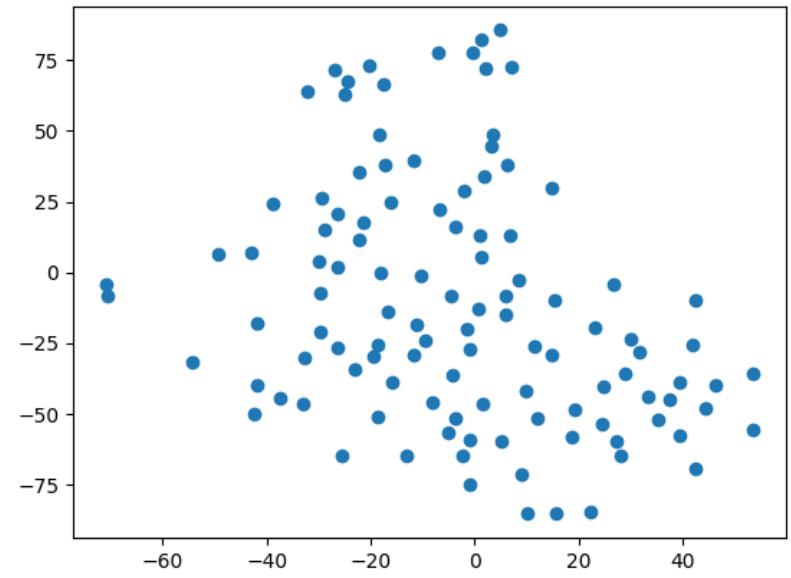

Figura 5.4: *Outro exemplo do processo PCA e t-SNE sendo aplicado.*
Fonte: Autor

Esse processo não trouxe resultados interessantes. A premissa era de que alguns grupos

se formassem antes de aplicar um algoritmo de clusterização e que se refinassem com a aplicação de tal algoritmo. Como o resultado ficou aquém do desejado, fez-se uma nova tentativa utilizando outro algoritmo de redução de dimensionalidade ao invés do t-SNE, o UMAP (*Uniform Manifold Approximation and Projection*). Uma explicação desse algoritmo, mesmo que simplificada como já foi feita para os outros, foge do escopo dessa tese por sua grande complexidade, podendo ser encontrada no artigo McInnes *et al.*, 2018, no qual o algoritmo foi proposto.

Abaixo, alguns dos resultados da execução, com o UMAP.

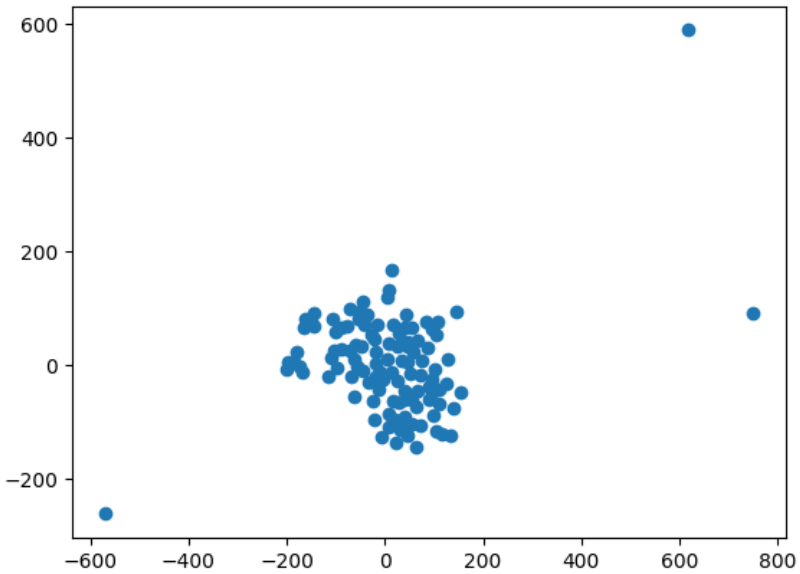

Figura 5.5: *Gráfico com execução do processo usando o PCA e o UMAP.*
Fonte: Autor

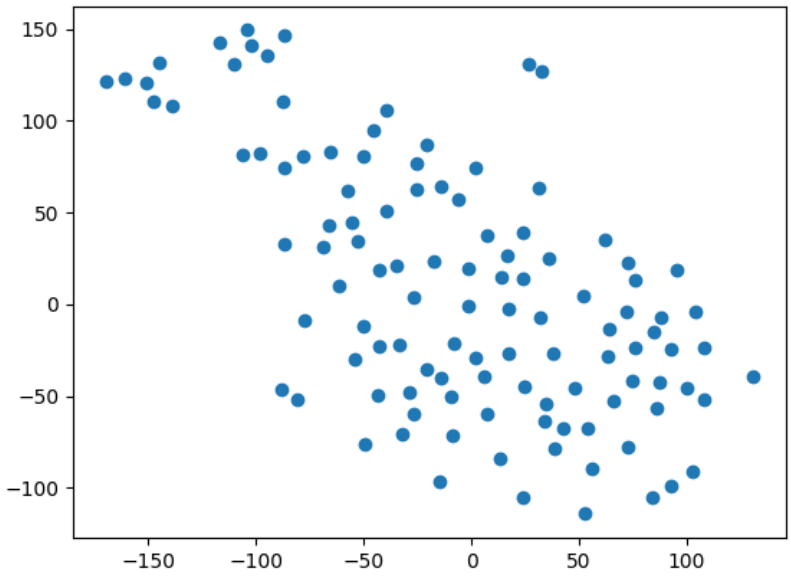

Figura 5.6: *Outro exemplo do processo usando o PCA e o UMAP.*
Fonte: Autor

Neste caso, evidencia-se que há um grupo deslocado do resto, portanto decidiu-se colorir os pontos usando como cor os valores das variáveis clínicas alvo do projeto, o grau do tumor e a sobrevida dos pacientes.

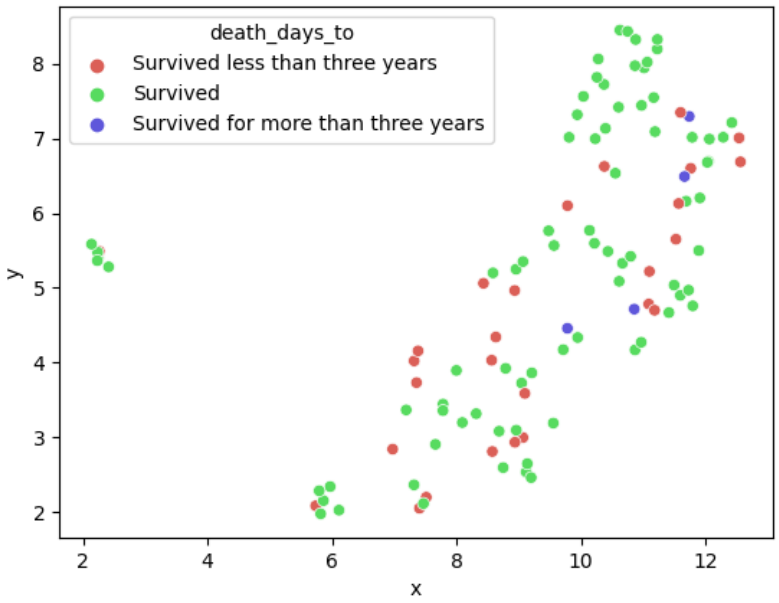

Figura 5.7: *Gráfico do processo com UMAP após colorir os pontos usando a sobrevida do paciente como cor. Nota-se que não há relação entre a sobrevida e o grupo de pontos separados.*
Fonte: Autor

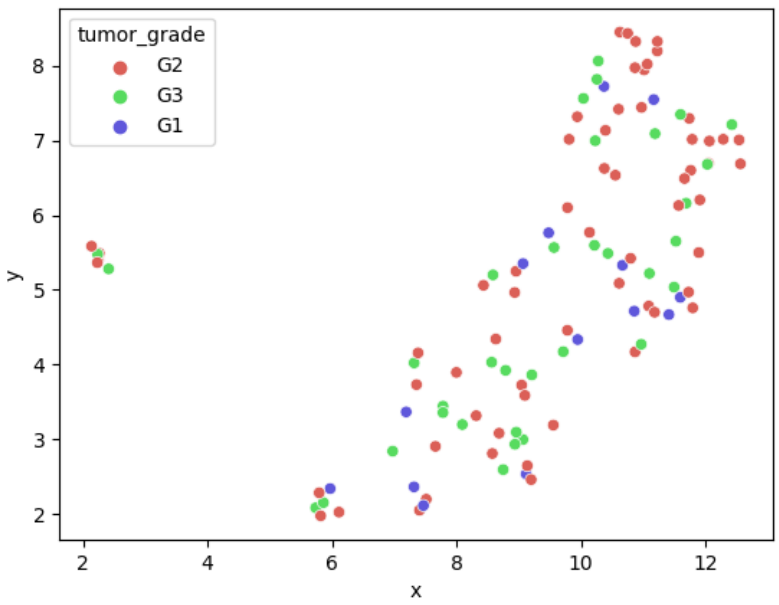

Figura 5.8: *Gráfico do processo com UMAP após colorir os pontos usando o grau do tumor como cor. Nota-se que não há relação entre o grau do tumor e o grupo de pontos separados.*
Fonte: Autor

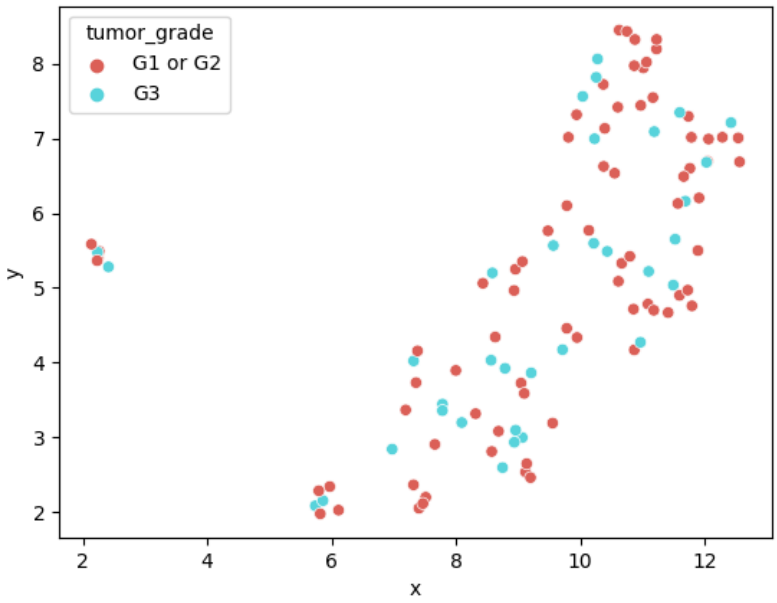

Figura 5.9: *Gráfico do processo com UMAP após colorir os pontos usando o grau do tumor como cor. Nesta situação, os pacientes estão separados em grau um e dois e grau três do tumor, pois os tumores de grau três são um caso mais extremo. Este gráfico é, portanto, semelhante ao anterior, porém os pacientes com tumores de grau um e dois estão com a mesma cor.*
Fonte: Autor

Resumindo: os resultados deste primeiro experimento não captaram uma relação significativa entre as variáveis clínicas de interesse e os dados radiômicos. Por isso, passou-se para uma segunda etapa de testes, aplicando-se análises mais diretas nos dados.

5.2 Segundo experimento

Iniciou-se esta etapa aplicando um teste Mann-Whitney nos dados radiômicos. O teste permitiu checar se as médias das *features* radiômicas têm valores parecidos entre os grupos criados pela variável de interesse, que, nesse caso, era o grau do tumor. Para cada variável, o teste atribui um p-valor. As variáveis com menor p-valor são as mais interessantes, porque uma diferença de valor nessas variáveis está relacionada a uma diferença no grau do tumor.

Abaixo, segue uma figura com os menores p-valores encontrados.

	pvalue
Histogram_Of_Gradient_245	0.001821
Local_Binary_Pattern_135	0.002042
Histogram_Bin_44	0.002042
Local_Binary_Pattern_60	0.002557
Local_Binary_Pattern_199	0.002794
GaborBank_F8GLCMVariance	0.003119
Local_Binary_Pattern_247	0.003187
Histogram_Of_Gradient_533	0.003552
Local_Binary_Pattern_192	0.003552
Histogram_Of_Gradient_311	0.005881
Local_Binary_Pattern_94	0.006252
GaborBank_D8NGTDMBusyness	0.006644
Local_Binary_Pattern_67	0.007200
Histogram_Bin_46	0.007493
Histogram_Of_Gradient_107	0.007953
GaborBank_D8GLCMVariance	0.008604
Histogram_Of_Gradient_160	0.008775
Local_Binary_Pattern_207	0.008948
Histogram_Of_Gradient_529	0.011061
Local_Binary_Pattern_98	0.011489
Local_Binary_Pattern_13	0.011489
Histogram_Of_Gradient_333	0.011708
Histogram_Bin_39	0.012158
GaborBank_D6GLSZMGLN	0.013104
Local_Binary_Pattern_149	0.013104
Local_Binary_Pattern_228	0.013350
Histogram_Of_Gradient_262	0.013600
GaborBank_D3GLRLMLGRE	0.013855
Local_Binary_Pattern_105	0.014914
GLRLM_SRHGE	0.015190
Local_Binary_Pattern_254	0.017247
Histogram_Of_Gradient_479	0.017559
Local_Binary_Pattern_74	0.019889
Local_Binary_Pattern_78	0.020967

Figura 5.10: *Variáveis com melhor resultado no teste Mann-Whitney. À esquerda estão os nomes das variáveis e à direita, os p-valores.*
Fonte: Autor

O próximo passo foi aplicar uma correção fdr (*False Discovery Rate*) nos p-valores. A ideia era manter as variáveis com p-valor menor do que 5% e descartar as outras.

	pvalue
Histogram_Of_Gradient_245	0.661098
Local_Binary_Pattern_135	0.661098
Histogram_Bin_44	0.661098
Local_Binary_Pattern_60	0.661098
Local_Binary_Pattern_199	0.661098
GaborBank_F8GLCMVariance	0.661098
Local_Binary_Pattern_247	0.661098
Histogram_Of_Gradient_533	0.661098
Local_Binary_Pattern_192	0.661098
Histogram_Of_Gradient_311	0.828808
Local_Binary_Pattern_94	0.828808
GaborBank_D8NGTDMBusyness	0.828808
Local_Binary_Pattern_67	0.828808
Histogram_Bin_46	0.828808
Histogram_Of_Gradient_107	0.828808
GaborBank_D8GLCMVariance	0.828808
Histogram_Of_Gradient_160	0.828808
Local_Binary_Pattern_207	0.828808
Histogram_Of_Gradient_529	0.828808
Local_Binary_Pattern_98	0.828808
Local_Binary_Pattern_13	0.828808
Histogram_Of_Gradient_333	0.828808
Histogram_Bin_39	0.828808
GaborBank_D6GLSZMGLN	0.828808
Local_Binary_Pattern_149	0.828808
Local_Binary_Pattern_228	0.828808
Histogram_Of_Gradient_262	0.828808
GaborBank_D3GLRLMLGRE	0.828808
Local_Binary_Pattern_105	0.838240
GLRLM_SRHGE	0.838240
Local_Binary_Pattern_254	0.838240
Histogram_Of_Gradient_479	0.838240
Local_Binary_Pattern_74	0.838240
Local_Binary_Pattern_78	0.838240
Local_Binary_Pattern_158	0.838240

Figura 5.11: *Variáveis com melhor resultado no teste Mann-Whitney após a correção fdr. À esquerda estão os nomes das variáveis e à direita, os p-valores.*
Fonte: Autor

Na imagem, percebe-se que nenhuma variável ficou com p-valor abaixo de 60% após a correção fdr. Decidiu-se, então, selecionar as 15 melhores variáveis (com menor p-valor) para o próximo passo.

Aplicou-se uma Regressão de Cox. Essa regressão é um modelo de sobrevivência, ou seja, um tipo de modelo utilizado para relacionar o tempo até um evento ocorrer com uma ou mais variáveis. Neste caso, o evento é a morte do paciente, portanto o tempo até esse evento ocorrer é a sobrevida, e as variáveis são as *features*. Além disso, foram escolhidas três covariáveis clínicas para usar no modelo, de modo a possibilitar uma análise comparativa do desempenho das radiômicas tendo as clínicas como base.

As três covariáveis escolhidas são:

1. Idade: os pacientes foram separados em dois grupos, um com mais de 40 anos e o outro com menos.

2. Sexo biológico
3. Tabagismo: aqui, usou-se um valor chamado *Pack years smoked*, que é o número de packs por dia multiplicado pelo número de anos em que a pessoa foi fumante.

As covariáveis foram selecionadas por sua alta correlação com a incidência do câncer aqui estudado. Outro fator relevante é ascendência asiática, porém essa variável não foi utilizada, pois no *dataset* havia apenas um paciente nesse grupo.

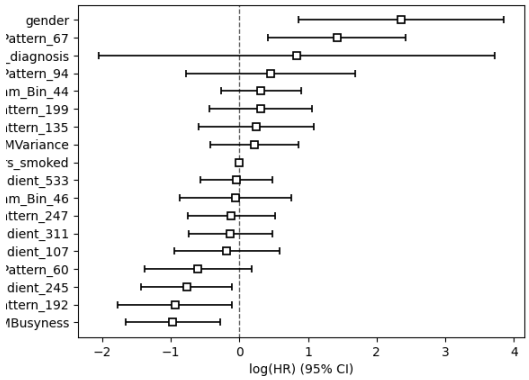

Figura 5.12: *Resultados da Regressão de Cox. Nota-se que algumas variáveis radiômicas, como a Local_Binary_Pattern_67 e o GaborBank_D8NGTDMBusyness, estão muito distantes do 0 na regressão, indicando uma alta correlação com a sobrevida do paciente.*
Fonte: Autor

Esse processo trouxe resultados promissores e interessantes. Algumas variáveis radiômicas têm forte correlação com a sobrevida dos pacientes, especialmente a Local_Binary_Pattern_67, que é uma *feature* de textura, e o GaborBank_D8NGTDMBusyness, uma *feature* de filtro, com desempenhos comparáveis ao tabagismo no conjunto de dados clínicos utilizado. Isso é interessante, porque se esperava, inicialmente, que as *features* mais relevantes fossem as de intensidade e as de tamanho e forma. Uma possível explicação é que as *features* de textura e filtro existem em uma quantidade muito maior do que as outras, então é mais provável que alguma delas tenha boa correlação com os dados clínicos.

Capítulo 6

Conclusão

O projeto tinha por objetivo relacionar dados radiômicos de tumores de boca e garganta com seus respectivos dados clínicos. Se encontrada uma relação entre os dois conjuntos de dados, seria possível prever fatores clínicos a partir dos dados radiômicos. A relevância dessa descoberta é a sua utilidade para a medicina. Entre pacientes reduziria a necessidade de uma biópsia, que, dependendo da parte do corpo onde é feita, pode ser um processo invasivo, caro e, perigoso. E, entre os médicos haveria uma ferramenta adicional para diagnosticar os tumores, por meio de interpretações dos dados radiômicos. Ou seja, se estabelecida essa relação entre as duas categorias de dados, haveria possibilidade de um diagnóstico mais barato, ágil e seguro.

As análises estatísticas realizadas neste projeto mostraram que existe uma relação grande entre os dados radiômicos e os dados clínicos. Ou seja, os resultados indicam que é possível prever alguns fatores clínicos a partir dos dados radiômicos.

As variáveis radiômicas que têm uma maior correlação com a sobrevida dos pacientes, neste conjunto de dados, são a Local_Binary_Pattern_67, que é uma *feature* de textura, e o GaborBank_D8NGTDMBusyness, uma *feature* de filtro. Essas duas variáveis tiveram um desempenho comparável ao tabagismo, uma das principais variáveis no diagnóstico do tipo de câncer aqui analisado.

Apesar desses resultados, algumas ressalvas devem ser feitas:

1. Os tumores aqui estudados são extremamente heterogêneos, indicando que talvez em outro *dataset*, outras variáveis tenham desempenhos melhores no processo descrito.

2. O *dataset* usado tem um universo estatístico limitado; para uma análise mais completa seria necessário uma quantidade maior de dados. A grande questão é que a disponibilidade de dados para projetos como este exige um processo muito trabalhoso, pois além da própria especificidade da extração dos dados radiômicos, também há questões legais envolvidas, já que os dados são extraídos de indivíduos.

Uma análise não realizada aqui, mas interessante como continuação deste projeto é experimentar como as variáveis se comportam em diferentes *datasets*. Como já mencionado, pode ser que diferentes *features* se comportem melhor em outros dados e as que se comportaram bem aqui sejam apenas pontos fora da curva. Isso é importante, pois

é falaciosa a conclusão de que as variáveis que tiveram bom desempenho neste projeto sempre serão bons preditores para a sobrevida do paciente. O que este projeto indicou é que é possível prever, mas são necessários mais estudos, com outros dados antes de se concluir quais variáveis são boas para prever os fatores.

Existem artigos que mencionam trabalhos parecidos com este, porém usando aprendizado supervisionado, que chegaram a conclusões semelhantes.

Haim *et al.*, 2022 fizeram um estudo sobre uma forma de prever, em pacientes com metástase de Câncer de Pulmão de Células Pequenas no cérebro, o status da mutação EGFR (um importante fator, checado por médicos ao tratar casos desse câncer) usando *Deep Learning* e outras técnicas de aprendizado supervisionado, como *Transfer Learning*. Eles concluíram que alguns fatores são possíveis de serem previstos a partir dos dados radiômicos. Ahn *et al.*, 2020 também estudaram esse mesmo tipo de previsão no mesmo tipo de câncer, com conclusões semelhantes.

Referências

[Ahn *et al.* 2020] Sung Jun Ahn *et al.* "Contrast-enhanced t1-weighted image radiomics of brain metastases may predict egfr mutation status in primary lung cancer". Em: *Scientific Reports* 10.1 (2020), pgs. 1–9 (citado na pg. 26).

[Bro e Smilde 2014] Rasmus Bro e Age K Smilde. "Principal component analysis". Em: *Analytical methods* 6.9 (2014), pgs. 2812–2831 (citado na pg. 14).

[Haim *et al.* 2022] Oz Haim *et al.* "Predicting egfr mutation status by a deep learning approach in patients with non-small cell lung cancer brain metastases". Em: *Journal of Neuro-Oncology* 157.1 (2022), pgs. 63–69 (citado na pg. 26).

[Mayerhoefer *et al.* 2020] Marius E Mayerhoefer *et al.* "Introduction to radiomics". Em: *Journal of Nuclear Medicine* 61.4 (2020), pgs. 488–495 (citado na pg. 5).

[McInnes *et al.* 2018] Leland McInnes, John Healy e James Melville. "Umap: uniform manifold approximation and projection for dimension reduction". Em: *arXiv preprint arXiv:1802.03426* (2018) (citado na pg. 17).

[Van Griethuysen *et al.* 2017] Joost JM Van Griethuysen *et al.* "Computational radiomics system to decode the radiographic phenotype". Em: *Cancer research* 77.21 (2017), e104–e107 (citado na pg. 5).