UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

Daniel Cesar Braz

Aprendizado de máquina aplicado em dados de biossensores para diagnóstico de câncer e COVID-19

São Carlos

2022

Daniel Cesar Braz

Aprendizado de máquina aplicado em dados de biossensores para diagnóstico de câncer e COVID-19

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências.

Área de concentração: Física Computacional

Orientador: Prof. Dr. Osvaldo Novais de Oliveira Junior

Versão Corrigida

(versão original disponível na Unidade que aloja o Programa)

São Carlos

2022

Braz, Daniel Cesar
Aprendizado de máquina aplicado em dados de biossensores para diagnóstico de câncer e COVID-19 / Daniel Cesar Braz; orientador Osvaldo Novais de Oliveira Junior - versão corrigida -- São Carlos, 2022.
175 p.

Tese (Doutorado - Programa de Pós-Graduação em Física Computacional) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2022.

1. Aprendizado de máquina. 2. Biossensores. 3. Câncer. 4. COVID-19. I. Oliveira Junior, Osvaldo Novais de, orient. II. Título.

Dedico este trabalho à minha esposa Isis, aos meus filhos Danilo, Bernardo e Vitória, e aos meus pais Izabel e Daniel (in memorian), que sempre cuidaram de mim com amor e fizeram o possível para me ajudar.

AGRADECIMENTOS

O reconhecimento e a valorização são práticas importantes em um trabalho de pesquisa científica que, por suas especificidades e nível de complexidade, seria de difícil realização ou não seria possível. A partir dessa assertiva e com alegria faço os agradecimentos a seguir.

Agradeço primeiramente a **Deus e a seu filho Jesus Cristo** pela vida, pela saúde, pela força, pela orientação e pelas pessoas que colocaram em meu caminho.

Agradeço profundamente e com amor à minha **esposa Isis e aos meus filhos Danilo, Bernardo e Vitória,** pela compreensão, por estarem juntos comigo enfrentando todas as dificuldades, motivando-me e ajudando-me sempre.

De forma muito especial, agradeço ao meu **orientador, Prof. Dr. Osvaldo Novais de Oliveira Jr.** por acreditar no meu potencial, dando-me a oportunidade de realizar esse trabalho, pelos direcionamentos e esclarecimentos, respeito, amizade, disponibilidade, paciência, compreensão, dedicação e espírito de colaboração.

Agradeço profundamente e com amor aos meus **pais, Izabel de Souza Braz *(in memorian)* e Daniel Braz *(in memorian)***, por terem me dado a oportunidade e as condições para buscar e trilhar o meu caminho na direção da realização dos meus sonhos. Também à minha **irmã, Paula de Souza Braz,** pelo apoio e pelas dificuldades que passamos juntos em nossa caminhada.

Agradeço aos demais professores, pós-docs e alunos por seus ensinamentos, pela amizade e pelas importantes contribuições seja na coleta, disponibilização e análise dos dados: Prof. Dr. Odemir Martinez Bruno, Prof. Dr. Fernando V. Paulovich, Profa. Dra. Maria Cristina F. Oliveira, Prof. Dr. Mário Popolin Neto, Prof. Dr. Alexandre S. Martinez, Prof. Dr. Zhao Liang, Prof. Dr. Moacir A. Ponti, Dra. Valquiria C. Rodrigues, Dr. Flavio M. Shimizu, Dr. Acelino C. Sá, Dra. Gisela Ibáñez-Redin, Dr. Wallance M. Pazin, Dra. Juliana C. Soares, Dr. Andrey C. Soares, Dr. Lucas C. Ribas, MSc. Leonardo F. S. Scabini, Lorrany V. G. Barros, Lucas B. Souza, Lucas V. Voltera.

Agradeço aos professores Dr. Luciano da Fontoura Costa (IFSC/USP), Dr. Hélio Pedrini (IC/UNICAMP) e Dr. João do Espirito Santo Batista Neto (ICMC/USP) por suas valorosas contribuições no momento da qualificação deste trabalho.

Agradeço ao meu **sogro e sogra, Geraldo Alves de Faria e Regina Izabel de Faria *(in memorian)*,** pelo importante apoio e aconselhamentos.

Agradeço a todos os familiares e amigos não citados que de alguma forma, ao longo do meu caminho até aqui, deram sua contribuição.

Agradeço à secretaria do Grupo de Polímeros, representada por **Rosângela Maria**

Marcondes de Oliveira e **Simone Ferreira dos Reis**, pelo pronto atendimento às minhas demandas enquanto aluno.

Agradeço à secretaria da Pós-graduação do Instituto de Física de São Carlos – USP, representada por **Silvio César Athayde** e **Ricardo Vital do Prado**, por sempre estarem à disposição e pronto a atender às minhas demandas enquanto aluno.

Agradeço à **Maria Cristina Cavarette Dziabas** e **Maria Neusa de Aguiar Azevedo** da biblioteca do Instituto de Física de São Carlos – USP, pelas contribuições na revisão desta tese e por sempre estarem à disposição a atender às minhas demandas enquanto aluno.

Agradeço ao Instituto de Física de São Carlos (IFSC) da USP, sua Pós-Graduação, seus funcionários e professores.

Agradeço ao Instituto de Ciências Matemáticas e de Computação (ICMC) e à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da USP.

Agradeço à Universidade de São Paulo (USP).

Agradeço à Universidade Estadual de Mato Grosso do Sul (UEMS), pelo Programa de Capacitação, que permitiu o afastamento integral das minhas atividades enquanto docente para realizar esse trabalho.

Agradeço às agências que financiaram direta e indiretamente esse projeto: CAPES, CNPq, FAPESP e FINEP.

Agradeço a todos que não foram mencionados e com quem tive contato ao longo da vida, por terem contribuído de alguma forma com a minha formação e com a realização desse trabalho.

“Filho, no fim tudo dará certo. Se não deu certo ainda é por que não chegou o fim. Eu te amo.”

Izabel de Souza Braz (in memorian)

RESUMO

BRAZ, D. C. **Aprendizado de máquina aplicado em dados de biossensores para diagnóstico de câncer e COVID-19**. 2022. 175p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2022.

Esta tese explora o conceito de sistemas computacionais semi-automatizados de diagnóstico baseados em Aprendizado de Máquina (AM), em que diferentes tipos de dados de biossensores e de outras fontes são empregados. A partir de um pipeline base de AM, foram desenvolvidas quatro aplicações e diversos métodos foram implementados para cada uma das etapas/tarefas do pipeline. Como foram selecionados problemas desafiadores, um alto desempenho na acurácia do diagnóstico em geral só foi alcançado com algoritmos de AM supervisionado. Três aplicações foram em diagnóstico de câncer, sendo a primeira a partir de imagens de microscopia eletrônica de genossensores que detectam o biomarcador de PCA3 para câncer de próstata. Essas imagens foram usadas como entrada para algoritmos de AM supervisionado. Com os atributos de textura *Local Complex Features and Neural Network (LCFNN)* e o algoritmo *Linear Discriminant Analysis (LDA)* obteve-se uma taxa de acerto de 99,9% para classificação binária (sim/não para PCA3) e 88,3% para a classificação multiclasse em que se determina a concentração do biomarcador de PCA3. As outras duas aplicações envolveram a detecção de biomarcadores de câncer a partir de medidas elétrica/eletroquímica. A concentração da proteína p53, importante marcador de diferentes tipos de câncer, em amostras de urina e saliva sintéticas, foi determinada a partir de medidas eletroquímicas com imunossensores, em que voltamogramas foram analisados com os algoritmos *Logistic Regression (LR)*, LDA, *Support Vector Machine-kernel linear (SVM-L)*, *Gaussian Naive Bayes (GNB)*, *K-Nearest Neighbors (KNN)* e *Decision Tree (DT)*. O imunossensor otimizado exibiu acurácia de 100% com todos os algoritmos na maioria dos conjuntos de atributos construídos a partir dos dados brutos. No diagnóstico de câncer de boca, a partir de medidas de impedância elétrica com uma língua eletrônica em amostras de saliva de pacientes e voluntários, a maior acurácia de 86.7% foi obtida com o algoritmo SVM-kernel radial. Nesta aplicação, a acurácia da classificação multiclasse aumentou quando foram adicionadas informações clínicas dos pacientes, indicando a importância de combinação de diferentes tipos de dados nos sistemas computacionais. A quarta aplicação foi o diagnóstico de COVID-19 com a detecção da proteína S do SARS-CoV-2 a partir de mapas hiperespectrais de Espectroscopia Raman com Amplificação de Superfície (SERS) obtidos de imunossensores. Usando algoritmo LDA obteve-se uma acurácia de 100% na distinção dos mapas para resultado positivo e negativo para SARS-CoV-2. Os resultados dessas quatro aplicações demonstram a possibilidade de se desenvolverem sistemas automatizados de diagnóstico, pois as várias etapas/tarefas dos pipelines de AM podem ser implementadas sem necessidade de intervenção humana, mesmo quando se combinam imagens, dados clínicos e de testes clínicos.

Palavras-chave: Aprendizado de máquina. Biossensores. Câncer. COVID-19.

ABSTRACT

BRAZ, D. C. **Machine learning applied to data of biosensors for diagnosis of cancer and COVID-19**. 2022. 175p. Thesis (Doctor in Science) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2022.

This thesis explores the concept of computer-assisted diagnosis based on machine learning (ML), in which different types of data from biosensors and other sources are employed. Using a ML pipeline, we developed four applications using different methods in the steps of the pipeline. Because the diagnostic problems addressed were all challenging, a high performance in accuracy was only achieved with supervised ML algorithms. Three applications involved cancer diagnosis, the first being from electron microscopy images of genosensors that detect the PCA3 biomarker for prostate cancer. These images were used as input for the ML algorithms, with texture features from Local Complex Features and Neural Network (LCFNN) and the algorithm Linear Discriminant Analysis (LDA) leading to a 99.9% accuracy for binary classification (yes/no for PCA3) and 88.3% accuracy for the multiclass classification where the PCA3 biomarker concentration is determined. The other two applications were related to detection of cancer biomarkers using electrical or electrochemical measurements. The concentration of p53 protein, an important marker of different types of cancer, in synthetic urine and saliva samples was determined from electrochemical measurements with immunosensors, and the voltammograms were analyzed with the Logistic Regression (LR), LDA, Support Vector Machine-kernel linear (SVM-L), Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN) and Decision Tree (DT) algorithms. The optimized immunosensor had 100% accuracy with all ML algorithms for most of the datasets with the raw voltammetric data. In the diagnosis of oral cancer using impedance measurements with an electronic tongue in saliva samples from volunteers and patients, the highest accuracy was 86.7% with SVM-kernel radial algorithm. In this application, the accuracy increased when patients' clinical information was added, indicating the importance of combining different types of data in computer-assisted diagnosis systems. The fourth application was the diagnosis of COVID-19 with detection of the SARS-CoV-2 S protein using Surface-Enhanced Raman Spectroscopy (SERS). Using the algorithm LDA an accuracy of 100% was achieved in distinguishing spectra for positive and negative result for SARS-CoV-2. The results of these four applications demonstrate the possibility of developing automated diagnostic systems, as the various stages/tasks in the ML pipeline can be implemented without the need for human intervention, even when combining images, clinical information and data from biosensors.

Keywords: Machine learning. Biosensors. Cancer. COVID-19.

LISTA DE FIGURAS

Figura 1.1 – Proposta de sistema computacional inteligente de diagnóstico baseado em diferentes fontes de dados. 30
Figura 1.2 – Esquema geral da composição e organização das aplicações de AM em dados de biossensores para diagnóstico de câncer e COVID-19. 32
Figura 2.1 – Organização de um conjunto de dados estruturado. 35
Figura 2.2 – Mapa conceitual do trabalho baseado em dados de biossensores. 36
Figura 2.3 – Exemplo de Curva de Aprendizagem. A linha vertical cinza indica o nível ótimo de complexidade do modelo aprendido. 39
Figura 2.4 – Exemplo de um dendrograma. A linha tracejada vermelha indica o limiar estipulado para determinar o número de clusters em que os objetos $Oi(i = 1...7)$ estão agrupados. 42
Figura 2.5 – Pipeline base de Aprendizado de Máquina. 43
Figura 2.6 – Exemplos de gráficos para visualização da distribuição de dados: (a) histograma e (b) boxplot. 44
Figura 2.7 – Matriz de confusão. 49
Figura 3.1 – Evolução anual da quantidade de registros nas bases de dados consultadas. 54
Figura 3.2 – Fluxograma baseado no modelo PRISMA para seleção dos estudos encontrados. 55
Figura 4.1 – Imagem de MEV do genossensor exposto à solução com o biomarcador PCA3 na concentração 10^{-2} (μmol.L^{-1}). 68
Figura 4.2 – Imagens de MEV de unidades do genossensor para as classes (a) negativa e (b) zero e positiva em ordem crescente das concentrações (μmol.L^{-1}) do biomarcador PCA3 (esquerda para direita): (c) 10^{-5}, (d) 10^{-4}, (e) 10^{-3}, (f) 10^{-2}, (g) 10^{-1}, (h) 1 68
Figura 4.3 – Recorte aplicado em regiões diferentes (sem sobreposição) de uma imagem gerando 9 janelas quadradas. 70
Figura 4.4 – Projeção t-SNE dos conjuntos de atributos (a) AHP, (b) CNRNN, (c) LCFNN e (d) GLDM para análise binária. 73
Figura 4.5 – Projeção t-SNE dos conjuntos de atributos (a) AHP, (b) CNRNN, (c) LCFNN e (d) GLDM para análise multiclasse. 75
Figura 4.6 – Projeção IDMAP dos conjuntos de atributos (a) Fourier e (b) CLBP para análise binária. 75
Figura 4.7 – Projeção IDMAP dos conjuntos de atributos (a) Fourier e (b) CLBP para análise multiclasse. 76
Figura 5.1 – Espectro de capacitâncias da língua eletrônica exposta à amostra de saliva de um paciente com câncer. 80

Figura 5.2 – Espectros de capacitâncias para as classes (a) cancer (yes/no) e (b) caso (control/cavity/floor). 80
Figura 5.3 – Conjunto dos espectros de capacitâncias para as classes (a) cancer (yes/no) e (b) caso (control/cavity/floor). 81
Figura 5.4 – Visualização com (a) PCA, (b) NCA, (c) t-SNE e (d) IDMAP para os dados de espectroscopia de impedância (only-sensor) para sim ou não para câncer. 83
Figura 5.5 – Visualização usando (a) PCA, (b) NCA, (c) t-SNE and (d) IDMAP para os espectros de impedância (only-sensor) para o tipo de câncer. 84
Figura 6.1 – Voltamogramas de pulso diferencial do imunossensor AA (a) e PEI (b) para amostras contendo biomarcador p53 para câncer. 92
Figura 6.2 – Voltamogramas de pulso diferencial do imunossensor AA (a) e PEI (b) para as classes controle (0) e com a presença (1) do biomarcador p53 para câncer. 93
Figura 6.3 – Visualização do conjunto de atributos AA-raw com (a) PCA, (b) NCA, (c) Isomap e (d) IDMAP. As classes 0 e 1 correspondem a amostras negativas e positivas, respectivamente. 95
Figura 6.4 – Visualização do conjunto de atributos PEI-raw com (a) PCA, (b) NCA, (c) Isomap e (d) IDMAP. As classes 0 e 1 correspondem às amostras negativas e positivas, respectivamente. 96
Figura 7.1 – Espectro de SERS do Imunossensor exposto à amostra de saliva de um paciente com COVID-19) 100
Figura 7.2 – Espectros de SERS do Imunossensor para as classes de amostras controle (0), antígeno da COVID-19 (1) e agente interferente (2). 100
Figura 7.3 – Espectros de SERS do Imunossensor exposto à amostras das classes positivo e negativo). 102
Figura 7.4 – Visualização da projeção IDMAP dos espectros médios (conjunto sers-mean) para amostras das classes positivo e negativo. 103
Figura A.1 – Histogramas das intensidades dos pixels das imagens do conjunto de dados agrupados por classe: 1 (a), 2 (c), 3 (e), 4 (g), 5 (a), 6 (c), 7 (e), 8 (g). 133
Figura A.2 – Gráfico de caixas (*boxplot*) das intensidades dos pixels das imagens do conjunto de dados agrupados por classe: 1 (a), 2 (b), 3 (c), 4 (d), 5 (e), 6 (f), 7 (g), 8 (h). O triângulo indica o valor da média das intensidades dos pixels. 142
Figura A.3 – Média das medidas de distância entre as imagens do conjunto de dados para cada classe: 1 (a), 2 (b), 3 (c), 4 (d), 5 (e), 6 (f), 7 (g), 8 (h). 144
Figura A.4 – Imagens de pontos (tamanho 1 pixel, intensidade 255) distribuídos uniformemente com diferentes frequências espaciais (ciclos/pixel): 8.049 (a), 16.098 (b), 32.105 (c) e 63.935 (d). 145
Figura A.5 – Medidas das frequências espaciais das imagens do conjunto de dados na direção das linhas (a), das colunas (b) e resultante (c). 147

Figura A.6 – Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 100 pixels do conjunto de dados agrupados por classe (1 a 4)..... 149
Figura A.7 – Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 100 pixels do conjunto de dados agrupados por classe (5 a 8)..... 150
Figura A.8 – Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 200 pixels do conjunto de dados agrupados por classe (1 a 4)..... 151
Figura A.9 – Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 200 pixels do conjunto de dados agrupados por classe (5 a 8)..... 152
Figura A.10–Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 300 pixels do conjunto de dados agrupados por classe (1 a 4)..... 153
Figura A.11–Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 300 pixels do conjunto de dados agrupados por classe (5 a 8)..... 154
Figura E.1 – Imagens da região de interesse (a) e pertencentes ao conjunto de dados de OA com níveis KL=0 (ausência)(b) e KL=2 (mínimo)(c)..... 164

LISTA DE TABELAS

Tabela 2.1 – Ferramentas computacionais utilizadas no trabalho. 52

Tabela 3.1 – Descrição dos estudos incluídos das bases de dados Pubmed, WoS e Scopus em termos do Objetivo [OB] pretendido, Conjunto de Dados [CD], tarefas de Pré-Processamento [PP], Engenharia de Atributos [EA], Modelagem [MO], Validação e Avaliação [VA], e Desempenho [DE] obtido pelo(s) modelo(s) de AM. 55

Tabela 3.2 – Síntese da descrição dos estudos incluídos na revisão. Símbolos utilizados: número de objetos (n), positivo (pos.) e controle (cont.). 64

Tabela 4.1 – Composição do conjunto de dados para uso do genossensor e MEV para diagnóstico de câncer de próstata. 69

Tabela 4.2 – Distribuição dos objetos entre as classes do conjunto de dados na aplicação genossensor e MEV para diagnóstico de câncer de próstata. Símbolo(s) utilizado(s): número de objetos (n). 69

Tabela 4.3 – Dimensões das janelas quadradas e números de exemplos gerados em cada classe de imagens do conjunto de dados. 71

Tabela 4.4 – Dimensionalidade dos conjuntos de atributos de textura extraídos das janelas com dimensões 300x300 pixels (90000 atributos). 72

Tabela 4.5 – Acurácia da análise de agrupamento com algoritmo KM para os casos binário e multiclasse para os conjuntos de atributos de textura extraídos das janelas com dimensões 300x300 pixels. 74

Tabela 4.6 – Acurácia média (desvio padrão) em (%) da análise de classificação computada com os algoritmos LDA, SVM e 1-NN para os casos binário e multiclasse para os conjuntos de atributos de textura extraídos das janelas com dimensões 300x300 pixels. 76

Tabela 5.1 – Composição do conjunto de dados utilizado na aplicação Língua Eletrônica e EIE para Diagnóstico de Câncer de Boca. 82

Tabela 5.2 – Distribuição dos objetos entre as classes no conjunto utilizado na aplicação Língua Eletrônica e EIE para Diagnóstico de Câncer de Boca. Símbolo utilizado: número de objetos (n). 82

Tabela 5.3 – Distribuição dos objetos em relação aos atributos clínicos e às classes. Símbolo utilizado: número de objetos (n). 82

Tabela 5.4 – Valor médio da largura da silhueta (Average Silhouette Width) (ASW) para o agrupamento com os algoritmos KM, HAC e SC com os dados de espectroscopia de impedância (only-sensor). 85

Tabela 5.5 – Média da largura da silhueta (Average Silhouette Width) (ASW) para o agrupamento com os algoritmos KM, HAC e SC para dados dos espectros (f-all) e com todos os atributos. 86
Tabela 5.6 – Acurácia média (desvio padrão) em (%) da análise de classificação computada com os algoritmos LDA, SVM-L e 1-NN sobre o conjunto de atributos selecionados/extraídos apenas dos espectros (only-sensor) para classes cancer (yes/no) e case (control, cavity, floor). 87
Tabela 5.7 – Acurácia média (desvio padrão) para a classificação com os algoritmos LR, LDA, GNB, KNN, SVM-L/R/P e RF para os conjuntos de atributos only-sensor e all-features. 88
Tabela 6.1 – Composição do conjunto de dados dos imunossensores AA e PEI. 93
Tabela 6.2 – Distribuição do número de objetos (n) entre as classes de dados dos imunossensores. 93
Tabela 6.3 – Largura média da silhueta (Average Silhouette Width) (ASW) para 2 agrupamentos com os algoritmos KM, HAC e SC para todos os conjuntos de dados. 95
Tabela 6.4 – Acurácia média (desvio padrão) (%) para a classificação usando os algoritmos LR, LDA, SVM-L, GNB, KNN e DT para todos os conjuntos de dados. 97
Tabela 7.1 – Composição do conjunto de dados utilizado na aplicação Imunossensor e SERS para Diagnóstico de COVID-19. 101
Tabela 7.2 – Distribuição dos objetos entre as classes no conjunto utilizado na aplicação Imunossensor e SERS para Diagnóstico de COVID-19. Símbolo utilizado: número de objetos (n). 101
Tabela 7.3 – Conjuntos de atributos extraídos dos mapas hiperespectrais. 103
Tabela 7.4 – Largura Média da Silhueta (Average Silhouette Width) (ASW) do agrupamento dos mapas das classes 0 e 1 com os algoritmos KM, HAC e SC para todos conjuntos de atributos. 104
Tabela 7.5 – Largura Média da Silhueta (Average Silhouette Width) (ASW) do agrupamento multiclasse com os algoritmos KM, HAC e SC para todos conjuntos de atributos 105
Tabela 7.6 – Acurácia média (desvio padrão) para a classificação das classes 0 e 1 com os algoritmos LR, LDA, GNB, KNN, SVM (L, R, P) e KNN. 106
Tabela 7.7 – Acurácia média (desvio padrão) para a classificação das classes 0 e (1-2) com os algoritmos LR, LDA, GNB, KNN, SVM (L, R, P) e KNN. 107
Tabela 7.8 – Acurácia média (desvio padrão) para a classificação multiclasse com os algoritmos LR, LDA, GNB, KNN, SVM (L, R, P) e KNN. 108
Tabela A.1 – Medidas de centralidade das intensidades dos pixels das imagens do conjunto de dados. 136

Tabela A.2 – Média (avg) e desvio padrão (std) das medidas de centralidade das imagens do conjunto de dados. 137
Tabela A.3 – Medidas de dispersão das intensidades dos pixels das imagens do conjunto de dados 138
Tabela A.4 – Média (avg) e desvio padrão (std) das medidas de dispersão das imagens do conjunto de dados. 139
Tabela A.5 – Medidas dos quartis das intensidades dos pixels das imagens do conjunto de dados 140
Tabela A.6 – Média (avg) e desvio padrão (std) das medidas dos quartis das imagens do conjunto de dados. 141
Tabela A.7 – Média (avg) e desvio padrão (std) das medidas de distância entre as imagens do conjunto de dados. 143
Tabela A.8 – Medidas das frequências espaciais (ciclos/pixel) das imagens do conjunto de dados 146
Tabela A.9 – Média (avg) e desvio padrão (std) das medidas das frequências espaciais (ciclos/pixel) das imagens do conjunto de dados. 148
Tabela A.10–Acurácia média (desvio padrão) em (%) da análise de classificação computada com os algoritmos LDA, SVM e 1-NN para os casos binário e multiclasse para os conjuntos de atributos de textura extraídos das janelas com dimensões 100x100 pixels. 155
Tabela A.11–Acurácia média (desvio padrão) em (%) da análise de classificação computada com os algoritmos LDA, SVM e 1-NN para os casos binário e multiclasse para os conjuntos de atributos de textura extraídos das janelas com dimensões 200x200 pixels. 156
Tabela E.1 – Análise estatística descritiva dos atributos das 344 imagens da classe KL=0.167
Tabela E.2 – Análise estatística descritiva dos atributos das 344 imagens da classe KL=2.167
Tabela E.3 – Experimentos para ajuste dos parâmetros *tmin e tmax* 168
Tabela E.4 – Experimentos para ajuste do parâmetro *tinc* 168
Tabela E.5 – Seleção dos Atributos. 169

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
AHP	Adaptative Hybrid Pattern
AM	Aprendizado de Máquina
ANN	Artificial Neural Network
AUC	Area Under the Curve
CNN	Convolutional Neural Network
CNRNN	Complex Network and Randomized Neural Network
CNTD	Complex Network Texture Descriptor
DT	Decision Tree
EIE	Espectroscopia de Impedância Elétrica
ERAS	Espectroscopia Raman com Amplificação de Superfície
FFT	Fast Fourier Transform
GLDM	Grey Level Difference Matrix
HAC	Hierarchical Agglomerative Clustering
ICMC	Instituto de Ciências Matemáticas e de Computação
IFSC	Instituto de Física de São Carlos
KCV	K-Fold Cross-Validation
KM	K-Means
KNN	K-Nearest Neighbors
LBP	Local Binary Patters
LCFNN	Local Complex Features and Neural Network
LDA	Linear Discriminant Analysis
LOOCV	Leave-One-Out Cross-Validation
LR	Logistic Regression

MeSH	Medical Subject Headings
MEV	Microscopia Eletrônica de Varredura
MH	Mapa Hiperespectral
MI	Multual Information
NKCV	Nested K-Fold Cross-Validation
PCA	Principal Component Analysis
PICO	Patient/Population, Intervention, Comparator, Outcome
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SC	Spectral Clustering
SERS	Surface-enhanced Raman Spectroscopy
SVM	Support Vector Machine
USP	Universidade de São Paulo
VPD	Voltametria de Pulso Diferencial
VPN	Valor Preditivo Negativo
VPP	Valor Preditivo Positivo

LISTA DE SÍMBOLOS

u.a.	unidade arbitrária

SUMÁRIO

1 INTRODUÇÃO 29
1.1 Contexto 29
1.2 Objetivos 31
1.3 Organização da tese 31

2 APRENDIZADO DE MÁQUINA 33
2.1 Introdução 33
2.2 Dados, atributos e objetos 33
2.3 Tipos e processos de análise 34
2.4 Algoritmos 37
2.5 Pipeline 42
2.6 Ferramentas computacionais 51

3 REVISÃO DE LITERATURA 53
3.1 Metodologia 53
3.2 Resultados 54
3.3 Discussão 64
3.4 Conclusão 66

4 GENOSSENSOR E MEV PARA DIAGNÓSTICO DE CÂNCER DE PRÓSTATA 67
4.1 Introdução 67
4.2 Conjunto de dados 67
4.3 Resultados 71
4.4 Conclusão 74

5 LÍNGUA ELETRÔNICA E ESPECTROSCOPIA DE IMPEDÂNCIA PARA DIAGNÓSTICO DE CÂNCER DE BOCA 79
5.1 Introdução 79
5.2 Conjunto de dados 79
5.3 Resultados 83
5.4 Conclusão 88

6 IMUNOSSENSOR E VOLTAMETRIA PARA DIAGNÓSTICO DE CÂNCER 91
6.1 Introdução 91
6.2 Conjunto de dados 92

6.3 Resultados 94
6.4 Conclusão 97

7 IMUNOSSENSOR E SERS PARA DIAGNÓSTICO DE COVID-19 . 99
7.1 Introdução 99
7.2 Conjunto de dados 99
7.3 Resultados 101
7.4 Conclusão 106

8 CONSIDERAÇÕES FINAIS 109
8.1 Síntese das contribuições 110
8.2 Perspectivas 113

REFERÊNCIAS 117

APÊNDICES 129

APÊNDICE A – GENOSSENSOR E MEV PARA DIAGNÓSTICO DE CÂNCER DE PRÓSTATA: INFORMAÇÕES COMPLEMENTARES 131
A.1 Análise exploratória 131
A.2 Análise comparativa entre as janelas e as imagens originais 135

APÊNDICE B – LÍNGUA ELETRÔNICA E ESPECTROSCOPIA DE IMPEDÂNCIA PARA DIAGNÓSTICO DE CÂNCER DE BOCA: INFORMAÇÕES COMPLEMENTARES 157

APÊNDICE C – IMUNOSSENSOR E VOLTAMETRIA PARA DIAGNÓSTICO DE CÂNCER: INFORMAÇÕES COMPLEMENTARES 159

APÊNDICE D – IMUNOSSENSOR E SERS PARA DIAGNÓSTICO DE COVID-19: INFORMAÇÕES COMPLEMENTARES 161

APÊNDICE E – AM E REDES COMPLEXAS APLICADAS AO DIAGNÓSTICO DE OSTEOARTRITE 163
E.1 Introdução 163
E.2 Conjunto de dados 164
E.3 Metodologia de análise 164

E.4 Resultados 166

APÊNDICE F – GUIA PARA TREINAMENTO DE ALUNOS DE IC SOBRE AM APLICADO A BIOSSENSORES . 171

F.1 Aprendizado de máquina 171

F.2 Biossensores 172

F.3 Metodologia e escrita científica 174

1 INTRODUÇÃO

1.1 Contexto

O diagnóstico precoce de muitas doenças, inclusive o câncer (1), é essencial para tratamentos mais adequados e maiores chances de cura.(2, 3) Tal diagnóstico requer o desenvolvimento de métodos baratos e de baixo custo para detectar biomarcadores das doenças, além de abordagens eficientes para processar os dados obtidos na detecção. A literatura recente é rica em métodos baratos para esse tipo de detecção, por exemplo empregando biossensores (4) para detectar biomarcadores em fluidos biológicos (*e.g* saliva, urina, sangue, secreção nasal). Biomarcadores são substâncias que podem indicar a presença de processos biológicos (normais ou anormais) ou a resposta farmacológica a intervenções terapêuticas.(5) A detecção de biomarcadores pode ser realizada em baixas concentrações e com seletividade.(6) Essas e outras características são obtidas através dos (bio)materiais empregados na organização estrutural e funcional dos dispositivos, o que também determina o tipo de resposta (*e.g.* elétrica, óptica, térmica).(7) Abordagens que integram diferentes tipos de dados têm se incrementado à medida que novas técnicas são usadas em diagnóstico. Por exemplo, o diagnóstico de muitos tipos de câncer é feito com técnicas variadas, desde análises clínicas até imagens, além da análise do histórico de pacientes que em princípio está na forma de texto. (8) Essa diversidade de possíveis dados inspirou propostas de sistemas computacionais para diagnóstico (9), principalmente com o uso de diferentes técnicas de inteligência artificial. (10)

Esta tese seguiu a perspectiva de que no futuro diagnósticos e monitoramentos serão realizados com apoio de sistemas computacionais inteligentes. Esses sistemas podem ser desenvolvidos segundo diferentes estratégias, dependendo do grau de automação pretendido, e possível. A Figura 1.1 mostra uma proposta feita há alguns anos quando se percebeu a necessidade de empregar diferentes técnicas de análises de dados com sensores e biossensores fabricados com várias abordagens de nanotecnologia. É interessante discutir o fluxograma da Figura 1.1 e seus módulos. Fica implícito que um sistema de diagnóstico médico mais sofisticado precisa incluir dados de diferentes naturezas. Na figura são ilustrados dados de textos, que podem vir de prontuários eletrônicos ou de buscas da literatura, dados de análises clínicas, de imagens. Como a previsão é de que tais conjuntos de dados serão volumosos, e com características diversas, será necessário realizar pré-processamentos. São mencionados filtros, procedimentos de limpeza e regularização de dados. Além disso, como o fluxograma foi concebido para um sistema semi-automático, de apoio ao profissional de saúde, prevê-se também mecanismos de visualização interativa. Ou seja, o usuário - no caso o profissional de saúde - poderia visualizar os dados com diferentes técnicas, inclusive para fazer seleção de atributos e dados. O módulo final para gerar o diagnóstico corresponde ao de mineração dos dados, que pode incluir diferentes técnicas de estatística, visualização de dados e de inteligência

artificial.

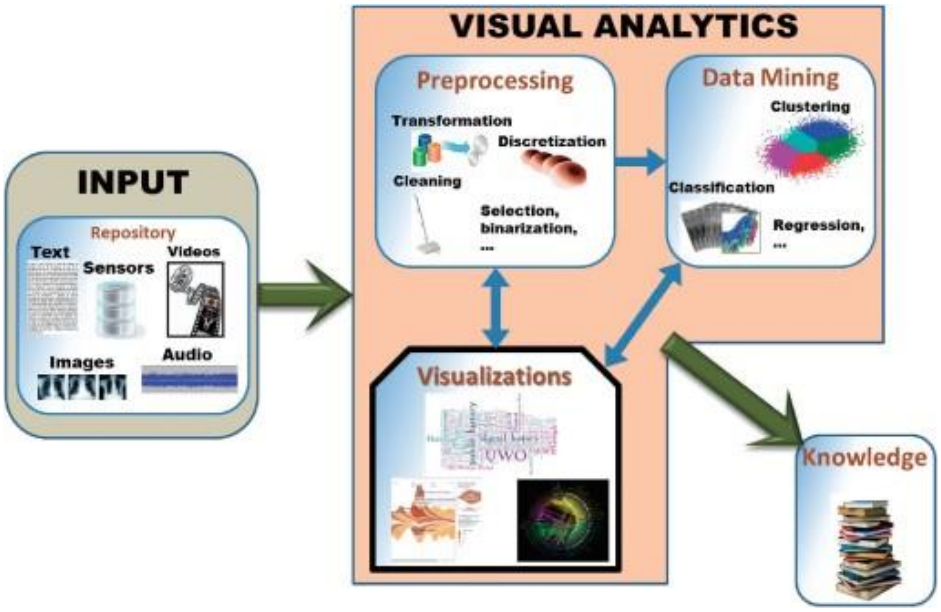


Figura 1.1 – Proposta de sistema computacional inteligente de diagnóstico baseado em diferentes fontes de dados.

Fonte: OLIVEIRA *et al.* (9)

O interesse em sistemas inteligentes advém da complexidade de alguns diagnósticos (11), não só porque dados de natureza diferentes precisam ser utilizados mas também porque algumas amostras – como as dos fluidos biológicos – têm muitas substâncias que podem interferir nas medidas com os biossensores.(12) De fato, é um grande desafio desenvolver um sistema de diagnóstico que considere dados de análises clínicas, imagens, e informações de saúde e socioeconômicas dos pacientes. (8, 13) Destaque-se também a busca por soluções computacionais de maior validade, confiabilidade, capacidade de generalização, e que exibam alto desempenho na classificação de amostras. Das metodologias usadas nesta tese, a mais essencial foi a do Aprendizado de Máquina (14), que pode satisfazer os requisitos mencionados acima.(15) Com AM é possível desenvolver sistemas que possam aprender relações e padrões em grandes volumes de dados, mesmo que diversos e complexos. Em particular, sabe-se que os algoritmos de Aprendizado de Máquina são adequados para tarefas de classificação, sendo portanto ideais para diagnóstico, pois a tarefa de diagnosticar é essencialmente de classificação.

O *Pipeline* de Aprendizado de Máquina compreende o conjunto organizado de tarefas a serem efetuadas sobre os dados para a treinamento/teste de modelos de AM.(16) A depender da aplicação, um determinado pipeline pode ser reutilizado ou pode ser necessário o desenvolvimento de um novo com a inclusão, exclusão ou modificação de uma ou várias tarefas, bem como da organização destas. Na seção 2.5 será apresentado um pipeline base para sistemas inteligentes de diagnóstico, com ênfase nas diferentes tarefas, para as quais diferentes técnicas foram implementadas nesta tese. Ressalte-se que esta tese foi desenvolvida para consolidar uma

colaboração do Grupo de Polímeros do IFSC/USP com vários outros grupo de pesquisa, numa iniciativa para integrar métodos de computação às análises de dados de sensores e biossensores (17–21).

1.2 Objetivos

O principal objetivo desta tese é:

Aplicar métodos de Aprendizado de Máquina em dados de biossensores visando ao diagnóstico de alguns tipos de câncer e COVID-19.

Para o cumprimento deste objetivo, é necessário:

- Conhecer os biossensores e os diferentes tipos de dados a serem processados;
- Conhecer conceitos, métodos, algoritmos e ferramentas de AM a serem empregados;
- Coletar e analisar trabalhos recentes de aplicação de AM em dados de biossensores;
- Conhecer os conjuntos de dados disponibilizados para processamento;
- Desenvolver modelos de AM a partir dos conjuntos de dados disponibilizados;
- Estimar e avaliar o desempenho dos modelos de AM para diagnóstico de câncer e COVID-19.

1.3 Organização da tese

A partir do interesse em sistemas inteligentes de diagnóstico, da disponibilidade de dados de biossensores para detecção de biomarcadores específicos de algumas doenças (câncer e COVID-19), e de um pipeline base (16) a ser apresentado na seção 2.5, este trabalho de doutorado foi estruturado possibilitando o desenvolvimento de 4 aplicações de AM. O esquema apresentado na Figura 1.2 ilustra a composição e organização geral das aplicações.

Esta tese está estruturada em oito capítulos. No **Capítulo 1** são apresentados o contexto, os desafios e a proposição do uso de dados de biossensores em conjunto com técnicas de AM para a realização de diagnóstico de câncer e COVID-19. São apresentados também a motivação e os objetivos do trabalho. Os conceitos, métodos e ferramentas computacionais de AM utilizados no trabalho são apresentados no **Capítulo 2**. Além de conceitos básicos como dado, atributos, tipos de análises, serão apresentados o processo de geração de aprendizado, os algoritmos de aprendizado supervisionado e não-supervisionado utilizados, e o pipeline base a partir do qual foram desenvolvidos os pipelines específicos de cada aplicação. Uma revisão

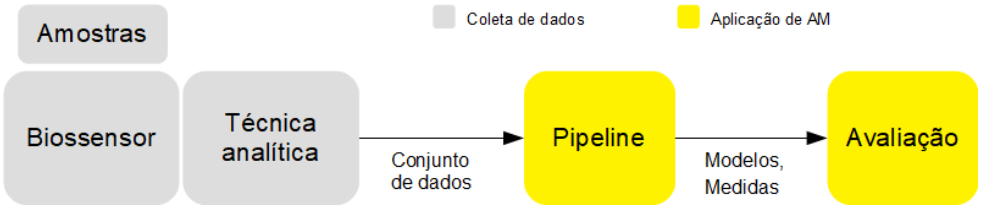


Figura 1.2 – Esquema geral da composição e organização das aplicações de AM em dados de biossensores para diagnóstico de câncer e COVID-19.

Fonte – Elaborada pelo autor.

de literatura do tipo sistemática integrativa foi realizada e está contida no **Capítulo 3**. São apresentados a metodologia empregada segundo o protocolo PRISMA (22), os resultados e uma discussão sobre os artigos selecionados nos últimos 5 anos (até 2021) extraídos das bases de dados Pubmed*, Web of Science (WoS)† e Scopus‡, e são relativos a este trabalho de doutorado.

As aplicações de AM em dados de biossensores para o diagnóstico de alguns tipos de câncer e de COVID-19 são apresentadas nos **capítulos 4 ao 7**. Cada aplicação está identificada através do tipo de biossensor, do conjunto de dados e da doença a ser diagnosticada, e estão ordenadas considerando-se o estágio de desenvolvimento, iniciando-se com as que já foram publicadas em periódicos científicos (4 e 5), seguindo com as aplicações cujos manuscritos estão em processo de escrita (6 e 7) para submissão em breve. O **Capítulo 8** traz o fechamento desta tese com uma síntese das principais contribuições e perspectivas para continuidade do trabalho.

Os **Apêndices** trazem informações complementares sobre cada aplicação (**A** a **D**) e os trabalhos adicionais realizados durante o doutorado e pertinentes à aplicação de AM, mas que não foram incluídos no corpo principal desta tese. O **Apêndice E** apresenta o aplicação de AM sobre imagens radiográficas do joelho para diagnóstico de Osteoartrite. No **Apêndice F** é apresentado um guia de estudos/preparação sobre aplicação de AM na análise de dados de biossensores para alunos de Iniciação Científica (IC).

* https://pubmed.ncbi.nlm.nih.gov/
† https://www-webofscience.ez67.periodicos.capes.gov.br/wos/woscc/basic-search
‡ https://www-scopus.ez67.periodicos.capes.gov.br/search/form.uri?display=basic#basic

2 APRENDIZADO DE MÁQUINA

Este capítulo apresenta os conceitos básicos, os métodos implementados e as ferramentas utilizadas para desenvolver as aplicações de aprendizado de máquina (AM) em dados de biossensores. Há farta literatura sobre esses assuntos, com abordagens mais detalhadas e abrangentes, sendo algumas delas citadas ao longo do texto e dos demais capítulos. Pretende-se aqui fornecer apenas uma breve introdução que viabilize a compreensão do trabalho.

2.1 Introdução

Na literatura, encontram-se diferentes definições sobre Aprendizado de Máquina (AM). Para MITCHEL (23), o AM envolve a questão de como construir programas de computador que melhoram automaticamente com a experiência. Segundo ALPAYDIN (24), o AM é programar computadores para otimizar um critério de desempenho usando dados de exemplo ou experiências anteriores. E conforme FLATCH (25), o AM é o estudo sistemático de algoritmos e sistemas que melhoram seu conhecimento ou desempenho com a experiência. A partir destas referências, pode-se inferir que o AM envolve dados, algoritmos adaptáveis aos dados, e realização de tarefa com elevado desempenho. Como área do conhecimento, o AM é uma subárea da Inteligência Artificial, hoje a mais empregada em ciência e tecnologia, comparativamente a outros tópicos de inteligência artificial. Grosso modo, os algoritmos de AM podem ser divididos em não-supervisionados e supervisionados. Como AM já vem sendo discutido em muitos outros trabalhos, aqui apenas serão sucintamente apresentadas definições e descrições dos métodos utilizados. Serão fornecidas informações de contexto para explicar os diferentes elementos, etapas e tarefas do pipeline base de AM usado nas aplicações de biossensores. Na última seção, encontram-se informações sobre as ferramentas computacionais utilizadas nas aplicações. Para aprofundamento teórico sobre AM, recomenda-se o estudo das obras *Pattern Recognition and Machine Learning* (26) e *The Elements of Statistical Learning*.(14) Para implementação dos métodos de AM em Python, as obras *Introduction to Machine Learning with Python* (27) e *Hands-on Machine Learning with Scikit-Learn and TensorFlow* (16) são referências iniciais úteis.

2.2 Dados, atributos e objetos

Dado é um dos elementos fundamentais para geração do aprendizado de máquina e pode ser definido como o registro do atributo de um objeto, ação ou evento. (28, 29) Um exemplo é a medida de temperatura corporal (dado) de uma pessoa (objeto). O **Atributo** de um dado, também conhecido como variável ou dimensão, pode ser definido como uma propriedade ou característica do dado. (30) A medida (dado) da altura (atributo) de uma pessoa (objeto) é um exemplo. O tipo de um atributo é determinado pelo conjunto de valores possíveis,

podendo ser qualitativos e quantitativos. (30) Os qualitativos expressam uma caraterística ou qualidade do objeto. São eles: nominal, quando não há ordenamento entre os valores (e.g. sexo, estado civil), e ordinal, quando há ordenamento (e.g. grau de instrução, classe social). Os quantitativos expressam quantidades ou medidas relacionadas ao objeto. Podem ser: discreto, quando os valores são enumeráveis ou pertencentes ao conjunto dos números inteiros (e.g. número de filhos), e contínuo, quando os valores pertencem ao conjunto dos números reais (e.g. peso e altura de uma pessoa).

O ciclo de vida de um dado compreende as fases de produção, armazenamento, transformação, análise e descarte. (31) Os dados podem ser produzidos por pessoas, dispositivos, equipamentos ou máquinas. Podem ser registrados e armazenados em formato físico (impresso) ou eletrônico (analógico/digital). Podem estar estruturados (organizados) ou não. Se não estiverem, ou se houver características indesejáveis, os dados poderão ser transformados para serem analisados. Após a análise, os dados podem ser descartados.

As etapas e tarefas de análise (30, 32, 33) são realizadas sobre um *conjuntos de dados* composto, por exemplo, de dados clínicos, laboratoriais, socio-econômicos de pacientes. Um conjunto de dados é compostos de objetos, também referidos como amostras, exemplos, instâncias ou pontos de dados. Em saúde, um objeto pode representar um paciente. Associados aos objetos pode haver um ou vários atributos, formando um *vetor de atributos*. Um conjunto de dados estruturado é organizado em formato de tabela conforme apresentado (Figura 2.1). Nas linhas, encontram-se os vetores de atributos de cada objeto, e nas colunas os atributos. Se o conjunto de dados for preparado para aprendizado supervisionado, então o último atributo do conjunto será a *classe* a que o objeto pertence.

Organizado dessa forma, um conjunto de dados está pronto para ser utilizado em um pipeline de AM para realização de análises em uma aplicação. Os conjuntos de dados utilizados neste trabalho são originados de biossensores/línguas eletrônicas: imagens de microscopias eletrônicas de varredura (MEV), espectros de impedância, voltamogramas e mapas hyperspectrais de espectros SERS. Suas diferenças são marcantes e complexas, sendo provenientes de diferentes características dos dispositivos (ópticas, elétricas, eletrônicas), coletados com diferentes equipamentos, representados com tipos/unidades de medida diferentes. Houve inclusive uma aplicação (Capítulo 5) que mesclou estes tipos com dados clínicos de pacientes. Isso compõe o que é chamado de domínio do problema. No caso deste trabalho, o domínio compreende a *análise de sinais com técnicas de aprendizado de máquina*. Para lidar com os vários problemas/análises e seus domínios, há que se ter um bom nível de compreensão sobre os conceitos listados na Figura 2.2.

2.3 Tipos e processos de análise

Neste trabalho foram empregados os tipos de análise dos dados (30, 32, 33) descritos a seguir:

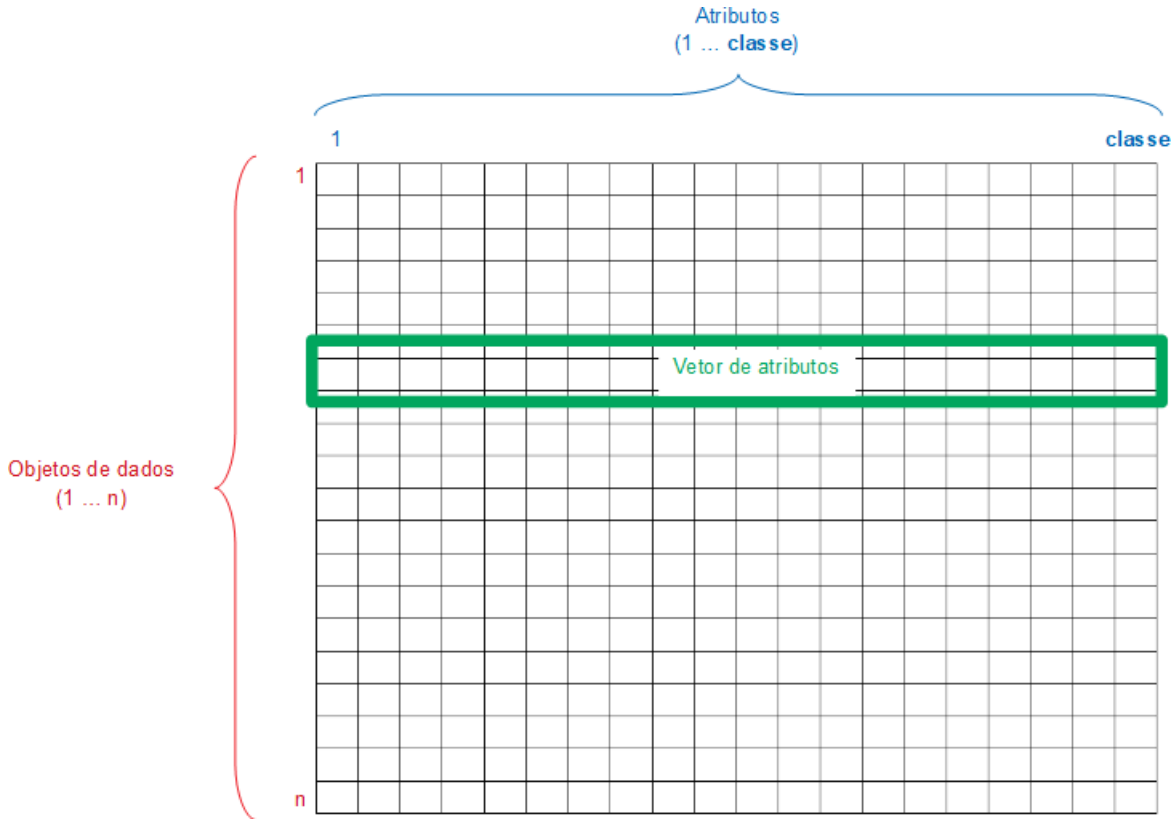


Figura 2.1 – Organização de um conjunto de dados estruturado.

Fonte – Elaborada pelo autor.

- **Análises Descritivas:** consistem em conhecer as características e encontrar/determinar padrões (correlações, tendências, agrupamentos, trajetórias e anomalias) nos dados. São exemplos as análises exploratórias (ou exploração), de associação e de agrupamento (clustering). A análise exploratória é usada para conhecer as características dos dados com técnicas estatísticas e de visualização. A análise de associação é usada para descobrir padrões que descrevam atributos fortemente associados. Já a análise de agrupamento é usada para encontrar grupos de dados mais relacionados/similares entre si que em relação a outros grupos.

- **Análises Preditivas:** consistem em prever o valor de um atributo específico, também

Figura 2.2 – Mapa conceitual do trabalho baseado em dados de biossensores.

Fonte – Elaborada pelo autor.

conhecido como alvo ou variável dependente, com base nos valores de outros atributos, as variáveis independentes. Como exemplos têm-se as análises de classificação e de regressão. A principal diferença entre as duas é o tipo de atributo a ser predito (alvo). Na classificação, o atributo alvo tem variação discreta, enquanto que na regressão esse atributo tem variação contínua.

Uma forma de aumentar a chance de sucesso e possibilitar a reprodução das análises de dados é estabelecendo e seguindo um processo sistemático. Existem diversas definições de processos de análise de dados na literatura, tais como o SEMMA (Sample, Explore, Modify, Model, and Assess) (34) e CRISP-DM (Cross Industry Standard Process for Data Mining).(35, 36) O processo CRISP-DM foi o adotado neste trabalho para construção dos pipelines de AM. Ele é composto das seguintes etapas e tarefas:

1. **Entendimento do negócio/pesquisa:** definição do objetivo da análise de dados e da metodologia a ser executada.

2. **Compreensão dos dados:** coleta e conhecimento dos dados sobre sua estrutura, atributos e contexto;
3. **Preparação dos dados:** execução de tarefas como limpeza, filtragem, estruturação, redução, integração dos dados. Pode incluir a construção e seleção dos atributos, e seleção dos dados.
4. **Modelagem dos dados:** definição e construção do modelo.
5. **Validação do modelo:** avaliação dos resultados gerados pelo modelo e verificação se os parâmetros de qualidade foram atingidos.
6. **Utilização do modelo:** modelo validado é utilizado e monitorado.

Os conjuntos de dados utilizados neste trabalho possibilitaram a realização de análises binárias e multiclasse (30, 32, 33) com algoritmos de aprendizado supervisionado e não-supervisionado. Na análise binária busca-se agrupar/classificar os objetos de dados em duas classes/categorias, por exemplo, positivo/negativo para o diagnóstico de uma doença, presente/ausente para presença de um biomarcador, etc. Na análise multiclasse, há a possibilidade de agrupamento/classificação em múltiplas classes, por exemplo, as várias concentrações de biomarcador em uma amostra analisada, ou a presença/ausência do biomarcador para a doença e amostras contendo um biomarcador interferente.

2.4 Algoritmos

As análises definidas anteriormente foram realizadas com algoritmos de aprendizado supervisionado e não-supervisionado bastante conhecidos.(37–40) Algumas aplicações empregaram algoritmos como LR, LDA, 1-NN, GNB e SVM-kernel linear (14, 41) com vistas a verificar, principalmente, a capacidade dos dados em evidenciar padrões separáveis/classificáveis. Em outras, aplicaram-se também algoritmos capazes de gerar modelos mais compatíveis com dados mais complexos. Serão apresentados os algoritmos supervisionados (24, 30), juntamente ao seu processo de aprendizagem, e os não-supervisionados (24, 30, 42) que foram utilizados nas análises de agrupamento.

Algoritmos Supervisionados

A *classificação* dos dados é uma tarefa que, de maneira geral, envolve duas etapas: aprendizagem e predição. Durante a *aprendizagem*, um algoritmo de AM gera um modelo analisando ou "aprendendo com"um conjunto de dados rotulados, isto é, cada objeto de dados possui o rótulo da classe a que pertence. Com o modelo gerado, denominado classificador, faz-se a *Predição* das classes de um novo conjunto de dados, os quais não foram utilizados durante a aprendizagem. Um classificador aprende de forma *supervisionada*, isto é, processando alguns exemplos de dados rotulados (entrada), cujas classes são conhecidas, e aprendendo um

modelo (função ou probabilidade) que mapeia da entrada x_i para a saída y_i. A entrada x_i é composta por m atributos. Formalmente, o processo de aprendizagem (23, 24, 43) se dá da seguinte maneira. Seja um conjunto de dados de treinamento contendo n exemplos de dados rotulados $(x_1, y_1)...(x_n, y_n)$, onde cada saída y_i foi gerada a partir de uma função desconhecida $y = f(x)$. E seja h uma *hipótese* para essa função desconhecida. A aprendizagem é alcançada através da busca, no espaço de possíveis *hipóteses* H, de uma *hipótese* h que tenha um bom desempenho mesmo em novos objetos de dados. Esse requisito é denominado de *generalização*. Para avaliar a capacidade de generalização de h, utiliza-se um conjunto de dados de teste, contendo dados que não pertençam ao conjunto de treinamento. A hipótese *h é consistente*, se ela se ajusta o mais próximo possível ao conjunto de dados de treinamento.

A busca por uma função h de maior complexidade pode proporcionar um melhor desempenho no conjunto de treinamento, mas também pode levar a uma diminuição da sua capacidade de generalização, devido ao elevado grau de ajuste ao conjunto de dados de treinamento. Esse é o chamado sobreajuste (*overfitting*). (44) Segundo o princípio de *Ockham's razor*, a *hipótese h consistente mais simples* é a que deve ser escolhida no espaço H. Além disso, h deve ser tal que minimize o erro de generalização. Portanto, trata-se de uma questão de otimização a ser solucionada pelos algoritmos de AM através do conjunto de treinamento. Uma forma de se fazer é através da *Minimização do Risco Empírico*. Definindo-se uma função de perda $L(h)$ para um determinado conjunto de treinamento, com $h(x_i) \neq y_i$, busca-se determinar uma função h que minimize a função de perda. Uma forma prática para a determinação do grau de complexidade de h, e portanto, do modelo a ser aprendido, é feita através da escolha do algoritmo de AM mais adequado, do ajuste dos seus parâmetros de configuração e depende do conjunto de treinamento. Para auxiliar nessa tarefa, uma forma prática é utilizar a curva de aprendizagem (Figura 2.3) para visualizar o comportamento do erro de h em ambos os conjuntos de dados. O nível ótimo de complexidade de h é atingido quando o erro de generalização no conjunto de teste é mínimo, indicado pela linha vertical cinza da Figura 2.3. Essa linha também evidencia as regiões em que estão ocorrendo o subajuste (*underfitting*) e o sobreajuste (*overfitting*) do modelo aprendido sobre o conjunto de dados de treinamento.

A seguir são apresentados sucintamente os algoritmos supervisionados utilizados neste trabalho: *Logistic Regression (LR)*, *Linear Discriminant Analysis (LDA)*, *Support Vector Machine (SVM)*, *Gaussian Naive Bayes (GNB)*, *K-Nearest Neighbors (KNN)* e *Decision Tree (DT)*. São algoritmos muito utilizados em aplicações de AM e são baseados em diferentes hipóteses e hiperparâmetros de configuração para modelagem dos classificadores. O estabelecimento de um conjunto de algoritmos para as aplicações de AM permite comparar e melhorar os classificadores obtidos sobre os mesmos/diferentes conjuntos de dados. Essas tarefas pertencem à etapa de modelagem do pipeline base de AM (seção 2.5).

O algoritmo *Logistic Regression (LR)* é usado quando a variável dependente é de

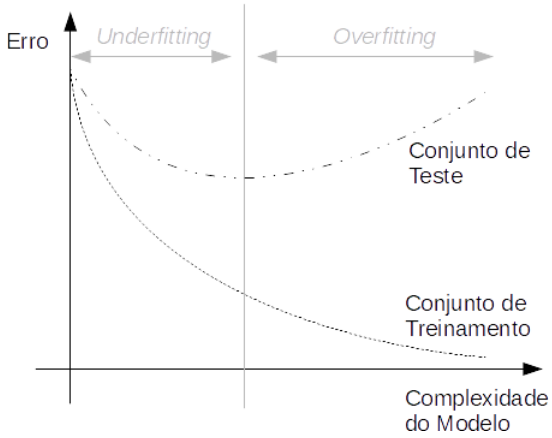


Figura 2.3 – Exemplo de Curva de Aprendizagem. A linha vertical cinza indica o nível ótimo de complexidade do modelo aprendido.

Fonte: Adaptada de LAROSE.(45)

natureza binária, ou seja, quando pode ser um dos dois valores (categorias) exemplo positivo ou negativo. Ele utiliza uma combinação linear dos pesos dos atributos de entrada (*e.g.* uma entrada x) aplicados à função logit ou sigmóide Equação 2.1, cuja saída está limitada ao intervalo de 0 a 1, podendo ser interpretada como a probabilidade $P(x)$ de um dado pertencer a uma das classes. Isso é definido impondo-se um limiar (*threshold*) mínimo de probabilidade a partir do qual considera-se o dado pertencente à classe.

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{2.1}$$

O algoritmo *Gaussian Naive Bayes (GNB)* usa o conjunto de dados de treinamento para calcular a probabilidade observada de cada classe, com base em todos os atributos dos objetos de dados. A probabilidade relaciona os valores dos atributos a cada classe como uma "função de mapeamento". A classificação de um novo objeto de dados é realizada utilizando-se os valores dos seus atributos $x_1, \ldots, x_n$ para, através das probabilidades determinadas na etapa de treinamento (lado direito da Equação 2.2), determinar a classe C_i com a maior probabilidade (lado esquerdo da Equação 2.2).

$$P(C_i \mid x_1, \ldots, x_n) = \frac{P(C_i) \prod_{j=1}^{n} P(x_j \mid C_i)}{\prod_{j=1}^{n} P(x_j)} \tag{2.2}$$

Baseando-se principalmente no conceito de similaridade entre objetos, o algoritmo *K-Nearest Neighbors (KNN)* classifica um novo objeto de dados com a classe que a maioria dos k vizinhos mais próximos possui. A proximidade, ou a similaridade, é calculada a partir de alguma métrica (*e.g* Euclidiana) para cálculo da distância entre os atributos do objeto de dados a ser classificado e dos demais objetos de dados já rotulados (conjunto de treinamento). A seleção do parâmetro k é realizada avaliando o desempenho do classificador em relação a

diferentes valores de $k = \{1, 2, ..., n\}$ e selecionando o valor que fornece o melhor desempenho segundo as métricas de validação escolhidas. O algoritmo 1-NN (k = 1) foi utilizado neste trabalho.

O algoritmo *Linear Discriminant Analysis (LDA)* usa o conjunto de dados de treinamento para determinar os parâmetros da função discrimante (modelo), a qual divide o domínio dos dados nas regiões de cada classe através de hiperplanos (*e.g.* linhas em duas dimensões, planos em três etc.). A classificação é realizada determinando-se de que lado do hiperplano o objeto de dados ficará através da utilização dos seus atributos na função discriminante.

O *Support Vector Machine (SVM)* é um algoritmo que cria um mapeamento não linear para transformar os dados de treinamento do seu espaço de dimensões originais para um espaço de dimensão superior. Dentro desse novo espaço de maior dimensionalidade, procura-se o hiperplano ótimo de separação linear (ou seja, um "limite de decisão" separando as tuplas de uma classe de outro). Com um mapeamento não linear apropriado para uma dimensão suficientemente alta, os dados de duas classes sempre podem ser separados por um hiperplano. O SVM encontra esse hiperplano usando vetores de suporte (tuplas de treinamento "essenciais") e margens (definidas por os vetores de suporte). A determinação das margens se faz maximizando a distância delas com o hiperplano e entre elas. A classe de um novo dado é determinada em relação a qual lado fora das margens esse dado está localizado. Esse é caso linear desse algoritmo, isto é, SVM-L. Para os casos de separação não linear, o algoritmo é similar mas ao invés de hiperplanos são buscadas hipersuperfícies com funções (ou *kernels*) do tipo polinomial (SVM-P), gaussian radial basis (SVM-R).

O treinamento de uma árvore de decisão gerada pelo algoritmo *Decision Tree (DT)* para a tarefa de classificação é realizado com dados rotulados. Uma árvore de decisão é uma estrutura de árvore semelhante a um fluxograma, onde cada nó interno (nó não folha) representa um teste em um atributo, cada ramo representa um resultado do teste, e cada nó folha (ou nó terminal) contém um rótulo de classe. O nó mais alto em uma árvore é o nó raiz. A construção da árvore é feita com base nos atributos de entrada mais significativos para formar grupos tão distintos quanto possível. A divisão dos dados de treinamento em grupos heterogêneos utiliza critérios como *Gini index*, *Information Gain*, *Chi-square* ou *entropy*. A divisão dos grupos é encerrada quando atinge-se o ponto de máxima otimização baseada em um desses critérios. A classificação de um novo dado é realizada testando-se os valores dos seus atributos em relação aos árvore de decisão. Um caminho é traçado da raiz até um nó folha, que contém a classe de previsão para esse novo dado.

O algoritmo *Random Forest (RF)* combina as previsões por um conjunto de árvores de decisão, onde cada árvore é construída com base nos valores de um conjunto independente de vetores aleatórios gerados com uma distribuição de probabilidade fixa a partir do conjunto de dados de treinamento. Cada árvore de decisão usa um desses vetores aleatórios em seu processo de crescimento e treinamento. Depois de construídas as árvores, as classificações

são combinadas usando um regime de votação por maioria. Esse esquema de combinação de classificadores é conhecido como *bagging*.(16)

Algoritmos Não-supervisionados

Neste trabalho foram utilizados os algoritmos não-supervisionados *K-Means (KM)* (41, 42), *Hierarchical Agglomerative Clustering (HAC)* (41, 42) e *Spectral Clustering (SC)* (46) para as análises de agrupamento.

K-Means (KM) é um algoritmo de agrupamento por partição que atribui os objetos de dados em um número predeterminado de grupos (*clusters*). O número de clusters k é o hyperparâmetro do algoritmo. Os grupos são formados escolhendo aleatoriamente centroides em uma distribuição de dados e designando os objetos aos centroides que estão mais próximos (e.g distância Euclidiana). O número de centróides é igual ao número de clusters k a serem determinados. Os centróides são então recalculados tomando a média das distâncias de todos objetos designados ao cluster desse centroide. O processo é iterativo. Os objetos são redesignados aos centróides aos quais estão agora mais próximos e os centroides são atualizados para cada cluster. A convergência é alcançada quando os pontos de dados não são mais reatribuídos a novos clusters.

No agrupamento utilizando o algoritmo *Hierarchical Agglomerative Clustering (HAC)* cada objeto de dados começa em um cluster próprio. Uma medida de distância (e.g distância Euclidiana) é usada para calcular a distância de entre todos os objetos. Os objetos com a menor distância (modo *single linkage*) são agrupados para formar um cluster próprio. O processo continua até que todos os objetos formem o maior cluster possível. A intuição é que os objetos de dados que são semelhantes são provavelmente separados por uma pequena distância no espaço de atributos. Uma forma de visualizar e analisar os agrupamentos formados é utilizando-se de um dendrograma (Figura 2.4). Nele um valor limite, representado pela linha tracejada vermelha, pode ser estipulado para determinar o número de clusters em que os objetos $Oi(i = 1...7)$ estão agrupados.

O agrupamento com o algoritmo *Spectral Clustering (SC)* é baseado na mudança de espaço dos dados. Os dados são modelados através de um grafo em que os k pontos mais próximos de um certo ponto formam arestas com este. O passo seguinte é determinar os autovalores e autovetores da matriz laplaciana (L) associada que é dada por $L = G - A$, onde A é a matriz de adjacências e G é uma matriz diagonal com os elementos da diagonal igual ao grau do nó correspondente. Assim, pode-se calcular a Laplaciana de um conjunto de dados e utilizar os k primeiros autovetores dessa matriz (com autovalores diferentes de 0) e gerar uma matriz n × k contendo a informação dos grupos, similar à matriz de decomposição obtida por PCA. Dessa forma, os dados encontram-se projetados em um espaço de menor dimensionalidade e basta aplicar a técnica de agrupamento por KM sobre essa representação para obtenção dos grupos.

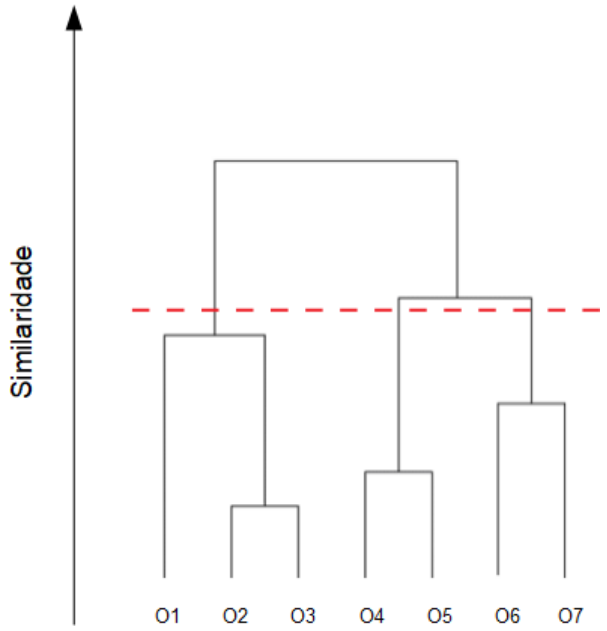


Figura 2.4 – Exemplo de um dendrograma. A linha tracejada vermelha indica o limiar estipulado para determinar o número de clusters em que os objetos $Oi(i = 1...7)$ estão agrupados.

Fonte – Elaborada pelo autor.

2.5 Pipeline

O produto do aprendizado realizado por uma máquina é um modelo gerado sobre um conjunto de dados de treinamento através de um determinado algoritmo de AM.(25) Para geração de aprendizado, é necessária a execução de um conjunto organizado de tarefas sobre os dados de treinamento conhecido como *pipeline*. (16) A Figura 2.5 apresenta as etapas e tarefas de um pipeline base de Aprendizado de Máquina. Ressalte-se que para cada conjunto de dados e análise pretendida, as etapas, tarefas e métodos podem ser diferentes, dando origem a pipelines específicos como é caso das aplicações em biossensores neste trabalho.

Preparação dos dados

Essa é a etapa inicial do pipeline de AM. Ela é composta por três tarefas principais: exploração, pré-processamento e engenharia de atributos. A depender das análises desejadas, todas ou qualquer uma delas pode ser executada. Os pipelines empregados nas aplicações combinaram diferentes tarefas (baseadas em diferentes métodos da literatura) na execução desta etapa. Importante ressaltar que essa etapa é altamente dependente do tipo de dado a ser processado.

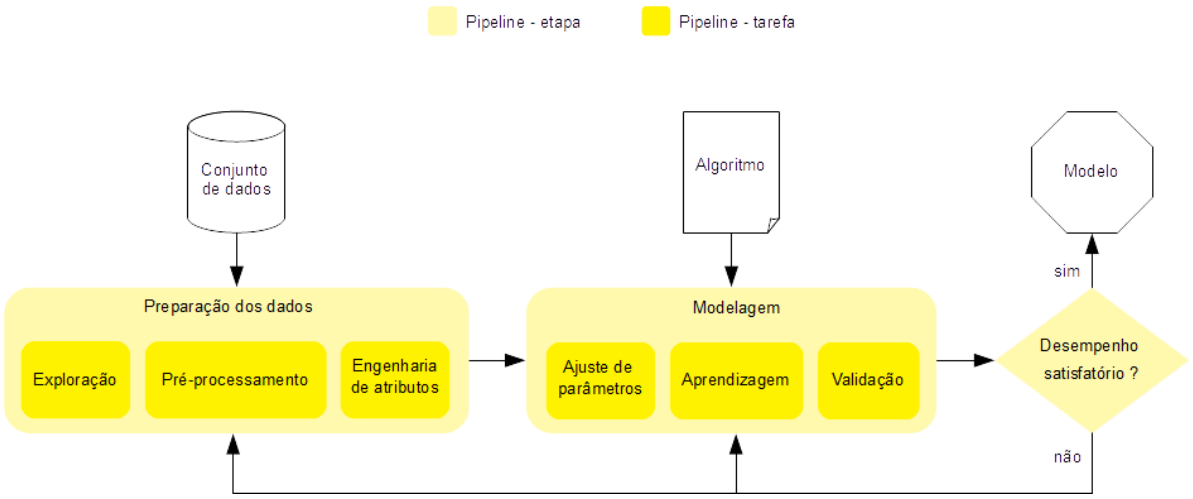


Figura 2.5 – Pipeline base de Aprendizado de Máquina.

Fonte – Elaborada pelo autor.

Exploração

Para uma boa preparação dos dados a serem analisados, é fundamental ter um conhecimento geral dos dados. (30,33) Descrições estatísticas básicas podem ser usadas para identificar propriedades dos dados e evidenciar características (*e.g.* distribuição, anomalias) que devem ser consideradas e/ou tratadas. No Apêndice A até o Apêndice D são apresentadas análises exploratórias para os conjuntos de dados utilizados nas aplicações de AM em biossensores.

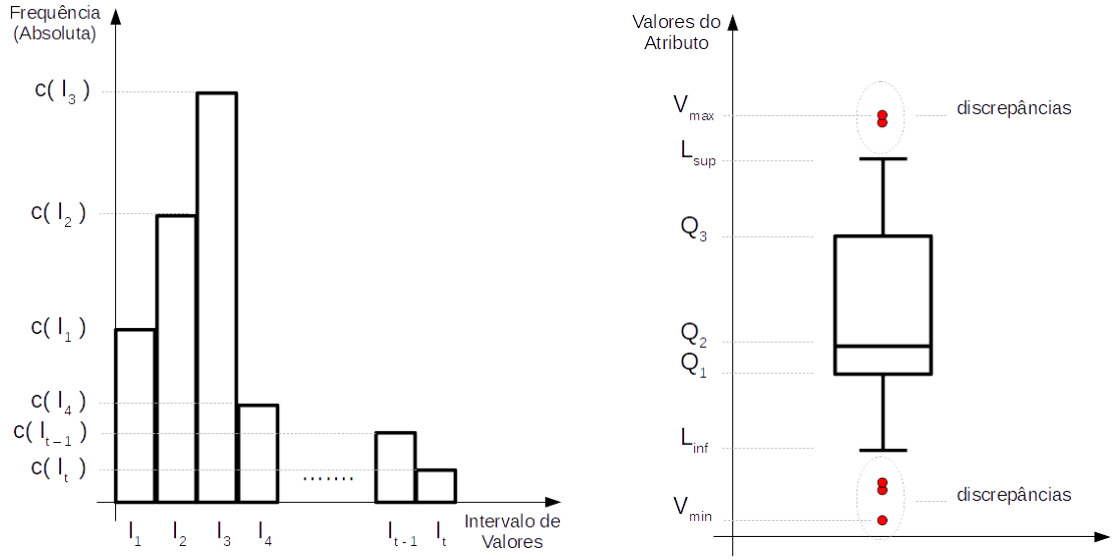
A *frequência* é uma medida da forma como os dados estão distribuídos entre os possíveis valores de um atributo. (47) Essa medida pode ser utilizada para caracterizar atributos qualitativos (nominais ou ordinais) ou quantitativos (discretos). Dado um conjunto de dados com m objetos e seja x um desses atributos, que pode assumir valores ($v_1, \ldots v_k$). Existem dois tipos de frequência: absoluta e relativa.(47) A *frequência absoluta* ($c(v_i)$) é definida pela contagem do número de vezes que um determinado valor v_i aparece dentre os valores possíveis do atributo. Já a *frequência relativa* $f(v_i) = c(v_i)/m$ evidencia a mesma contagem só que em relação à quantidade total de objetos. A *moda* desse atributo é o valor que aparece com a maior frequência (absoluta ou relativa). Uma forma de visualizar as medidas de frequência é através do gráfico denominado *histograma* (2.6a).(47) Para atributos quantitativos contínuos, a contagem é feita a partir da definição de intervalos de valores (I) de mesmo tamanho. O tamanho e o número de intervalos (t) são parâmetros a serem escolhidos. Para atributos quantitativos discretos ou qualitativos, a contagem é feita sobre os k possíveis valores do atributo. O histograma pode ser apresentar a frequência absoluta ou relativa.

Para dados ordenados, pode-se considerar os *percentis* de um conjunto de valores. Em particular, dado um atributo quantitativo (discreto ou contínuo) ordenado x e um número p entre 0 e 100, o p-ésimo percentil x_p é um valor de x tal que p% dos valores do atributo x

são menores que x_p. Por exemplo, o 50º-percentil é o valor x_{50}, para o qual 50% de todos os valores de x são menores que x_{50}. Os principais percentis utilizados são os intervalos quartis: 25% (Q_1), 50% (Q_2, *mediana*) e 75% (Q_3). A partir desses intervalos podem ser definidos a distância interquartil (DIQ), e os limites inferior (L_{inf}) e superior (L_{sup}) (Equação 2.3).(47) A detecção de valores discrepantes ou anômalos pode ser feita para os dados com valores menores que L_{inf} e maiores que L_{sup}.

$$DIQ = Q_3 - Q_1, L_{inf} = Q_1 - 1,5 \times DIQ, L_{sup} = Q_3 + 1,5 \times DIQ \qquad (2.3)$$

Uma forma de sumarizar os dados de um atributo é através das seguintes medidas estatísticas: *valor mínimo* (V_{min}) , Q_1, *mediana* ou (Q_2), Q_3, *valor máximo* (V_{max}).(30) A partir dessas medidas, pode-se construir o gráfico *boxplot* (2.6b) e utilizá-lo para visualizar a distribuição dos dados de um atributo.



(a) (b)

Figura 2.6 – Exemplos de gráficos para visualização da distribuição de dados: (a) histograma e (b) boxplot.

Fonte: Elaborada pelo autor.

A *média* e a *mediana* são medidas de centralidade dos dados associados a atributos quantitativos.(47) Seja um conjunto de dados de m objetos. Tome-se um atributo x desses objetos, com seus valores (v_1, ...v_m) ordenados de forma crescente. Assim, $v_1 = min(x)$ e $v_m = max(x)$. A *média* ($\bar{x}$) (Equação 2.4) e a *mediana* (Equação 2.5) desse atributo são definidas da seguinte forma:

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} v_i \qquad (2.4)$$

$$mediana(x) = \begin{cases} v_{(\frac{m+1}{2})}, & m\ impar \\ \frac{v_{(\frac{m}{2})} + v_{(\frac{m}{2}+1)}}{2}, & m\ par \end{cases} \quad (2.5)$$

A média pode ser considerada como uma boa medida de centralidade dos dados para um determinado atributo quanto mais simétrica for a distribuição dos dados em relação a essa medida. Se a distribuição dos dados é distorcida, a mediana é uma melhor medida de centralidade. Além disso, a média é sensível à presença de valores discrepantes ou anômalos. Neste caso, também a mediana é uma medida mais representativa da centralidade dos dados do atributo. (30, 47)

O *intervalo de variação IV* (x), a *variância* (σ^2) e o *desvio padrão* (σ) são medidas de dispersão dos dados associados a atributos quantitativos (discretos ou contínuo), evidenciando se estão amplamente distribuídos ou concentrados em torno de um ponto qualquer, por exemplo, a média.(47) Seja um atributo x de um conjunto de dados de m objetos, com seus valores ($v_1, \ldots v_m$). O *intervalo de variação* é definido como $IV(x) = max(x) - min(x)$. A *variância* (σ^2) (Equação 2.6) pode ser definida como:

$$\hat{\sigma}^2(x) = \frac{1}{m-1} \sum_{i=1}^{m} (v_i - \bar{x})^2 \quad (2.6)$$

E o *desvio padrão* (σ) é dado pela raiz quadrada da variância.(47) O desvio padrão fornece uma medida de dispersão na mesma unidade dos dados.

Pré-processamento

Os dados a serem analisados podem conter características indesejáveis, por exemplo, discrepâncias, ausências, inconsistências ou duplicação.(45) A presença dessas características diminui a qualidade dos dados. Há também casos em que pode ser necessário ou desejável a transformação dos dados, geração de novos atributos e seleção dos dados e atributos existentes. Essas questões podem interferir na qualidade das análises, seja piorando, seja melhorando os resultados. Por isso, é importante conhecer as técnicas básicas de preparação dos dados. Há técnicas para discretização, binarização, transformação, integração, agregação, amostragem e redução da dimensionalidade dos dados. (30) Algumas delas foram utilizadas em pipelines específicos e serão apresentadas nos capítulos das aplicações de AM. Aqui serão apresentadas apenas as técnicas básicas que foram empregadas.

A *integração de dados* consiste na mesclagem de dados produzidos por várias fontes para formarem o conjunto de dados. Por exemplo, os dados de um determinado paciente (objeto) podem ser produzidos por médicos, via prontuário eletrônico, por exames laboratoriais, e por exames de imagem. Esses dados podem ter sido produzidos com organização e estrutura

diferentes. Assim, a integração cuidadosa pode ajudar a evitar ou reduzir problemas na identificação, redundâncias e inconsistências dos dados de um ou vários objetos do conjunto de dados resultante.

A *limpeza dos dados* consiste em tratar os dados que contenham características indesejáveis tais como repetições e inconsistências, ou que estejam faltando. Essas características indesejáveis podem ocorrer em um ou vários atributos de um ou vários objetos. Sobre os dados repetidos, se for detectada a repetição de todos os atributos de um mesmo objeto, caberá ao analista decidir sobre a exclusão ou não desses valores. Em geral, esses são excluídos do conjunto de dados.

As *inconsistências dos dados* podem ser verificadas em um atributo pela presença de valores muito diferentes dos demais. A inconsistência pode ser, por exemplo, em função dos dados estarem em formatos (ou unidades) diferentes, quando será necessária a uniformização do formato. Outro exemplo é a presença de valores discrepantes ou anômalos, e caberá ao analista decidir sobre a exclusão ou não, a substituição ou não, desses valores.

Em relação ao *dado faltante*, há algumas formas de tratá-lo.(45) Pode-se remover todos os dados do objeto, caso o dado faltante seja a classe a que esse objeto pertence (análise de classificação). Pode-se também preenchê-lo com uma das seguintes possibilidades: valor constante, a média ou a moda do atributo, valor gerado aleatoriamente a partir da distribuição de probabilidades observada no atributo, ou valor baseado em outros atributos do objeto.

Engenharia de atributos

Para a geração de um modelo de AM com bom nível de desempenho e generalização na tarefa para a qual foi concebido, é necessário construí-lo a partir de um conjunto de dados que, tanto quanto possível: (a) possua alta qualidade;(30,48) (b) seja representativo do domínio a que os dados pertençam; (c) revele padrões separáveis, identificáveis e classificáveis dos dados. Para tanto, o conjunto de dados deve possuir uma quantidade de exemplos (objetos) e de informações (atributos) sobre estes, reduzindo tanto quanto possível os redundantes e irrelevantes, de forma a evitar ou minimizar o subajuste (*underfitting*) ou sobreajuste (*overfitting*) dos modelos aos dados de treinamento.(24, 30, 44) Essa quantidade deve ser suficiente também para permitir o treinamento, a seleção e avaliação do modelo obtido.(14) Estudos da literatura tratam da relação entre o tamanho do conjunto de dados e a qualidade de classificadores.(49–52)

De forma prática, busca-se atender a esses requisitos cuidando-se da aquisição dos dados e, se necessário, empregando-se métodos de pré-processamento(30) e/ou de engenharia de atributos(53), neste caso para seleção ou extração de novos atributos dos conjuntos de dados. Para a seleção de atributos são encontradas as seguintes abordagens e exemplos de métodos (54):

- ***Embbeded****:* A seleção dos melhores atributos é inerente ao algoritmo de aprendizagem

de máquina utilizado para o treinamento do modelo a ser utilizado na análise. Os algorítmos de árvore de decisão, exemplo ID3 e C4.5, podem ser utilizados em análises de classificação para selecionar os atributos mais relevantes.

- ***Filtering*** **:** A seleção de atributos é realizada através de um processamento prévio dos atributos que determina quais atributos serão removidos baseando-se em critérios de filtragem. Por exemplo, pode-se calcular a correlação entre cada atributo e o atributo alvo (classe) e filtrar os atributos que ficam abaixo de um limite.

- ***Wrapper****:* A seleção de atributos é realizada testando combinações dos atributos sobre o algoritmo de aprendizagem de máquina a ser utilizado na análise, e utilizando algum método para avaliação da qualidade dos resultados. Um exemplo de método wrapper é o *Recursive Feature Elimination (RFE)*. (16)

Há situações também em que seja necessário criar novos atributos a partir dos atributos existentes.(30) Nesse caso, o número de novos atributos pode ser maior ou menor (redução da dimensionalidade) que o número de atributos originais. Para isso, são apresentadas as seguintes abordagens básicas:

- ***Extração****:* obtenção de um novo conjunto de dados, contendo novos atributos, a partir dos dados originais. Por exemplo, a extração de características de forma e textura (novo conjunto de dados) a partir dos pixels de um conjunto de imagens (dados originais).

- ***Construção****:* adição de novos atributos calculados a partir dos atributos já existentes no conjunto de dados. Por exemplo, o cálculo do índice de massa corporal (IMC) (novo atributo) a partir dos atributos de massa e altura presentes no conjunto de dados.

- ***Mapeamento*****:** obtenção de um novo conjunto de dados, a partir de alguma transformação aplicada aos dados originais, levando-os para um novo domínio ou espaço de dados, e fornecendo novas possibilidades de atributos. Por exemplo, o mapeamento dos dados de séries temporais para o espaço de frequências através da Transformada de Fourier. Outro exemplo é a aplicação da *Principal Component Analysis* (PCA) (41) que realiza o mapeamento dos atributos originais em um novo espaço de atributos, os quais maximizam a variação entre as classes de dados.

Modelagem

Finalizada a etapa de preparação, os dados estão prontos para serem utilizados para a geração dos modelos na etapa de modelagem. Em termos simples e no contexto do aprendizado de máquina, um modelo pode ser descrito como uma relação entre os atributos de um objeto de dados (atributos de entrada, independentes) e a classe a que o objeto pertence (atributo de saída, dependente). Às vezes, essa relação pode ser apenas entre os atributos de entrada

(no caso de conjuntos de dados sem saída definida ou variáveis dependentes) em análises não-supervisionadas. (30) Este relacionamento entre os atributos pode ser expresso em termos de equações matemáticas, funções ou regras. (14,41)

Na tarefa de *ajuste de parâmetros* são escolhidos os valores de configuração dos hiperparâmetros do algoritmo. Essa escolha pode ser feita de forma manual ou com busca automática utilizando-se métodos como grid search ou random search. (16) Após a definição de uma determinada configuração dos hiperparâmetros, inicia-se a tarefa de *aprendizagem*, buscando-se determinar de forma automática os valores associados às equações matemáticas, funções ou regras do modelo a ser gerado. (14)

A *validação* é a tarefa que afere o desempenho dos modelos sobre os dados de treinamento. Neste trabalho foram empregadas algumas métricas e métodos para essa aferição em modelos de agrupamento e classificadores. Das métricas para modelos de agrupamento,(42, 55) foi utilizada nesta tese a métrica de validação interna *Average Silhouette Width (ASW)* (56), que varia entre -1 e $+1$. Quanto mais próximo de $+1$, maior a qualidade do agrupamento.

Nos modelos classificadores, o emprego das métricas requer a definição da *classe de dados positiva e a negativa*, caso a classificação seja binária.(30) Em relação às classificações para testes-diagnósticos de doenças, define-se a *classe positiva = presença de uma condição particular ou doença*, e a *classe negativa = ausência de uma condição particular ou doença*. Para determinar a correção do novo teste-diagnóstico, comparam-se os resultados do teste com os de um padrão (*padrão-ouro*). Esse pode ser o verdadeiro estado do indivíduo, se a informação está disponível, um conjunto de exames julgados mais adequados, ou uma outra forma de diagnóstico que sirva de referência. O teste-diagnóstico ideal deveria fornecer a resposta correta, ou seja, um resultado positivo nos indivíduos com a condição particular (doença) (dado positivo) e um resultado negativo nos indivíduos sem a condição particular (doença) (dado negativo). As possibilidades de resultados fornecidos por um classificador (novo teste-diagnóstico) podem ser:(30,57,58)

- *Verdadeiros-Positivos (VP):* dados positivos que foram rotulados como positivos. Por exemplo, um indivíduo é *corretamente* classificado como tendo uma condição particular (doença).
- *Verdadeiros-Negativos (VN):* dados negativos que foram rotulados como negativos. Por exemplo, um indivíduo é *corretamente* classificado como não tendo uma condição particular (doença).
- *Falsos-Positivos (FP):* dados negativos que foram rotulados como positivos. Por exemplo, um indivíduo é *incorretamente* classificado como tendo uma condição particular (doença).
- *Falsos-Negativos (FN):* dados positivos que foram rotulados como positivos. Por exemplo, um indivíduo é *incorretamente* classificado como não tendo uma condição particular

(doença).

O número total de dados positivos P e negativos N é denominado por T $(= P + N)$. A sua soma, ou seja, o número total de dados classificados como positivos e negativos são denominados por P' e N' respectivamente. A partir desses valores, pode-se construir a matriz de confusão (30) visualizada na Figura 2.7 e definir outras métricas de avaliação do classificador.

		Teste (Classificador) (Classe Predita)		
		Positivo	Negativo	
Padrão-Ouro (Classe Verdadeira)	Positivo	**VP**	**FN**	P
	Negativo	**FP**	**VN**	N
		P'	N'	P+N

Figura 2.7 – Matriz de confusão.

Fonte: Adaptada de HAN. (30)

- *Acurácia (Acu):* taxa de dados corretamente classificados calculada por $Ac = (VP + VN)/T$.
- *Erro (Err):* taxa de dados classificados incorretamente. Calculada por $Er = (FP + FN)/T$ ou $Er = 1 - Ac$.
- *Sensibilidade (S):* taxa de dados corretamente classificados como positivos. Calculada por $S = VP/P$. Também conhecida como *Recall*.
- *Especificidade (E):* taxa de dados corretamente classificados como negativos. Calculada por $E = VN/N$.
- *Precisão (Pre):* taxa de dados classificados como positivos que verdadeiramente são positivos. Calculada por $Pre = VP/P'$.
- *Valor Preditivo Positivo (VPP):* probabilidade de um dado classificado como positivo ser verdadeiro positivo. Calculado por $VPP = VP/P'$
- *Valor Preditivo Negativo (VPN):* probabilidade de um dado classificado como negativo ser verdadeiro negativo. Calculado por $VPN = VN/N'$
- *F1-score (F1):* média harmônica entre *Precisão* e *Recall*. Calculada por $F1 = (2 \times Precisão \times Recall)/(Precisão + Recall)$.

Os métodos mais conhecidos para avaliação de desempenho de classificadores são 4: *Holdout, Random Subsampling, K-Fold Cross-Validation e Bootstrap*.(30) Em todos esses métodos, o conjunto de dados é dividido em duas partes para validação: conjunto de treinamento e conjunto de teste. No método *Holdout*, o conjunto de dados é dividido, aleatoriamente e uma única vez, em um conjunto de treinamento e um conjunto de teste em alguma proporção definida. As proporções mais utilizadas são 30/70% e 20/80% a depender da quantidade de dados disponível. Realizada essa divisão, um classificador é gerado a partir do conjunto de treinamento e as métricas de desempenho são calculadas sobre os resultados da classificação dos dados de teste. Esse método traz consigo duas dificuldades. Ao dividir o conjunto de dados em duas partes, o conjunto de treinamento fica menor e p classificador a ser gerado pode não ser tão bom quanto se fosse gerado sobre o conjunto de dados inteiro. (59) A outra dificuldadeé sobre a definição da proporção.

O método *Random Subsampling* é uma variação do *Holdout* com a divisão aleatória dos conjuntos de treinamento e de teste sendo repetida um determinado número de vezes. Isso pode melhorar o desempenho do classificador. (60) O desafio da redução do conjunto de treinamento continua presente. Um desafio adicional é que não há controle sobre o número de vezes que cada objeto de dado é usado para teste e treinamento. Consequentemente, alguns objetos de dados podem ser usados com maior frequência para treinamento ou teste.

No método *Bootstrap*(59, 60), o conjunto de treinamento é amostrado com reposição. Um objeto/exemplo já escolhido para treinamento é colocado de volta em no conjunto de treinamento para que tenha a mesma probabilidade de ser amostrado novamente. Os objetos que não são incluídos na amostra bootstrap tornam-se parte do conjunto de teste. O modelo induzido a partir do conjunto de treinamento é então aplicado ao conjunto de teste para obter uma estimativa de desempenho da amostra bootstrap $i(i = 1 \ldots n)$. O procedimento de amostragem é então repetido n vezes para gerar n amostras bootstrap. Em geral, ao final do processo, calcula-se a média dos desempenhos dos n testes realizados.

O método *K-Fold Cross-Validation (KCV)* (59, 60) é usado quando o conjunto de dados é relativamente pequeno ou quando as métricas de desempenho são determinadas a partir de alguns objetos de dados e não de um único. (60) Nesse método, o conjunto de dados é dividido em k partes iguais (*e.g.* $k = 10$) e k testes são feitos. Em cada teste, um dos subconjuntos serve como um conjunto de testes, enquanto os outros $k - 1$ subconjuntos são utilizados para treinamento. Cada subconjunto é utilizada apenas uma vez. As métricas de desempenho podem ser calculadas a cada iteração, ou a partir da média calculada sobre todos os k testes. Há duas variações bastante conhecidas desse método. Uma é o *Leave-one-out Cross-Validation (LOOCV)* (30) em que o número de divisões ou testes $k = N$, onde N é igual ao número de objetos de dados do conjunto de dados. Esse método possui um custo computacional relativamente elevado por ser repetido N vezes. Além disso, como cada conjunto de teste contém apenas um objeto de dado, as métricas de desempenho tendem a ter alta variância.

A segunda variação do método KCV é o *Nested K-Fold Cross-Validation (NKCV)* (61) que tem sido útil quando poucos objetos de dados (amostras) estão disponíveis. (62, 63) É preferível ao KCV, sendo um método de estimativa de desempenho robusto (64) e excessivamente zeloso. (61) No NKCV dois procedimentos de KCV estão incluídos. O KCV interno é executado pelo laço interno para a seleção do modelo (ajuste do hiperparâmetros do modelo) (65), enquanto o desempenho do modelo é realizado pelo laço externo do outro KCV.(62, 63) Aqui há os parâmetros de configuração *kouter* e *kinner* para os laços externo (avaliação) e interno (ajuste), respectivamente.

2.6 Ferramentas computacionais

Conforme listadas na Tabela 2.1, neste trabalho foram utilizadas ferramentas computacionais para(a) desenvolvimento dos programas necessários às aplicações, (b) análises de dados e (c) apoio à pesquisa. Todas são de acesso gratuito ou possuem autorização de uso como aluno do programa de pós-graduação do IFSC/USP. A implementação das diferentes tarefas do pipeline de AM foi feita com a programas em linguagem Python e bibliotecas/módulos, principalmente Numpy, Pandas, Matplotlib, Seaborn, SciPy, Scikit-image e Scikit-learn. O ambiente utilizado para escrita dos programas em formato de notebooks foi o VSCode em associação com o gerenciador de bibliotecas/módulos Anaconda. A indicação das ferramentas visa a fornecer aos usuários que pretenderem ingressar na área um conjunto básico para realização de trabalhos similares.

Tabela 2.1 – Ferramentas computacionais utilizadas no trabalho.

Nome	Tarefa	Link de acesso
Numpy	Manipulação de matrizes	https://numpy.org/
Pandas	Manipulação e análise de dados	https://pandas.pydata.org/
Matplotlib	Visualização de dados	https://matplotlib.org/
Seaborn	Visualização de dados	https://seaborn.pydata.org/
SciPy	Processamento e análise de sinais	https://scipy.org/
Scikit-image	Processamento de imagens	https://scikit-image.org/
Scikit-learn	Aprendizado de máquina	https://scikit-learn.org/stable/
VSCode	Ambiente de programação	https://code.visualstudio.com/
Anaconda	Gerenciador de bibliotecas/módulos	https://www.anaconda.com/
Orange3	Mineração de dados	https://orangedatamining.com/
Origin	Processamento e análise de dados	https://www.originlab.com/
PexSensors	Visualização de dados	http://vicg.icmc.usp.br/vicg
ImageJ	Processamento de imagens	https://imagej.nih.gov/ij/
Google Drive	Armaz./compart. de dados	https://www.google.com
Github	Armaz./compart. de dados	https://github.com/
Mendeley	Gerenciador de referências	https://www.mendeley.com/search/
TexStudio	Escrita de textos	https://www.texstudio.org/
Overleaf	Escrita de textos (online)	https://pt.overleaf.com/project
Libreoffice	Pacote de ferramentas de escritório	https://www.libreoffice.org

Fonte: Elaborada pelo autor.

3 REVISÃO DE LITERATURA

O uso de aprendizado de máquina (AM) para diagnóstico tem se intensificado enormemente nos últimos anos, de maneira que decidiu-se fazer uma revisão de literatura específica para esse tópico. Além disso, empregou-se uma metodologia consolidada para revisões sistemáticas. A revisão compreendeu estudos recentes (últimos 5 anos) sobre AM aplicado a biossensores para diagnóstico de câncer e COVID-19. A seção 3.1 apresenta a metodologia para busca, triagem e seleção dos estudos. A seção 3.2 apresenta a quantidade e informações sobre os estudos incluídos. Uma discussão sobre esses estudos é feita na seção 3.3.

3.1 Metodologia

Foi realizada uma Revisão Bibliográfica Integrativa (66) dos estudos selecionados para responder à pergunta de pesquisa formulada a partir do acrônimo PICO (Patient/Population, Intervention, Comparator, Outcome)(22):

> *Quais são os métodos, técnicas e abordagens de Aprendizado de Máquina aplicados na análise de dados de biossensores visando ao diagnóstico de câncer ou COVID-19?*

A identificação de estudos seguiu o protocolo de construção de revisões sistemáticas Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)(22). Foram consultadas as bases de dados Pubmed*, Web of Science (WoS)† e Scopus‡. A partir de termos controlados do sistema Medical Subject Headings (MeSH)(67) e de termos comumente utilizados, foi construída a seguinte expressão de busca:

> *("machine learning") AND (biosensor* OR immunosensor* OR genosensor* OR immunoassay* OR genoassay* OR assay* OR "electronic tongue"OR e-tongue* OR etongue*) AND (diagnos* OR detect* OR predict* OR screen*) AND (neoplasm* OR cancer* OR covid* OR sars*)*

Essa expressão foi aplicada aos campos título/resumo/palavras-chave e sobre os resultados foram aplicados os seguintes filtros: artigo original, texto completo acessível, publicado em periódico a partir de 2016, idioma inglês. Ainda na fase de identificação, foram removidos os registros duplicados. Na fase de triagem, foram examinados os títulos/resumos e foram excluídos os registros considerados inadequados para responder à pergunta de pesquisa. Na

* https://pubmed.ncbi.nlm.nih.gov/
† https://www-webofscience.ez67.periodicos.capes.gov.br/wos/woscc/basic-search
‡ https://www-scopus.ez67.periodicos.capes.gov.br/search/form.uri?display=basic#basic

fase de elegibilidade, os documentos elegíveis foram lidos na íntegra, as razões de exclusão foram aplicadas e foram incluídos os estudos considerados pertinentes à pergunta de pesquisa.

3.2 Resultados

Da estratégia de busca, foram retornados 569 registros de estudos distribuídos em: Pubmed (142 registros), WoS (165 registros) e Scopus (262 registros). A evolução anual dos estudos (Figura 3.1) apresenta tendência de crescimento, evidenciando o aumento do interesse científico.

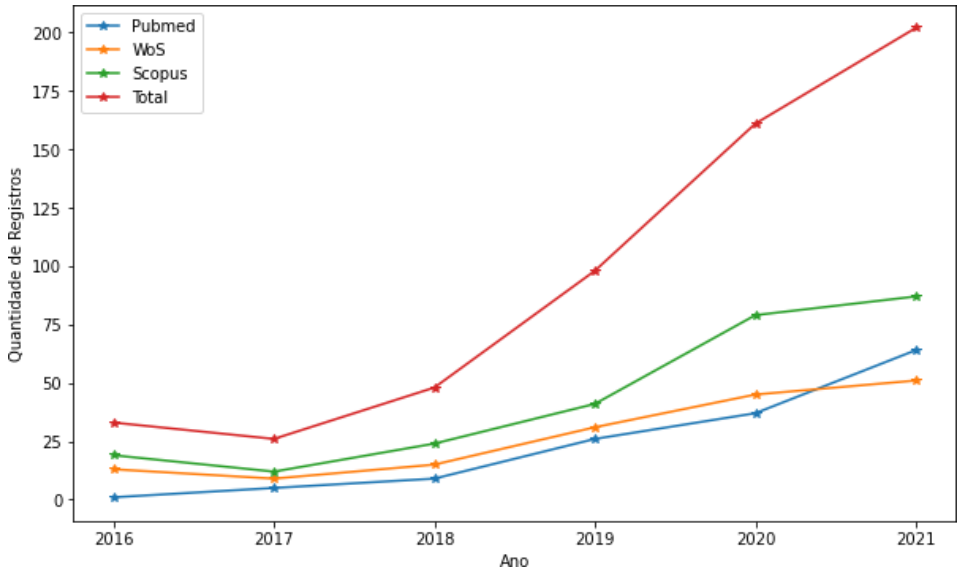


Figura 3.1 – Evolução anual da quantidade de registros nas bases de dados consultadas.

Fonte: Elaborada pelo autor.

A organização e o número de estudos processados em cada fase de seleção estão no fluxograma da Figura 3.2. Durante a fase identificação, foram removidos 257 registros repetidos (45% do total). A fase de triagem foi realizada sobre 312 registros, sendo elegíveis apenas 33. Após a leitura completa, foram aplicadas as seguintes razões de exclusão: (1) aplica AM em tarefas diferentes de diagnóstico; (2) dados utilizados não provenientes de biossensor/detecção molecular ou são provenientes de sequenciamento genético; (3) artigo de revisão/artigo de apresentação com dados insuficientes. Ao fim, foram incluídos 9 estudos que estão listados e descritos na Tabela 3.1. Na descrição de cada estudo incluído são apresentados o Objetivo **[OB]** pretendido, o Conjunto de Dados **[CD]** utilizado, as tarefas executadas de Pré-Processamento **[PP]**, Engenharia de Atributos **[EA]**, Modelagem **[MO]**, Validação e Avaliação **[VA]**, e o Desempenho **[DE]** obtido pelo(s) modelo(s) de AM construído(s).

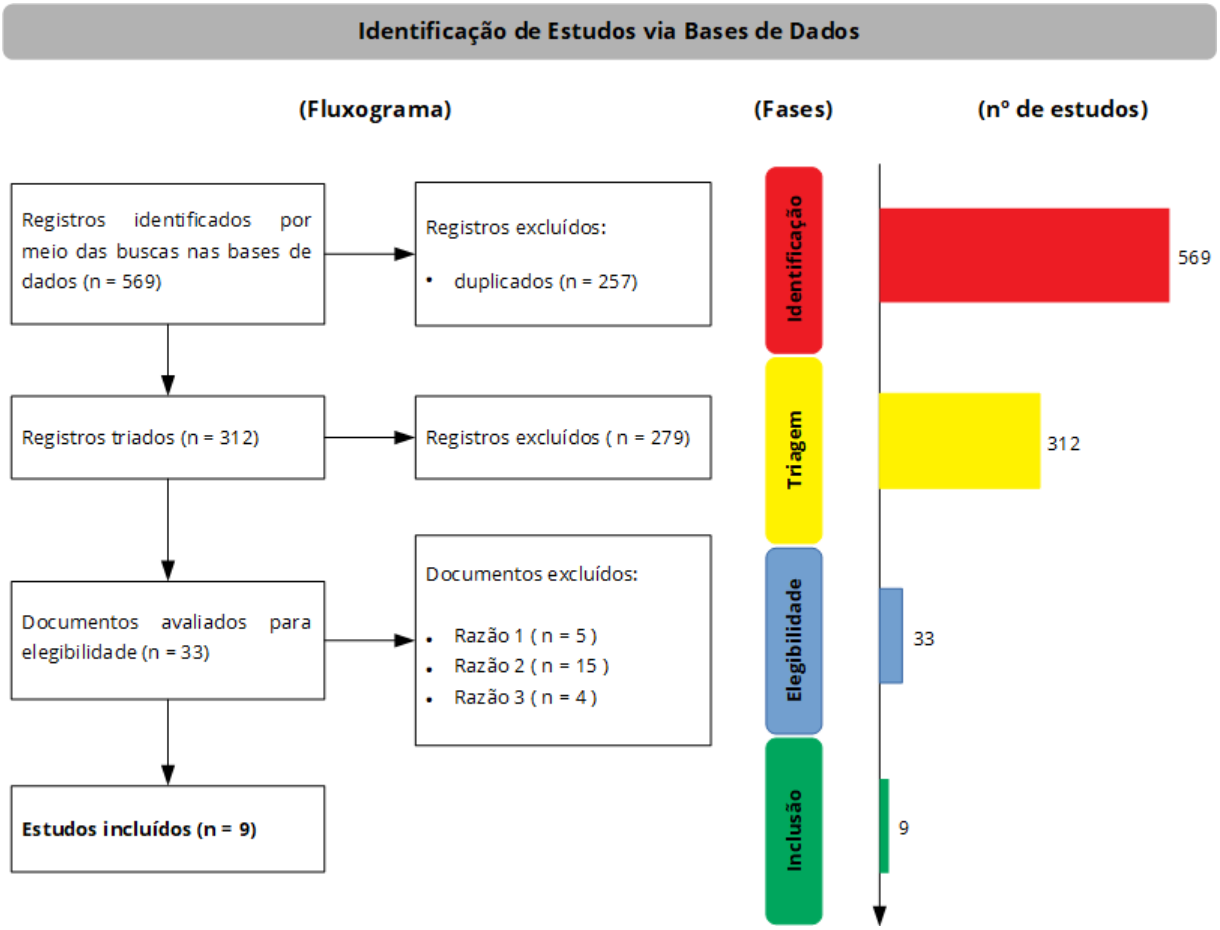


Figura 3.2 – Fluxograma baseado no modelo PRISMA para seleção dos estudos encontrados.

Fonte – Adaptada de PAGE *et al.* (22)

Tabela 3.1 – Descrição dos estudos incluídos das bases de dados Pubmed, WoS e Scopus em termos do Objetivo [OB] pretendido, Conjunto de Dados [CD], tarefas de Pré-Processamento [PP], Engenharia de Atributos [EA], Modelagem [MO], Validação e Avaliação [VA], e Desempenho [DE] obtido pelo(s) modelo(s) de AM.

Ref.	Descrição
(68)	**[OB]** Avaliação de desempenho do imunoensaio multiplex Oncuria para detecção de câncer de bexiga através de amostras de urina. **[CD]** Objetos: 362 amostras de urina (n = 362). Classes (objetos): positivo (n = 46), negativo (n = 316). Atributos: medidas de concentrações (pg/mL) de 10 biomarcadores (proteínas A1AT, APOE, ANG, CA9, IL8, MMP9, MMP10, PAI1, SDC1, VEGFA) e 3 fatores demográficos (idade, sexo, cor). **[PP]** Escala: transformação log10 sobre as concentrações dos biomarcadores.

(continua)

Tabela 3.1 – Descrição dos estudos incluídos das bases de dados Pubmed, WoS e Scopus em termos do Objetivo [OB] pretendido, Conjunto de Dados [CD], tarefas de Pré-Processamento [PP], Engenharia de Atributos [EA], Modelagem [MO], Validação e Avaliação [VA], e Desempenho [DE] obtido pelo(s) modelo(s) de AM

(continuação)

Ref.	Descrição
	[EA] Seleção de atributos: na análise de assinatura molecular, somente atributos dos biomarcadores. Na análise híbrida, todos os atributos foram utilizados. **[MO]** Supervisionada: algoritmo Logistic Regression (LR) na análise individual de cada biomarcador no diagnóstico de câncer; algoritmo Linear Discriminant Analysis (LDA) com 10 e 13 atributos na análise preditiva. **[VA]** Hold-out Treinamento/Teste (80/20%). No conjunto de treinamento, validação cruzada Leave-one-out (LOO). Medidas de desempenho: sensibilidade, especificidade, valor de predição positiva (VPP)/negativa(VPN), curva Receiver Operating Characteristic (ROC). Classificadores foram avaliados no conjunto de teste $p < 0.05$ (p bilaterais). **[DE]** Com todos os 10 biomarcadores usando corte de 50% de probabilidade resultou em uma Area Under de Curve (AUC) de 0,93 (intervalo de confiança de 95%, 0,87-0,98), sensibilidade de 87%, especificidade de 92%, VPN 98% e VPP 61%. A AUC melhorou para 0,95 (IC de 95%, 0,90-1,00) com a adição dos três fatores demográficos na assinatura híbrida com sensibilidade 93%, especificidade 93%, VPN 99% e VPP 65%.

(continua)

Tabela 3.1 – Descrição dos estudos incluídos das bases de dados Pubmed, WoS e Scopus em termos do Objetivo [OB] pretendido, Conjunto de Dados [CD], tarefas de Pré-Processamento [PP], Engenharia de Atributos [EA], Modelagem [MO], Validação e Avaliação [VA], e Desempenho [DE] obtido pelo(s) modelo(s) de AM

(continuação)

Ref.	Descrição
(21)	**[OB]** Diagnóstico de câncer de próstata por um genossensor através da detecção eletroquímica e ótica do biomarcador PCA3 e da aplicação de AM sobre imagens do genossensor. **[CD]** Objetos: 32 imagens de microscopia eletrônica do genossensor (n = 32). Classes (objetos): negativa (n = 2), controle (n = 3), concentrações de PCA3 (μmol.L^{-1}) - 10^{-5} (n = 4), 10^{-4} (n = 4), 10^{-3} (n = 4), 10^{-2} (n = 6), 10^{-1} (n = 5), 1 (n = 4). Atributos: 1024x768 pixels das imagens. **[PP]** Augmentação: segmentação de 3 regiões de 300x300 pixels de cada imagem. **[EA]** Extração de atributos: foram comparados os extratores de texturas das imagens Gray Level Difference Matrix (GLDM), Fourier descriptors, Complex Network Texture Descriptor (CNTD), Fractal descriptors, Adaptative Hybrid Pattern (AHP), Local Binary Patters (LBP), Complex Network and Randomized Neural Network (CNRNN) and Local Complex Features and Neural Network (LCFNN). **[MO]** Supervisionada: análises binária e multiclasse com algorítmos Linear Discriminant Analysis (LDA), Support Vector Machine (SVM-linear), k-Nearest Neighboors (KNN, k=1). Não-supervisionada: análises binária e multiclasse com algoritmo k-Means. **[VA]** Supervisionada: 100 rodadas do experimento amostragem-treinamento-validação. Medidas de desempenho acurácia média e desvio padrão sobre as 100 rodadas. Amostragem com balanceamento e estratificação. Validação cruzada 10-fold. Não-supervisionada: acurácia sobre o agrupamento obtido de todos os objetos de dados. **[DE]** Supervisionada: nos casos binário e multiclasse, o melhor classificador foi LDA com extrator LCFNN obtendo acurácias médias de 99,9(0,3)% e 88,3(3,4)% respectivamente. Não-supervisionada: acurácia nos casos binário e multiclasse foi de 95,83% e 70,83% respectivamente.

(continua)

Tabela 3.1 – Descrição dos estudos incluídos das bases de dados Pubmed, WoS e Scopus em termos do Objetivo [OB] pretendido, Conjunto de Dados [CD], tarefas de Pré-Processamento [PP], Engenharia de Atributos [EA], Modelagem [MO], Validação e Avaliação [VA], e Desempenho [DE] obtido pelo(s) modelo(s) de AM

(continuação)

Ref.	Descrição
(69)	**[OB]** Detecção da proteína spike dos vírus SARS-COV e SARS-COV-2 em amostras de urina por biossensor funcionalizado com a proteína ACE2 com aplicação de AM sobre espectros de Surface-Enhanced Raman Spectroscopy (SERS). **[CD]** Objetos: 1094 espectros SERS (n = 1094). Classes (balanceadas): proteína SARS-CoV-2 S, proteína SARS-CoV S, VS (variante da proteína S do SARS-CoV-2) e VN (variante da proteína nucleocapsídeo SARS-CoV-2). Atributos: pontos do espectro no intervalo de bandas 550 a 1750 cm^{-1}. **[PP]** Não há. **[EA]** Extração de atributos: aplicação da transformação Principal Component Analysis (PCA) para extração dos scores e loadings. Seleção de atributos: scores das componentes da PCA foram utilizadas para classificação visual das diferentes classes de amostras. Os loadings da PCA foram utilizados para seleção das bandas mais importantes dos espectros SERS, chamado padrão de identificação. **[MO]** Supervisionada: algoritmo Discriminant Analysis (DA) para classificação de novas amostras utilizando-se as bandas do padrão de identificação. **[VA]** Hold-out. Medida de desempenho: contagem e a taxa das detecções dos vírus e suas variantes em amostras de teste mais complexas. **[DE]** 100% de taxa de detecção em todas as amostras de teste.

(continua)

Tabela 3.1 – Descrição dos estudos incluídos das bases de dados Pubmed, WoS e Scopus em termos do Objetivo [OB] pretendido, Conjunto de Dados [CD], tarefas de Pré-Processamento [PP], Engenharia de Atributos [EA], Modelagem [MO], Validação e Avaliação [VA], e Desempenho [DE] obtido pelo(s) modelo(s) de AM

(continuação)

Ref.	Descrição
(70)	**[OB]** Detecção de citocinas em amostras de soro sanguíneo para monitoramento do estado do sistema imunológico em pacientes com COVID-19 através de imunoensaio digital multiplexado (12 citosinas) baseado no método Enzyme-Linked Immunosorbent Assay (ELISA). **[CD]** Objetos: 120 imagens de microscopia de fluorescência (n = 120) de cada uma das unidades da matriz do imunoensaio (10 amostras e 12 citosinas). Classes (objetos): não há. Atributos: 6000x4000 pixels das imagens. **[PP]** Recorte da imagem, filtragem de ruídos e elevação de contraste. Todos esses processos só foram citados. **[EA]** Extração de atributos: Convolutional Neural Network (CNN) pré-treinada de duas vias para a segmentação e a marcação dos pixels das imagens com as seguintes etiquetas: (I) micropoços com fluorescência "On" (Canal Qred), (II) gotas codificados por cor Alexa Fluor® 488 (Canal AF488), (III) defeitos de imagem e (IV) imagem de fundo. Após a marcação, foi realizada a contagem de gotas totais (Ntot) e de gotas com enzimas ativas por uma dada citosina (Nactive). Foi calculada a fração de gotas com enzima ativa (OnRate) através da divisão de Nactive/Ntot. Finalmente, foi calculado o Número Médio de Complexos-imunes Formados por Gota (AIB). **[MO]** Supervisionada: Regressão linear para obtenção da curva de calibração do imunoensaio, isto é, da curva AIB x concentração de cada citocina. **[VA]** Medida de desempenho: coeficiente R2 para avaliação da concordância da curva obtida com a fornecida pelo teste LEGENDPlex. **[DE]** Para o grupo de citosinas com grande abundância obteve-se R2=0,916, enquanto para o grupo com menor abundância R2= 0,873.

(continua)

Tabela 3.1 – Descrição dos estudos incluídos das bases de dados Pubmed, WoS e Scopus em termos do Objetivo [OB] pretendido, Conjunto de Dados [CD], tarefas de Pré-Processamento [PP], Engenharia de Atributos [EA], Modelagem [MO], Validação e Avaliação [VA], e Desempenho [DE] obtido pelo(s) modelo(s) de AM

(continuação)

Ref.	Descrição
(71)	**[OB]** Classificador de imagens de smartphone de testes rápidos de diagnóstico (TRD) para COVID-19. **[CD]** Objetos: treinamento com 1000 imagens (n = 1000) e teste com 3344 imagens (n = 3344) provenientes de 11 modelos de TRDs. Classes (objetos): treinamento com positivo (n = 500), negativo (n = 500). Atributos: pixels das imagens coloridas. **[PP]** Escala: conversão para escala de cinza e normalização. Recorte da região de interesse (ROI). Redução da resolução espacial para 256x256 pixels. **[EA]** Extração de atributos: realização automaticamente pelo algoritmo Convolutional Neural Network (CNN). **[MO]** Supervisionada: algorítmo Artificial Neural Network (ANN). Foi gerado um classificador para cada modelo de TRD. **[VA]** Hold-out Treinamento/Teste (1000/3344 imagens). A avaliação foi realizada separadamente para cada TRD. Medidas de desempenho: sensibilidade, especificidade, número de verdadeiro-positivos, número de verdadeiro-negativos, número de falso-positivos, número de falso-negativos. **[DE]** Resultado (objetos de teste/sensibilidade %/especificidade %) global (3344/98.9/99.7). Resultados por modelo de TRD: Alco (24/100.0/100.0), Avioq (317/97.6/99.0), Biolidics (336/97.9/100.0), Biotime (25/100.0/100.0), Biosynex (190/98.4/100.0), NGBiotech Cassette (1254/99.4/99.8), NGBiotech All in one (852/99.7/99.4), Nova (270/98.0/100.0), Realy (27/100.0/100.0), Solo Lab (25/100.0/100.0), Vedalab (24/100.0/100.0).

(continua)

Tabela 3.1 – Descrição dos estudos incluídos das bases de dados Pubmed, WoS e Scopus em termos do Objetivo [OB] pretendido, Conjunto de Dados [CD], tarefas de Pré-Processamento [PP], Engenharia de Atributos [EA], Modelagem [MO], Validação e Avaliação [VA], e Desempenho [DE] obtido pelo(s) modelo(s) de AM

(continuação)

Ref.	Descrição
(72)	**[OB]** Detecção em estágio inicial de Non-Small Cell Lung Cancer (NSCLC) utilizando biomarcadores (DNA, proteínas) e AM sobre dados de teste simples de sangue. **[CD]** Objetos: treinamento com 258 amostras de sangue (n = 258). Classes (objetos): positivo (n = 79), outros (n = 179). Validação com 228 amostras de sangue (n = 228). Classes (objetos): positivo (n = 55), outros (n = 173). Atributos: medidas de concentrações (pg/mL) de 21 biomarcadores. **[PP]** Remoção de dados anômalos com a filtragem de dados cujo coeficiente de variação (dp/mediana) fosse maior ou igual a 20%. **[EA]** Seleção de atributos: o conjunto todos atributos (21) foi comparado com o conjunto (LCDT1) que continha os atributos mais importantes selecionados pelo algoritmo Random Forest (RF) (com Gini Index). **[MO]** Supervisionada: algoritmo LR (5 parâmetros). **[VA]** Supervisionada: Hold-out Treinamento/Teste (258/228 objetos). A avaliação foi realizada separadamente para cada um dos dois conjuntos de atributos. Medidas de desempenho: acurácia, sensibilidade, especificidade, AUC. **[DE]** Resultado de (acurácia %, sensibilidade %, especificidade %, AUC) dos algoritmos: Algorithm 21 (94.3, 89.1, 96.0, 0.960), Algoritmo LCDT1 (95.6, 89.1, 97.7, 0.966).
(73)	**[OB]** Detecção de um tipo de RNA expressado pelo virus SARS-COV-2 através de AM aplicada a imagens coletadas por smartphone de biossensor de cristal líquido (LC). **[CD]** Objetos: 88 imagens (n = 88) do biossensor contendo amostras sintéticas. Classes (objetos): positivo (n = 29), negativo (n = 59). Atributos: 48 (4 x 4 grids x 3 cores). **[PP]** Escala: Conversão dos pixels das imagens para a escala CIELAB (separa a iluminação das componentes de cores). Recorte da região de interesse (ROI). **[EA]** Extração de atributos: a partir da distribuição espacial de intensidades dos pixels das imagens. A ROI (128x128 pixels) foi subdividida em um arranjo espacial em forma de grade (4x4 grids). Os atributos foram normalizados (0,1). **[MO]** Supervisionada: algoritmo classificador SVM. **[VA]** Não foi apresentada. **[DE]** O classificador conseguiu distinguir perfeitamente os 88 objetos.

(continua)

Tabela 3.1 – Descrição dos estudos incluídos das bases de dados Pubmed, WoS e Scopus em termos do Objetivo [OB] pretendido, Conjunto de Dados [CD], tarefas de Pré-Processamento [PP], Engenharia de Atributos [EA], Modelagem [MO], Validação e Avaliação [VA], e Desempenho [DE] obtido pelo(s) modelo(s) de AM

(continuação)

Ref.	Descrição
(74)	**[OB]** Teste baseado em microchip para aprisionamento de microbolhas visando a quantificação de proteína biomarcadora de prostate-specific antigen (PSA) para diagnóstico de câncer de próstata com auxílio de AM, smartphone e pequeno microscópio acoplado. **[CD]** Objetos: para treinamento 493 imagens (n = 493) do microchip (3x3 mm) contendo amostras sintéticas com diferentes concentrações (pg/mL) da proteína biomarcadora. Classes (balanceado): positivo, negativo. Atributos: pixels das imagens. **[PP]** Não há. **[EA]** Extração de atributos: automática por duas CNNs. A primeira é treinada para localizar a região da imagem que contem o microchip com as microbolhas. A segunda é treinada para reconhecer as microbolhas e contá-las. **[MO]** Regressão linear. **[VA]** A contagem realizada pela aprendizadem de máquina foi confrontada com a contagem manual realizada com o software de processamento de imagens ImageJ. **[DE]** Foi verificada uma correlação linear entre a contagem de microbolhas por aprendizado de máquina e a contagem manual. Foi verificado também um comportamento linear entre a contagem de microbolhas e a concentração da proteína biomarcadora de PSA nas amostras, com R2 = 0.994 para um limite de detecção (LOD) de 0.06 pg/mL.

(continua)

Tabela 3.1 – Descrição dos estudos incluídos das bases de dados Pubmed, WoS e Scopus em termos do Objetivo [OB] pretendido, Conjunto de Dados [CD], tarefas de Pré-Processamento [PP], Engenharia de Atributos [EA], Modelagem [MO], Validação e Avaliação [VA], e Desempenho [DE] obtido pelo(s) modelo(s) de AM

(continuação)

Ref.	Descrição
(75)	**[OB]** Detecção de múltiplos biomarcadores associados ao câncer de pâncreas, ovário e pancreatite, através de imunoensaio baseado em Surface-Enhanced Raman Spectroscopy (SERS) em conjunto com aplicação de AM. **[CD]** Objetos: 1000 espectros SERS (n = 1000) de 20 amostras de soro sanguíneo. Cada um dos 5 biomarcadores (CA19-9, HE4, mesothelin, MMP7, MUC4) possuía 4 amostras sendo uma para cada classe. Classes (amostras): controle (n = 5), câncer no pâncreas (n = 5), câncer no ovário (n = 5) e pancreatite (n = 5). Atributos: 1783 pontos do espectro x 10 posições aleatórias de medição. **[PP]** Filtragem de ruído e alisamento de cada espectro SERS com Fast Fourier Transform (FFT). **[EA]** Extração de atributos: para a classificação com algoritmo Decision Tree (DT), foi extraída a intensidade do espectro na banda 1336 cm^{-1} e calculada a média de cada biomarcador por paciente. Para a classificação com o algoritmo KNN (distância euclidiana), foram concatenados os espectros completos dos biomarcadores de cada paciente. **[MO]** Supervisionada (multiclasse): algoritmo Decision Tree (DT) (depth=2) e algoritmo KNN (distância euclidiana, k=3 e 5, todas as combinações possíveis dos biomarcadores). **[VA]** Com algoritmo DT foi utilizada somente a validação cruzada 5-fold. Com algoritmo KNN, pela forma como os atributos foram arranjados, foi aplicado o método Hold-out treinamento/teste (80/20%). Medidas de desempenho: acurácia, sensibilidade, especificidade, AUC. **[DE]** Para o algoritmo DT, os melhores resultados foram obtidos no diagnóstico das 3 doenças e o controle com a combinação dos dados dos 5 biomarcadores com sensibilidade (0.8-1.0), especificidade (0.93-1.0). Para o algoritmo KNN (k=5), sensibilidade 86%, especificidade 93% e acurácia 91%. Com esse mesmo algoritmo, mas k=3 e as várias combinações de dados dos 5 biomarcadores, o melhor resultado de AUC=0.89 foi obtido com a combinação de todos os 5 biomarcadores.

Fonte - Elaborada pelo autor.

3.3 Discussão

A expressão de busca delimita o que se está buscando e impacta na capacidade de encontrar estudos que contribuirão à resposta da pergunta de pesquisa. A definição dos termos e seus conectivos foi feita num processo iterativo em que foram necessários 3 ciclos de identificação/triagem. Durante essa tarefa, verificou-se que autores usam termos com diferentes significados para designar o mesmo conceito/ideia. O período da busca foi delimitado de 2016 a 2021 para verificar os conceitos e métodos mais recentes de AM empregadas no diagnóstico de Câncer e COVID-19 utilizando dados de biossensores. No período pesquisado, 7 estudos foram publicados a partir de 2020. Os 9 estudos incluídos foram descritos a partir de tópicos pertinentes ao AM: doença (Câncer, COVID-19) objetivo do estudo, conjuntos de dados utilizados, técnicas empregadas de pré-processamento, de engenharia de atributos, de modelagem, de validação/avaliação, e o desempenho (acurácia, sensibilidade, especificidade, etc) obtido na tarefa de diagnóstico. A Tabela 3.2 apresenta uma síntese dessa descrição.

Tabela 3.2 – Síntese da descrição dos estudos incluídos na revisão. Símbolos utilizados: número de objetos (n), positivo (pos.) e controle (cont.).

Ref.	Diagnóstico	n (pos./cont.)	Dados	Algorítimos	Máx. Desemp.
(68)	Câncer (bexiga)	362 (46/316)	concentrações de biomarcadores	LR, LDA	0,93 (AUC)
(21)	Câncer (Próstata)	32 (29/3)	imagens MEV	LDA, SVM KNN	99.9% (acurácia)
(69)	COVID-19	1094 (547/547)	espectros SERS	DA	100% (acurácia)
(70)	COVID-19	120 (10/citosina)	imagens Mic. Fluorescência	Regressão linear	0.916 (R^2)
(71)	COVID-19	1000 (500/500)	imagens Mic. Óptica	ANN	98.9% (sensib.) 99.7% (especif.)
(72)	Câncer (Pulmão)	486 (134/352)	concentrações de biomarcadores	LR	95.6% (acurácia)
(73)	COVID-19	88 (29/59)	imagens Mic. Óptica	SVM	100% (acurácia)
(74)	Câncer (Próstata)	493 (247/246)	imagens Mic. Óptica	Regressão linear	0.994 (R^2)
(75)	Câncer (Múltiplos)	1000 (250/classe)	espectros SERS	DT, KNN	1.0 (sensib.) 1.0 (especif.)

Fonte – Elaborada pelo autor.

Em termos dos Objetivos [OB], 5 estudos visam ao diagnóstico de câncer de bexiga

(68), próstata (21, 74), pulmão (72), pâncreas/ovário (75) a partir de amostras de urina e soro de sangue. Para a COVID-19, foram encontrados 4 estudos (69–71, 73) utilizando amostras de secreção nasal e soro de sangue. Os Conjuntos de Dados [CD] apresentam a diversidade oferecida pelas diferentes composições dos biossensores e das técnicas analíticas. Foram utilizados conjuntos de imagens de microscopias óptica (71, 73, 74), eletrônica (21) e fluorescência (70), espectros de SERS (69, 75) e medidas de concentração de biomarcadores em escalas muito baixas (pg/mL) (68, 72). Em relação aos tamanhos, o número de objetos (n) variou de 32 a 1094. As técnicas de Pré-Processamento [PP] aplicadas às imagens foram recorte da região de interesse (region of interest) (ROI), transformação para escala de cinza e normalização dos pixels para o intervalo de valores [0, 1]. (76) Sobre os espectros SERS foram aplicadas técnicas de filtragem da linha de base, alisamento e transformação Fast Fourier Transform (FFT). (77–79) Nas medidas de concentração, houve também a aplicação de ajustes nas escalas e normalização dos atributos para o intervalo de valores [0, 1]. (77–79)

Na Engenharia de Atributos [EA], foram empregados métodos de seleção e de extração de atributos a partir dos conjuntos de dados brutos. Sobre as imagens, algoritmos Convolutional Neural Network (CNN) (80, 81) foram utilizados como extrator automático de atributos.(70, 71, 74) Extratores de textura (82) como Gray Level Difference Matrix (GLDM) (83), Fourier descriptors (84), Complex Network Texture Descriptor (CNTD) (85), Fractal descriptors (86), Adaptative Hybrid Pattern (AHP) (87), Local Binary Patters (LBP) (88), Complex Network and Randomized Neural Network (CNRNN) (89) and Local Complex Features and Neural Network (LCFNN) (89) também foram utilizados. (21) Nos espectros SERS, extraíram-se novos atributos com a transformação por Principal Component Analysis (PCA) (14, 69) e seleção da linha espectral referente ao biomarcador de interesse (75), chamado de *repórter*. Nas medidas de concentração, foi aplicado o algoritmo Random Forest (RF) (14) para seleção dos atributos mais importantes. (72)

Os modelos de AM foram construídos e aplicados na Engenharia de Atributos para extração/seleção de atributos utilizando algoritmos Convolutional Neural Network (CNN) (80, 81) e RF (14). Na Modelagem [MO] para classificação e regressão foram utilizados os algoritmos Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighboors (KNN), Linear Discriminant Analysis (LDA), Logistic Regression (LR) e ensembles como RF. (14) Para aglomeração foi utilizado o algoritmo KMeans. (14) Quanto às estratégias de Validação e Avaliação [VA] dos modelos de classificação, foram usadas a aplicação de Hold-out do conjunto de dados, e validação cruzada k-Fold e leave-one-out (LOO). (14) Dentro dessas estratégias, o Desempenho [DE] foi mensurado com métricas de acurácia, sensibilidade, especificidade, VPP, VPN, AUC, R^2. (30) Em termos das acurácias, os desempenhos evidenciados no diagnóstico de câncer e COVID-19 ficam acima de 95%. As sensibilidades e especificidades são muito próximas aos seus valores máximos.

3.4 Conclusão

Da revisão relatada neste capítulo, identificaram-se os elementos necessários para localizar/verificar a originalidade e contribuições deste trabalho de doutorado em relação à literatura mais recente. Inclusive, um dos trabalhos incluídos (21) está entre as aplicações de AM apresentadas no próximo capítulo. Chama a atenção o pequeno número de artigos diretamente relacionados a esta tese. Embora o uso de aprendizado de máquina esteja se popularizando rapidamente, seu emprego para analisar dados de sensores e biossensores ainda não é tão disseminado. Como será mencionado nos capítulos de aplicação, em alguns casos as primeiras utilizações - como em imagens de biossensores - foram as realizadas em nosso Grupo, e parte desta tese. Por essa razão, consideramos importante apresentar e discutir o pipeline para AM, para facilitar o emprego de AM por pesquisadores da área de sensores e biossensores.

4 GENOSSENSOR E MEV PARA DIAGNÓSTICO DE CÂNCER DE PRÓSTATA

4.1 Introdução

O diagnóstico de doenças já é feito há muito tempo por análise de imagens. Entretanto, essas imagens são obtidas de amostras biológicas, como as de tecido e tumores, empregando-se diferentes tipos de técnicas para gerar as imagens. Até recentemente não havia sido tentado fazer diagnóstico a partir de imagens dos sensores (ou biossensores) e não das amostras biológicas. A vantagem de se fazer diagnóstico com imagens de sensores está na possibilidade de sensoriamento sem instrumentos (com exceção do aparelho para obter a imagem). A estratégia de usar imagens de biossensores para diagnóstico foi utilizada pela primeira vez por nosso Grupo de Pesquisa.(20) Nesta tese apresentamos uma nova utilização, desta vez incluindo aprendizado de máquina. O problema científico escolhido foi o diagnóstico de câncer de próstata a partir da detecção do biomarcador PCA3, que é um marcador genético específico para câncer de próstata. O biossensor desenvolvido é um genossensor fabricado com a imobilização de uma sequência de DNA que reconhece o PCA3, gerando hibridização. Essa hibridização altera propriedades elétricas do genossensor, além de sua morfologia. São essas alterações de morfologia que podem ser capturadas com análise de imagens, com classificação via algoritmos de aprendizado de máquina.

Nesta seção serão analisadas imagens para genossensores expostos a soluções com diferentes concentrações de PCA3, além de interferentes para verificar a seletividade. Assim, foi possível realizar classificação binária e multiclasse.

4.2 Conjunto de dados

O conjunto de dados para esta aplicação é formado por 31 imagens de microscopia eletrônica de varredura (MEV) (n = 31) do genossensor. As imagens possuem resolução de intensidade de 8 bits (escala de cinza) e foram geradas com magnificação para janelas de 200 nm e diferentes resoluções espaciais (1024x768 e 2048x1536 pixels) utilizando-se o microscópio eletrônico de varredura modelo DSM 960 (Zeiss, Alemanha). A geração e disponibilização das imagens foi feita por Valquíria C. Rodrigues, pesquisadora do Grupo de Pesquisa em Polímeros (IFSC/USP).(21) A Figura 4.1 apresenta um exemplo de imagem do conjunto de dados.

Para avaliar a capacidade em detectar o biomarcador PCA3, 31 unidades do genossensor foram usadas com soluções de tampão fosfato-salino (sigla PBS em inglês) contendo diferentes analitos. Algumas soluções continham o biomarcador PCA3 nas seguintes concentrações (μmol.L^{-1}): 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1. Outras continham a sequência negativa (não-complementar) do biomarcador. Algumas aquisições foram realizadas com o genossensor em solução de PBS (branco) para controle. Exemplos das imagens dos genossensores obtidas após

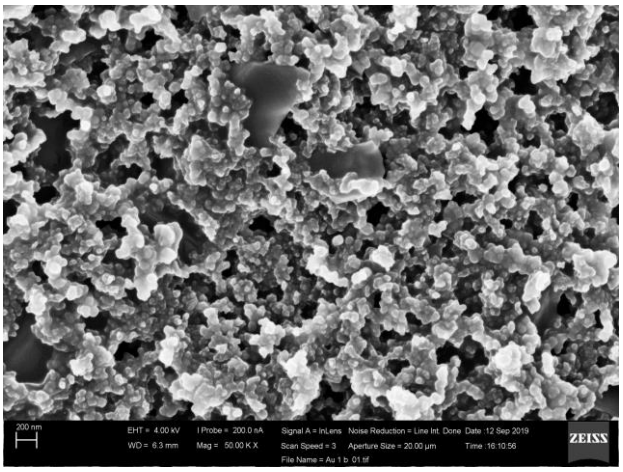


Figura 4.1 – Imagem de MEV do genossensor exposto à solução com o biomarcador PCA3 na concentração 10^{-2} (μmol.L^{-1}).

Fonte: RODRIGUES *et al.*(21)

exposição a cada uma das soluções são apresentados na Figura 4.2.

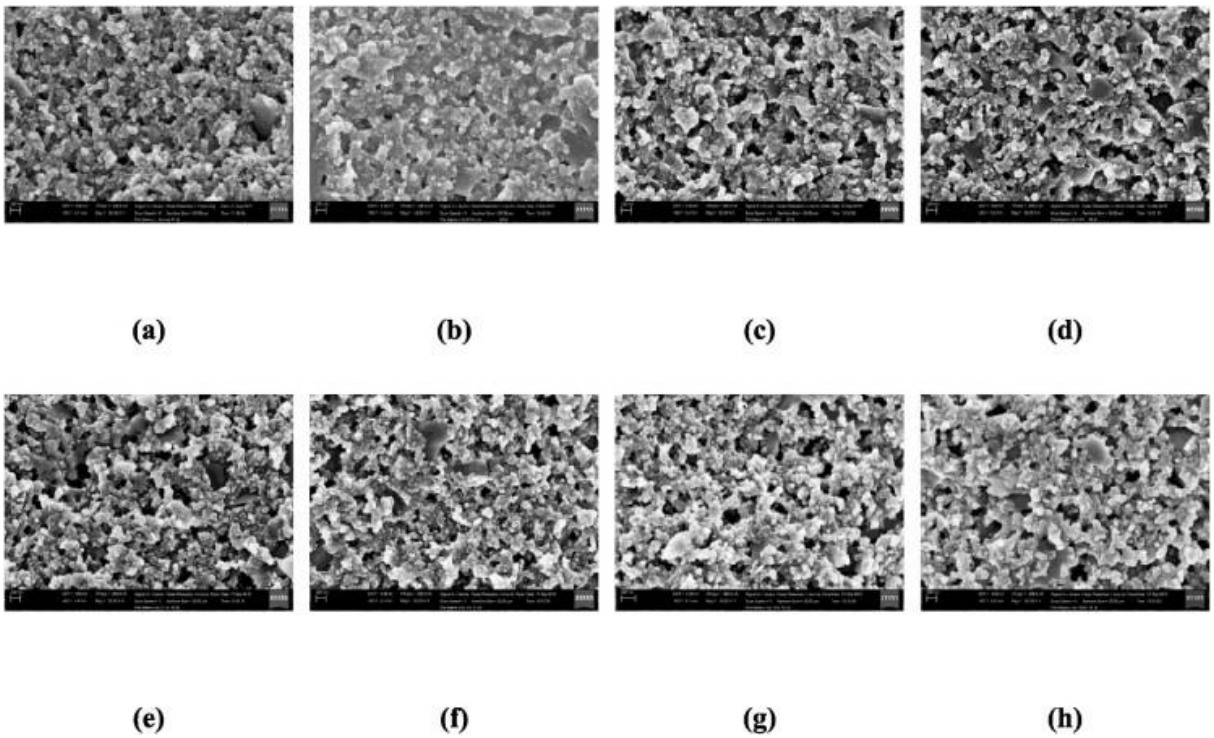


Figura 4.2 – Imagens de MEV de unidades do genossensor para as classes (a) negativa e (b) zero e positiva em ordem crescente das concentrações (μmol.L^{-1}) do biomarcador PCA3 (esquerda para direita): (c) 10^{-5}, (d) 10^{-4}, (e) 10^{-3}, (f) 10^{-2}, (g) 10^{-1}, (h) 1.

Fonte: RODRIGUES *et al.*(21)

A composição do conjunto de dados está sintetizada na Tabela 4.1, incluindo a quantidade de objetos (n = 31) e imagens com duas resoluções espaciais (1024x768 e 2048x1536).

Os atributos que caracterizam os objetos são as intensidades dos pixels em cada posição das imagens. Quanto à distribuição dos dados em relação às classes (Tabela 4.2), há uma quantidade maior de objetos nas classes com o biomarcador PCA3 (positiva). Essas características foram consideradas nas análises.

Tabela 4.1 – Composição do conjunto de dados para uso do genossensor e MEV para diagnóstico de câncer de próstata.

Elemento	Descrição	Qtde.
Objetos	Imagens de MEV (1024x768)	19
	Imagens de MEV (2048x1536)	12
Atributos	Intensidade dos pixels	1024x768
		2048x1536
Classes	negative, zero	8
	positive 10^{-5}, positive 10^{-4}	
	positive 10^{-3}, positive 10^{-2}	
	positive 10^{-1}, positive 1	

Fonte: RODRIGUES *et al.*(21)

Tabela 4.2 – Distribuição dos objetos entre as classes do conjunto de dados na aplicação genossensor e MEV para diagnóstico de câncer de próstata. Símbolo(s) utilizado(s): número de objetos (n).

Analito	Classe	Marca	n
Sequência negativa (não-complementar)	negative	n (1)	2
Branco (para controle)	zero	o (2)	3
[PCA3] (μmol.L^{-1})			
10^{-5}	positive 10^{-5}	p_0p00001 (3)	4
10^{-4}	positive 10^{-4}	p_0p0001 (4)	4
10^{-3}	positive 10^{-3}	p_0p001 (5)	4
10^{-2}	positive 10^{-2}	p_0p01 (6)	6
10^{-1}	positive 10^{-1}	p_0p1 (7)	4
1	positive 1	p_1p0 (8)	4
		Total =	31

Fonte: RODRIGUES *et al.*(21)

Pré-processamento

Sobre o conjunto de dados foi necessário realizar algumas tarefas para posterior análise. Foi feita uma "limpeza"nas imagens, com remoção do rodapé das imagens contendo informações do microscópio sobre as configurações para coleta das imagens. Foi removida uma única imagem da classe 0p1 que possuía 300 nm de escala de magnificação. Diferentes escalas de magnificação representam experimentos físicos diferentes, e com isso o conjunto de dados ficou com n = 31. No caso de conjuntos de dados com quantidade reduzida de exemplos, podem ser necessários métodos de pré-processamento para ampliar essa quantidade e melhorar o desempenho dos classificadores. *Métodos de aumento (augmentation*) fornecem meios para complementar um conjunto de dados com dados semelhantes gerados a partir de ações de processamento sobre os dados originais do conjunto.(48, 90, 91) Existem várias ações de processamento que podem ser aplicadas às imagens para gerar novos exemplos, a saber: rotação (*rotation*), espelhamento (*flipping*), recorte (*cropping*), entre outras. Nesta tese empregou-se o recorte(91–93), que é uma ação de processamento para gerar uma cópia dos pixels em uma região de interesse, aqui denominada *janela*. As janelas podem ter diversos formatos e tamanhos, e serem geradas a partir de regiões sobrepostas ou não da imagem. A Figura 4.3 ilustra o recorte aplicado em regiões diferentes (sem sobreposição) de uma imagem gerando 9 janelas quadradas.

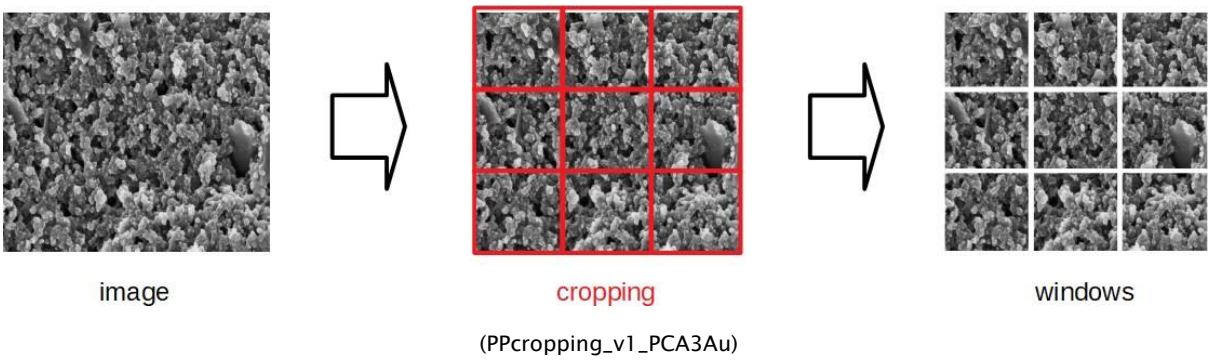


Figura 4.3 – Recorte aplicado em regiões diferentes (sem sobreposição) de uma imagem gerando 9 janelas quadradas.

Fonte: Elaborada pelo autor.

O conjunto de dados possui 31 imagens distribuídas de forma desbalanceada em 8 classes. Para o aumento da quantidade de exemplos, foi aplicado o recorte de cada imagem em janelas quadradas (Figura 4.3) gerando 3 conjuntos de imagens de dimensões 100, 200 e 300 pixels. A tabela apresenta as dimensões das janelas e o números de exemplos gerados em cada classe de imagem. Uma análise comparativa dos histogramas das imagens originais e das janelas apresentada no seção A.2 mostra que, dentre as dimensões recortadas, as janelas com dimensão 300 x 300 pixels são as que mais representativas das originais, sendo utilizadas na sequência de execução do pipeline de AM.

Tabela 4.3 – Dimensões das janelas quadradas e números de exemplos gerados em cada classe de imagens do conjunto de dados.

		Dimensão (pixel)		
Classe	Imagens	100	200	300
1	2	120	30	12
2	3	180	45	18
3	4	240	60	24
4	4	240	60	24
5	4	240	60	24
6	6	360	90	36
7	4	240	60	24
8	4	240	60	24
Total	31	1860	465	186

Fonte: Elaborada pelo autor.

Engenharia de atributos

A partir das janelas de 300x300 pixels, foram extraídos outros conjuntos de atributos empregando-se técnicas de análise de textura, que é um elemento essencial na percepção visual de humanos. Atributos de textura são empregados em muitos sistemas de visão computacional. (82) Foram usadas as seguintes técnicas: *Gray Level Difference Matrix (GLDM)* (83), *Fourier descriptors* (84), *Complex Network Texture Descriptor (CNTD)* (85), *Fractal descriptors* (86), *Adaptative Hybrid Pattern (AHP)* (87), *Local Binary Patters (LBP)* (88), *Complex Network and Randomized Neural Network (CNRNN)* (89) e *Local Complex Features and Neural Network (LCFNN)* (89). Essas técnicas lidam com informações de textura de maneiras diferentes, usando modelos, estatísticas, espectros, sendo adequadas para um número pequeno de amostras num conjunto de dados. São executadas rapidamente em termos de tempo de computação. Uma vantagem adicional dessas técnicas é a menor dimensionalidade dos conjuntos extraídos como pode ser observado na Tabela 4.4 para o caso das janelas com 300x300 pixels que possuem 90000 atributos.

4.3 Resultados

Foram empregadas imagens de microscopia eletrônica de varredura (MEV), pois sabe-se que as interações entre os genossensores e as amostras ocorrem na escala nanoscópica. O ideal no futuro é utilizar imagens de microscopia óptica ou mesmo fotos obtidas com câmeras comuns. Entretanto, era importante verificar se as abordagens de processamento de imagens funcionariam nas condições ideais, com imagens com alta resolução. Por isso o uso de imagens

Tabela 4.4 – Dimensionalidade dos conjuntos de atributos de textura extraídos das janelas com dimensões 300x300 pixels (90000 atributos).

Conjuntos de atributos de textura	Qtde. de atributos
GLDM	60
Fourier	149
CNTD	108
Fractal	69
AHP	120
LBP	648
CNRNN	180
LCFNN	330

Fonte: Elaborada pelo autor.

de microscopia eletrônica. Para avaliar a capacidade dos conjuntos extraídos em evidenciar padrões separáveis dos dados, foram aplicadas técnicas de projeção *t-Distributed Stochastic Neighbor Embedding (t-SNE)* (94) e *Interactive Document map (IDMAP)* (95) e o algoritmo não-supervisionado *K-Means (KM)* (30, 42) para análises binária e multiclasse. No caso do algoritmo KM foi utilizada a acurácia como métrica de validação externa dos agrupamentos.

A Figura 4.4 mostra as projeções dos atributos do quatro melhores descritores de textura, com as amostras (pontos) rotuladas usando o caso de classe binária apenas para fins de visualização. Este experimento revela uma estrutura de dados para os descritores GLDM que formam dois clusters, um compreendendo as amostras positivas (em azul), enquanto o outro teve uma maioria de amostras negativas (em vermelho). Os outros descritores também mostram uma estrutura de agrupamento em que as amostras positivas e negativas podem ser facilmente separadas, corroborando os resultados da análise binária com o algoritmo KM (Tabela 4.5). As projeções para os atributos de textura com amostras coloridas usando rótulos multiclasse são mostradas na Figura 4.5, onde um cluster é observado para as classes negativas ('n' e 'o') e 'p1p0'. O agrupamento das outras classes positivas é menos claro, como se poderia esperar dos resultados do experimento de aprendizado supervisionado que serão relatados a seguir.

Uma visualização dos descritores das imagens foi feita com a técnica de projeção multidimensional IDMAP (95), em que cada descritor de imagem é uma instância de dado projetada num gráfico 2D. Com IDMAP, procura-se preservar a relação de similaridade entre os objetos (descritores de imagens neste caso) no espaço multidimensional na projeção no espaço 2D. O objetivo é verificar se há algum agrupamento (clustering) dos dados. A Figura 4.6 mostra as visualizações mais eficientes para os descritores Fourier e CLBP no caso de classificação binária. Nota-se que há formação de dois agrupamentos correspondentes às classes positiva

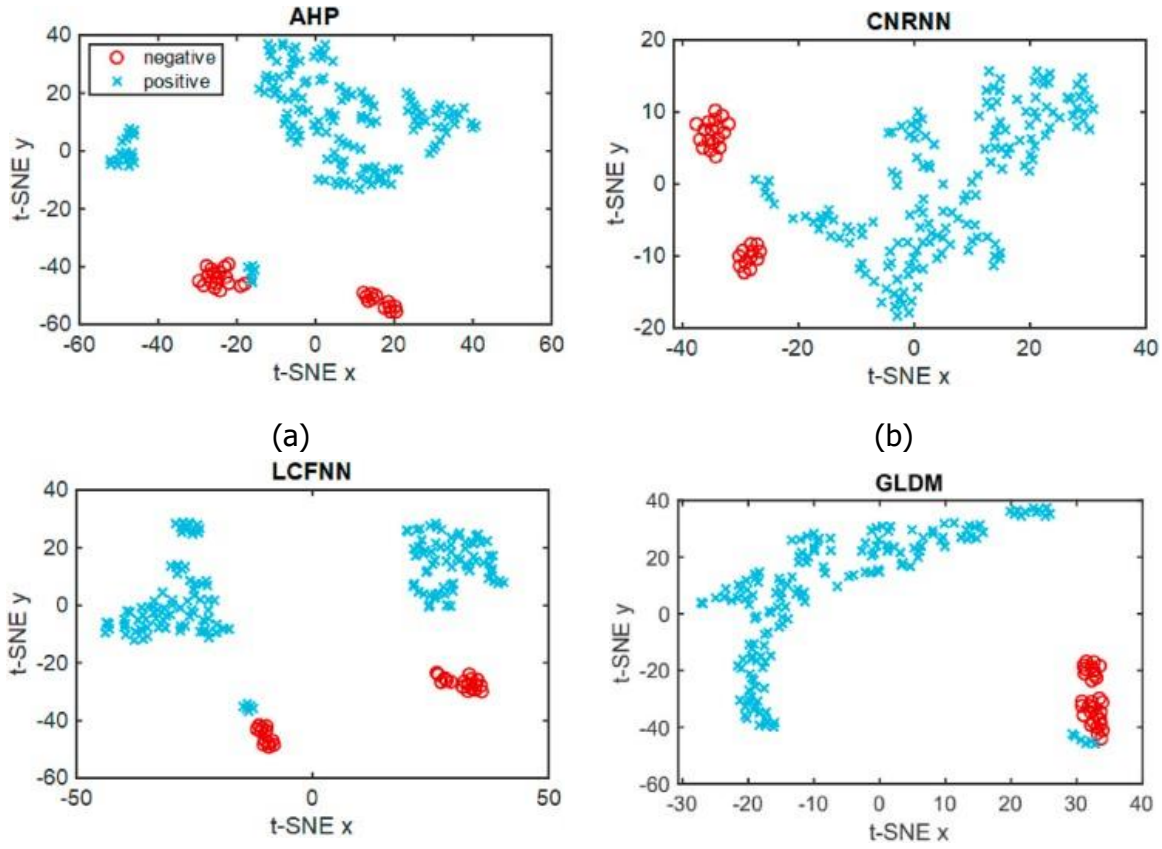


Figura 4.4 – Projeção t-SNE dos conjuntos de atributos (a) AHP, (b) CNRNN, (c) LCFNN e (d) GLDM para análise binária.

Fonte: Elaborada pelo autor.

e negativa, principalmente para o descritor de Fourier. Para a classificação multiclasse, a Figura 4.7 mostra as melhores visualizações para os mesmos descritores, Fourier e CLBP. Não é possível separar todas as classes, o que indica que técnicas de visualização de informação podem não ser suficientes para distinção das várias classes. De fato, uma distinção eficiente só foi obtida com algoritmos AM supervisionados, como será discutido a seguir.

A tarefa de classificação foi realizada em 3 conjuntos de dados, que diferem no número de exemplos e na quantidade de atributos (veja Tabela 4.3) dependendo do tamanho de recorte da janela. Comparando com os resultados dos conjuntos para as janelas de de 100x100 e 200x200 pixels (seção A.2), os melhores desempenhos para classificação binária e multiclasse foram obtidos com as janelas de 300x300 pixels. Na Tabela 4.6 são fornecidos a acurácia e desvio padrão (entre parênteses) para cada combinação de técnica de análise de textura e algoritmo AM. Foram usados os algoritmos *Support Vector Machine (SVM) (linear kernel)*, *Linear Discriminant Analysis (LDA)* e *1-Nearest Neighbors (1-NN)*. A acurácia máxima foi 99,9% (0,3) com o descritor LCFNN e os algoritmos SVM e LDA para a classificação binária, em que a distinção foi feita entre as imagens para os genossensores expostos a soluções

Tabela 4.5 – Acurácia da análise de agrupamento com algoritmo KM para os casos binário e multiclasse para os conjuntos de atributos de textura extraídos das janelas com dimensões 300x300 pixels.

Conjuntos de atributos de textura	Acurácia (%)	
	Binário	Multiclasse
GLDM	86.98	52.10
Fourier	87.50	62.50
CNTD	84.37	52.60
Fractal	90.10	48.44
AHP	80.21	46.35
LBP	82.81	51.56
CNRNN	95.83	52.10
LCFNN	87.50	70.83

Fonte: Elaborada pelo autor.

contendo PCA3 (com todas as concentrações) e aquelas de genossensores expostas ao tampão e marcadores diferentes de PCA3. A alta acurácia indica uma grande capacidade de separação das duas classes. Na classificação multiclasse, a Tabela 4.6 mostra valores menores de acurácia, como esperado. De fato, uma inspeção visual permite antever a dificuldade em separar o branco (sem PCA3) das imagens com baixas concentrações de PCA3, pois atinge-se o limite de detecção. A combinação mais eficiente é LCFNN como descritor de textura e o algoritmo LDA, com acurácia de 88,3% (3,4). Da Tabela 4.6 também se percebe que os maiores desempenhos nas classificações são alcançados com a combinação dos algoritmos SVM ou LDA com os descritores de textura LCFNN, CNRNN ou GLDM.

4.4 Conclusão

Os resultados deste capítulo permitiram concluir que a análise de imagens de sensores combinada com AM é adequada para realizar diagnóstico, no caso específico para um bio-marcador de câncer de próstata (PCA3). Essa abordagem pode ser complementar, ou mesmo combinada, com medidas elétricas, eletroquímicas ou ópticas com os sensores. De qualquer forma, deve-se mencionar que a acurácia na distinção de todas as concentrações de PCA3 foi no máximo 88,3% para AM supervisionado e 70,83% para não-supervisionado, ao passo que nas medidas eletroquímicas e ópticas obteve-se separação completa das diferentes concentrações. (21)

Portanto, há ainda muito espaço para investigar outras técnicas de análise de imagens, embora de nossa experiência talvez o mais relevante seja ter um conjunto maior e mais representativo de imagens. Quanto ao desafio de usar imagens adquiridas a partir de equipamentos

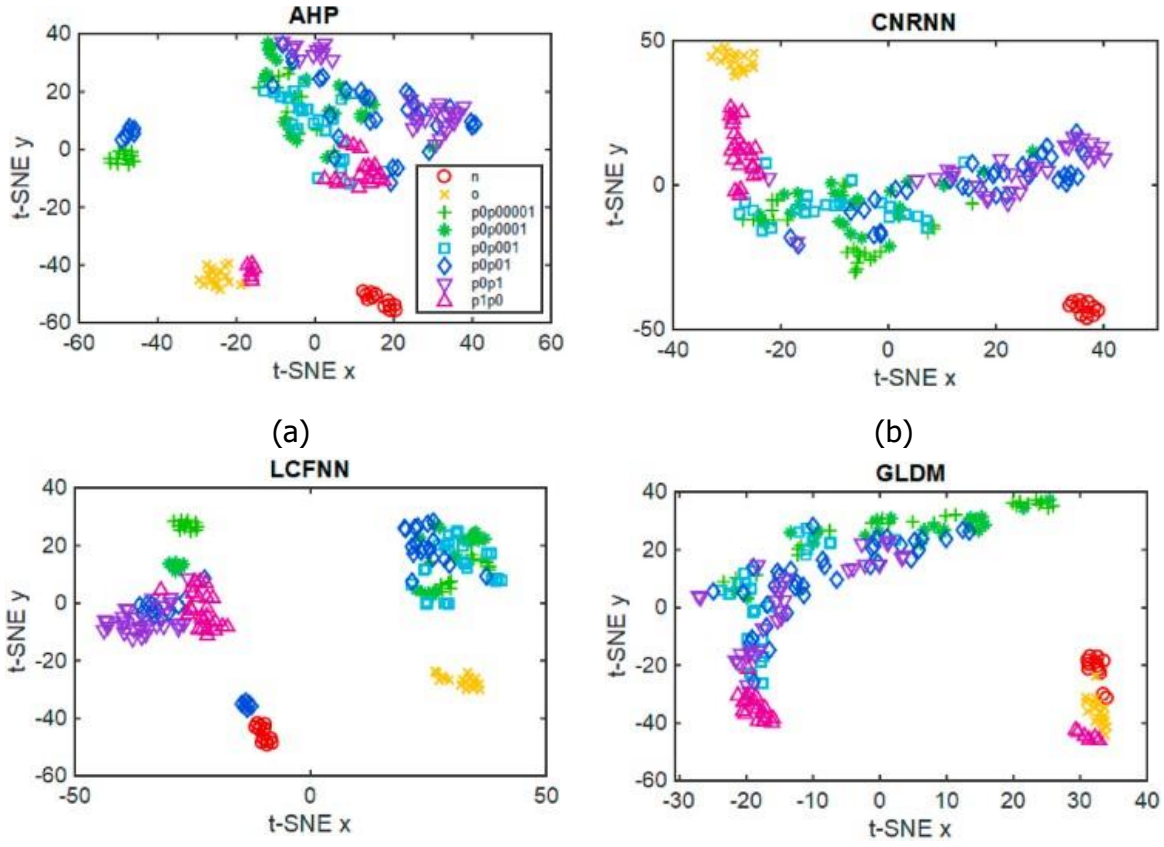


Figura 4.5 – Projeção t-SNE dos conjuntos de atributos (a) AHP, (b) CNRNN, (c) LCFNN e (d) GLDM para análise multiclasse.

Fonte: Elaborada pelo autor.

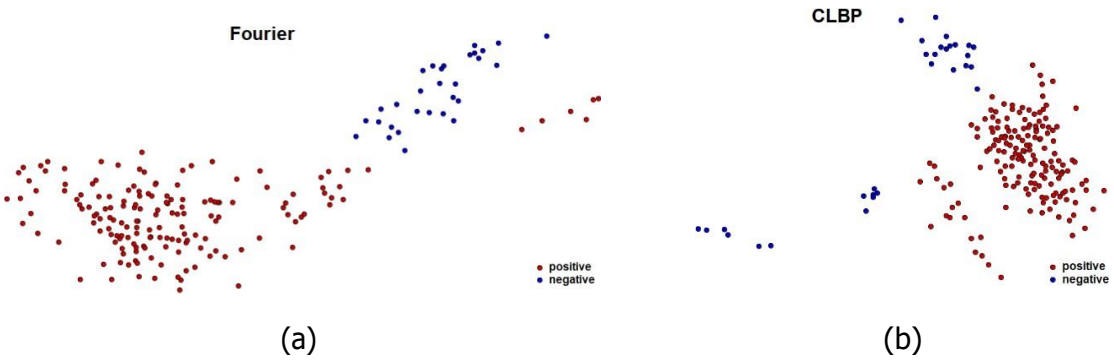
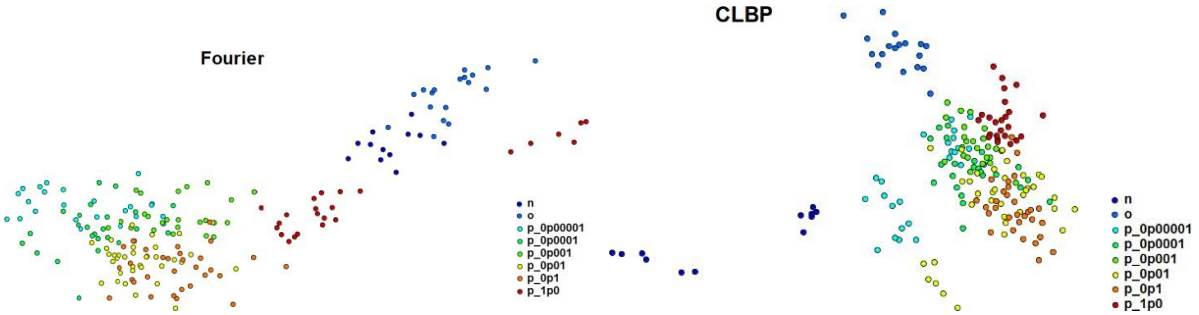


Figura 4.6 – Projeção IDMAP dos conjuntos de atributos (a) Fourier e (b) CLBP para análise binária.

Fonte: Elaborada pelo autor.

menos caros, e não microscópios eletrônicos, há evidências de outro trabalho do Grupo (96) que imagens com resolução de um microscópio óptico também servem para diagnóstico com genossensores. Como há possibilidade de se usarem sensores com padrões de topografia (ou morfologia) pré-determinados, como nos que há eletrodos interdigitados, é provável que mesmo



(a) (b)

Figura 4.7 – Projeção IDMAP dos conjuntos de atributos (a) Fourier e (b) CLBP para análise multiclasse.

Fonte: Elaborada pelo autor.

Tabela 4.6 – Acurácia média (desvio padrão) em (%) da análise de classificação computada com os algoritmos LDA, SVM e 1-NN para os casos binário e multiclasse para os conjuntos de atributos de textura extraídos das janelas com dimensões 300x300 pixels.

Conjuntos de atributos de textura	Binário			Multiclasse		
	LDA	SVM	1-NN	LDA	SVM	1-NN
GLDM	92.8	98.7	98.5	75.7	79.3	62.8
	(6.7)	(1.3)	(1.3)	(4.4)	(3.3)	(3.8)
Fourier	92.4	97.1	98.6	56.3	75.9	67.2
	(5.5)	(1.9)	(1.2)	(5.7)	(3.9)	(5.2)
CNTD	98.1	96.9	97.3	68.6	75.5	62.6
	(1.8)	(1.9)	(2.0)	(5.3)	(3.4)	(4.1)
Fractal	92.9	97.9	96.3	59.5	64.2	52.7
	(5.8)	(1.5)	(2.9)	(5.2)	(4.0)	(4.8)
AHP	98.9	98.8	98.4	66.7	78.2	68.4
	(1.3)	(1.3)	(1.3)	(6.2)	(3.7)	(4.9)
LBP	99.0	98.6	94.6	76.5	74.7	72.5
	(1.3)	(1.4)	(3.3)	(4.7)	(3.5)	(4.7)
CNRNN	98.1	98.9	98.9	72.7	82.4	70.4
	(1.8)	(1.0)	(1.0)	(4.3)	(3.5)	(3.7)
LCFNN	99.9	99.9	99.5	88.3	86.9	80.3
	(0.3)	(0.3)	(0.9)	(3.4)	(3.1)	(3.8)

Fonte: Elaborada pelo autor.

imagens de câmeras fotográficas forneçam informação suficiente para o processamento. Além disso, se o sensor for colorimétrico, deve ser suficiente processar imagens de um telefone celular, o que foi demonstrado por Elsa Materon no Grupo de Polímeros (resultados ainda não publicados). Essa é a expectativa de uma revolução na indústria de diagnóstico, pois fazer a detecção com imagens de sensores e biossensores, sem outro tipo de instrumento, permitirá diagnóstico em qualquer lugar e de maneira rápida. Obviamente que isso só se efetivará com o uso de processamento de imagens e AM como se fez neste capítulo para uma aplicação específica.

Não é possível fazer uma comparação direta da análise feita neste capítulo com outras da literatura, pois este é o primeiro trabalho em que se empregam imagens de biossensores para diagnóstico. Sabe-se, também, que a abordagem pode ser estendida a outros tipos de sensores e biossensores, o que já está sendo feito no Grupo de Polímeros Bernhard Gross do IFSC/USP para biossensores colorimétricos e sensores mecanocrômicos, em trabalhos cujos resultados ainda não estão publicados. A metáfora que consiste em fazer diagnóstico ou monitoramento a partir de imagens (ou vídeos) de sensores, com processamento de imagens e AM, é, portanto, promissora.

Quanto aos métodos computacionais empregados nas diversas tarefas do pipeline de AM desta aplicação, o pré-processamento com o recorte de janelas possibilitou a redução de dimensionalidade do problema com consequências para o aumento do número de exemplos de treinamento dos modelos. O uso de extratores de textura (93) na tarefa de engenharia de atributos trouxe uma melhora significativa para o desempenho dos classificadores, evidenciando que esse tipo de atributo das imagens é efetivamente revelador de padrões separáveis das classes de dados. Os extratores de textura CNRNN e LCNRNN (89), baseados na transformação de imagens em redes complexas em associação com redes neurais artificiais (ANN) (41), e sendo invariantes à rotação das imagens, conferem aos modelos maior robustez considerando-se que as imagens a serem classificadas podem sofrer com esse tipo de perturbação no momento da aquisição. Extratores de cor e forma em imagens, como utilizado no trabalho de Soares et. al (96) também devem ser considerados em aplicações envolvendo imagens. Também neste tipo de aplicação, vale ressaltar que a etapa de preparação dos dados é crucial para obtenção de classificadores de elevado desempenho.

Sobre a etapa de modelagem do pipeline, os algoritmos supervisionados lineares foram fixados com seus hiperparâmetros em valores iniciais (97), buscando ressaltar a capacidade dos conjuntos de dados em evidenciar padrões. Essa é uma abordagem típica da área de reconhecimento de padrões (98) sendo interessante quando se está conhecendo os dados em novas aplicações e desenvolvendo/aplicando novos extratores de atributos para os conjuntos de dados.

5 LÍNGUA ELETRÔNICA E ESPECTROSCOPIA DE IMPEDÂNCIA PARA DIAGNÓSTICO DE CÂNCER DE BOCA

5.1 Introdução

Uma das maiores vantagens da abordagem empregada nesta tese está na possibilidade de fazer diagnóstico com a identificação de padrões em condições bastante difíceis. De fato, num estudo pioneiro empregamos algoritmos de AM para diagnóstico de câncer de boca, a partir de medidas de espectroscopia de impedância com uma língua eletrônica. A língua eletrônica (99–101) é um dispositivo que mimetiza a língua humana, em que a percepção de sabor é dada por processamento de sinais de diferentes sensores na língua. Esses sensores não são específicos para um determinado sabor. O paladar é então definido por uma combinação de sabores básicos, ou seja, doce, amargo, azedo, salgado e umami.(102, 103) Cada líquido, portanto, pode ser definido por uma espécie de "impressão digital", que é o seu sinal elétrico característico. O funcionamento da língua eletrônica é, assim, baseado no conceito de seletividade global e não específica como num biossensor.(100) Por isso, uma língua eletrônica é constituída de um arranjo de diferentes sensores, cujas respostas elétricas são combinadas para gerar a tal "impressão digital". Algumas características das línguas eletrônicas as tormam interessantes para diferentes aplicações: elas são extremamente sensíveis, muito mais sensíveis que a língua humana para a detecção dos sabores básicos, por exemplo.(103) Além disso, como são eletrônicas, podem ser usadas com quaisquer tipos de líquidos, não só alimentos. Essa última possibilidade tem sido explorada para monitoramento de qualidade de águas e gosto amargo de remédios, com línguas eletrônicas. A ideia de usar uma língua eletrônica para diagnóstico já foi sugerida há tempos, mas somente agora identificaram-se métodos adequados, como será descrito aqui.

5.2 Conjunto de dados

O conjunto de dados utilizado nesta aplicação é composto de dados de 27 pacientes (n = 27) do Hospital de Câncer de Barretos (Barretos, SP/Brasil) com aprovação do seu comitê de ética (proc. 468/2011). Para cada paciente foram coletados 6 espectros de capacitância com uma língua eletrônica exposta à respectiva amostra de saliva, e também dados clínicos: fumante (não, ex, sim), alcoolismo (não, ex, sim), gênero (masculino, feminino) e idade. Os espectros de capacitância (Figura 5.1) têm valores de 19 capacitâncias (expressa na unidade nF) medidas no intervalo de frequências de 1 HZ a 1 MHz utilizando-se o Analisador de Impedâncias modelo 1260A em conjunto com interface dielétrica modelo 1296A (Solartron Analytical, Inglaterra). A língua eletrônica utilizada é microfluídica e de resposta única.(101) O dispositivo é constituído de 4 pares de microfios de aço recobertos com filmes de óxido. Esses sensores foram colocados em curto-circuito, formando-se assim um arranjo de capacitores em paralelo, e por isso há uma única resposta elétrica de impedância (ou capacitância). A aquisição dos espectros foi realizada

por Flávio M. Shimizu e Acelino C. Sá, pesquisadores do Grupo de Pesquisa em Polímeros (IFSC/USP).

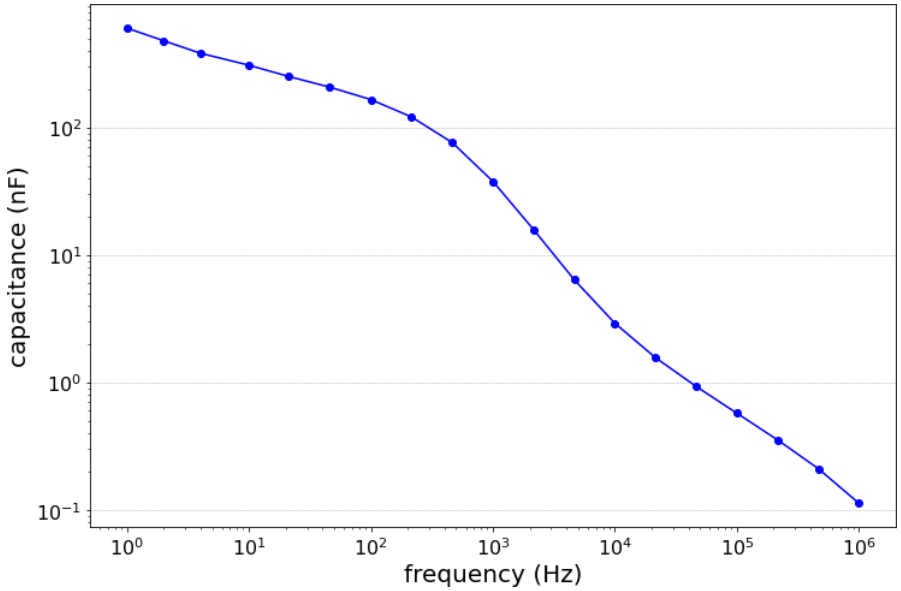
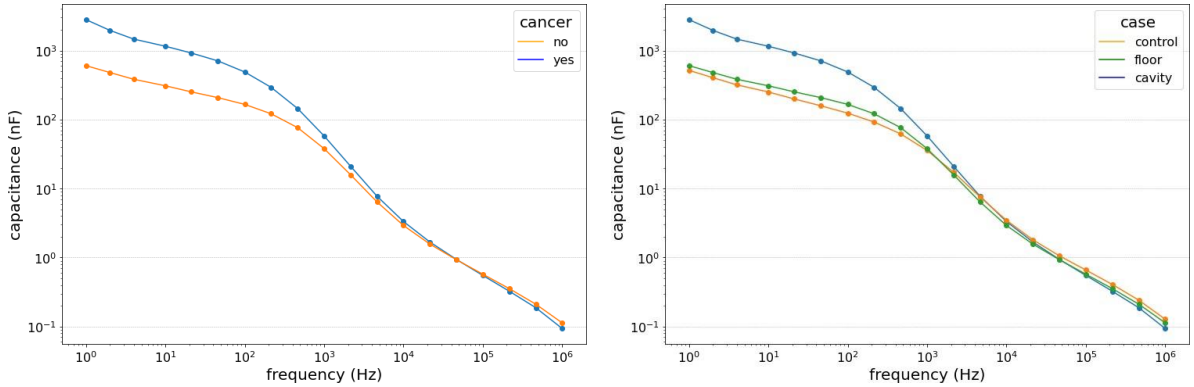


Figura 5.1 – Espectro de capacitâncias da língua eletrônica exposta à amostra de saliva de um paciente com câncer.

Fonte: Elaborada pelo autor.

Para avaliar a capacidade em diagnosticar câncer de boca em diferentes regiões, a língua eletrônica foi exposta a amostras de saliva de 14 pacientes sem câncer (não/controle) e 13 com câncer. Nos pacientes com câncer, os casos foram estudados para amostras coletadas no assoalho e em outros subsítios da cavidade da boca. Exemplos dos espectros para cada uma destas classes são apresentados na Figura 5.2.

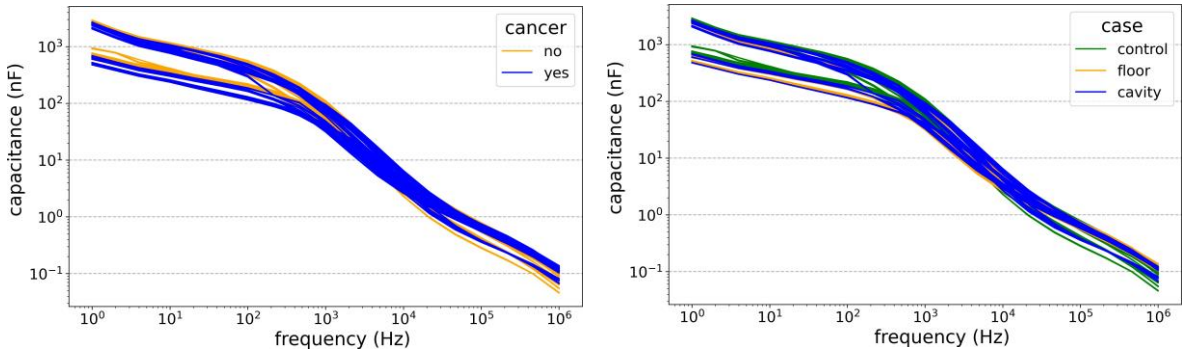


(a) (b)

Figura 5.2 – Espectros de capacitâncias para as classes (a) cancer (yes/no) e (b) caso (control/cavity/floor).

Fonte: Elaborada pelo autor.

Houve um problema na aquisição de dados, que foi feita em duas semanas diferentes. As medidas da segunda semana forneceram espectros de capacitância deslocados com relação às medidas da primeira semana, e isso pode ser verificado na Figura 5.3. Em situações normais com o uso de línguas eletrônicas, esses resultados teriam que ser descartados, pois esse deslocamento (drift) é um artefato experimental. Entretanto, esse problema foi considerado nesta tese como uma feliz coincidência, pois permite testar a viabilidade de um sistema inteligente de diagnóstico de lidar com erros sistemáticos de medida. Os resultados com AM supervisionado, discutidos neste capítulo, atestam essa viabilidade.



(a) (b)

Figura 5.3 – Conjunto dos espectros de capacitâncias para as classes (a) cancer (yes/no) e (b) caso (control/cavity/floor).

Fonte: Elaborada pelo autor.

A composição do conjunto de dados está sintetizada na Tabela 5.1, em que se mostra a quantidade de objetos (n = 27) e sua organização nas classes do diagnóstico de câncer (não, sim) e caso da amostra (controle, assoalho, cavidade) permitindo análises binária e multiclasse dos dados. Os objetos possuem atributos clínicos e originados do espectro de capacitâncias permitindo analisar separadamente as suas contribuições para o diagnóstico. Quanto à distribuição dos dados em relação às classes (Tabela 5.2), há uma quantidade equilibrada ou balanceada entre os objetos em relação às classes pertinentes ao diagnóstico de câncer. O mesmo não ocorre para as classes referentes ao caso da amostra. Para os atributos clínicos, a distribuição dos objetos (Tabela 5.3) está desbalanceada e não há objetos para algumas classes. Todas essas características devem ser consideradas no desenvolvimento do pipeline de AM para esta aplicação.

Para verificar a influência dos dados clínicos, foram feitas análises só com os valores de capacitância (only-sensor) - capacitâncias em 19 frequências - e considerando os dados clínicos (all-features) - fumante, alcoolismo, gênero e idade. Para análises só com dados da língua eletrônica (only-sensor), foram selecionados 5 conjuntos de atributos, como segue:

- f-all: todas as 19 frequências

Tabela 5.1 – Composição do conjunto de dados utilizado na aplicação Língua Eletrônica e EIE para Diagnóstico de Câncer de Boca.

Elemento	Descrição	Qtde.
Objetos	Dados de Pacientes	27
Atributos	Medidas de capacitância dos espectros	19
	Clínicos (fumante, alcoolismo, gênero e idade)	4
Classes	Câncer (não, sim)	2
	Caso da amostra (controle, assoalho, cavidade)	3

Fonte: BRAZ *et al.*(104)

Tabela 5.2 – Distribuição dos objetos entre as classes no conjunto utilizado na aplicação Língua Eletrônica e EIE para Diagnóstico de Câncer de Boca. Símbolo utilizado: número de objetos (n).

Classe		
Cancer	Caso da amostra (marca)	n
Não	Controle (controle)	14
Sim	Assoalho da boca (assoalho)	4
	Outros subsítios da cavidade da boca (cavidade)	9
	Total =	27

Fonte: BRAZ *et al.*(104)

Tabela 5.3 – Distribuição dos objetos em relação aos atributos clínicos e às classes. Símbolo utilizado: número de objetos (n).

Classe		n	Fumante			Alcoolismo			Gênero	
Câncer	Caso da amostra		Não	Ex	Sim	Não	Ex	Sim	Masc.	Fem.
Não	Controle	14	3	3	8	9	4	1	11	3
Sim	Assoalho	4	0	2	2	0	4	0	4	0
	Cavidade	9	1	3	5	1	2	6	7	2
	Total =	27	4	8	15	10	10	7	22	5

Fonte: BRAZ *et al.*(104)

- k-best: as 10 melhores frequências (1, 2, 4, 10, 21, 46, 100, 215 e 464 Hz, e 1 MHz) escolhidas nos testes com estatística

- f-low: frequências de 1 Hz a 100 Hz

- f-med: frequências de 100 Hz a 10 kHz
- f-high: frequências de 10 kHz a 1 MHz

5.3 Resultados

As Figura 5.4 e Figura 5.5 mostram os melhores resultados com redução de dimensionalidade e os métodos de projeção aplicados ao conjunto f-all contendo dados só de sensores (only-sensor). Os gráficos foram organizados em 2 colunas para facilitar a comparação para as classificações binárias e multiclasse. Os métodos usados foram *Principal Components Analysis (PCA)* (98), *Neighborhood Components Analysis (NCA)* (105), *t-distributed Stochastic Neighbor Embedding (t-SNE)* (94) e *Interactive Document Mapping (IDMAP)* (95). Como se observa, nenhum dos métodos é capaz de fazer uma separação razoável entre indivíduos saudáveis e com câncer.

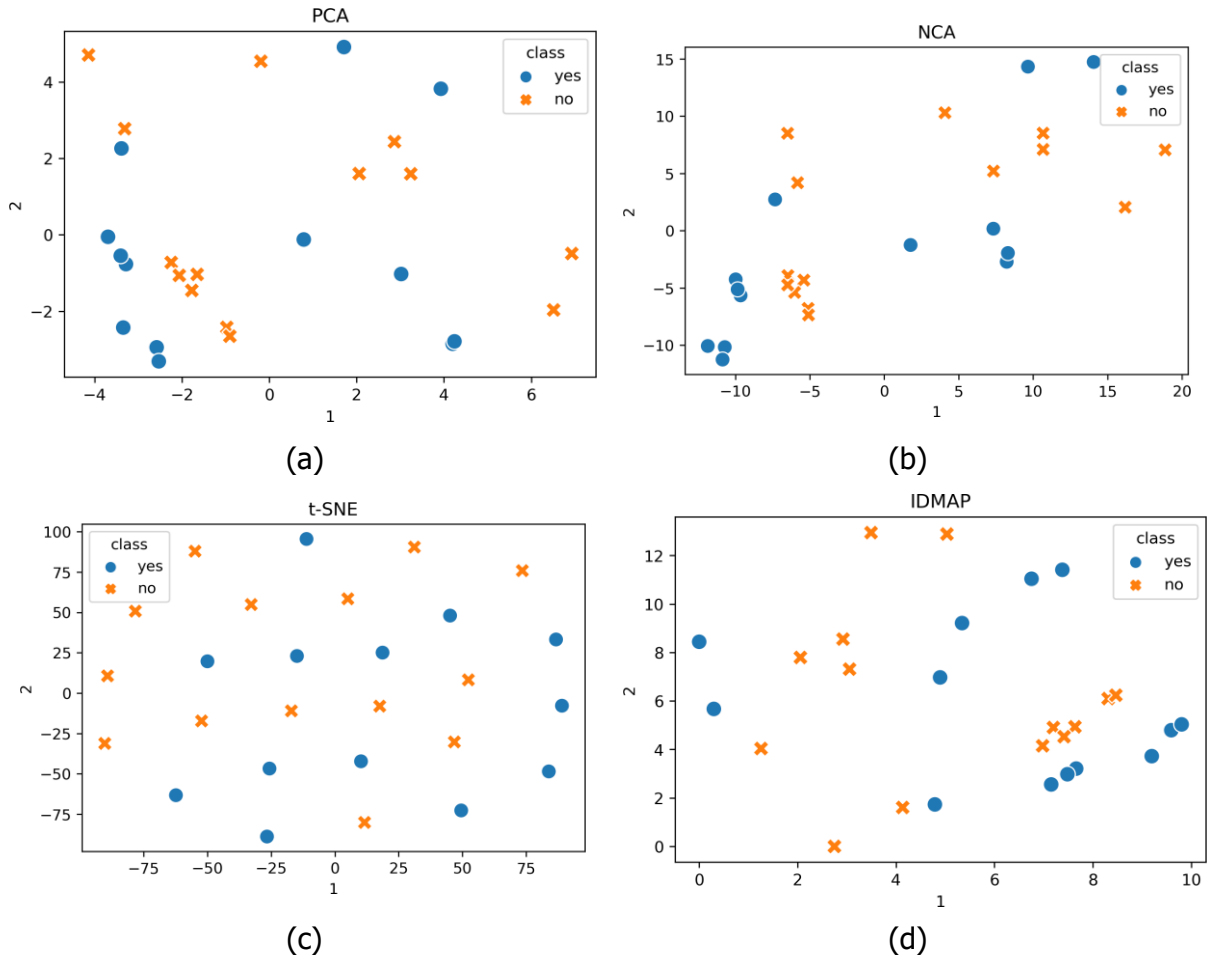


Figura 5.4 – Visualização com (a) PCA, (b) NCA, (c) t-SNE e (d) IDMAP para os dados de espectroscopia de impedância (only-sensor) para sim ou não para câncer.

Fonte: Elaborada pelo autor.

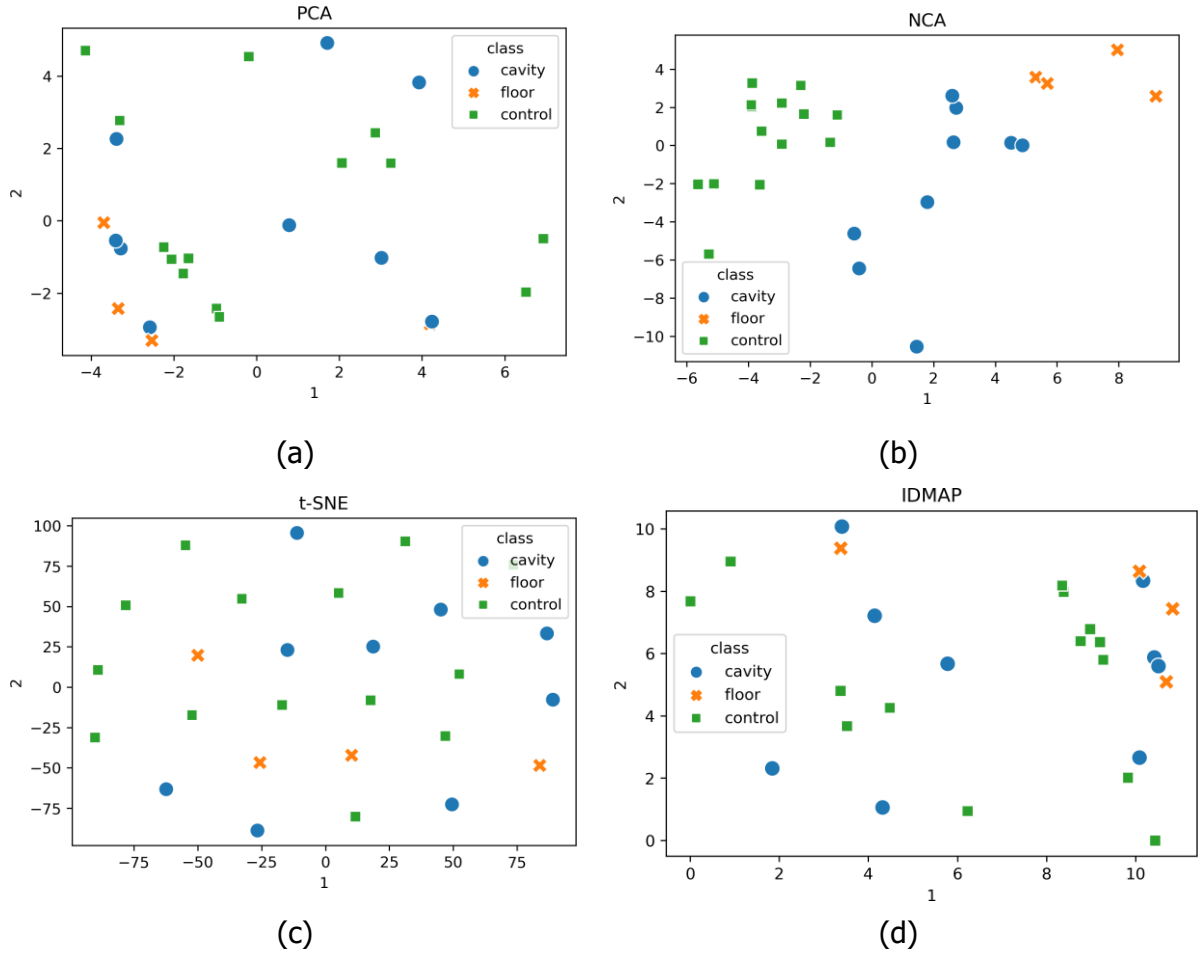


Figura 5.5 – Visualização usando (a) PCA, (b) NCA, (c) t-SNE and (d) IDMAP para os espectros de impedância (only-sensor) para o tipo de câncer.

Fonte: Elaborada pelo autor.

Agrupamento

Foram realizadas análises de agrupamento: (a) atributos selecionados ou extraídos apenas dos espectros, (b) atributos dos espectros vs todos atributos (espectro + clínicos).

Atributos only-sensor

O protocolo dos experimentos consiste nas seguintes etapas: 1) seleção aleatória (com reposição) de um espectro para cada um dos 27 pacientes; 2) agrupamento dos dados; 3) obter a *Average Silhouette Width (ASW)* (56). O experimento é repetido 100 vezes, obtendo-se então o valor médio e o desvio padrão. A ASW varia entre -1 e +1, sendo que quanto mais próximo de +1, maior a qualidade do agrupamento. A Tabela 5.4 mostra os valores de ASW para o agrupamento dos dados usando os algoritmos *K-Means (KM)* (41), *Hierarchical Agglomerative Clustering (HAC)* (41) e *Spectral Clustering (SC)* (46) com os conjuntos de atributos selecionados dos espectros (only-sensor), ou seja, f-all, k-best, f-low, f-med, f-high. Os valores na tabela representam a média (e desvio padrão) de 100 experimentos. Os maiores

valores de ASW ocorrem para 2 e 3 agrupamentos (*clusters*), respectivamente. Isso poderia ser um bom resultado, pois as amostras poderiam pertencer a 2 classes (yes/no) ou 3 classes (no, floor, cavity). Entretanto, esses altos valores são enganosos, como se pode visualizar nas figuras 5.4 e 5.5 que apresentam as projeções realizadas sobre o conjunto de atributos f-all. Dois grandes agrupamentos são formados, mas as classes yes/no estão misturadas, pois provavelmente a separação se deve às medidas de capacitância com drift e sem drift. De fato, esta hipótese é confirmada verificando-se os rótulos de todas as amostras dos pacientes. Em suma, os algoritmos de aprendizado não supervisionado conseguiram distinguir os grupos de amostras, mas foram "enganados"pelo artefato experimental do drift. Mencione-se que altos ASW significam em grande capacidade de separar os dados em grupos, o que pode não se traduzir em grande acurácia na previsão das classes corretas.

Tabela 5.4 – Valor médio da largura da silhueta (Average Silhouette Width) (ASW) para o agrupamento com os algoritmos KM, HAC e SC com os dados de espectroscopia de impedância (only-sensor).

Conjuntos de atributos	Algoritmos	Qtde. de agrupamentos	
		2	3
f-all	KM	0.873 (0.005)	0.722 (0.018)
	HAC	0.873 (0.005)	0.719 (0.022)
	SC	0.872 (0.005)	0.639 (0.023)
k-best	KM	0.873 (0.005)	0.723 (0.018)
	HAC	0.873 (0.005)	0.721 (0.021)
	SC	0.872 (0.005)	0.641 (0.023)
f-low	KM	0.876 (0.005)	0.731 (0.019)
	HAC	0.876 (0.005)	0.729 (0.022)
	SC	0.875 (0.005)	0.647 (0.022)
f-med	KM	0.705 (0.012)	0.666 (0.020)
	HAC	0.704 (0.014)	0.657 (0.033)
	SC	0.687 (0.028)	0.557 (0.044)
f-high	KM	0.580 (0.054)	0.528 (0.027)
	HAC	0.596 (0.058)	0.520 (0.034)
	SC	0.436 (0.033)	0.398 (0.058)

Fonte: Elaborada pelo autor.

Atributos all-features

Os resultados de agrupamento com os algoritmos de aprendizado não supervisionado KM, HAC, e SC são mostrados na Tabela 5.5, sendo a qualidade do agrupamento avaliada

com ASW. Diferentemente da análise anterior, cada paciente foi representado pela média dos espectros (média de 6 spectros) e os atributos foram padronizados para minimizar os efeitos do drift já mencionado. O desempenho dos algoritmos foi baixo, mesmo para a classificação binária, havendo uma pequena melhora quando todos os atributos são utilizados (dos espectros e de dados clínicos).

Tabela 5.5 – Média da largura da silhueta (Average Silhouette Width) (ASW) para o agrupamento com os algoritmos KM, HAC e SC para dados dos espectros (f-all) e com todos os atributos.

Conjuntos de atributos	Algoritmos	Qtde. de agrupamentos	
		2	3
only-sensor	KM	0.431	0.459
	HAC	0.443	0.452
	SC	0.443	0.394
all-features	KM	0.462	0.426
	HAC	0.462	0.428
	SC	0.453	0.401

Fonte: Elaborada pelo autor.

Classificação

Foram realizadas análises de classificação: (a) atributos selecionados ou extraídos apenas dos espectros (only-sensor), (b) atributos dos espectros vs todos atributos (espectro + clínicos) (all-features).

Atributos only-sensor

A análise realizada para verificar a capacidade dos atributos dos espectros em evidenciar padrões separáveis das classes cancer (yes/no) e caso (control, cavity, floor). O protocolo dos experimentos consiste nas seguintes etapas: 1) seleção aleatória (com reposição) de um espectro para cada um dos 27 pacientes; 2) classificação (com validação leave-one-out) dos dados; 3) obter as acurácias. O experimento é repetido 100 vezes, obtendo-se então o valor médio e o desvio padrão das acurácias. A Tabela 5.6 mostra a acurácia obtidas com os algoritmos *(41) 1-Nearest Neighbors (1-KNN), Linear Discriminant Analysis (LDA) e Support Vector Machine - with kernels linear (SVM-L)* aplicados aos vários conjuntos de atributos, como na análise com algoritmos de aprendizado não supervisionado. Valores de acurácia acima de 83% foram obtidos para a classificação binária para câncer (yes/no) nos conjuntos de atributos f-all, k-best e f-low com o algoritmo 1-NN. Para os dois últimos, em que se usam apenas frequências intermediárias (f-med) e altas (f-high), poderíamos esperar uma acurácia menor

porque a capacidade de distinção é menor nas frequências mais altas. A acurácia para a classificação multiclasse (control/cavity/floor) foi bem menor, na faixa de 50-65%, o que não é surpreendente porque o número de amostras de pacientes é pequeno.

Tabela 5.6 – Acurácia média (desvio padrão) em (%) da análise de classificação computada com os algoritmos LDA, SVM-L e 1-NN sobre o conjunto de atributos selecionados/extraídos apenas dos espectros (only-sensor) para classes cancer (yes/no) e case (control, cavity, floor).

Conjuntos de atributos only-sensor	Classificação					
	cancer (yes/no)			case (control/floor/cavity)		
	1-NN	SVM-L	LDA	1-NN	SVM-L	LDA
f-all	84.3	66.6	64.1	62.5	49.3	46.3
	(4.3)	(5.0)	(10.2)	(4.6)	(4.0)	(11.8)
k-best	84.5	57.5	70.2	63.0	46.4	49.0
	(4.2)	(4.1)	(7.1)	(4.5)	(4.8)	(9.1)
f-low	83.9	63.7	75.5	63.6	46.0	50.9
	(4.3)	(5.3)	(5.7)	(4.6)	(2.0)	(7.1)
f-med	79.2	51.3	58.6	55.0	42.9	43.9
	(7.2)	(5.0)	(7.0)	(8.4)	(3.4)	(6.6)
f-high	54.7	71.7	70.0	42.0	58.7	57.4
	(9.3)	(4.6)	(7.1)	(8.3)	(4.3)	(6.2)

Fonte: Elaborada pelo autor.

Atributos all-features

A Tabela 5.7 mostra a acurácia média (desvio padrão) para os algoritmos (41) *Logistic Regression (LR), Linear Discriminant Analysis (LDA), Gaussian Naive-Bayes (GNB), K-Nearest Neighbors (KNN), Support Vector Machine - with kernels linear (SVM-L), polynomial (SVM-P), radial (SVM-R)* e *Random Forest (RF)* aplicados nos conjuntos de atributos só dos espectros de impedância (only-sensor) e com todos os atributos (all-features) para as classificações binárias e multiclasse. Na tarefa do pipeline de pré-processamento, foi calculada a média dos espectros de cada paciente, resultando em 27 espectros. Para engenharia de atributos, o método *label encoding* (106) foi aplicado para codificação de cada atributo clínico. Em todos os modelos, a acurácia média foi calculada com a validação 10 × 5 *Nested K-Fold Cross-Validation* (kouter = 10 and kinner = 5). (61) Como nos experimentos de agrupamento, os atributos dos espectros foram padronizados, com exceção dos modelos de RF que empregaram os dados sem pré-processamento. Altos valores de acurácia foram obtidos na classificação binária com os algoritmos SVM-R e RF (acurácias consideradas equivalentes em virtude da dispersão dos

valores). Como esperado, na classificação multiclasse a acurácia foi bem menor, o que se vê na última coluna da Tabela 5.7. O algoritmo mais eficiente foi RF, com um pequeno aumento de acurácia quando se incluíam os dados clínicos.

Tabela 5.7 – Acurácia média (desvio padrão) para a classificação com os algoritmos LR, LDA, GNB, KNN, SVM-L/R/P e RF para os conjuntos de atributos only-sensor e all-features.

Conjuntos de atributos	Algoritmos	Classificação	
		cancer (yes/no)	case (control/floor/cavity)
only-sensor	LR	0.700 (0.221)	0.481 (0.105)
	LDA	0.717 (0.248)	0.519 (0.105)
	GNB	0.433 (0.249)	0.370 (0.052)
	KNN	0.783 (0.224)	0.630 (0.105)
	SVM-L	0.667 (0.197)	0.481 (0.052)
	SVM-P	0.717 (0.248)	0.519 (0.052)
	SVM-R	0.867 (0.208)	0.519 (0.052)
	RF	0.800 (0.256)	0.630 (0.052)
all-features	LR	0.650 (0.252)	0.556 (0.091)
	LDA	0.633 (0.306)	0.556 (0.240)
	GNB	0.567 (0.291)	0.556 (0.157)
	KNN	0.617 (0.373)	0.519 (0.139)
	SVM-L	0.733 (0.238)	0.556 (0.091)
	SVM-P	0.733 (0.186)	0.519 (0.139)
	SVM-R	0.767 (0.200)	0.519 (0.052)
	RF	0.800 (0.256)	0.667 (0.091)

Fonte: Elaborada pelo autor.

5.4 Conclusão

A confirmação de que é possível fazer diagnóstico com uma língua eletrônica, sem detectar um biomarcador específico, traz grandes perspectivas para sistemas automatizados de diagnósticos. É uma demonstração da adequação da estratégia de busca por padrões, que pode ser feita de maneira sistemática. Essa adequação talvez pudesse ser esperada porque as línguas eletrônicas funcionam segundo o conceito de seletividade global, justamente baseada no estabelecimento de padrões. Entretanto, encontrar padrões pode ser difícil, como foi evidenciado aqui com a dificuldade em distinguir os espectros de impedância obtidos de salivas de pacientes

voluntários com e sem câncer de boca, empregando-se técnicas de visualização ou algoritmos de AM não supervisionados.

No caso dos dados de impedância analisados nesta tese, havia a complicação do drift num dos conjuntos de medidas, que gerava uma separação artificial de amostras de saliva que deveriam ter o mesmo comportamento elétrico nas medidas com a língua eletrônica. O emprego de AM supervisionado sobre dados pré-processados resolveu esse problema, o que é marcante, pois o sistema conseguiu identificar diferenças intrínsecas das amostras ou causadas pelo drift. Esse último resultado é promissor por um motivo adicional.

Sabe-se que aplicações de línguas eletrônicas não são rotineiras devido à dificuldade de se substituir unidades sensoriais de um conjunto de sensores, sem ter que repetir todo o processo de calibração. Ou seja, se uma língua eletrônica é utilizada para construir uma biblioteca de padrões, por exemplo de diferentes tipos de vinho ou café, quando o dispositivo é substituído (um ou mais sensores da língua), a resposta elétrica para aqueles objetos da biblioteca será alterada. Significa dizer que todo o trabalho de determinação de padrões precisa ser repetido usando a nova língua eletrônica para o mesmo conjunto de amostras. Percebe-se que é impraticável ter uma aplicação comercial com línguas eletrônicas cujos resultados não são reprodutíveis. Conseguir alta reprodutibilidade não é factível para muitas línguas eletrônicas, principalmente as fabricadas com materiais orgânicos. É o preço a pagar pela alta sensibilidade que essas línguas apresentam.

A partir deste trabalho, pode-se antever a possibilidade de fazer calibração de nova língua com um número limitado de experimentos, pois um sistema computacional poderia "aprender"o comportamento geral da língua. Mais especificamente, poder-se-ia usar a metodologia de espaços multidimensionais de calibração (19), que foi explorada complementarmente no artigo publicado.(104) Uma constatação importante sobre a classificação binária para câncer (yes/no) dos espectros (conjunto only-sensor) foi a de que os desempenhos dos atributos f-all e f-low são similares e acima de 83% de acurácia, podendo-se inferir que medidas em frequências baixas são tão representativas das classes quanto medidas no espectro completo. Como consequência, pode-se trabalhar com um número de atributos menor no modelo de AM. Esse resultado é consistente com a experiência com sensores e biossensores baseados em impedância elétrica, pois o sensoriamento depende das interações moleculares que afetam a dupla-camada elétrica, cujos mecanismos são os mais relevantes para frequências baixas.

Outro resultado deste capítulo pode ter implicações relevantes. Trata-se da observação de que a qualidade do diagnóstico pode ser melhorada ao combinar dados de impedância da língua eletrônica com dados clínicos de pacientes. Embora isso seja intuitivo, e até tratado como premissa na concepção dos sistemas inteligentes de diagnóstico, não é simples ter uma demonstração inequívoca. Neste trabalho (104) identificou-se, inclusive, qual fator clínico pode ter correlação positiva com câncer de boca. Como a quantidade de amostras era pequena, e a acurácia do diagnóstico limitada (abaixo de 90%), não foi possível alcançar uma conclusão

definitiva. O resultado, porém, indica que a abordagem é promissora e pode ser melhorada se forem empregados algoritmos de AM baseados em regras, como árvores de decisão (DT) e florestas (RF). Entre outras coisas, porque a relevância dos diferentes tipos de informação pode ser inferida, principalmente se for empregado o conceito de espaço de calibração multidimensional.(19)

Provavelmente, para se obter melhores resultados na classificação das amostras, será necessário gerar um conjunto de dados maior e mais abrangente no sentido de incorporar, tanto quanto possível, a diversidade de fatores que fazem parte do experimento real de detecção. Esta é uma conclusão que pode ser estendida para todas as aplicações desta tese: a aquisição dos dados precisa ser bem planejada para a obtenção de bons resultados com o emprego de AM. Nesta tese, apenas a aplicação relacionada à detecção do biomarcador de câncer p53 com medidas voltamétricas foi implementada com essa preocupação. Ainda na etapa do pipeline de preparação dos dados, há possibilidades de explorar novos métodos de pré-processamento e engenharia de atributos sobre as curvas espectrais. Os algoritmos SVM-R kernel radial e RF geraram os melhores desempenhos na classificação. Modelos gerados por esses algoritmos possuem características de não-linearidade (41) que favorecem a separação de classes em dados com organização complexa como é o caso. Especificamente sobre RF, conforme já apresentado, seu resultado é produto da combinação (*ensemble*) das classificações de múltiplas árvores de decisão.

Na literatura, o uso de AM para analisar dados de língua eletrônica já é relativamente antigo. Em 2004, Riul e colaboradores (107) usaram redes neurais para classificar diferentes tipos de vinho. Em 2007, Ferreira e co-autores (108) usaram AM para correlacionar dados de impedância de uma língua eletrônica com o sabor segundo percepção de degustadores de café. Entretanto, nesta tese é a primeira vez que se combinam diferentes tipos de informação para o diagnóstico de uma doença, como o câncer. De certa forma, a contribuição aqui apresentada tem semelhanças com as de Nicoliche *et al.* (109), em que AM é usado para processar dados de sensores.

6 IMUNOSSENSOR E VOLTAMETRIA PARA DIAGNÓSTICO DE CÂNCER

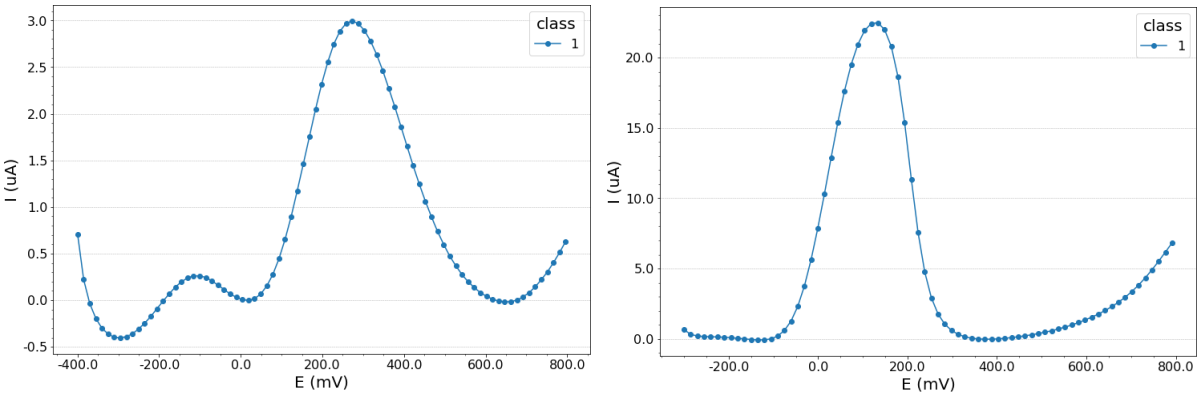
6.1 Introdução

Técnicas eletroquímicas são amplamente empregadas em biossensores, para detecção de uma grande diversidade de analitos, incluindo biomarcadores de câncer. O método de medida pode ser voltametria, de diferentes modos, amperometria ou espectroscopia de impedância eletroquímica. Em todos esses métodos, busca-se medir a variação de uma corrente ou impedância com a concentração do analito, sendo comportamentos típicos um aumento linear ou com o logaritmo da concentração. A sensibilidade é geralmente garantida com uma matriz ativa no biossensor que interage forte e seletivamente com o analito da amostra sendo considerada. Quando se utiliza voltametria, observam-se as correntes de pico de redução e oxidação, e variações nessas correntes ou nos potenciais em que os picos ocorrem são usadas no sensoriamento. Esse tipo de sensoriamento pode enfrentar problemas quando o analito está numa matriz complexa, como num fluido biológico (por exemplo, urina, saliva, sangue). Pois além de possíveis interferências nos picos, causadas por outras substâncias no fluido além do analito, alterações adicionais podem aparecer nos voltamogramas. Normalmente essas alterações não são levadas em conta, pois podem depender de múltiplos fatores. Assim, parte da informação dos voltamogramas é desprezada nos tratamentos tradicionais de química analítica e de eletroquímica. A abordagem desta tese foi concebida para estender a capacidade de análise e permitir o aproveitamento de todo tipo de informação dos sinais eletroquímicos. Por isso, foram concebidos experimentos de voltametria com um imunossensor para detectar o biomarcador de câncer p53. Este biomarcador é um antígeno, relacionado a diferentes tipos de câncer. Por ser um antígeno, sua detecção pode ser feita com um imunossensor, em que se imobilizam anticorpos anti-p53. A interação específica antígeno-anticorpo gera uma alteração no sinal eletroquímico, empregada como detecção do biomarcador.

Ao contrário das outras aplicações nesta tese, em que a decisão de utilizar AM no processamento de dados se deu após a aquisição dos dados, neste capítulo relatamos resultados de experimentos planejados. Por isso, ao invés de fazer medidas em triplicadas com os imunossensores, foram realizadas muito mais medidas. Isso também permitiu comparar o desempenho de diferentes tipos de imunossensores. Foram usados imunossensores com eletrodos de carbono impresso, um dos quais recebeu uma camada de um polímero, o poli(etileno imina) (PEI). As sondas redox foram ácido ascórbico (AA) para o sensor de eletrodos impressos de carbono (SPCE) sem recobrimento com PEI e hexacianoferrato de potássio para o sensor com PEI. Para facilitar, faremos referência aos dois tipos de sensores como AA e PEI. Na análise dos resultados, repetimos a estratégia geral: os dados são visualizados com técnicas de projeções multidimensionais, e depois são usados como entrada para os algoritmos de AM.

6.2 Conjunto de dados

Nesta aplicação foram utilizados dois conjuntos de dados gerados da aplicação da técnica de voltametria sobre os imunossensores AA e PEI. Ambos os conjuntos são compostos de 40 voltamogramas (n = 40) contendo medidas de corrente elétrica (μA) em função da tensão elétrica (mV) aplicada conforme exemplos da Figura 6.1. Para o imunossensor AA, os voltamogramas possuem 81 medidas no intervalo de -399.933 a 796.356 mV. Para o imunossensor PEI, os voltamogramas possuem 74 medidas no intervalo de -299.988 a 791.626 mV. A preparação e disponibilização dos conjuntos foi feita por Gisela Ibáñez-Redín, pesquisadora do Grupo de Pesquisa em Polímeros (IFSC/USP).



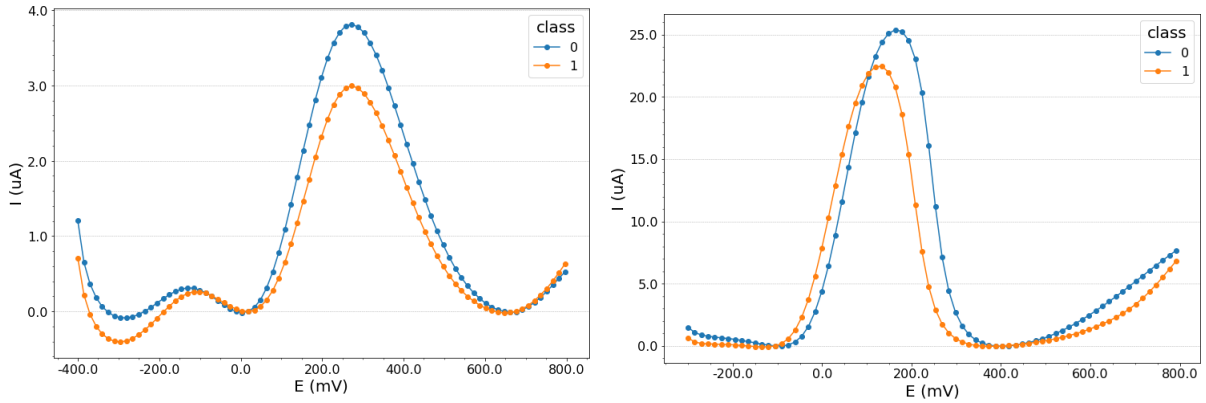
(a) (b)

Figura 6.1 – Voltamogramas de pulso diferencial do imunossensor AA (a) e PEI (b) para amostras contendo biomarcador p53 para câncer.

Fonte: Elaborada pelo autor.

A capacidade dos imunossensores de detectar o biomarcador p53 em matrizes complexas foi avaliada usando amostras de saliva e urina artificiais. Uma parte das amostras foi modificada com a adição do biomarcador p53 e a outra foi preservada para ser utilizada como controle. Exemplos de voltamogramas desses tipos de amostras e imunossensores são apresentados na Figura 6.2.

Cada voltamograma no conjunto de dados AA possui 81 medidas de corrente elétrica realizadas no intervalo de -400 a 800 mV. No conjuto PEI, os voltamogramas possuem 76 pontos medidos no intervalo de -300 a 800 mV. As curvas já foram fornecidas com as respectivas linhas de base removidas, não necessitando de pré-processamentos adicionais. As formas das curvas evidenciam regiões de picos cujas características podem ser aproveitadas para diferenciação das classes de amostras. A Tabela 6.1 mostra uma síntese dos conjuntos de dados utilizados nas análises. A quantidade de atributos é maior que a de objetos implicando na necessidade de aplicar técnicas de engenharia de atributos para redução de dimensionalidade visando evitar o sobreajuste dos modelos de AM. Verifica-se pela Tabela 6.2 que a distribuição dos objetos entre as classes está balanceada.



(a) (b)

Figura 6.2 – Voltamogramas de pulso diferencial do imunossensor AA (a) e PEI (b) para as classes controle (0) e com a presença (1) do biomarcador p53 para câncer.

Fonte: Elaborada pelo autor.

Tabela 6.1 – Composição do conjunto de dados dos imunossensores AA e PEI.

Elemento	Descrição	Qtde.	
		AA	PEI
Objetos	Curvas de VPD	40	40
Atributos	Medidas de corrente elétrica	81	76
Classes	Controle, Biomarcador p53	2	2

Fonte: Elaborada pelo autor.

Tabela 6.2 – Distribuição do número de objetos (n) entre as classes de dados dos imunossensores.

Analito	Classe	Marca	n
Ausência do biomarcador p53	Controle	0	20
Presença do biomarcador p53	Biomarcador p53	1	20
		Total =	40

Fonte: Elaborada pelo autor.

Engenharia de atributos

Para cada tipo de biossensor foram gerados três conjuntos de dados adicionais a partir dos dados originais (AA-raw, PEI-raw). Os procedimentos e métodos utilizados são apresentados a seguir.

- AA-mi, PEI-mi: com o método de seleção de atributos por filtragem e a estatística *Multual Information (MI)* (54, 110) aplicados ao conjunto AA-raw, foram selecionados

5 atributos com melhor desempenho: intensidades de corrente nos potenciais 317,841, 272,98, 302,887, 287,933, e 258,026 mV. Aplicando-se o mesmo procedimento ao conjunto PEI-raw, foram selecionados os seguintes atributos: correntes elétricas nos potenciais 761,719, 776,672, 791,626, 746,765, 731,812 mV.

- AA-stats, PEI-stats: formados com a extração de 7 atributos de estatística dos voltamogramas: média, desvio padrão, máximo, mínimo, e quartis com 25%, 50%, e 75% das intensidades de corrente.

- AA-peak, PEI-peak: formados com a extração de 5 atributos do principal pico dos voltamogramas, alguns dos quais são comumente usados na análise de sinais eletroquímicos (111, 112): posição do pico, proeminência, largura à meia altura, largura na base e fator de qualidade (= proeminência/largura à meia altura).

6.3 Resultados

A partir dos objetivos desta aplicação e seguindo o pipeline de AM, foram realizadas análises exploratórias e de visualização de dados para verificação. As análises iniciaram com a exploração dos dados através de gráficos e medidas estatísticas apresentados no Apêndice C. A Figura 6.3 mostra os gráficos 2D em que os voltamogramas foram projetados como pontos a partir de diferentes técnicas de projeção para o sensor AA. Nota-se que, apesar da dispersão dos resultados, há boa separação entre as amostras positivas (para p53) e negativas. Da mesma forma, há separação - embora não tão evidente - dos dados nos gráficos gerados para o sensor PEI na Figura 6.4.

Agrupamento

A distinção entre amostras positivas e negativas para p53 foi testada com algoritmos não supervisionados, KM, HAC e SC. A Tabela 6.3 mostra os valores de silhueta (ASW) para os dois tipos de sensores (AA e PEI) e diferentes conjuntos de dados. Quando utilizados os voltamogramas completos (AA-raw e PEI-raw), sem seleção de atributos, o desempenho dos algoritmos foi baixo, provavelmente devido à dispersão dos resultados, como indicado pelas projeções das figuras 6.3 e 6.4. O máximo valor de ASW foi 0,413. O desempenho melhorou muito considerando-se apenas as características do voltamograma mais definidoras da interação antígeno-anticorpo. Não é surpreendente que usar o valor da corrente de pico gere o maior valor de ASW, 0,841, para o sensor AA. Como também já era esperado, a capacidade de distinção é maior para o sensor AA do que para o PEI.

Classificação

A Tabela 6.4 mostra os resultados dos experimentos de classificação com os algoritmos de AM supervisionados LR, LDA, SVM-L, GNB, KNN e DT. Os hyperparâmetros foram mantidos

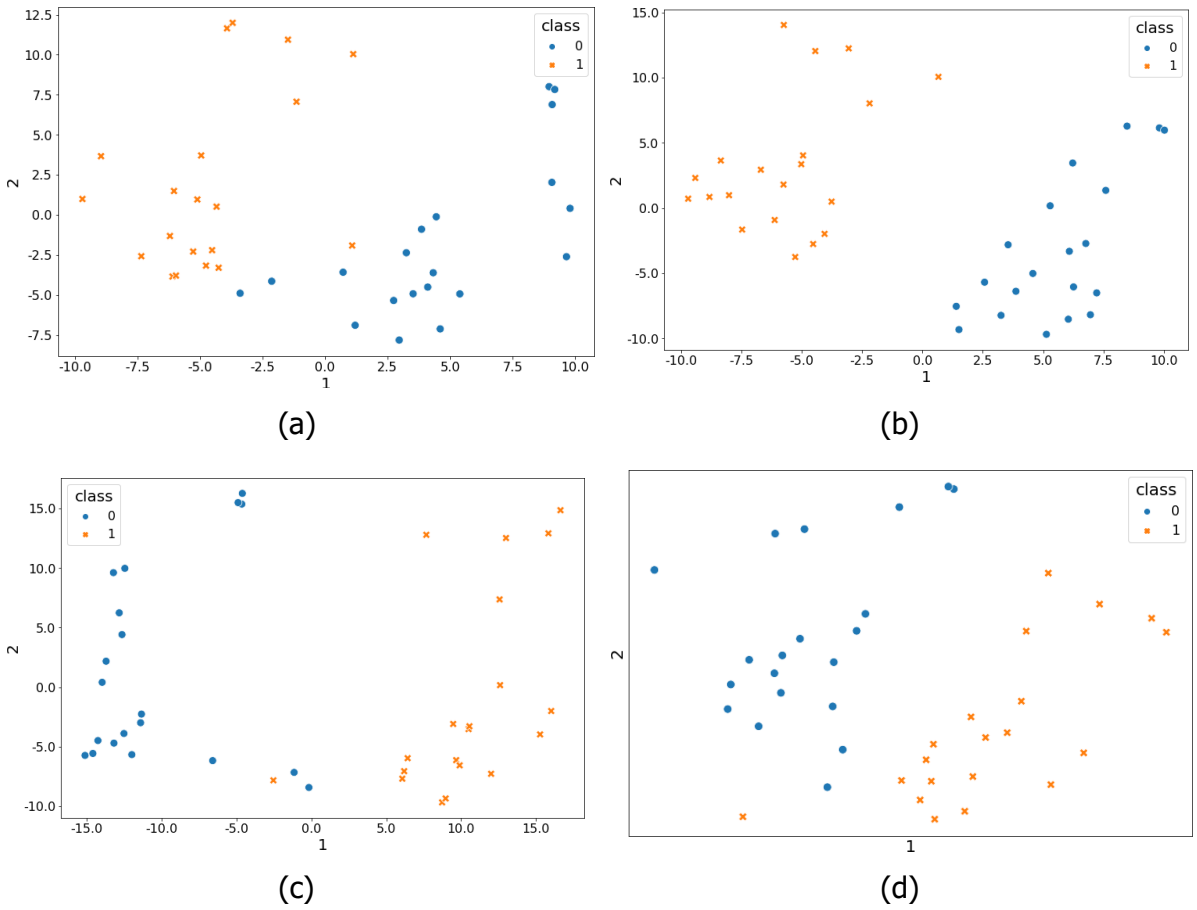


Figura 6.3 – Visualização do conjunto de atributos AA-raw com (a) PCA, (b) NCA, (c) Isomap e (d) IDMAP. As classes 0 e 1 correspondem a amostras negativas e positivas, respectivamente.

Fonte: Elaborada pelo autor.

Tabela 6.3 – Largura média da silhueta (Average Silhouette Width) (ASW) para 2 agrupamentos com os algoritmos KM, HAC e SC para todos os conjuntos de dados.

Conjunto	KM	HAC	SC
AA-raw	0.395	0.413	0.359
AA-mi	0.796	0.796	0.796
AA-peak	0.841	0.841	0.133
AA-stats	0.523	0.527	0.527
PEI-raw	0.391	0.391	0.369
PEI-mi	0.643	0.642	0.642
PEI-peak	0.429	0.344	0.429
PEI-stats	0.501	0.460	0.514

Fonte: Elaborada pelo autor.

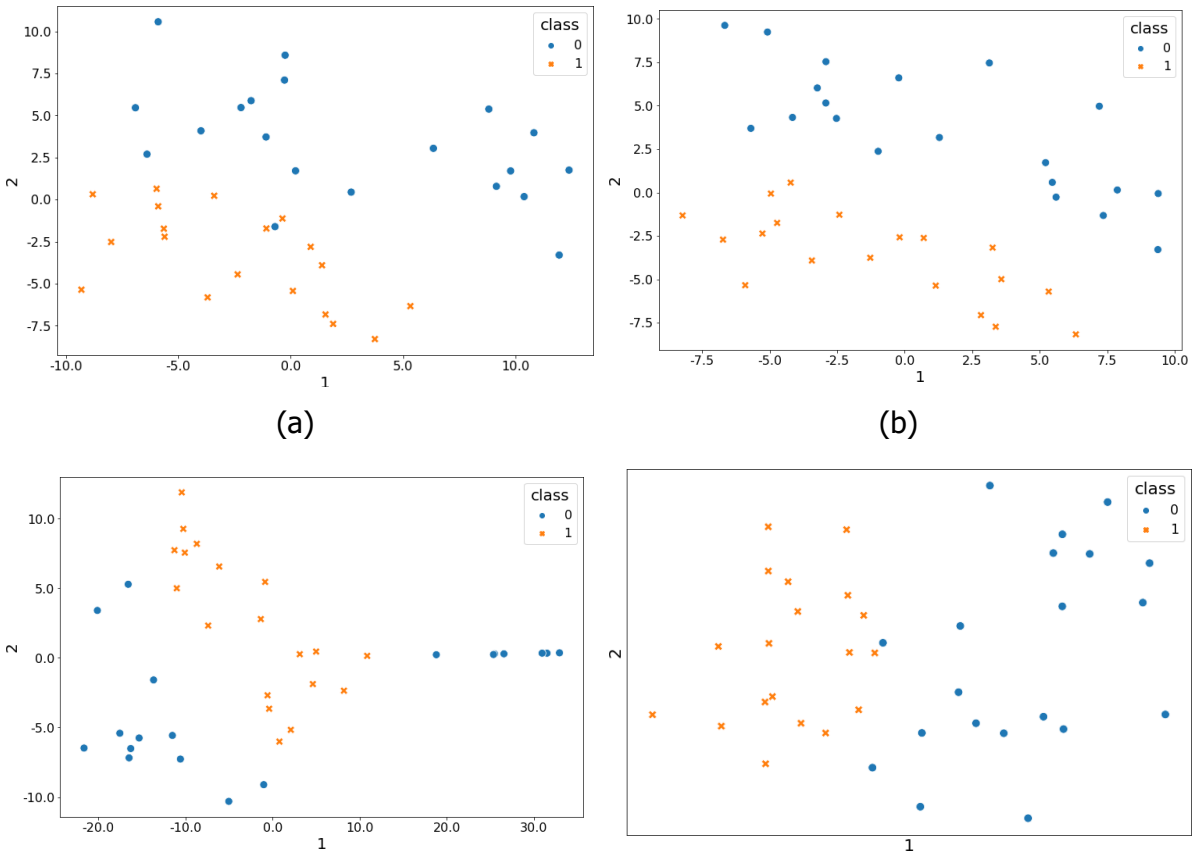


Figura 6.4 – Visualização do conjunto de atributos PEI-raw com (a) PCA, (b) NCA, (c) Isomap e (d) IDMAP. As classes 0 e 1 correspondem às amostras negativas e positivas, respectivamente.

Fonte: Elaborada pelo autor.

em seus valores iniciais dados pela biblioteca sklearn. (97) O alto desempenho com acurácia de 100% com os vários algoritmos indica que essa tarefa de detecção do biomarcador p53 é mais simples que as outras estudadas nesta tese, principalmente porque foi considerada apenas a classificação binária. Os resultados, de toda forma, mostram a utilidade da abordagem com AM, pois permitiram identificar inequivocamente o imunossensor com desempenho otimizado, e indicaram como explorar os voltamogramas para além do que normalmente é feito em química analítica e eletroquímica. De fato, esse alto desempenho na detecção de p53 a partir do processamento dos voltamogramas confirma a premissa inicial de que há informações relevantes nas características dos voltamogramas, além dos picos de oxidação e redução. Inferir conhecimento a partir do voltamograma completo não é simples devido aos vários fatores que o afetam. Com a abordagem desta tese, esse trabalho fica facilitado pois não há restrição quanto à quantidade de dados que podem ser processados. Pelo contrário, quanto maior a quantidade de dados, mais fácil é para se encontrar padrões com os algoritmos de AM. Ressalte-se, neste aspecto, que os experimentos de voltametria foram planejados em número muito maior do que

normalmente se faz. Ao invés de fazer experimentos em triplicata o mesmo sensor, como é habitual, para cada tipo de sensor (AA e PEI) foram feitos 20 experimentos para cada condição com e sem p53. Cada experimento utilizou uma unidade diferente do sensor, o que permitiu obter resultados mais representativos por AM. Foi possível, também, verificar que a detecção de p53 pode se dar igualmente em urina ou saliva, resultado que não era conhecido na literatura.

Tabela 6.4 – Acurácia média (desvio padrão) (%) para a classificação usando os algoritmos LR, LDA, SVM-L, GNB, KNN e DT para todos os conjuntos de dados.

Conjunto	LR	LDA	SVM-L	GNB	KNN	DT
AA-raw	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
AA-mi	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
AA-peak	95.0 (2.2)	100.0 (0.0)	95.0 (2.2)	100.0 (0.0)	97.5 (1.6)	100.0 (0.0)
AA-stats	97.5 (1.6)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
PEI-raw	95.0 (2.2)	97.5 (1.6)	97.5 (1.6)	95.0 (2.2)	92.5 (2.6)	87.5 (3.3)
PEI-mi	90.0 (3.0)	87.5 (3.3)	92.5 (2.6)	90.0 (3.0)	92.5 (2.6)	80.0 (4.0)
PEI-peak	80.0 (4.0)	80.0 (4.0)	80.0 (4.0)	80.0 (4.0)	72.5 (4.5)	62.5 (4.8)
PEI-stats	92.5 (2.6)	100.0 (0.0)	97.5 (1.6)	92.5 (2.6)	95.0 (2.2)	92.5 (2.6)

Fonte: Elaborada pelo autor.

6.4 Conclusão

O planejamento dos experimentos com a obtenção de uma grande quantidade de dados (voltamogramas), muito acima do que se costuma fazer com imunossensores eletroquímicos, surtiu excelente resultado para a análise com AM. De fato, uma alta acurácia na classificação das amostras contendo p53 foi obtida. Ressalta-se que todos os cuidados necessários foram adotados para evitar o sobreajuste dos modelos durante a etapa de modelagem do pipeline. Mesmo os algoritmos de AM lineares (LR, LDA e SVM-L) obtiveram elevado desempenho, evidenciando a qualidade dos conjuntos de dados em evidenciar padrões separáveis das classes.

Ainda em relação aos conjuntos de dados processados, um resultado importante sobre a classificação dos voltamogramas dentro do espaço original de atributos (*i.e.* intensidade de corrente x potencial elétrico) contidos nos conjuntos com -raw e -mi foi a verificação de que os desempenhos obtidos são similares mesmo tendo quantidades de atributos bem diferentes. Os conjuntos -mi possuem apenas 5 atributos enquanto que os conjuntos -raw possuem acima de 75 atributos, representando uma dimensionalidade 15 vezes menor e uma maior simplificação dos modelos de AM. Há oportunidades ainda não exploradas na etapa do pipeline de preparação dos dados, como novos métodos de pré-processamento e engenharia de atributos sobre voltamogramas.

Também é relevante destacar que os resultados de classificação foram consistentemente melhores para os sensores AA do que para o PEI. Essa conclusão guiará estudos futuros de sensores eletroquímicos, que é um dos objetivos do uso de sistemas de processamento de dados com AM, ou seja, a de otimização de materiais e condições experimentais para melhorar a qualidade do diagnóstico. Como nas outras aplicações desta tese, bons resultados são gerados com os algoritmos de AM supervisionado já disponíveis em pacotes de software como o sklearn.(97)

A contribuição de que tratou este capítulo não tem paralelo na literatura, pois até onde sabemos não foram feitos experimentos planejados de voltametria (ou outras medidas de eletroquímica) para aplicação de algoritmos de AM no diagnóstico. Por outro lado, há trabalhos na literatura em que dados eletroquímicos foram tratados com AM, principalmente para línguas eletrônicas baseadas em métodos eletroquímicos.(113) Há trabalhos que aplicam métodos de Quimiometria (77) sobre dados de línguas eletrônicas voltamétricas para análise de alimentos (114–116) e de seus contaminantes (117), e para detecção de fármacos.(118)

7 IMUNOSSENSOR E SERS PARA DIAGNÓSTICO DE COVID-19

7.1 Introdução

Sensores plasmônicos são utilizados para detecção de analitos e em diagnóstico (119–122), em que se aproveita a possibilidade de amplificação de sinal devido à ressonância de plásmons em superfícies metálicas.(123–125) Um dos efeitos utilizados é o efeito Raman com amplificação de superfície (SERS, do inglês *Surface-Enhanced Raman Scattering*) (126), usado neste capítulo para diagnóstico de COVID-19. A espectroscopia SERS tem sido utilizada como uma técnica altamente sensível para detectar várias classes de compostos de interesse, incluindo proteínas (biomarcadores de câncer e antígenos virais) (127, 128), poluentes emergentes (129), metais pesados (130) e neurotransmissores.(131)

O biossensor plasmônico foi constituído de nanopartículas de ouro sobre as quais foram imobilizados anticorpos correspondentes à proteína S do vírus SARS-CoV-2. Trata-se, portanto, de um imunossensor. Como os sinais Raman das biomoléculas usadas não são grandes, usou-se a estratégia de incluir no biossensor uma molécula repórter, o 4-aminotiofenol (4-ATP). O sinal Raman do 4-ATP deve ser modificado quando a proteína S interagir com o anticorpo do imunossensor plasmônico, e é essa alteração no espectro Raman que será utilizada no sensoriamento (diagnóstico).

7.2 Conjunto de dados

O conjunto de dados para esta aplicação é formado por mapas hiperespectrais (132) obtidos com a técnica SERS (126) sobre 46 unidades (n = 46) do imunossensor. Os mapas foram adquiridos através do equipamento in Via Raman microscope (Renishaw Inc., Hoffman Estates, IL). A preparação e disponibilização dos mapas foi feita por Wallance Pazin, pesquisador do Laboratório de Filmes Nanoestruturados e Espectroscopia (Unesp-Presidente Prudente) e membro do nosso grupo de pesquisa. Cada mapa possui 441 espectros de SERS coletados em 441 posições (21 linhas x 21 colunas) com espaçamento de 1 μm em uma área de dimensões 20 μm x 20 μm do imunossensor. Os espectros de SERS têm 1014 medidas de intensidade da radiação eletromagnética espalhada (expressa em u.a.) em função do deslocamento Raman no intervalo de 1611.19 a 471.042 cm^{-1}. O exemplo apresentado na Figura 7.1 evidencia a presença do pico do 4-ATP (repórter) na região de 1079.12 cm^{-1}.

Para avaliar a capacidade de detecção do antígeno biomarcador para COVID-19, o imunossensor foi exposto a amostas de três classes, sendo uma classe de controle, outra com a presença do antígeno, e a terceira com a presença de agente interferente. Exemplos dos espectros para cada uma destas classes são apresentados na Figura 7.2.

A composição do conjunto de dados está sintetizada na Tabela 7.1, incluindo a

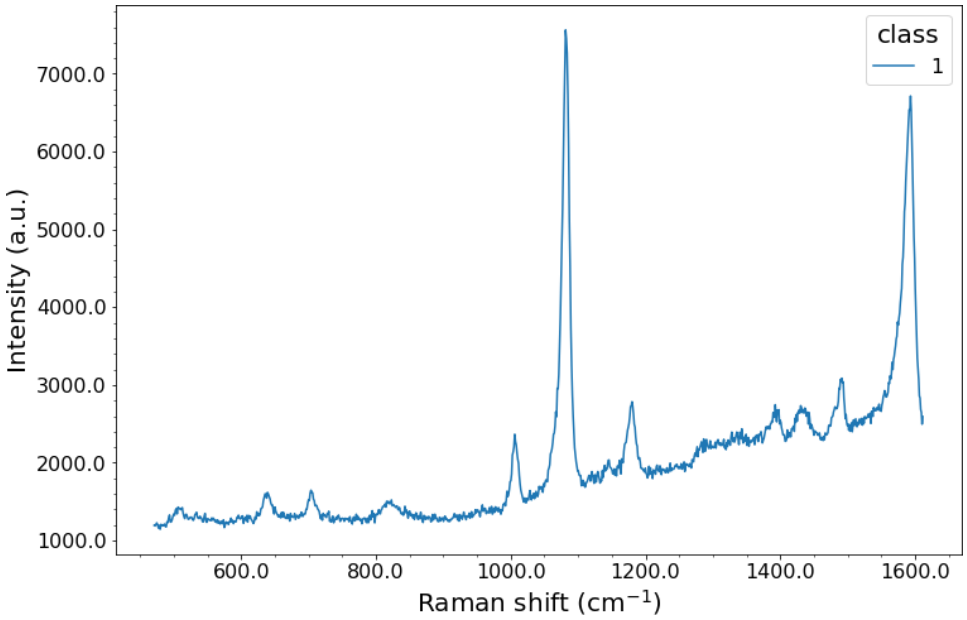


Figura 7.1 – Espectro de SERS do Imunossensor exposto à amostra de saliva de um paciente com COVID-19).

Fonte: Elaborada pelo autor.

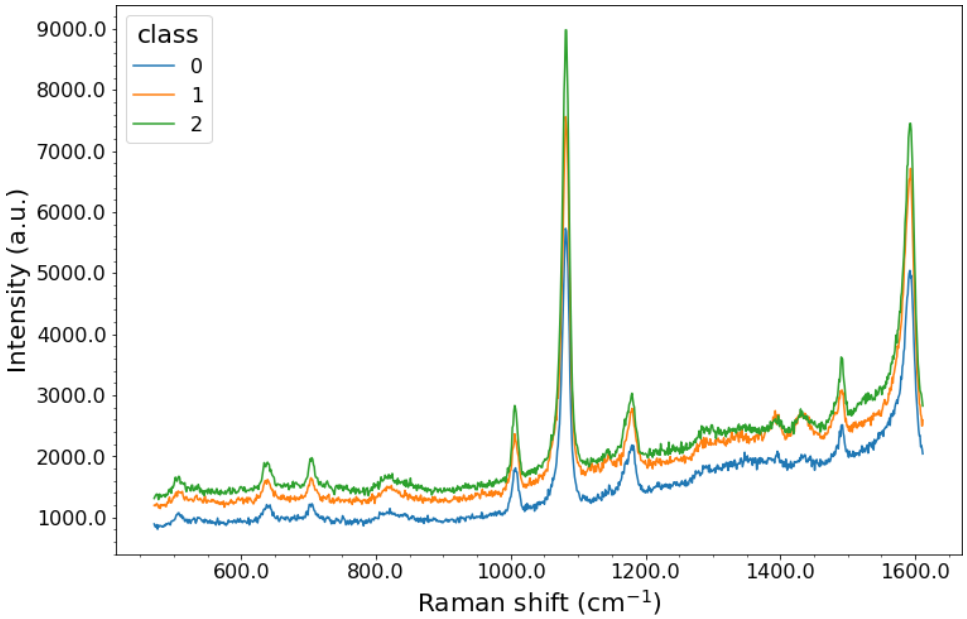


Figura 7.2 – Espectros de SERS do Imunossensor para as classes de amostras controle (0), antígeno da COVID-19 (1) e agente interferente (2).

Fonte: Elaborada pelo autor.

quantidade de objetos e evidenciando a elevada dimensionalidade. A presença de três classes possibilita análises binárias do tipo OvO (classe 0 v 1) e OvR (classe 0 v 1+2). Conforme a Tabela 7.2, a distribuição dos objetos entre as classes está desbalanceada.

As características do conjunto de dados e o conhecimento mais detalhado sobre os

Tabela 7.1 – Composição do conjunto de dados utilizado na aplicação Imunossensor e SERS para Diagnóstico de COVID-19.

Elemento	Descrição	Qtde.
Objetos	Mapas hiperespectrais	46
Atributos	Medidas de intensidade da radiação eletromagnética espalhada dos espectros (= 441 espectros x 1014 medidas)	447174
Classes	Controle, Antígeno, Interferente	3

Fonte: Elaborada pelo autor.

Tabela 7.2 – Distribuição dos objetos entre as classes no conjunto utilizado na aplicação Imunossensor e SERS para Diagnóstico de COVID-19. Símbolo utilizado: número de objetos (n).

Analito	Classe	Marca	n
Ausência de antígeno COVID-19	Controle	0	20
Presença de antígeno COVID-19	Antígeno	1	16
Presença de interferente	Interferente	2	10
		Total =	46

Fonte: Elaborada pelo autor.

objetos e seus atributos (Apêndice D) são fundamentais para o desenvolvimento do pipeline de AM para esta aplicação. Na etapa de preparação, o pré-processamento incluiu a reorganização dos espectros em formato *Band Interleave Pixels (BIP)*(132) e a remoção da linha de base.(133) Na engenharia de atributos, foram selecionados atributos próximos $(+/- 100 cm^{-1})$ do pico da molécula repórter e, a partir desta seleção foram aplicados extratores de atributos. A etapa de modelagem será apresentada conjuntamente aos resultados gerados.

7.3 Resultados

A grande quantidade de espectros e as pequenas alterações perceptíveis por inspeção visual dos espectros tornaram essencial fazer uma análise exploratória. Os resultados dessa análise estão no Apêndice D, em que se verificaram fatores como espectros anômalos, separação entre as classes da posição dos picos referente à molécula repórter (4-ATP). A Figura 7.3 mostra a grande similaridade entre os espectros de uma amostra positiva e outra negativa para COVID-19. Uma observação sobre essa figura é a de que o imunossensor não parece ser seletivo, o que representa um desafio para o diagnóstico. Não há por exemplo deslocamento do pico referente à molécula repórter (4-ATP). Usar técnicas mais avançadas de processamento do sinal Raman será essencial para se ter bom diagnóstico.

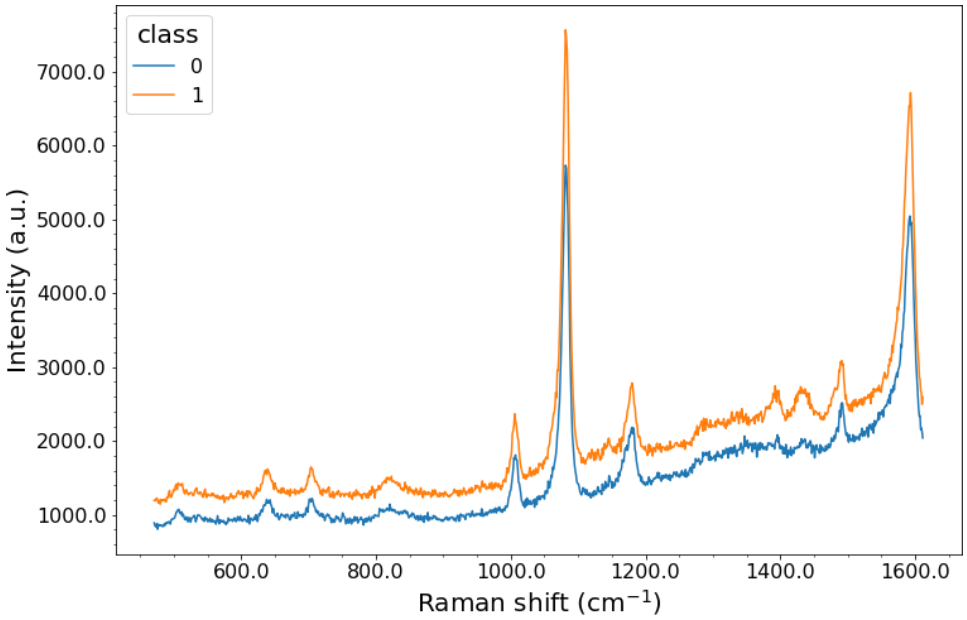


Figura 7.3 – Espectros de SERS do Imunossensor exposto à amostras das classes positivo e negativo).

Fonte: Elaborada pelo autor.

Para obter dados que possibilitem a separação das classes de dados, foram aplicados sobre os mapas hiperespectrais métodos de seleção e extração de atributos que geraram os conjuntos apresentados na Tabela 7.3. Uma vantagem adicional desses conjuntos é a redução da dimensionalidade original dos mapas conforme Tabela 7.1 reduzindo a chance de obtenção de sobreajuste dos modelos de AM.

Empregando-se técnicas de projeção aos conjuntos de atributos gerados, não foi possível separar amostras positivas das negativas conforme exemplificado na projeção com IDMAP (Figura 7.4). Mencione-se que outras técnicas de projeção também não permitiram a separação, como mostram os resultados do Apêndice D.

Agrupamento

Com os métodos de aglomeração não-supervisionados (KM, HAC e SC), foram obtidos valores de coeficiente médio de silhouette (métrica interna escolhida para avaliar a qualidade dos resultados) na Tabela 7.4. O objetivo era aglomerar os dados em 2, 3 e 4 classes; nas análises 0x1 há apenas duas classes conhecidas e nas análises multi há 3 classes conhecidas. Os maiores valores do coeficiente de silhueta para as análises dos espectros médios do mapa obtido para cada amostra das classes estudadas são inferiores a 0.800, ou seja, os atributos não evidenciam fortemente as classes de dados que sabemos estarem presentes. Isso indica a necessidade de empregar modelos supervisionados que utilizam a informação das classes dos dados para aprenderem e realizarem a classificação de novos dados.

Tabela 7.3 – Conjuntos de atributos extraídos dos mapas hiperespectrais.

Conjunto de atributos	Atributos	Qtde.
vox	Voxels dos mapas (intervalo 979 to 1179 cm^{-1})	88641
vox-stats	Medidas estatisticas dos voxels (media, desvio padrão, mínimo, 25%, 50%, 75%, máximo, assimetria/skew, curtose)	9
sers-mean	Média de cada deslocamento Raman dos espectros	201
sers-median	Média de cada deslocamento Raman dos espectros	201
rep-stats	Medidas estatisticas das intensidades dos picos em 1079 cm^{-1}) (media, desvio padrão, mínimo, 25%, 50%, 75%, máximo, assimetria/skew, curtose)	9
peak-mean	Valores médios de características do pico principal (intensidade, largura 1/2 altura, largura na base e fator de qualidade)	5
peak-median	Valores médios de características do pico principal (intensidade, largura 1/2 altura, largura na base e fator de qualidade)	5
map-pca	Componente principal 1 da PCA	201

Fonte: Elaborada pelo autor.

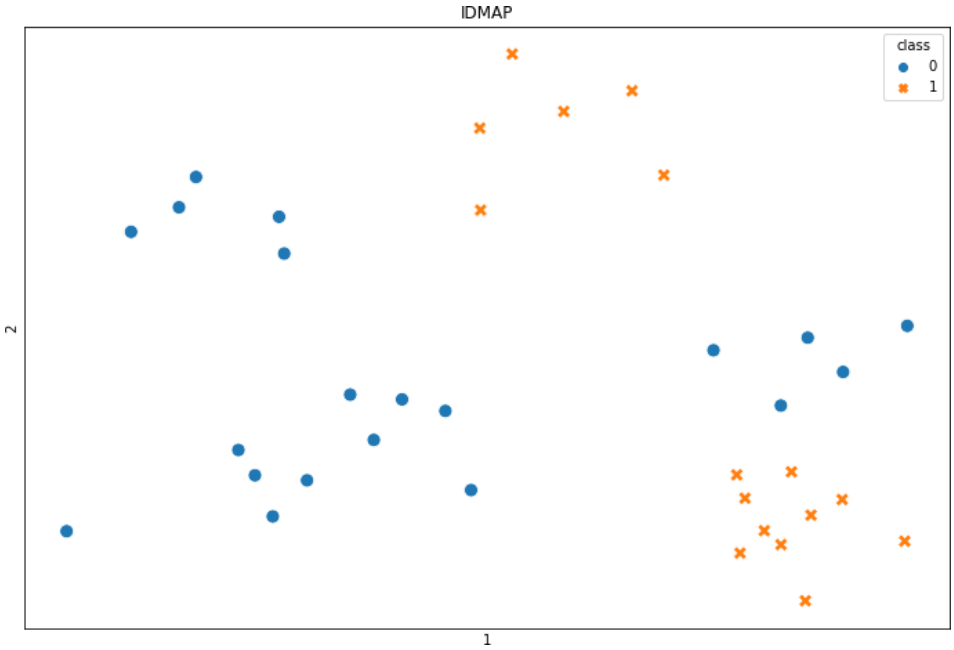


Figura 7.4 – Visualização da projeção IDMAP dos espectros médios (conjunto sers-mean) para amostras das classes positivo e negativo.

Fonte: Elaborada pelo autor.

Tabela 7.4 – Largura Média da Silhueta (Average Silhouette Width) (ASW) do agrupamento dos mapas das classes 0 e 1 com os algoritmos KM, HAC e SC para todos conjuntos de atributos.

Conjuntos de atributos	Algoritmos	Qtde. de agrupamentos		
		2	3	4
vox-0v1	KM	0.591	0.514	0.485
	HAC	0.595	0.506	0.485
	SC	0.595	0.506	0.320
vox-stats-0v1	KM	0.771	0.717	0.590
	HAC	0.771	0.727	0.607
	SC	0.218	0.449	0.457
rep-stats-0v1	KM	0.640	0.631	0.631
	HAC	0.630	0.608	0.631
	SC	0.630	0.608	0.525
sers-mean-0v1	KM	0.643	0.590	0.591
	HAC	0.650	0.587	0.591
	SC	0.650	0.587	0.416
sers-median-0v1	KM	0.647	0.591	0.595
	HAC	0.654	0.590	0.595
	SC	0.654	0.590	0.419
peak-mean-0v1	KM	0.706	0.676	0.671
	HAC	0.704	0.676	0.691
	SC	0.704	0.676	0.567
peak-median-0v1	KM	0.711	0.681	0.664
	HAC	0.711	0.681	0.697
	SC	0.711	0.681	0.562
map-pca-0v1	KM	0.563	0.542	0.578
	HAC	0.563	0.562	0.594
	SC	0.543	0.204	0.214

Fonte: Elaborada pelo autor.

Classificação

Uma boa distinção das classes foi obtida com AM supervisionado, o que demonstra ser possível utilizar o imunosensor plasmônico, a despeito de sua falta de seletividade. Os resultados para os algoritmos LR, LDA, SVM (L, P and R), GNB e KNN estão na Tabela 7.6.

Tabela 7.5 – Largura Média da Silhueta (Average Silhouette Width) (ASW) do agrupamento multiclasse com os algoritmos KM, HAC e SC para todos conjuntos de atributos.

Conjuntos de atributos	Algoritmos	Qtde. de agrupamentos		
		2	3	4
vox-multi	KM	0.6	0.472	0.427
	HAC	0.594	0.472	0.427
	SC	0.602	0.472	0.321
vox-stats-multi	KM	0.743	0.691	0.611
	HAC	0.7	0.691	0.611
	SC	0.273	0.448	0.46
rep-stats-multi	KM	0.644	0.569	0.576
	HAC	0.62	0.565	0.576
	SC	0.632	0.569	0.43
sers-mean-multi	KM	0.649	0.555	0.53
	HAC	0.643	0.555	0.53
	SC	0.651	0.555	0.417
sers-median-multi	KM	0.653	0.556	0.529
	HAC	0.647	0.556	0.529
	SC	0.656	0.556	0.423
peak-mean-multi	KM	0.706	0.613	0.575
	HAC	0.706	0.59	0.569
	SC	0.702	0.613	0.473
peak-median-multi	KM	0.71	0.611	0.572
	HAC	0.71	0.589	0.565
	SC	0.709	0.611	0.474
map-pca-multi	KM	0.53	0.548	0.572
	HAC	0.53	0.548	0.577
	SC	0.506	0.211	0.199

Fonte: Elaborada pelo autor.

A validação dos modelos foi feita com o método de *K-Fold Cross-Validation (KCV)* (k=5). O conjunto de atributos mais importante para a classificação foi sers-mean, ou seja, média de cada deslocamento Raman dos espectros. A acurácia foi de 100% para o algoritmo LDA na análise 0v1 que diferencia amostras negativas das positivas. Mesmo nas análises 0vR (Tabela 7.7) e multiclasse (Tabela 7.8), foram obtidas acurácias de 100% para esse conjunto permitindo verificar uma baixa influência do agente interferente como elemento de confusão da

classificação.

Tabela 7.6 – Acurácia média (desvio padrão) para a classificação das classes 0 e 1 com os algoritmos LR, LDA, GNB, KNN, SVM (L, R, P) e KNN.

Conjuntos de atributos	LR	LDA	SVM-L	SVM-P	SVM-R	GNB	KNN
vox-0v1	0.861	0.832	0.804	0.721	0.721	0.807	0.718
	(0.128)	(0.141)	(0.116)	(0.181)	(0.181)	(0.211)	(0.160)
vox-stats-0v1	0.771	0.971	0.943	0.557	0.721	0.779	0.668
	(0.171)	(0.057)	(0.070)	(0.095)	(0.203)	(0.192)	(0.141)
rep-stats-0v1	0.443	0.668	0.750	0.750	0.721	0.804	0.643
	(0.204)	(0.190)	(0.189)	(0.189)	(0.181)	(0.116)	(0.156)
sers-mean-0v1	1.000	1.000	1.000	0.721	0.721	0.807	0.746
	(0.000)	(0.000)	(0.000)	(0.181)	(0.181)	(0.142)	(0.111)
sers-median-0v1	1.000	0.971	1.000	0.721	0.721	0.861	0.693
	(0.000)	(0.057)	(0.000)	(0.181)	(0.181)	(0.091)	(0.062)
peak-mean-0v1	0.750	0.746	0.636	0.721	0.721	0.664	0.664
	(0.189)	(0.111)	(0.176)	(0.181)	(0.181)	(0.077)	(0.119)
peak-median-0v1	0.693	0.775	0.579	0.721	0.721	0.750	0.664
	(0.142)	(0.075)	(0.102)	(0.181)	(0.181)	(0.139)	(0.119)
map-pca-0v1	0.586	0.557	0.557	0.611	0.586	0.693	0.639
	(0.114)	(0.183)	(0.131)	(0.166)	(0.171)	(0.168)	(0.192)

Fonte: Elaborada pelo autor.

7.4 Conclusão

A análise de espectros Raman (SERS ou não) já vem sendo feita com métodos computacionais há vários anos (134–136), especialmente porque os espectrômetros modernos podem fornecer dezenas de milhares de espectros em uma hora. Essa grande quantidade de dados é importante para análise de material biológico, ou de biossensores como os apresentados aqui, em virtude da variabilidade das amostras, dos dispositivos/sistemas e dos dados adquiridos.(137)

Os resultados deste capítulo indicaram que o processamento de dados usando AM supervisionado consegue superar problemas (75, 137) do planejamento e fabricação de biossensores. Em outras palavras, mesmo biossensores com mau desempenho podem ser utilizados - fornecendo diagnóstico correto - desde que os dados recebam tratamento adequado com AM.(78,79) No pipeline aplicado, o pré-processamento incluiu a reorganização dos espectros em formato *Band Interleave Pixels (BIP)*(132) e a remoção da linha de base.(133) Na engenharia de atributos, foram selecionados atributos próximos do pico da molécula reporter e, a partir

Tabela 7.7 – Acurácia média (desvio padrão) para a classificação das classes 0 e (1-2) com os algoritmos LR, LDA, GNB, KNN, SVM (L, R, P) e KNN.

Conjuntos de atributos	LR	LDA	SVM-L	SVM-P	SVM-R	GNB	KNN
vox-0vR	0.958	0.807	0.849	0.698	0.698	0.764	0.696
	(0.052)	(0.158)	(0.165)	(0.099)	(0.121)	(0.196)	(0.129)
vox-stats-0vR	0.804	0.829	0.738	0.611	0.567	0.764	0.609
	(0.043)	(0.126)	(0.152)	(0.070)	(0.054)	(0.168)	(0.149)
rep-stats-0vR	0.564	0.740	0.827	0.784	0.698	0.762	0.653
	(0.123)	(0.148)	(0.087)	(0.093)	(0.099)	(0.160)	(0.125)
sers-mean-0vR	0.978	1.000	0.933	0.720	0.698	0.764	0.653
	(0.044)	(0.000)	(0.054)	(0.124)	(0.121)	(0.168)	(0.075)
sers-median-0vR	1.000	0.956	0.956	0.698	0.698	0.807	0.631
	(0.000)	(0.054)	(0.054)	(0.121)	(0.121)	(0.158)	(0.046)
peak-mean-0vR	0.758	0.802	0.671	0.698	0.698	0.653	0.631
	(0.132)	(0.086)	(0.144)	(0.121)	(0.121)	(0.125)	(0.046)
peak-median-0vR	0.698	0.780	0.693	0.698	0.698	0.676	0.653
	(0.121)	(0.102)	(0.166)	(0.121)	(0.121)	(0.152)	(0.075)
map-pca-0vR	0.616	0.456	0.502	0.671	0.671	0.718	0.582
	(0.178)	(0.074)	(0.120)	(0.105)	(0.126)	(0.087)	(0.137)

Fonte: Elaborada pelo autor.

dessa seleção foram aplicados extratores de atributos. Na aplicação específica deste capítulo, esperava-se que o biossensor plasmônico fosse muito sensível e seletivo para a interação entre seus anticorpos e os antígenos das amostras que representavam infecção por COVID-19. Com base em resultados da literatura para outros biossensores desse tipo (69)(136), esperavam-se deslocamentos de bandas importantes do espectro SERS ou mesmo alterações significativas de intensidades. Ao contrário, os espectros SERS explorados no Apêndice D diferiam muito pouco entre si para amostras positivas e negativas para o antígeno do SARS-CoV-2. Sem uma análise adequada dos dados, ou empregando-se apenas uma análise manual, não teria sido possível usar o biossensor para o diagnóstico.

Com os algoritmos AM supervisionados, entretanto, a análise dos conjuntos de atributos gerados a partir dos espectros permitiram grande acurácia no diagnóstico, chegando a 100% para os algoritmos LR, LDA e SVM-L. Aqui, como nas aplicações anteriores, na etapa de modelagem do pipeline, os hiperparâmetros foram fixados em seus valores iniciais (97) para avaliar a capacidade dos dados em evidenciar os padrões das classes identificadas. Os resultados comprovaram essa capacidade. O método de validação foi o KCV, com os cuidados para evitar

Tabela 7.8 – Acurácia média (desvio padrão) para a classificação multiclasse com os algoritmos LR, LDA, GNB, KNN, SVM (L, R, P) e KNN.

Conjuntos de atributos	LR	LDA	SVM-L	SVM-P	SVM-R	GNB	KNN
vox-multi	0.782	0.651	0.804	0.564	0.564	0.698	0.607
	(0.122)	(0.111)	(0.082)	(0.142)	(0.142)	(0.186)	(0.154)
vox-stats-multi	0.671	0.827	0.891	0.436	0.564	0.676	0.480
	(0.225)	(0.112)	(0.122)	(0.072)	(0.158)	(0.152)	(0.154)
rep-stats-multi	0.411	0.589	0.567	0.609	0.564	0.611	0.480
	(0.120)	(0.156)	(0.089)	(0.165)	(0.142)	(0.211)	(0.095)
sers-mean-multi	0.933	1.000	0.933	0.564	0.564	0.698	0.696
	(0.089)	(0.000)	(0.054)	(0.142)	(0.142)	(0.121)	(0.147)
sers-median-multi	0.956	0.978	0.933	0.564	0.564	0.696	0.718
	(0.089)	(0.044)	(0.054)	(0.142)	(0.142)	(0.082)	(0.111)
peak-mean-multi	0.673	0.716	0.607	0.564	0.564	0.587	0.547
	(0.211)	(0.093)	(0.095)	(0.142)	(0.142)	(0.044)	(0.110)
peak-median-multi	0.673	0.758	0.560	0.564	0.564	0.676	0.569
	(0.141)	(0.087)	(0.130)	(0.142)	(0.142)	(0.115)	(0.140)
map-pca-multi	0.458	0.504	0.436	0.478	0.458	0.560	0.476
	(0.228)	(0.174)	(0.212)	(0.227)	(0.228)	(0.216)	(0.208)

Fonte: Elaborada pelo autor.

o sobreajuste dos modelos. Na continuidade deste trabalho, serão incluídos os resultados com o método *Multidimensional Calibration Spaces (MCS)* (19) que possibilita a compreensão da importância de cada atributo dos modelos classificadores.

8 CONSIDERAÇÕES FINAIS

A hipótese inicial que guiou a proposta desta tese foi amplamente comprovada, ou seja, a de que é possível desenvolver sistemas computacionais com inteligência artificial para melhorar o diagnóstico de diferentes doenças. Para além das demonstrações aqui apresentadas, houve tanta evolução na literatura que a confirmação da hipótese hoje parece quase trivial. Para a análise de espectros Raman ou de outras técnicas de espectroscopia vibracional, já está ficando rotineiro o uso de AM supervisionado.(75, 137–139) Assim, acreditamos que esta tese consolida estratégias para análise de dados de sensores e biossensores, que esperamos sejam disseminadas para pesquisadores dessas áreas, ainda que não especialistas em ciências de dados ou computação. Para tanto, foi desenvolvido um pipeline com sistematização de procedimentos, o que deve facilitar não apenas o emprego dos diferentes métodos mas também o treinamento de usuários não especialistas. O treinamento de alunos de Iniciação Científica, seguindo os passos do pipeline, serviu para testar a adequação dos procedimentos e identificar dificuldades com os diferentes tipos de dados. Como conclusões gerais das contribuições, a serem especificadas abaixo, podemos enumerar:

1. Conhecer o domínio do problema. Nesta tese o domínio compreende a *análise de sinais físicos e físico-químicos com aprendizado de máquina*, implicando em mesclar conhecimentos de Física, Instrumentação e Computação;

2. É essencial seguir um pipeline com metodologia sistemática. Nesta tese usamos o CRISP-DM;

3. No pipeline de AM, a tarefa de análise exploratória é relevante para um usuário não especialista em ciência de dados, para que possa perceber os desafios em prover conjuntos de dados adequados para as tarefas de agrupamento e de classificação;

4. Os métodos correntes para AM supervisionado e não-supervisionado são suficientes para fornecer alto desempenho no diagnóstico, mesmo em problemas desafiadores, como alguns abordados na tese;

5. O maior desafio para o uso de AM está na aquisição de dados para formar conjuntos adequados, principalmente porque pode ser necessário obter muito mais dados do que normalmente se faz em experimentos de sensores e biossensores. Em alguns casos, pode ser simplesmente inviável obter a quantidade de dados necessária, devido ao alto custo dos experimentos. Além de uma maior quantidade de dados, é importante planejar experimentos para que o conjunto de dados seja balanceado quanto às classes das amostras.

6. Apesar da limitação mencionada no item anterior, é digno de nota que diagnósticos de qualidade foram obtidos em alguns casos nesta tese, a despeito de a quantidade de dados ser pequena. Esses diagnósticos não requereram uso de métodos de redes neurais profundas (deep learning), que formam hoje os algoritmos mais eficientes para classificação.;

7. Pode-se, assim, antever grandes perspectivas do emprego da metodologia desenvolvida aqui para aplicações em que haja grandes volumes de dados. Isso inclui os casos em que dados de aplicações semelhantes possam ser aproveitados em estratégias de transferência de conhecimento (transfer learning), como já é feito para diagnóstico a partir de imagens.

8.1 Síntese das contribuições

A premissa fundamental desta tese foi comprovada com quatro aplicações, em que diferentes tipos de dados foram usados. Dos resultados também ficou claro que é necessário estabelecer um pipeline de AM para guiar a análise dos dados. O pipeline é importante para a sistematização do processo de análise, e serve também para auxiliar pesquisadores não especialistas em sistemas computacionais. A implementação do pipeline adotado nesta tese foi construída a partir de uma ideia geral inicial, e adaptada à medida que novas aplicações foram desenvolvidas. O pipeline atual é suficientemente genérico para ser usado para qualquer tipo de diagnóstico, não só médico, e com qualquer tipo de dado. Exploraram-se dados científicos de sensores e imagens nas aplicações, mas não texto. Portanto, módulos adicionais podem ter que ser incluídos no processamento para um sistema completo de diagnóstico. A composição e organização do pipeline não se alterariam.

A análise exploratória é uma tarefa fundamental do pipeline de AM, pois pode trazer informações sobre os atributos dos dados, as quais são importantes para o planejamento das tarefas da etapa de preparação do pipeline de AM, e para respostas a questões iniciais sobre a capacidade dos dados em evidenciar por si, isto é, sem o auxílio de aprendizado de máquina, padrões separáveis das classes de dados. Os resultados destas análises nas aplicações demonstraram que isso não foi possível. Por exemplo, imaginava-se inicialmente que a posição dos picos presentes nos espectros SERS da molécula repórter separassem as classes positivo e negativo. Fato que não foi comprovado pela exploração dos dados.

Das 6 tarefas principais do pipeline base (seção 2.5), as mais dependentes do tipo de dados são o pré-processamento e a engenharia de atributos. Os conjuntos de dados utilizados nas aplicações são originados de diferentes configurações de biossensores/língua eletrônica: imagens de microscopias eletrônicas de varredura (MEV), espectros de impedância, voltamogramas e mapas hyperspectrais de espectros SERS. Tratam-se de sinais provenientes de diferentes respostas dos dispositivos (ópticas, elétricas, eletrônicas), coletados com diferentes equipamentos, representados com tipos/unidades de medida diferentes. Possuem diferentes características de forma, intensidade, nível de ruído. Dessa forma, trazendo um grande desafio

para a preparação dos dados com técnicas adequadas para o tratamento e a extração dos atributos mais relevantes para o objetivo de diagnóstico. Por exemplo, na aplicação envolvendo imagens de MEV, foram empregas técnicas de recorte, seleção de atributos com abordagem wrapper e extração de atributos de textura. Os algoritmos de AM, supervisionados ou não, são os usuais e contidos em muitos pacotes de software, fornecendo altos desempenhos para todas as aplicações, sendo quase sempre os mesmos. Portanto, o componente final de classificação não deve sofrer grande variação de uma aplicação para outra. Igualmente, as métricas de avaliação de desempenho do diagnóstico podem ser as mesmas para as diferentes aplicações.

Mencione-se, também, a necessidade de seleção de um conjunto de dados (atributos) extraídos dos espectros para uma classificação efetiva. Nesse aspecto, o estabelecimento de um pipeline como o proposto, e implementado nesta tese, é essencial para trabalhos sistemáticos de diagnóstico. Como acreditamos que o processamento de dados de sensores e biossensores com métodos computacionais será muito frequente no futuro, pode ser interessante automatizar o pipeline, com consultas ao usuário em alguns passos. Por exemplo, o usuário pode ser consultado sobre a necessidade de fazer visualização com técnicas de projeção, quais técnicas utilizar em alguns passos. Pode-se, também, conceber um pipeline completamente automatizado com as decisões sendo tomadas pelo sistema inteligente com base nas métricas de desempenho. Para o pipeline inteiramente automatizado as técnicas de visualização não são úteis, e os passos de análise incluíriam apenas agrupamento com algoritmos de AM não supervisionado, e classificação com algoritmos de AM supervisionado. A análise exploratória inicial, limpeza de dados e a seleção de atributos seriam guiadas pelas medidas de desempenho, silhueta ou acurácia, por exemplo, num processo iterativo envolvendo os diferentes passos do pipeline.

Em relação às aplicações, até onde sabemos, este trabalho trouxe no Capítulo 4 o primeiro exemplo de aprendizado de máquina aplicado no diagnóstico usando análise de imagens das unidades de detecção (não de amostras biológicas). A acurácia na distinção de todas as concentrações de PCA3 foi no máximo 88,3% para AM supervisionado e 70,83% para não-supervisionado a partir de engenharia de atributos com extratores de textura. Essa nova abordagem parece promissora, pois pode permitir chegar ao diagnóstico simplesmente tirando fotos das unidades de detecção antes e depois do uso. Demonstrou-se que a análise pode ser feita com imagens de microscopia eletrônica de varredura (MEV). Utilizamos esse tipo de imagem porque era mais provável de fornecer bons resultados, uma vez que as mudanças na morfologia do filme ocorrem no nível nanoscópico. Tentaremos outros tipos de imagens em um futuro próximo, inclusive de câmeras de smartphones e microscópios ópticos que podem ser acoplados a smartphones. Isso será muito mais desafiador e exigirá um novo estudo completo. Destaque-se que os resultados dessa aplicação foram publicados no seguinte artigo:

RODRIGUES, V. C.; SOARES, J. C.; SOARES, A. C.; BRAZ, D. C.; MELENDEZ, M. E.; RIBAS, L. C.; SCABINI, F. S.; BRUNO, O. B.; CARVALHO, A. L.;REIS, R. M.; SANFELICE, R. C.; OLIVEIRA JUNIOR, O. N. Electrochemical and optical detection and machine learning

applied to images of genosensors for diagnosis of prostate cancer with the biomarker PCA3, **Talanta**, v. 222, p. 121444, 2021.

A aplicação para diagnóstico de câncer de boca (Capítulo 5) representou o primeiro exemplo da literatura em que uma língua eletrônica foi usada para o diagnóstico de câncer, sem a necessidade de detecção de um biomarcador. Uma constatação importante sobre a classificação binária para câncer (yes/no) dos espectros (conjunto only-sensor) foi a de que os desempenhos dos atributos f-all e f-low são similares e acima de 83% de acurácia, podendo-se inferir que medidas em frequências baixas são tão representativas das classes quanto medidas no espectro completo. Destaque também deve ser dado ao uso de saliva de pacientes voluntários para as medidas de espectroscopia de impedância com a língua eletrônica. Empregar saliva, cuja coleta é muito menos invasiva do que em outros exames, é uma tendência para diagnósticos modernos. Talvez o mais relevante para a estratégia adotada nesta tese tenha sido o emprego de informações de sensores aliados a informações clínicas. A combinação de diferentes tipos de dados só é possível com técnicas, como as utilizadas aqui. Além disso, demonstramos que aliar informações clínicas aos dados de sensores (ou de análises de laboratório) pode melhorar a capacidade de diagnóstico. Com aprendizado de máquina também se pode obter diagnóstico com algum grau de explicação, como verificado em trabalho publicado com resultados desta aplicação.(104) Para os dados de impedância em saliva e dados clínicos, obtiveram-se regras com algoritmos de aprendizado de máquina, inclusive com o conceito de calibração multidimensional.(19) Com este último conceito observou-se que problemas de alcoolismo podem estar correlacionados positivamente com câncer de boca. O número de amostras de pacientes não era suficiente para uma conclusão definitiva, mas os resultados demonstram o poder da abordagem com aprendizado de máquina de descobrir informações que não são passíveis de fazê-lo de outra maneira. Destaque-se também que os resultados dessa aplicação foram publicados no seguinte artigo:

BRAZ, D. C.; POPOLIN NETO, M.; SHIMIZU, F. M.; SÁ, A. C.; LIMA, R. S.; GOBBI, A. L.; MELENDEZ, M. E.; ARANTES, L. M. R. B.; CARVALHO, A. L.; PAULOVICH, F. V.; OLIVEIRA JUNIOR, O. N. Using machine learning and an electronic tongue for discriminating saliva samples from oral cavity cancer patients and healthy individuals, **Talanta**, v. 243, p.123327, 2022.

Na aplicação para detecção do biomarcador p53 para diversos tipos de câncer apresentada no Capítulo 6, foi possível demonstrar com o pipeline de AM desenvolvido que o uso de ácido ascórbico, que possui alta afinidade com proteínas, leva a imunossensores com melhor desempenho. O alto desempenho com acurácia de 100% com os vários algoritmos indica que essa tarefa de detecção do biomarcador p53 é mais simples que as outras estudadas nesta tese, principalmente porque foi considerada apenas a classificação binária. Outro ponto de destaque foi, com base

em questões intrínsecas do experimento que gerou os dados, das caraterísticas dos sinais e de computação, a demonstração que usar mais informações (atributos) dos voltamogramas, e não apenas a intensidade da corrente de pico, leva a uma melhor acurácia nesta detecção. Além de possíveis interferências nos picos, causadas por outras substâncias no fluido além do analito, o que poderia introduzir informações irrelevantes, alterações adicionais com potencial relevância para classificação podem aparecer nos voltamogramas. Um aspecto importante desta aplicação foi o planejamento dos experimentos de voltametria com o imunossensor visando a geração de conjuntos de dados mais adequados ao pipeline de AM e para geração de melhores classificadores. As classes de objetos de dados, que estão diretamente ligadas às análises de interesse, e a quantidade de objetos total e por classe orientaram a realização dos experimentos. Essa abordagem experimental vem sendo cada vez mais no grupo fazendo parte do protocolo das pesquisas que pretendem fazer aplicação de AM. Um artigo com os resultados desta aplicação está sendo escrito para submissão.

Uma boa distinção das classes nas análises 0v1 (controle vs antígeno) e 0vR (controle vs antígeno+interferente) foi obtida com AM supervisionado para o diagnóstico de COVID-19 com imunosensor plasmônico (Capítulo 7). As acurácias obtidas por vários algoritmos chegaram a 100% sobre o conjunto de atributos sers-mean (média de cada deslocamento Raman dos espectros). Mesmo na análise multiclasse foram obtidas acurácias similares para esse conjunto permitindo verificar uma baixa influência do agente interferente como elemento de confusão da classificação. Um ponto importante é que os experimentos de coleta de dados SERS já foram orientados também para obtenção de melhores resultados com aprendizado de máquina. Outro ponto relevante é que através da etapa de preparação de dados do pipeline desenvolvido pode-se superar problemas do planejamento e fabricação de imunossensores, e detalhes dos experimentos de coleta dos espectros. Resultados originais desta aplicação fazem parte do artigo que está sendo preparado e será submetido em breve.

Para além das contribuições fornecidas pelas aplicações, a oportunidade de trabalhar a formação de alunos de iniciação científica para aplicação de AM em dados biossensores possibilitou a incorporação de novos alunos junto ao grupo de pesquisa e geração de um guia básico para treinamento (Apêndice F). Com isso, o grupo ganha tanto em recursos humanos como materiais para o desenvolvimento de novos trabalhos nesta temática, por exemplo, dentro das perspectivas que serão apresentadas a seguir.

8.2 Perspectivas

Os resultados alcançados nesta tese demonstram que há muito futuro para sistemas computacionais inteligentes em diagnóstico. Destaque-se, de toda forma, que esses resultados - assim como de outros trabalhos da literatura - ainda são modestos diante das possibilidades advindas da inteligência artificial. Ainda estamos muito longe de sistemas autônomos de diagnóstico, que se beneficiarão de progressos de várias áreas, inclusive a de descoberta aumotática

de conhecimento (*knowledge discovery*). Há assim, grandes perspectivas de continuidade de trabalhos como os desta tese. Considerando-se os resultados e os desafios científicos e tecnológicos que surgiram no decorrer do trabalho, propomos como perspectivas para continuidade:

1. Ampliação do conhecimento sobre os tópicos conceituais, métodos e ferramentas abrangidos nesta tese, destacando:
    a) Biossensores e os diferentes tipos de resposta.
    b) Técnicas analíticas para geração de dados.
    c) Processamento de sinais.
    d) Aprendizado de máquina.
    e) Aprendizado profundo (*Deep learning*).
    f) Reconhecimento de padrões, Visão computacional, Processamento de linguagem natural, etc.
2. Em relação às aplicações realizadas:
    a) Ampliação do tamanho e da representatividade dos conjuntos de dados.
    b) Emprego de diferentes técnicas de análise e processamento de sinais.
    c) Novos extratores de atributos dos objetos gerados pelas técnicas analíticas.
    d) Utilizar diferentes algoritmos e técnicas de AM, incluindo técnicas de Ensemble e Aprendizado profundo.
    e) Colocar os modelos desenvolvidos em ambiente de produção e acompanhar seus desempenhos.
3. Pesquisa, desenvolvimento e aplicação de AM na análise de dados de (bios)sensores:
    a) Trabalhar com novos conjuntos de dados e aplicações.
    b) Verificar as potenciais contribuições da fusão de diferentes atributos e conjuntos de dados (dados clínicos, microscopias, espectroscopias e demais técnicas analíticas) no diagnóstico de doenças.
    c) Propor novas etapas/tarefas de modelagem para melhoria dos modelos de AM.
    d) Utilizar diferentes algoritmos e técnicas de AM, incluindo técnicas de Ensemble e Aprendizado profundo.
4. Pesquisa, desenvolvimento e implementação de modelos de AM em sistemas embarcados visando ao desenvolvimento de testes diagnósticos tipo POC (Point-of-care).
5. Pesquisa e desenvolvimento de plataforma web de dados de (bios)sensores visando:

a) Facilitação da coleta, organização, integração e compartilhamento de dados.
b) Ampliação e organização do acervo de dados.
c) Curadoria do acervo de dados.
d) Infraestrutura para o desenvolvimento de novos modelos de AM e geração de conhecimento a partir do acervo de dados.

6. Formação de novos pesquisadores para aplicação de AM e outras técnicas de IA na análise de dados de (bios)sensores.

REFERÊNCIAS

1 NATIONAL CANCER INSTITUTE (NCI). **What is cancer?** 2021. Disponível em: https://www.cancer.gov/about-cancer/understanding/what-is-cancer. Acesso em: 1 jun. 2021.

2 TOBORE, T. O. On the need for the development of a cancer early detection, diagnostic, prognosis, and treatment response system. **Future Science OA**, v. 6, n. 2, 2020. ISSN 20565623.

3 VALENCIA, D. N. Brief review on COVID-19: the 2020 pandemic caused by SARS-CoV-2. **Cureus**, v. 12, n. 3, 2020.

4 MOHANTY, S. P.; KOUGIANOS, E. Biosensors: a tutorial review. **IEEE Potentials**, v. 25, n. 2, p. 35–40, 2006.

5 RUSLING, J. F. *et al.* Measurement of biomarker proteins for point-of-care early detection and monitoring of cancer. **Analyst**, v. 135, n. 10, p. 2496–2511, 2010.

6 LIU, R.; YE, X.; CUI, T. Recent progress of biomarker detection sensors. **Research**, AAAS, v. 2020, p. 1–26, 2020. ISSN 26395274.

7 BANERJEE, A.; MAITY, S.; MASTRANGELO, C. H. Nanostructures for biosensing, with a brief overview on cancer detection, IoT, and the role of machine learning in smart biosensors. **Sensors**, v. 21, n. 4, p. 1–34, 2021. ISSN 14248220.

8 RODRIGUES, J. F. *et al.* On the convergence of nanotechnology and Big Data analysis for computer-aided diagnosis. **Nanomedicine**, v. 11, n. 8, p. 959–982, 2016. ISSN 1743-5889.

9 OLIVEIRA, O. N. *et al.* Where chemical sensors may assist in clinical diagnosis exploring big data. **Chemistry Letters**, v. 43, n. 11, p. 1672–1679, 2014. ISSN 13480715.

10 RUSSEL, S.; NORVIG, P. **Artificial intelligence**: a modern approach. 3rd ed. New Jersey: Prentice Hall, 2013. 1016 p.

11 YANASE, J.; TRIANTAPHYLLOU, E. A systematic survey of computer-aided diagnosis in medicine: past and present developments. **Expert Systems with Applications**, v. 138, p. 112821, 2019.

12 JIN, X. *et al.* Artificial intelligence biosensors: challenges and prospects. **Biosensors and Bioelectronics**, v. 165, p. 112412, 2020. ISSN 18734235.

13 OLIVEIRA, O. N.; OLIVEIRA, M. C. F. Sensing and biosensing in the world of autonomous machines and intelligent systems. **Frontiers in Sensors**, v. 2, n. 752754, p. 1–7, 2021.

14 HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. 2nd ed. New York: Springer Science & Business Media, 2009. 745 p.

15 ZHANG, K. *et al.* Machine learning-reinforced noninvasive biosensors for healthcare. **Advanced Healthcare Materials**, v. 10, n. 17, p. 2100734, 2021.

16 GERON, A. **Hands-on machine learning with scikit-learn & tensor flow**. Beijing: O'Reilly Media, 2017. 760 p. ISSN 1662-5196.

17 PAULOVICH, F. V. *et al.* Using multidimensional projection techniques for reaching a high distinguishing ability in biosensing. **Analytical and Bioanalytical Chemistry**, v. 400, n. 4, p. 1153–1159, 2011. ISSN 16182642.

18 PAULOVICH, F. V.; OLIVEIRA, M. C. F.; OLIVEIRA, O. N. A future with ubiquitous sensing and intelligent systems. **ACS Sensors**, v. 3, n. 8, p. 1433–1438, 2018. ISSN 23793694.

19 POPOLIN, M. *et al.* Machine learning used to create a multidimensional calibration space for sensing and biosensing data. **Bulletin of the Chemical Society of Japan**, v. 94, n. 5, p. 1553–1562, 2021. ISSN 13480634.

20 RODRIGUES, V. D. C. *et al.* Analysis of scanning electron microscopy images to investigate adsorption processes responsible for detection of cancer biomarkers. **ACS Applied Materials and Interfaces**, v. 9, n. 7, p. 5885–5890, 2017. ISSN 19448252.

21 RODRIGUES, V. C. *et al.* Electrochemical and optical detection and machine learning applied to images of genosensors for diagnosis of prostate cancer with the biomarker PCA3. **Talanta**, v. 222, p. 121444, 2021. ISSN 00399140.

22 PAGE, M. J. *et al.* PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. **The BMJ**, v. 372, n. 160, 2021. ISSN 17561833.

23 MITCHELL, T. M. **Machine learning**. New York: McGraw-Hill, 1997. 432 p.

24 ALPAYDIN, E. **Introduction to machine learning**. 3rd ed. Boston: Massachusetts Institute of Technology, 2014. 613 p. ISBN 9780262028189.

25 FLATCH, P. **Machine learning -** the art and science of algorithms that make sense of data. Cambridge: Cambridge University Press, 2012. 396 p.

26 BISHOP, C. M. **Pattern recognition and machine learning**. Heidelberg: Springer-Verlag, 2006. ISBN 0387310738.

27 MÜLLER, A. C.; GUIDO, S. **Introduction to machine learning with Python**. Sebastopol: O'Reilly Media, 2016.

28 ACKOFF, R. L. From data to wisdom. **Journal of Applied Systems Analysis**, v. 16, n. 1, p. 3–9, 1989.

29 ROWLEY, J. The wisdom hierarchy: representations of the DIKW hierarchy. **Journal of Information Science**, v. 33, n. 2, p. 163–180, 2007.

30 HAN, J.; PEI, J.; KAMBER, M. **Data mining**: concepts and techniques. 3rd ed. Waltham:Elsevier, 2012. 744 p.

31 AMARAL, F. **Introdução à ciência de dados**: mineração de dados e big data. Rio de Janeiro: Alta Books Editora, 2016.

32 GOLDSCHMIDT, R.; PASSOS, E. **Data mining**: um guia prático. Rio de Janeiro: Elsevier Editora, 2005. 256 p. ISBN 9788535218770.

33 MARQUESONE, R. **Big Data**: técnicas e tecnologias para extração de valor dos dados. São Paulo: Casa do Código, 2016. ISBN 9788555192326.

34 AZEVEDO, A. I. R. L.; SANTOS, M. F. KDD, SEMMA and CRISP-DM: a parallelloverview. *In*: IADS-DM IADIS EUROPEAN CONFERENCE ON DATA MINING 2008, Amsterdam, The Netherlands, July 24-26, 2008. **Proceedings[...]** Amsterdam: Netherlands,2008.

34 CHAPMAN, P. **CRISP-DM 1.0**: step-by-step data mining guide. Chicago: SPSS, 2000. 78 p.

35 WIRTH, R.; HIPP, J. CRISP-DM: Towards a standard process model for data mining. *In*: CITESEER INTERNATIONAL CONFERENCE ON THE PRACTICAL APPLICATIONS OF KNOWLEDGE DISCOVERY AND DATA MINING, 4th,2000. **Proceedings[...]** Manchester, 2000. p. 29–39.

36 MAHESH, B. Machine learning algorithms - a review. **International Journal of Science and Research**, v. 9, n. 1, p. 381–386, 2020.

37 RAY, S. A quick review of machine learning algorithms.INTERNATIONAL CONFERENCE ON MACHINE LEARNING, BIG DATA, CLOUD AND PARALLEL COMPUTING: Trends, Perspectives and Prospects. **Proceedings[...].** [*S.l.*]: IEEE, 2019.p. 35–39.

38 SHINDE, P. P.; SHAH, S. A review of machine learning and deep learning applications. INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION CONTROL AND AUTOMATION, 4th, 2018. **Proceedings [...]**. [*S. l.*]: IEEE, 2018.

39 QIU, J. *et al.* A survey of machine learning for big data processing. **Eurasip Journal on Advances in Signal Processing**, v2016, n. 1, 2016. ISSN 16876180.

40 WEBB, A. R.; COPSEY, K. D.; CAWLEY, G. **Statistical pattern recognition**. 3rd ed.Hoboken: Wiley-Blackwell, 2011. 642 p. ISBN 978-0-470-68227-2.

41 GAN, G.; MA, C.; WU, J. **Data clustering**: theory, algorithms, and applications. Philadelphia: Society for Industrial and Applied Mathematics, 2007. 466 p.

42 VAPNIK, V. N. **The nature of statistical learning theory**. 2nd ed. New York: Springer-Verlag, 2000. 314 p.

43 HAWKINS, D. M. The problem of overfitting. **Journal of Chemical Information and Computer Sciences**, v. 44, n. 1, p. 1–12, 2004. ISSN 00952338.

44 LAROSE, D. T.; LAROSE, C. D. **Discovering knowledge in data**: an introduction to data mining. Hoboken: Wiley, 2014. 336 p. (Wiley series on methods and applications in data mining). ISBN 9780470908747.

45 NG, A. Y.; JORDAN, M. I.; WEISS, Y. **On spectral clustering**: analysis and an algorithm. 2002. Disponível em: https://ai.stanford.edu/~ang/papers/nips01-spectral.pdf. Acesso em: 23 jan. 2021.

46 MORETTIN, P. A.; BUSSAB, W. d. O. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017. 453 p. ISBN 9788547220235.

47 HOWARD, A. G. **Some improvements on deep convolutional neural network based image classification**. 2014. Disponível em: https://arxiv.org/ftp/arxiv/papers/1312/1312.5402.pdf. Acesso em: 23 jan. 2020.

48 BELEITES, C. *et al.* Sample size planning for classification models. **Analytica Chimica Acta**, v. 760, p. 25–33, 2013. ISSN 00032670.

49 RAUDYS, S. J.; JAIN, A. K. Small sample size effects in statistical pattern recognition: recommendations for practitioners. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 13, n. 3, p. 252–264, 1991.

50 FIGUEROA, R. L. *et al.* Predicting sample size required for classification performance. **BMC Medical Informatics and Decision Making**, v. 12, n. 1, p. 8, 2012. ISSN 14726947.

51 BARBEDO, J. G. A. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. **Computers and Electronics in Agriculture**, v. 153, p. 46–53, 2018. ISSN 01681699.

52 ZHENG, A.; CASARI, A. **Feature engineering for machine learning**: principles and techniques for data scientists. Sebastopol: O'Reilly, 2018. 200 p. ISBN 9781491953242.

53 GUYON, I. An introduction to variable and feature selection. **Journal of Machine Learning Research**, v. 3, p. 1157–1182, 2003.

54 HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, n. 2-3, p. 107–145, 2001. ISSN 09259902.

55 ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, n. C, p. 53–65, 1987. ISSN 03770427.

56 FLORKOWSKI, C. M. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. **Clinical Biochemist Reviews**, v. 29, p. S83–7, 2008. ISSN 0159-8090.

57 KAWAMURA, T. Interpretação de um teste sob a visão epidemiológica: eficiência de um teste. **Arquivos Brasileiros de Cardiologia**, v. 79, n. 4, p. 437–441, 2002. ISSN 0066-782X.

58 KIM, J. H. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. **Computational Statistics and Data Analysis**, v. 53, n. 11, p. 3735–3745, 2009. ISSN 01679473.

59 KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and modelselection**. 1995. Disponível em: https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf. Acesso em: 23 jan. 2021.

60 WAINER, J.; CAWLEY, G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. **Expert Systems with Applications**, v. 182, p. 115222, 2021. ISSN 09574174.

61 VARMA, S.; SIMON, R. Bias in error estimation when using cross-validation for model selection. **BMC Bioinformatics**, v. 7, p. 1–8, 2006. ISSN 14712105.

62 TSAMARDINOS, I.; RAKHSHANI, A.; LAGANI, V. Performance-estimation propertiesof cross-validation-based protocols with simultaneous hyper-parameter optimization. **International Journal on Artificial Intelligence Tools**, v. 24, n. 5, p. 1–29, 2015. ISSN 17936349.

63 ELLINGER, M. *et al.* Error propagation in spectrometric functions of soil organic carbon. **SOIL**, v. 5, n. 2, p. 275–288, 2019.

64 JAMES, G. *et al.* **An introduction to statistical learning**. New York:Springer, 2013. ISBN 978-1-4614-7137-0.

65 MENDES, K. D. S.; SILVEIRA, R. C. C. P.; GALVÃO, C. M. Revisao integrativa: método de pesquisa para a incorporacao de evidências na saúde e na enfermagem. **Texto Contexto Enfermagem**, v. 17, n. 4, p. 758–764, 2008.

66 NATIONAL LIBRARY OF MEDICINE. **Medical subject headings (MeSH)**. 2021. Disponível em: https://www.nlm.nih.gov/mesh/meshhome.html. Acesso em: 1 jun. 2021.

67 HIRASAWA, Y. *et al.* Diagnostic performance of Oncuria™, a urinalysis test for bladder cancer. **Journal of Translational Medicine**, v. 19, n. 1, p. 1–10, 2021. ISSN 14795876.

68 YANG, Y. *et al.* Human ACE2-functionalized gold "Virus-Trap" nanostructures for accurate capture of SARS-CoV-2 and single-virus SERS detection. **Nano-Micro Letters**, v. 13, n. 1, 2021. ISSN 21505551.

69 SONG, Y. *et al.* Machine learning-based cytokine microarray digital immunoassay analysis. **Biosensors and Bioelectronics**, v. 180, p. 113088, 2021. ISSN 18734235.

70 MENDELS, D. A. *et al.* Using artificial intelligence to improve COVID-19 rapid diagnostic test result interpretation. **Proceedings of the National Academy of Sciences of the United States of America**, v. 118, n. 12, p. 3–5, 2021. ISSN 10916490.

71 GOEBEL, C. *et al.* Blood test shows high accuracy in detecting stage i non-small cell lung cancer. **BMC Cancer**, v. 20, n. 1, p. 1–11, 2020. ISSN 14712407.

72 XU, Y. *et al.* Ultrasensitive and selective detection of SARS-CoV-2 using thermotropic liquid crystals and image-based machine learning. **Cell Reports Physical Science**, v. 1, n. 12, 2020. ISSN 26663864.

73 CHEN, H. *et al.* Quantitation of femtomolar-level protein biomarkers using a simple microbubbling digital assay and bright-field smartphone imaging. **Angewandte Chemie International Edition in English**, v. 58, n. 39, p. 13922–13928, 2019. ISSN 1521-3773.

74 BANAEI, N. *et al.* Machine learning algorithms enhance the specificity of cancer biomarker detection using SERS-based immunoassays in microfluidic chips. **RSC Advances**, v. 9, n. 4, p. 1859–1868, 2019. ISSN 20462069.

75 GONZALEZ, R. C.; WOODS, R. E. **Digital image processing**. 4th ed. Upper SaddleRiver: Pearson, 2018. 1192 p. ISBN 9780133356724.

76 BEEBE, K. R.; PELL, R. J.; SEASHOLTZ, M. B. **Chemometrics**: a practical guide. Nashville: John Wiley & Sons, 1999. 360 p. ISBN 978-0-471-12451-1.

77 BOCKLITZ, T. *et al.* How to pre-process Raman spectra for reliable and stable models? **Analytica Chimica Acta**, v. 704, n. 1-2, p. 47–56, 2011. ISSN 00032670.

78 GAUTAM, R. *et al.* Review of multidimensional data processing approaches for Raman and infrared spectroscopy. **EPJ Techniques and Instrumentation**, v. 2, n. 1, 2015. ISSN 2195-7045.

79 SKANSI, S. **Introduction to deep learning**. Basel, Switzerland: Springer International Publishing, 2018. 191 p.

80 LIU, W. *et al.* A survey of deep neural network architectures and their applications. **Neurocomputing**, v. 234, p. 11–26, 2017. ISSN 18728286.

81 HUMEAU-HEURTIER, A. Texture feature extraction methods: a survey. **IEEE Access**, v. 7, p. 8975–9000, 2019. ISSN 21693536.

82 WESZKA, J. S.; DYER, C. R.; ROSENFELD, A. A comparative study of texture measures for terrain classification. **IEEE Transactions on Systems, Man, and Cybernetics**, v. SMC-6,n. 4, p. 269–285, 1976.

83 JOURNAUX, L. *et al.* Texture classification with generalized fourier descriptors in dimensionality reduction context: an overview exploration. **Lecture Notes in Computer Science**, v. 5064, p. 280–291, 2008. ISSN 03029743.

84 BACKES, A. R.; CASANOVA, D.; BRUNO, O. M. Texture analysis and classification: a complex network-based approach. **Information Sciences**, v. 219, p. 168–180, 2013. ISSN 00200255.

85 BACKES, A. R.; CASANOVA, D.; BRUNO, O. M. Plant leaf identification based on volumetric fractal dimension. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 23, n. 6, p. 1145–1160, 2009. ISSN 02180014.

86 ZHU, Z. *et al.* An adaptive hybrid pattern for noise-robust texture analysis. **Pattern Recognition**, v. 48, n. 8, p. 2592–2608, 2015. ISSN 00313203.

87 OJALA, T.; PIETIKÄINEN, M.; MÄENPÄÄ, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 7, p. 971–987, 2002. ISSN 01628828.

88 RIBAS, L. C. *et al.* Fusion of complex networks and randomized neural networks for texture analysis. **Pattern Recognition**, v. 103, p. 107189, 2020. ISSN 00313203.

89 LEMLEY, J.; BAZRAFKAN, S.; CORCORAN, P. Smart augmentation learning an optimal data augmentation strategy. **IEEE Access**, v. 5, p. 5858–5869, 2017. ISSN 21693536.

90 KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. **ImageNet classification with deep convolutional neural networks**. 2012. Disponível em: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf. Acesso em: 23 jan. 2021.

91 TYAGI, V. **Understanding digital image processing**. Boca Raton: CRC Press, 2018. 374 p.

92 GONZALEZ, R. C.; WOODS, R. C. **Processamento digital de imagens**. 3. ed. SãoPaulo: Pearson Prentice Hall, 2010. 624 p.

93 MAATEN, L. van der; HINTON, G. Visualizing Data using t-SNE. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008.

94 PAULOVICH, F. V. *et al.* Information visualization techniques for sensing and biosensing. **Analyst**, v. 136, n. 7, p. 1344–1350, 2011.

95 SOARES, J. C. *et al.* Detection of a SARS-CoV-2 sequence with genosensors using data analysis based on information visualization and machine learning techniques. **Materials Chemistry Frontiers**, v. 5, n. 15, p. 5658–5670, 2021.

96 PEDREGOSA, F. *et al.* Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

97 DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. 2nd ed. Nashville,TN: John Wiley & Sons, 2000. (A wiley-interscience publication).

98 VLASOV, Y. G. *et al.* 'Electronic tongue' - new analytical tool for liquid analysis on the basis of non-specific sensors and methods of pattern recognition. **Sensors and Actuators B:** chemical, v. 65, n. 1, p. 235–236, 2000. ISSN 09254005.

99 VLASOV, Y. *et al.* Nonspecific sensor arrays ("electronic tongue") for chemical analysis of liquids (IUPAC Technical Report). **Pure and Applied Chemistry**, v. 77, n. 11, p. 1965–1983, 2005.

100 RIUL, A. *et al.* Recent advances in electronic tongues. **Analyst**, v. 135, n. 10, p. 2481–2495, 2010. ISSN 13645528.

101 TAHARA, Y.; TOKO, K. Electronic tongues – a review. **IEEE Sensors Journal**, v. 13, n. 8, p. 3001–3011, Aug. 2013. ISSN 1530-437X.

102 PODRAZKA, M. *et al.* Electronic tongue - a tool for all tastes? **Biosensors**, v. 8, n. 1, p. 1–24, 2017. ISSN 20796374.

103 BRAZ, D. C. *et al.* Using machine learning and an electronic tongue for discriminating saliva samples from oral cavity cancer patients and healthy individuals. **Talanta**, v. 243, p. 123327, 2022. ISSN 00399140.

104 GOLDBERGER, J. *et al.* Neighbourhood components analysis. INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 17th, 2004. **Proceedings [...]**. Cambridge: MIT Press, 2004. p. 513–520.

105 KUHN, M.; JOHNSON, K. **Feature engineering and selection:** a practical approach for predictive models. Boca Raton: CRC Press, 2019. 1–297 p. ISBN 9781351609470.

106 RIUL, A. *et al.* Wine classification by taste sensors made from ultra-thin films and using neural networks. **Sensors and Actuators B:** chemical, v. 98, n. 1, p. 77–82, 2004. ISSN 0925-4005.

107 FERREIRA, E. J. *et al.* Random subspace method for analysing coffee with electronic tongue. **Electronics Letters**, v. 43, n. 21, p. 1138–1140, 2007. ISSN 0013-5194.

108 NICOLICHE, C. Y. N. *et al.* Converging multidimensional sensor and machine learning toward high-throughput and biorecognition element-free multidetermination of extracellular vesicle biomarkers. **ACS Sensors**, v. 5, n. 7, p. 1864–1871, 2020. ISSN 2379-3694.

109 CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers and Electrical Engineering**, v. 40, n. 1, p. 16–28, 2014. ISSN 00457906.

110 BRERETON, R. G. **Chemometrics**. 2nd ed. Nashville: John Wiley & Sons, 2018.

111 VIRTANEN, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature Methods**, v. 17, p. 261–272, 2020.

112 ORTIZ-AGUAYO, D.; WAEL, K. D.; DEL VALLE, M. Voltammetric sensing using an array of modified SPCE coupled with machine learning strategies for the improved identification of opioids in presence of cutting agents. **Journal of Electroanalytical Chemistry**, v. 902, p. 115770, 2021. ISSN 15726657.

113 PÉREZ-RÀFOLS, C. *et al.* Voltammetric electronic tongues in food analysis. **Sensors**, v. 19, n. 19, Sept. 2019. ISSN 1424-8220 (Electronic).

114 WEI, Z. *et al.* The measurement principles, working parameters and configurations of voltammetric electronic tongues and its applications for foodstuff analysis. **Journal of Food Engineering**, v. 217, p. 75–92, 2018. ISSN 02608774.

115 MORAIS, T. C. B. de *et al.* A simple voltammetric electronic tongue for the analysis of coffee adulterations. **Food Chemistry**, v. 273, p. 31–38, 2019. ISSN 18737072.

116 KIANI, H. *et al.* Application of a voltammetric electronic tongue combined with chemometric approaches for the early classification of heavy metals in sunflower oil. **Journal of Food Processing and Preservation**, v. 45, n. 9, p. 0–1, 2021. ISSN 17454549.

117 HERRERA-CHACÓN, A. *et al.* Voltammetric electronic tongue for the simultaneous determination of three benzodiazepines. **Sensors**, v. 19, n. 22, p. 1–12, 2019. ISSN 14248220.

118 YASLI, A. Cancer detection with surface plasmon resonance-based photonic crystal fiber biosensor. **Plasmonics**, v. 16, n. 5, p. 1605–1612, 2021. ISSN 15571963.

119 BELLASSAI, N. *et al.* Surface plasmon resonance for biomarker detection: advances in non-invasive cancer diagnosis. **Frontiers in Chemistry**, v. 7, p. 1–16, 2019. ISSN 22962646.

120 HOMOLA, J. Surface plasmon resonance sensors for detection of chemical and biological species. **Chemical Reviews**, v. 108, n. 2, p. 462–493, 2008. ISSN 00092665.

121 ZENG, S. *et al.* Nanomaterials enhanced surface plasmon resonance for biological and chemical sensing applications. **Chemical Society Reviews**, v. 43, n. 10, p. 3426–3452, 2014. ISSN 14604744.

122 HUTTER, E.; FENDLER, J. H. Exploitation of localized surface plasmon resonance. **Advanced Materials**, v. 16, n. 19, p. 1685–1706, 2004. ISSN 09359648.

123 PATTNAIK, P. Surface plasmon resonance. **Applied Biochemistry and Biotechnology**,v. 126, n. 2, p. 79–92, 2005. ISSN 1559-0291.

124 MIYAZAKI, C. M.; SHIMIZU, F. M.; FERREIRA, M. 6 - Surface plasmon resonance (SPR) for sensors and biosensors. *In*: DA RÓZ, A. L. *et al.* (ed.). **Nanocharacterization techniques**. [*S.l.*]: William Andrew Publishing, 2017. p. 183-200. (Micro and nano technologies). ISBN 978-0-323-49778-7.

125 CAMPION, A.; KAMBHAMPATI, P. Surface-enhanced Raman scattering. **Chemical Society Reviews**, v. 27, n. 4, p. 241–250, 1998.

126 CHOI, N. *et al.* SERS biosensors for ultrasensitive detection of multiple biomarkers expressed in cancer cells. **Biosensors and Bioelectronics**, v. 164, p. 112326,2020. ISSN 18734235.

127 SAVIÑON-FLORES, F. *et al.* A review on Sers-based detection of human virus infections: influenza and coronavirus. **Biosensors**, v. 11, n. 3, 2021. ISSN 20796374.

128 RUBIRA, R. J. G. *et al.* Increasing the sensitivity of surface-enhanced Raman scattering detection for s-triazine pesticides by taking advantage of interactions with soil humic substances. **Journal of Raman Spectroscopy**, v. 53, n. 1, p. 40–48, 2022.

129 GUO, Z. *et al.* Detection of heavy metals in food and agricultural products by surface-enhanced Raman spectroscopy. **Food Reviews International**, p.1–22, 2021. ISSN 15256103.

130 CHOI, J. H. *et al.* In situ detection of neurotransmitters from stem cell-derived neural interface at the single-cell level via graphene-hybrid SERS nanobiosensing. **Nano Letters**, v. 20, n. 10, p. 7670–7679, 2020. ISSN 15306992.

131 GELADI, P.; GRAHN, H. F. Multivariate image analysis. *In*: **Encyclopedia of analytical chemistry**. New York: John Wiley, 2006. ISBN 9780470027318.

132 LIEBER, C. A.; MAHADEVAN-JANSEN, A. Automated method for subtraction of fluorescence from biological Raman spectra. **Applied Spectroscopy**, v. 57, n. 11, p. 1363–1367, 2003. ISSN 0003-7028.

133 FENG, S. *et al.* Esophageal cancer detection based on tissue surface-enhanced Raman spectroscopy and multivariate analysis. **Applied Physics Letters**, v. 102, n. 4, 2013. ISSN 00036951.

134 HEO, N. S. *et al.* Simple diagnosis of HbA1c using the dual-plasmonic platform integrated with LSPR and SERS. **Journal of Crystal Growth**, v. 469, p.154–159, 2017. ISSN 00220248.

135 CHEN, H. *et al.* Sensitive detection of SARS-CoV-2 using a SERS-based aptasensor. **ACS Sensors**, v. 6, n. 6, p. 2378–2385, 2021.

136 TADESSE, L. F. *et al.* Toward rapid infectious disease diagnosis with advances in surface-enhanced Raman spectroscopy. **Journal of Chemical Physics**, v152, n. 24, 2020. ISSN 10897690.

137 PARK, S. *et al.* SERSNet: surface-enhanced Raman spectroscopy-based biomolecule detection using deep neural network. **Biosensors**, v. 11, n. 12, p. 490, 2021. ISSN 20796374.

138 ALBUQUERQUE, C. D. L. *et al.* Digital protocol for chemical analysis at ultralow concentrations by surface-enhanced Raman scattering. **Analytical Chemistry**, v. 90, n. 2, p. 1248–1254, 2018. ISSN 15206882.

139 PINHEIRO, J. I. D. *et al.* **Estatística básica**: a arte de trabalhar com dados. Rio de Janeiro: Elsevier, 2009. 312 p. ISBN 9788535230307.

140 DEZA, M.-M.; DEZA, E. **Dictionary of distances**. Oxford: Elsevier, 2006. ISBN 978-0-444-52087-6.

141 GOSHTASBY, A. A. **Image registration**: principles, tools and methods. New York: Springer Science & Business Media, 2012. 441 p. ISBN 978-1-4471-2457-3.

142 CHA, S.-H. Comprehensive survey on distance/similarity measures between probability density functions. **City**, v. 1, n. 2, p. 1, 2007.

143 PRATT, W. K. **Digital image processing**: Piks scientific inside. 4th ed. Hoboken: Wiley-Interscience, 2007. 807 p. ISBN 9780471767770.

144 DOUGHERTY, G. **Digital image processing for medical applications**. Cambridge: Cambridge University Press, 2009. 447 p. ISBN 978-0-511-53343-3.

145 GOLDSTEIN, E. B. **Sensation and perception**. 8th ed. Belmont: Cengage Learning, 2010. 459 p. ISBN 9780495601494.

146 ESKICIOGLU, A. M.; FISHER, P. S. Image quality measures and their performance. **IEEE Transactions on Communications**, v. 43, n. 12, p. 2959–2965, 1995. ISSN 1558-0857.

147 LI, S.; KWOK, J. T.; WANG, Y. Combination of images with diverse focuses using the spatial frequency. **Information Fusion**, v. 2, n. 3, p. 169–176, 2001. ISSN 15662535.

148 LOESER, R. F. *et al.* Osteoarthritis: a disease of the joint as an organ. **Arthritis & Rheumatism**, v. 64, n. 6, p. 1697–1707, 2012. ISSN 0004-3591.

149 WOOLF, A. D.; PFLEGER, B. Burden of major musculoskeletal conditions. **Bulletin of the World Health Organization**, v. 81, n. 9, p. 646–656, 2003.ISSN 0042-9686.

150 KOHN, M. D.; SASSOON, A. A.; FERNANDO, N. D. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. **Clinical Orthopaedics and Related Research**, v. 474, n. 8, p. 1886–1893, 2016. ISSN 1528-1132.

151 ROEMER, F. W. *et al.* Best practice & research clinical rheumatology: the role of imaging in osteoarthritis. **Best Practice & Research Clinical Rheumatology**, v. 28, n. 1, p. 31–60, 2014. ISSN 1521-6942.

152 ALTMAN, R. D.; GOLD, G. E. Atlas of individual radiographic features in osteoarthritis,revised. **Osteoarthritis and Cartilage**, v. 15, p. 1–56, 2007. ISSN 10634584.

153 KELLGREN, J. H.; LAWRENCE, J. S. Radiological assessment of osteo-arthrosis. **Annals of the Rheumatic Diseases**, v. 16, n. 4, p. 494–502, 1957. ISSN 0003-4967.

154 WOLOSZYNSKI, T.; PODSIADLO, P.; STACHOWIAK, G. W. A signature dissimilaritymeasure for trabecular bone texture in knee radiographs. **Medical Physics**, v. 37, n. 5, p. 2030–2042, 2010.

155 PODSIADLO, P.; WOLSKI, M.; STACHOWIAK, G. W. Automated selection of trabecular bone regions in knee radiographs. **Medical Physics**, v35, n. 5, p. 1870–1883, 2008. ISSN 00942405.

156 JANVIER, T. *et al.* Subchondral tibial bone texture analysis predicts knee osteoarthritis progression: data from the Osteoarthritis Initiative Tibial bone texture & knee OA progression. **Osteoarthritis and Cartilage**, v. 25, n. 2, p. 259–266, 2017. ISSN 1063-4584.

157 RIAD, R. *et al.* Texture analysis using complex wavelet decomposition for knee osteoarthritis detection: data from the osteoarthritis initiative. **Computers and Electrical Engineering**, v. 68, p. 181–191, 2018. ISSN 00457906.

158 BACKES, A. R. **Estudo de métodos de análise de complexidade em imagens**. 2001. 161 p. Tese (Doutorado em Ciências da Computação) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2001.

159 GONÇALVES, W. N. **Análise de texturas estáticas e dinâmicas e suas aplicações em biologia e nanotecnologia**. 2013. 187 p. Tese (Doutorado em Ciências) — Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2013.

160 CASANOVA, D. **Redes complexas em visão computacional com aplicações em bioinformática**. 2013. 192 p. Tese (Doutorado em Ciências) — Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2013.

APÊNDICES

APÊNDICE A – GENOSSENSOR E MEV PARA DIAGNÓSTICO DE CÂNCER DE PRÓSTATA: INFORMAÇÕES COMPLEMENTARES

A.1 Análise exploratória

Esta seção apresenta os resultados dos experimentos para conhecimento, avaliação e comparação de atributos básicos das imagens, fornecendo subsídios para geração de modelos de AM. Foram realizadas medidas estatísticas, de distâncias e de frequência espacial cujos resultados serão apresentados nas seções a seguir.

Medidas estatísticas

Para entender o comportamento das intensidades dos pixels das imagens do conjunto de dados, foram realizadas medidas estatísticas de centralidade, dispersão e quartis.(47, 140) A distribuição das intensidades dos pixels foi analisada também através de histogramas e do gráfico de caixas (*boxplot*) agrupados por classe.

As tabelas A.1 e A.2 apresentam as medidas de centralidade das intensidades do pixels das imagens e um resumo contendo seus valores médios por classe e do conjunto. A média e a mediana possuem valores próximos entre si e pouco abaixo do centro ($= 128$) da escala de intensidades. Os valores por imagem não exibem uma relação monotônica (crescente ou decrescente) com a concentração do analito. A moda evidencia que há grande número de imagens cuja intensidade mais frequente é 0 (zero), especialmente nas classes 3 a 5.

As tabelas A.3 e A.4 apresentam as medidas de dispersão das intensidades do pixels das imagens e um resumo contendo seus valores médios por classe e do conjunto. Nota-se que as intensidades mínimas das imagens da classe 2 diferenciam-se das demais classes. Também pode-se verificar que a extensão da variação das intensidades ocupa quase toda a escala de intensidades ($0 - 255$). O desvio padrão (std) e o coeficiente de variação(140) (vc) indicam razoável dispersão das intensidades em torno da média. As medidas *skew (skewness)* e *kur (kurtosis)* são, respectivamente, medidas de assimetria e curtose da distribuição das intensidades em relação à distribuição normal. (47, 140) Quanto ao nível de assimetria, temos as seguintes categorias: baixa ($|skew| \approx 0$), moderada ($|skew| \leq 1$) e alta ($|skew| > 1$). Uma assimetria positiva evidencia que $moda < mediana < média$, significando que os pixels estão mais concentrados em intensidades menores que a média. Uma assimetria negativa tem significado contrário. Em relação à curtose, temos as seguintes categorias: achatamento ($kur < 0$) e elevação ($kur > 0$). O nível é maior quanto maior for a diferença em relação a 0 (zero). Os valores obtidos mostram distribuições moderadamente assimétricas positivas e negativas, e relativamente achatadas. Os histogramas da figura A.1 permitem visualizar essas características mais evidentes nas classes 2 (A.10c) a 8 (A.11g) em que há a presença do analito. Também é possível visualizar uma redistribuição dos pixels para as intensidades mais elevadas com o

aumento da concentração do analito, especialmente para as classes 2 (A.10c) a 7 (A.11e).

As tabelas A.5 e A.6 apresentam as medidas dos quartis das intensidades do pixels das imagens e um resumo contendo seus valores médios por classe e do conjunto. As imagens da classe 2 exibem limites inferiores (linf) diferentes das demais classes e também, a partir da medida noutliers, uma pequena presença de intensidades anômalas/discrepantes (*outliers*). Estas superam as intensidades dos demais pixels em 1.5 vezes o intervalo interquartil (iqr) a partir do quartil-3 (q3).(47, 140) A partir da medida iqr e dos gráficos de caixa apresentados na figura A.2, verifica-se que distribuição das intensidades em torno do quartil-2 (q2), também conhecido como mediana, é razoavelmente similar entre as classes com a presença do analito (2-8). Esse fato é importante, pois essa região da distribuição concentra 50% da quantidade de intensidades dos pixels.

Medidas de distâncias

As medidas de distância(141,142) podem ser utilizadas para avaliação da dissimilaridade (diferença) entre pontos num sistema de coordenadas, entre pontos num espaço de dados e também entre conjuntos de valores. Nesta tese, as imagens (2D) podem ter seus pixels arranjados na forma de vetores (1D) e diversas medidas de distância podem ser utilizadas para essa finalidade. A tabela A.7 apresenta o resumo das medidas de distância entre as imagens de cada classe do conjunto de dados. Para comparação, foram realizadas 5 medidas de distância(143): Euclideana (Euc) (eq. A.1), Canberra (Can) (eq. A.3), Tanimoto (Tan) (eq. A.3), Jaccard (Jac) (eq. A.4) e Kumar-Johnson (KJ) (eq. A.5). Essas foram escolhidas por representarem um método de cada família de distâncias.(143) Todas as medidas tiveram suas escalas ajustas para o intervalo [0, 1].

$$d_{Euc} = \sqrt{\sum_{i=1}^{I} |P_i - Q_i|^2} \tag{A.1}$$

$$d_{Can} = \sum_{i=1}^{I} \frac{|P_i - Q_i|}{P_i + Q_i} \tag{A.2}$$

$$d_{Tan} = \frac{\sum_{i=1}^{I} (max(P_i, Q_i) - min(P_i, Q_i))}{\sum_{i=1}^{I} max(P_i, Q_i)} \tag{A.3}$$

$$d_{Jac} = \frac{\sum_{i=1}^{I} P_i^2 + \sum_{i=1}^{I} Q_i^2 - \sum_{i=1}^{I} P_i \times Q_i}{\sum_{i=1}^{I} P_i^2 + \sum_{i=1}^{I} Q_i^2 - \sum_{i=1}^{I} P_i \times Q_i} \tag{A.4}$$

$$d_{KJ} = \sum_{i=1}^{I} \frac{(P_i^2 - Q_i^2)^2}{2\,(P_i Q_i)^{3/2}} \tag{A.5}$$

De forma geral, as diferentes medidas apresentam elevados valores, evidenciando a existência de grande dissimilaridade entre as imagens. A medida KJ apresenta valores mais

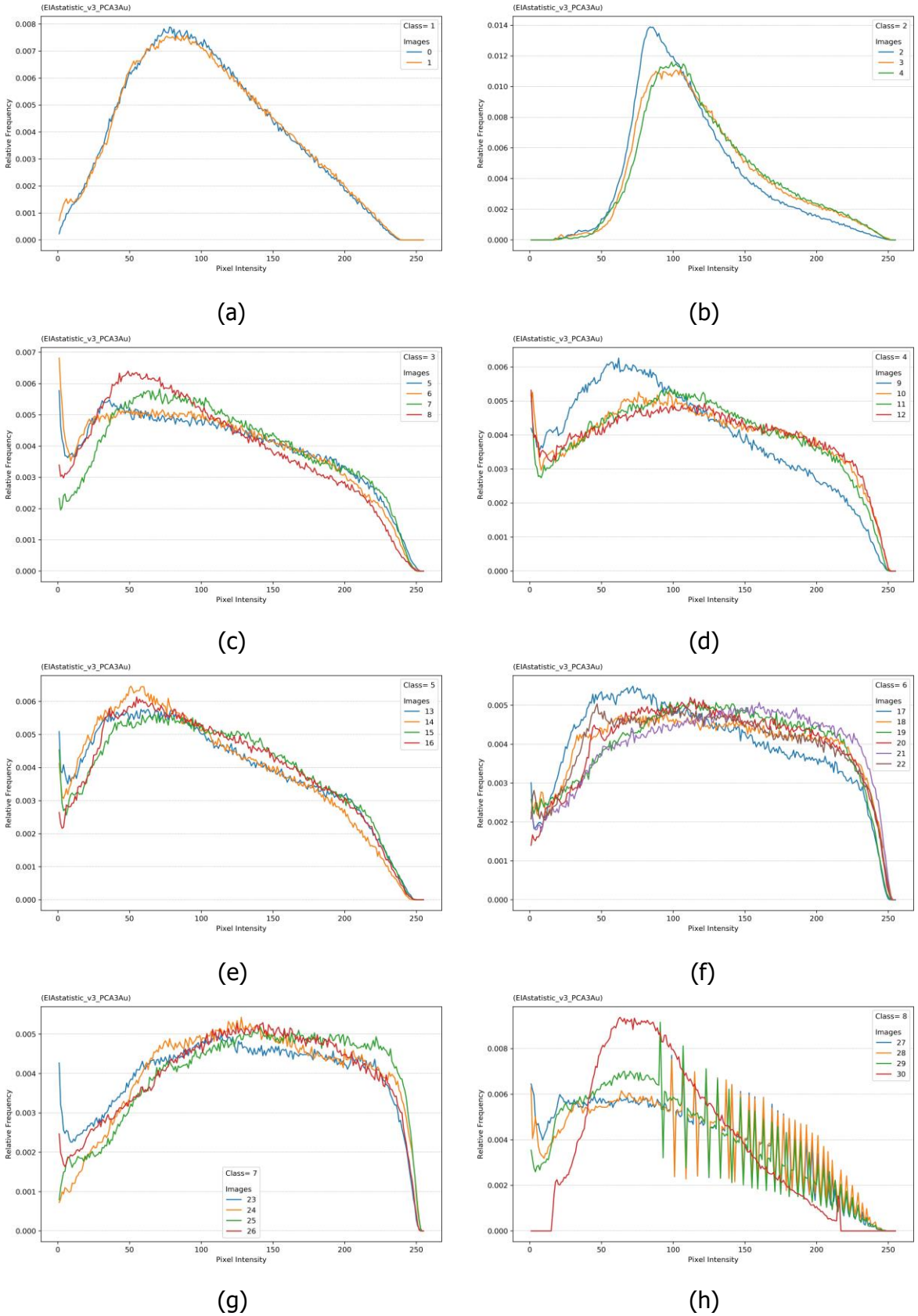


Figura A.1 – Histogramas das intensidades dos pixels das imagens do conjunto de dados agrupados por classe: 1 (a), 2 (c), 3 (e), 4 (g), 5 (a), 6 (c), 7 (e), 8 (g).
Fonte: Elaborada pelo autor.

baixos para a maioria das classes. Não há uma dispersão elevada em relação às médias das medidas. Quando se avaliam as medidas conjuntamente para cada classe, como na figura A.3, verifica-se uma diferença significativa entre as classes sem analito (A.3a e A.3b) e com analito (A.3c a A.3h). Essa evidência pode indicar um bom desempenho em análises binárias com relação à presença do analito que sejam baseadas nessas medidas.

Medidas de frequência espacial

A frequência espacial(144–146) é uma medida global da imagem que permite avaliar a periodicidade espacial das mudanças de intensidade dos pixels. Uma imagem com alta frequência espacial possui predominância de transições abruptas entre as intensidades dos pixels, isto é, as transições ocorrem em pequenos intervalos de pixels e entre grandes diferenças de intensidades. Esse tipo de comportamento está associado à existência de bordas nítidas e finas nos vários elementos da imagem, portanto conferindo maior contraste entre eles. Já uma imagem com baixa frequência espacial possui majoritariamente transições suaves, isto é, as transições ocorrem em intervalos maiores de pixels e entre menores diferenças de intensidades. Esse tipo de comportamento está associado à não-existência ou existência de bordas mais indefinidas e de longo alcance nos vários elementos da imagem, evidenciando baixo contraste entre eles. A figura A.4 fornece exemplos de imagens contendo elementos pontuais para o oferecimento de algumas referências de valores de frequências espaciais.

ESKICIOGLU(147) e LI(148) apresentam as expressões para medição da frequência espacial (*spatial_freq*) de uma imagem $M \times N$, onde M é o número de linhas e N é o número de colunas presentes na imagem. A unidade de medida é *ciclos/pixel*. Considerando-se $F(m, n)$ como sendo a intensidade de um pixel na posição (m, n), temos:

$$row\_freq = \sqrt{\frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=1}^{N-1} [F(m,n) - F(m,n-1)]^2} \tag{A.6}$$

$$col\_freq = \sqrt{\frac{1}{MN} \sum_{n=0}^{N-1} \sum_{m=1}^{M-1} [F(m,n) - F(m-1,n)]^2} \tag{A.7}$$

$$spatial\_freq = \sqrt{row\_freq^2 + col\_freq^2} \tag{A.8}$$

As tabelas A.8 e A.9 apresentam as medidas de frequência espacial das imagens e um resumo contendo seus valores médios por classe e do conjunto. Os valores das frequências espaciais (*spatial_freq*), em ciclos/pixel, estão distribuídos em um intervalo de baixos valores ($13.956 - 28.399$) com uma média de $21.826(\pm 2.816)$.

A partir da figura A.5, especificamente, em A.5c, verifica-se uma clara separação entre as classes sem analito (1) e com analito (2 a 7), favorecendo a classificação binária das amostras baseada nesse descritor. Entretanto, verifica-se também que as medidas da classe 0

(negativo) sobrepõem-se às das classes 2 a 7 (com analito), podendo, a depender da técnica de classificação empregada, levar à classificação de falsos-positivos.

A.2 Análise comparativa entre as janelas e as imagens originais

Para verificar a representatividade dos novos conjuntos, foram determinados os histogramas das intensidades dos pixels das janelas com dimensões 100 (Figuras A.6, A.7), 200 (Figuras A.8, A.9) e 300 (Figuras A.10, A.11), e comparadas aos histogramas das imagens originais. Para isso, foram escolhidas aleatoriamente 10 janelas de cada classe. Verifica-se que os histogramas das janelas são cada vez mais similares e representativos das imagens originais à medida que o tamanho da janela aumenta. Isso permite utilizar os conjuntos de janelas ao invés do conjunto original de imagens, e pode-se comparar o desempenho dos modelos gerados a partir de cada um. A tabelas A.10 e A.11 apresentam os resultados da acurácia na classificação binária e multiclasse dos conjuntos de atributos de textura extraídos dos conjuntos de dados de janelas 100x100 e 200x200 pixels respectivamente. Outra consequência é a redução da dimensionalidade do problema de análise das imagens do biossensor.

Tabela A.1 – Medidas de centralidade das intensidades dos pixels das imagens do conjunto de dados.

class	id	mean	median	mode	support
1	0	103.280	98	78	706560
1	1	103.818	99	90	706560
2	2	112.918	105	84	706560
2	3	121.962	114	105	706560
2	4	123.988	115	99	706560
3	5	107.365	103	0	706560
3	6	104.666	100	0	706560
3	7	111.497	106	0	706560
3	8	101.264	93	0	706560
4	9	102.147	95	0	706560
4	10	115.414	112	0	706560
4	11	114.387	112	0	706560
4	12	116.348	115	0	706560
5	13	102.172	95	0	706560
5	14	98.694	91	0	706560
5	15	107.215	103	0	706560
5	16	105.619	100	55	706560
6	17	115.681	111	0	706560
6	18	123.997	123	91	706560
6	19	125.587	127	0	706560
6	20	125.954	125	112	706560
6	21	131.320	134	160	706560
6	22	123.162	122	47	706560
7	23	127.118	128	0	706560
7	24	134.259	134	128	706560
7	25	138.342	141	141	706560
7	26	131.776	134	143	706560
8	27	95.674	90	99	706560
8	28	101.207	96	99	706560
8	29	95.631	88	91	706560
8	30	94.386	89	0	706560

Fonte: Elaborada pelo autor.

Tabela A.2 – Média (avg) e desvio padrão (std) das medidas de centralidade das imagens do conjunto de dados.

class	id	mean	median	mode	support
1	avg	103.549	98.500	84.000	2
1	std	0.380	0.707	8.485	2
2	avg	119.623	111.333	96.000	3
2	std	5.894	5.508	10.817	3
3	avg	106.198	100.500	0.000	4
3	std	4.325	5.568	0.000	4
4	avg	112.074	108.500	0.000	4
4	std	6.666	9.110	0.000	4
5	avg	103.425	97.250	13.750	4
5	std	3.792	5.315	27.500	4
6	avg	124.283	123.667	68.333	6
6	std	5.086	7.528	64.214	6
7	avg	132.874	134.250	103.000	4
7	std	4.696	5.315	68.988	4
8	avg	96.725	90.750	72.250	4
8	std	3.048	3.594	48.314	4
all	avg	113.447	109.613	52.323	31
all	std	12.634	15.422	55.814	31

Fonte: Elaborada pelo autor.

Tabela A.3 – Medidas de dispersão das intensidades dos pixels das imagens do conjunto de dados.

class	id	min	max	minmax	var	std	vc	skew	kur	support
1	0	0	239	239	2452.109	49.519	0.479	0.328	-0.593	706560
1	1	0	240	240	2549.520	50.493	0.486	0.304	-0.617	706560
2	2	17	251	234	1524.584	39.046	0.346	0.866	0.479	706560
2	3	19	252	233	1771.168	42.085	0.345	0.685	-0.028	706560
2	4	17	252	235	1766.860	42.034	0.339	0.657	-0.119	706560
3	5	0	253	253	4245.247	65.156	0.607	0.198	-1.042	706560
3	6	0	252	252	4127.562	64.246	0.614	0.203	-1.004	706560
3	7	0	250	250	3852.611	62.069	0.557	0.211	-0.953	706560
3	8	0	250	250	3710.781	60.916	0.602	0.341	-0.871	706560
4	9	0	250	250	3827.341	61.866	0.606	0.321	-0.863	706560
4	10	0	252	252	4341.898	65.893	0.571	0.083	-1.046	706560
4	11	0	251	251	4146.152	64.391	0.563	0.070	-0.995	706560
4	12	0	251	251	4437.182	66.612	0.573	0.051	-1.066	706560
5	13	0	250	250	3917.092	62.587	0.613	0.279	-0.952	706560
5	14	0	247	247	3628.327	60.236	0.610	0.328	-0.883	706560
5	15	0	250	250	3743.474	61.184	0.571	0.185	-0.950	706560
5	16	0	250	250	3615.680	60.131	0.569	0.246	-0.936	706560
6	17	0	251	251	4162.505	64.517	0.558	0.165	-1.044	706560
6	18	0	253	253	4381.984	66.197	0.534	0.022	-1.101	706560
6	19	0	252	252	4120.862	64.194	0.511	-0.079	-1.007	706560
6	20	0	252	252	4082.501	63.894	0.507	-0.003	-1.035	706560
6	21	0	253	253	4253.841	65.221	0.497	-0.121	-1.029	706560
6	22	0	252	252	4249.062	65.185	0.529	0.034	-1.067	706560
7	23	0	253	253	4353.378	65.980	0.519	-0.071	-1.040	706560
7	24	0	254	254	3933.672	62.719	0.467	-0.049	-0.993	706560
7	25	0	253	253	4094.561	63.989	0.463	-0.182	-0.968	706560
7	26	0	252	252	4066.190	63.767	0.484	-0.135	-0.955	706560
8	27	0	248	248	3522.581	59.351	0.620	0.290	-0.902	706560
8	28	0	249	249	3612.851	60.107	0.594	0.258	-0.906	706560
8	29	0	248	248	3162.271	56.234	0.588	0.407	-0.675	706560
8	30	0	216	216	1929.672	43.928	0.465	0.439	-0.302	706560

Fonte: Elaborada pelo autor.

Tabela A.4 – Média (avg) e desvio padrão (std) das medidas de dispersão das imagens do conjunto de dados.

class	id	min	max	minmax	var	std	vc	skew	kur	support
1	avg	0.000	239.500	239.500	2500.814	50.006	0.483	0.316	-0.605	2
1	std	0.000	0.707	0.707	68.880	0.689	0.005	0.018	0.018	2
2	avg	17.667	251.667	234.000	1687.537	41.055	0.343	0.736	0.111	3
2	std	1.155	0.577	1.000	141.138	1.740	0.004	0.114	0.322	3
3	avg	0.000	251.250	251.250	3984.050	63.097	0.595	0.238	-0.967	4
3	std	0.000	1.500	1.500	245.474	1.947	0.026	0.069	0.074	4
4	avg	0.000	251.000	251.000	4188.143	64.690	0.578	0.131	-0.993	4
4	std	0.000	0.816	0.816	269.322	2.098	0.019	0.127	0.091	4
5	avg	0.000	249.250	249.250	3726.143	61.034	0.591	0.260	-0.930	4
5	std	0.000	1.500	1.500	139.681	1.138	0.024	0.060	0.032	4
6	avg	0.000	252.167	252.167	4208.459	64.868	0.523	0.003	-1.047	6
6	std	0.000	0.753	0.753	109.059	0.838	0.022	0.100	0.033	6
7	avg	0.000	253.000	253.000	4111.950	64.114	0.483	-0.109	-0.989	4
7	std	0.000	0.816	0.816	175.563	1.362	0.026	0.061	0.038	4
8	avg	0.000	240.250	240.250	3056.844	54.905	0.567	0.348	-0.696	4
8	std	0.000	16.174	16.174	776.249	7.508	0.069	0.088	0.284	4
all	avg	1.710	249.226	247.516	3599.468	59.476	0.529	0.204	-0.821	31
all	std	5.318	7.008	8.290	862.885	8.011	0.079	0.246	0.362	31

Fonte: Elaborada pelo autor.

Tabela A.5 – Medidas dos quartis das intensidades dos pixels das imagens do conjunto de dados.

class	id	linf	q1	q2	q3	iqr	lsup	noutliers	support
1	0	0	65	98	138	73	239.000	0	706560
1	1	0	65	99	140	75	240.000	0	706560
2	2	17	85	105	134	49	207.500	19384	706560
2	3	19	90	114	147	57	232.500	6457	706560
2	4	17	93	115	150	57	235.500	4418	706560
3	5	0	52	103	160	108	253.000	0	706560
3	6	0	51	100	156	105	252.000	0	706560
3	7	0	61	106	160	99	250.000	0	706560
3	8	0	51	93	148	97	250.000	0	706560
4	9	0	52	95	149	97	250.000	0	706560
4	10	0	62	112	170	108	252.000	0	706560
4	11	0	63	112	166	103	251.000	0	706560
4	12	0	62	115	171	109	251.000	0	706560
5	13	0	50	95	151	101	250.000	0	706560
5	14	0	49	91	145	96	247.000	0	706560
5	15	0	57	103	155	98	250.000	0	706560
5	16	0	56	100	152	96	250.000	0	706560
6	17	0	62	111	168	106	251.000	0	706560
6	18	0	69	123	180	111	253.000	0	706560
6	19	0	75	127	179	104	252.000	0	706560
6	20	0	74	125	179	105	252.000	0	706560
6	21	0	79	134	186	107	253.000	0	706560
6	22	0	69	122	177	108	252.000	0	706560
7	23	0	74	128	182	108	253.000	0	706560
7	24	0	84	134	186	102	254.000	0	706560
7	25	0	88	141	192	104	253.000	0	706560
7	26	0	82	134	184	102	252.000	0	706560
8	27	0	46	90	141	95	248.000	0	706560
8	28	0	52	96	147	95	249.000	0	706560
8	29	0	51	88	136	85	248.000	0	706560
8	30	0	61	89	123	62	216.000	0	706560

Fonte: Elaborada pelo autor.

Tabela A.6 – Média (avg) e desvio padrão (std) das medidas dos quartis das imagens do conjunto de dados.

class	id	linf	q1	q2	q3	iqr	lsup	noutliers	support
1	avg	0.000	65.000	98.500	139.000	74.000	239.500	0.000	2
1	std	0.000	0.000	0.707	1.414	1.414	0.707	0.000	2
2	avg	17.667	89.333	111.333	143.667	54.333	225.167	10086.333	3
2	std	1.155	4.041	5.508	8.505	4.619	15.373	8116.301	3
3	avg	0.000	53.750	100.500	156.000	102.250	251.250	0.000	4
3	std	0.000	4.856	5.568	5.657	5.123	1.500	0.000	4
4	avg	0.000	59.750	108.500	164.000	104.250	251.000	0.000	4
4	std	0.000	5.188	9.110	10.231	5.500	0.816	0.000	4
5	avg	0.000	53.000	97.250	150.750	97.750	249.250	0.000	4
5	std	0.000	4.082	5.315	4.193	2.363	1.500	0.000	4
6	avg	0.000	71.333	123.667	178.167	106.833	252.167	0.000	6
6	std	0.000	5.955	7.528	5.845	2.483	0.753	0.000	6
7	avg	0.000	82.000	134.250	186.000	104.000	253.000	0.000	4
7	std	0.000	5.888	5.315	4.320	2.828	0.816	0.000	4
8	avg	0.000	52.500	90.750	136.750	84.250	240.250	0.000	4
8	std	0.000	6.245	3.594	10.210	15.564	16.174	0.000	4
all	avg	1.710	65.484	109.613	159.742	94.258	246.661	976.097	31
all	std	5.318	13.682	15.422	18.390	17.367	10.740	3685.174	31

Fonte: Elaborada pelo autor.

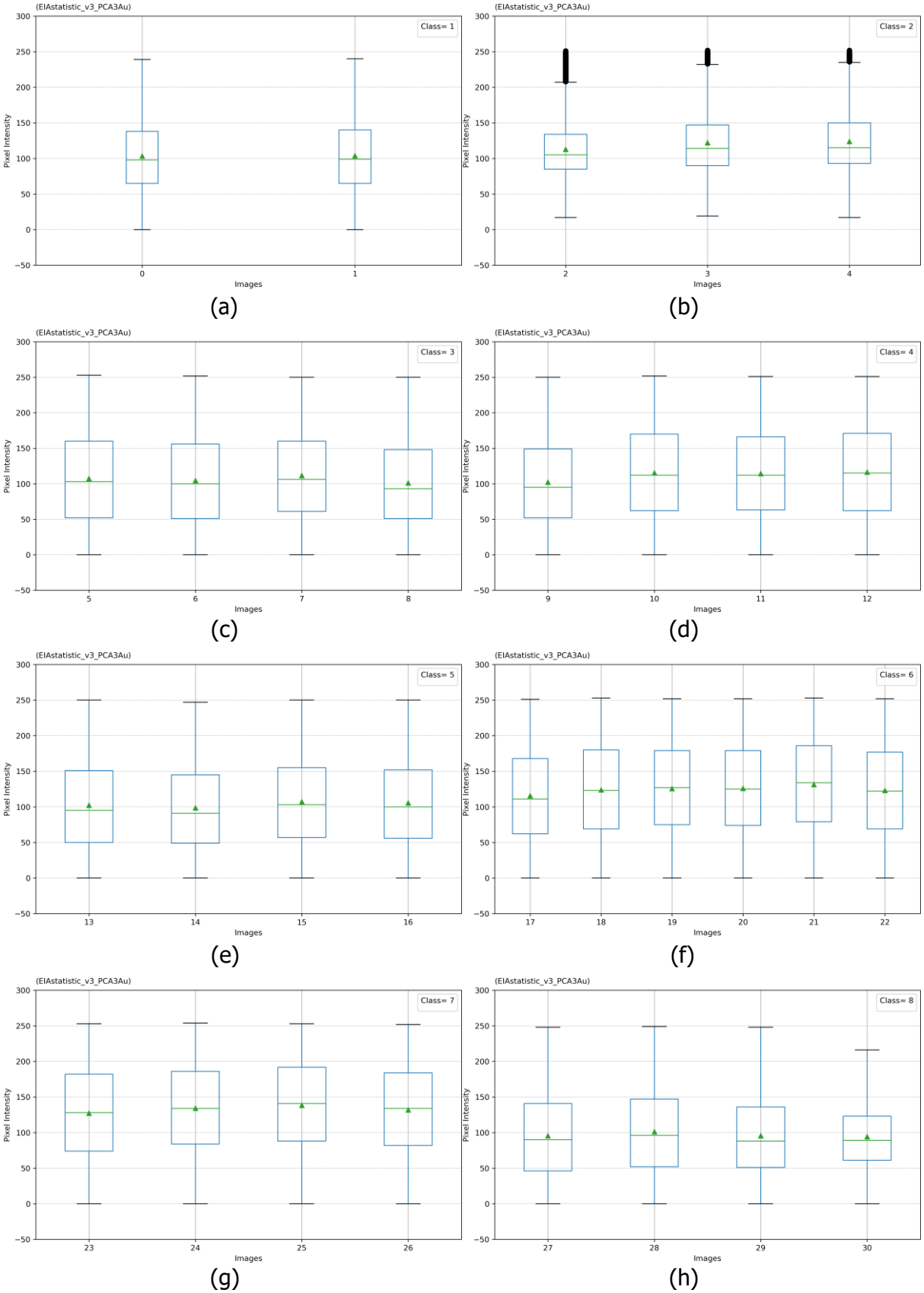


Figura A.2 – Gráfico de caixas (*boxplot*) das intensidades dos pixels das imagens do conjunto de dados agrupados por classe: 1 (a), 2 (b), 3 (c), 4 (d), 5 (e), 6 (f), 7 (g), 8 (h). O triângulo indica o valor da média das intensidades dos pixels.
Fonte: Elaborada pelo autor.

Tabela A.7 – Média (avg) e desvio padrão (std) das medidas de distância entre as imagens do conjunto de dados.

class	id	Euc	Can	Tan	Jac	KJ	support
1	avg	0.726	0.728	0.817	0.722	0.269	1
1	std	0.000	0.000	0.000	0.000	0.000	1
2	avg	0.600	0.468	0.610	0.437	0.427	3
2	std	0.011	0.003	0.004	0.009	0.033	3
3	avg	0.919	0.957	0.966	0.941	0.467	6
3	std	0.013	0.026	0.017	0.027	0.030	6
4	avg	0.944	0.951	0.949	0.913	0.563	6
4	std	0.012	0.010	0.014	0.024	0.052	6
5	avg	0.880	0.936	0.952	0.919	0.392	6
5	std	0.009	0.019	0.014	0.024	0.021	6
6	avg	0.940	0.847	0.874	0.797	0.765	15
6	std	0.008	0.016	0.016	0.025	0.046	15
7	avg	0.929	0.786	0.822	0.717	0.926	6
7	std	0.006	0.024	0.015	0.023	0.048	6
8	avg	0.798	0.888	0.926	0.886	0.274	6
8	std	0.051	0.060	0.039	0.059	0.035	6
all	avg	0.887	0.856	0.887	0.821	0.587	49
all	std	0.092	0.119	0.088	0.127	0.218	49

Fonte: Elaborada pelo autor.

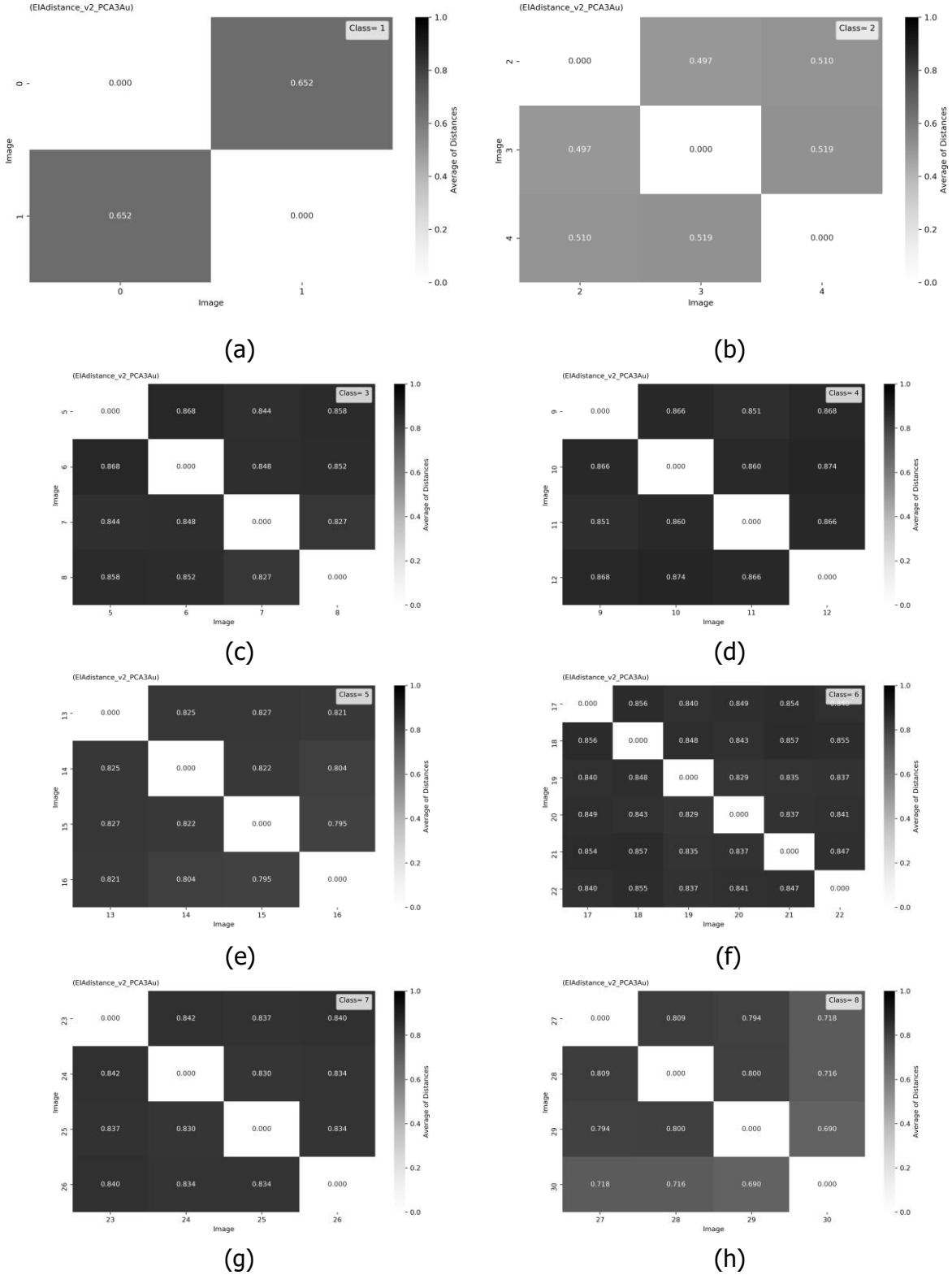
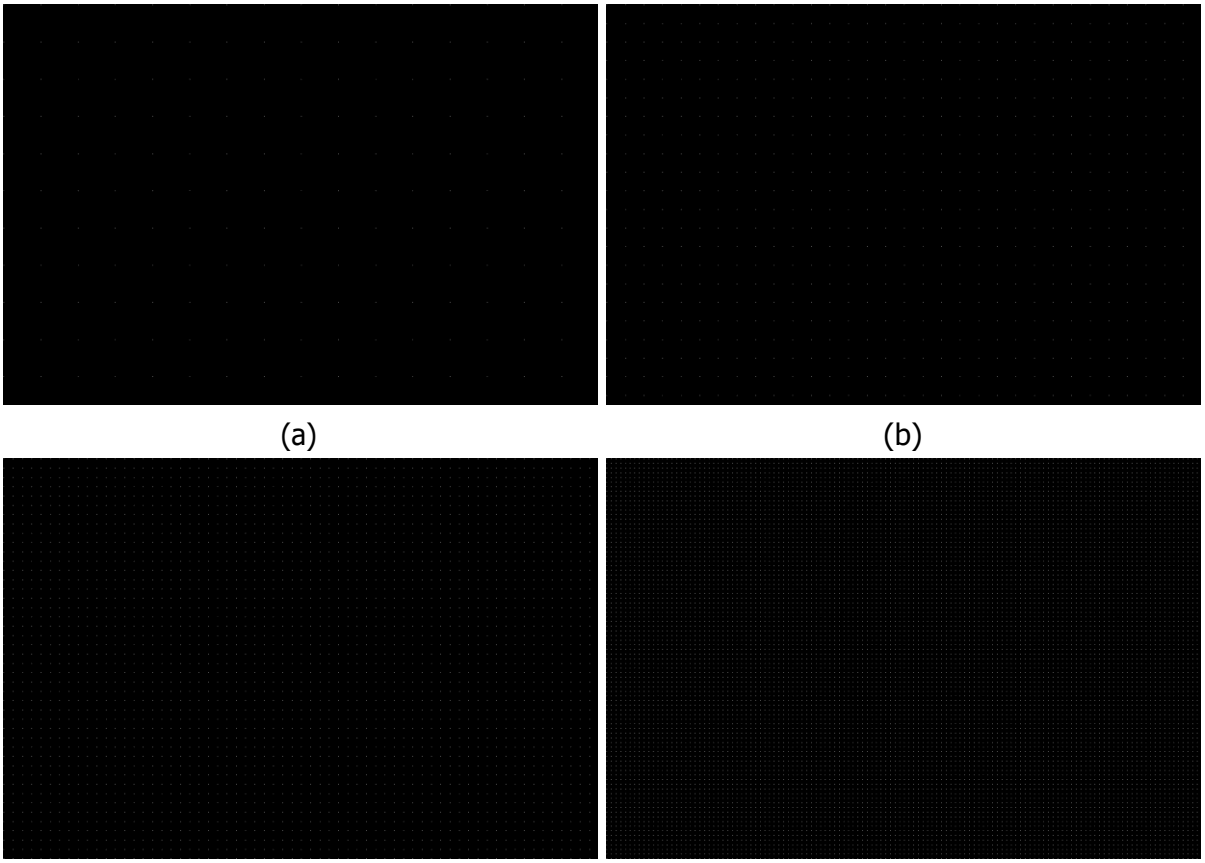


Figura A.3 – Média das medidas de distância entre as imagens do conjunto de dados para cada classe: 1 (a), 2 (b), 3 (c), 4 (d), 5 (e), 6 (f), 7 (g), 8 (h).
Fonte: Elaborada pelo autor.



(a) (b)

(c) (d)

Figura A.4 – Imagens de pontos (tamanho 1 pixel, intensidade 255) distribuídos uniformemente com diferentes frequências espaciais (ciclos/pixel): 8.049 (a), 16.098 (b), 32.105 (c) e 63.935 (d).

Fonte: Elaborada pelo autor.

Tabela A.8 – Medidas das frequências espaciais (ciclos/pixel) das imagens do conjunto de dados.

class	id	row_freq	col_freq	spatial_freq	support
1	0	13.707	15.587	20.757	706560
1	1	16.093	16.788	23.256	706560
2	2	12.119	13.219	17.933	706560
2	3	12.898	12.859	18.213	706560
2	4	12.86	12.705	18.077	706560
3	5	20.18	18.716	27.523	706560
3	6	20.225	19.936	28.399	706560
3	7	16.852	16.625	23.673	706560
3	8	16.086	16.455	23.011	706560
4	9	16.709	16.887	23.757	706560
4	10	18.015	17.99	25.46	706560
4	11	15.828	16.316	22.732	706560
4	12	16.269	16.604	23.246	706560
5	13	15.373	15.551	21.867	706560
5	14	15.623	15.979	22.347	706560
5	15	14.798	15.33	21.307	706560
5	16	15.694	15.372	21.968	706560
6	17	15.648	15.434	21.979	706560
6	18	16.023	16.221	22.8	706560
6	19	15.24	14.959	21.355	706560
6	20	15.106	14.796	21.145	706560
6	21	17.113	17.177	24.247	706560
6	22	15.763	15.926	22.408	706560
7	23	15.983	16.062	22.659	706560
7	24	14.871	15.021	21.137	706560
7	25	15.556	15.41	21.897	706560
7	26	15.421	15.515	21.875	706560
8	27	13.888	13.5	19.368	706560
8	28	13.649	13.71	19.346	706560
8	29	12.87	13.844	18.903	706560
8	30	9.645	10.086	13.956	706560

Fonte: Elaborada pelo autor.

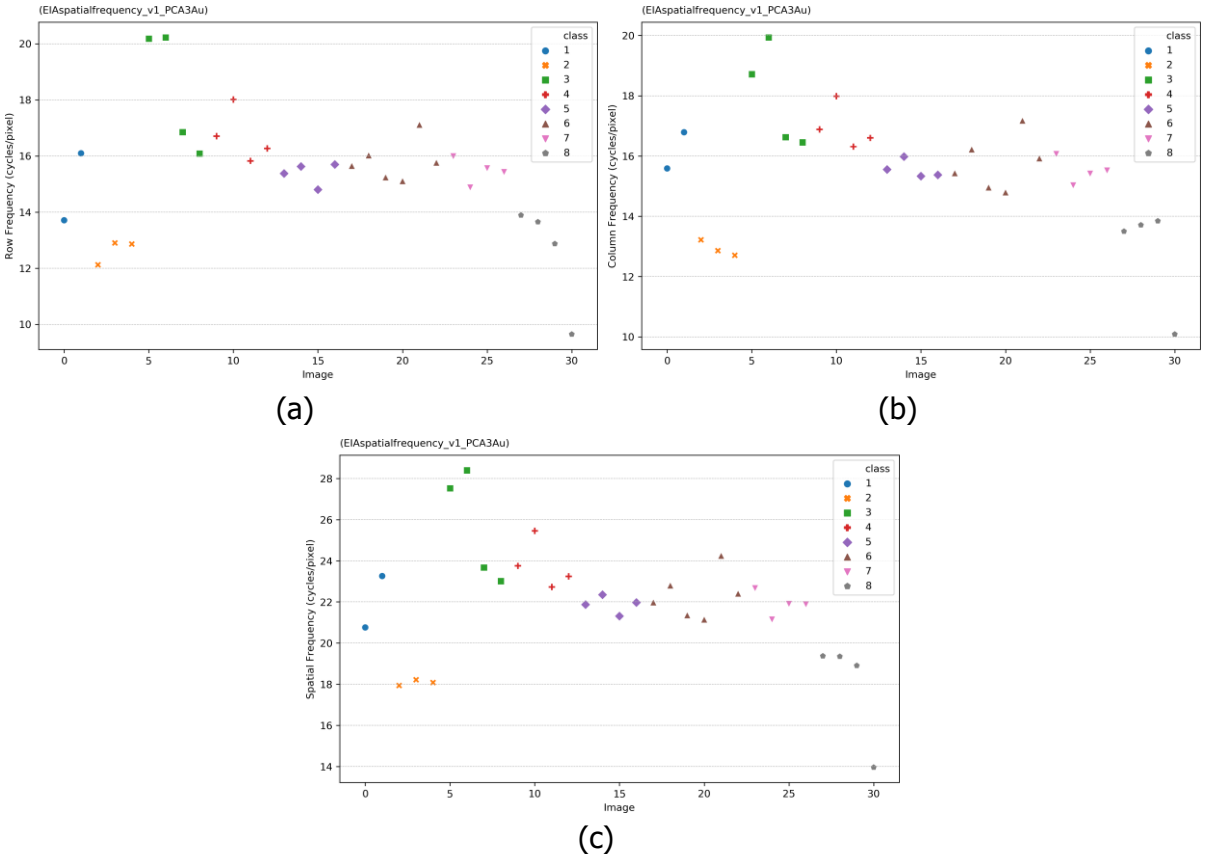


Figura A.5 – Medidas das frequências espaciais das imagens do conjunto de dados na direção das linhas (a), das colunas (b) e resultante (c).
Fonte: Elaborada pelo autor.

Tabela A.9 – Média (avg) e desvio padrão (std) das medidas das frequências espaciais (ciclos/pixel) das imagens do conjunto de dados.

class	id	row_freq	col_freq	spatial_freq	support
1	avg	14.9	16.188	22.007	2
1	std	1.687	0.849	1.767	2
2	avg	12.626	12.927	18.075	3
2	std	0.439	0.264	0.14	3
3	avg	18.336	17.933	25.651	4
3	std	2.178	1.685	2.704	4
4	avg	16.705	16.949	23.799	4
4	std	0.944	0.732	1.184	4
5	avg	15.372	15.558	21.872	4
5	std	0.407	0.297	0.43	4
6	avg	15.816	15.752	22.322	6
6	std	0.72	0.886	1.13	6
7	avg	15.458	15.502	21.892	4
7	std	0.459	0.429	0.621	4
8	avg	12.513	12.785	17.893	4
8	std	1.961	1.805	2.634	4
all	avg	15.358	15.503	21.826	31
all	std	2.118	1.9	2.816	31

Fonte: Elaborada pelo autor.

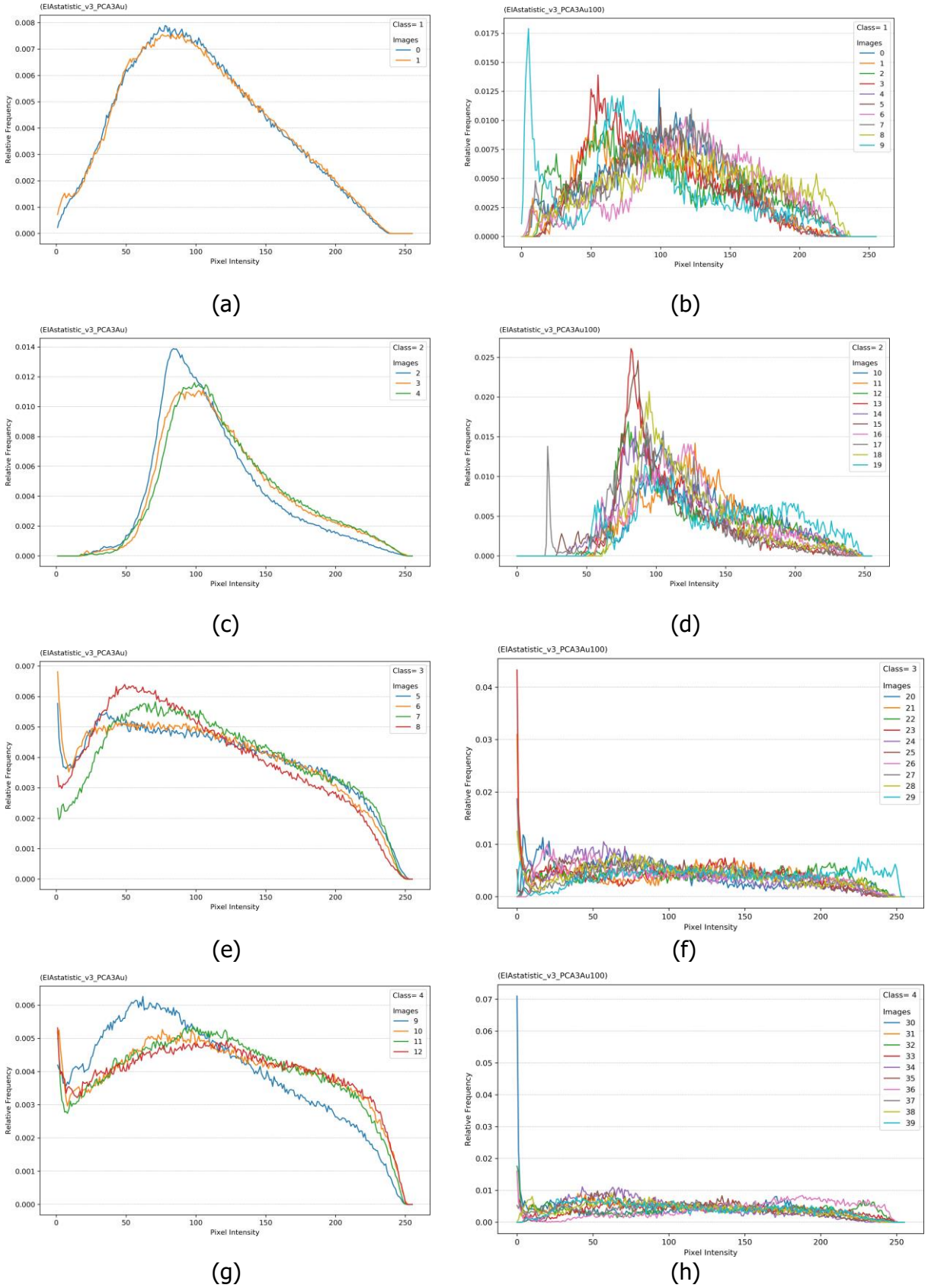


Figura A.6 – Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 100 pixels do conjunto de dados agrupados por classe (1 a 4).

Fonte: Elaborada pelo autor.

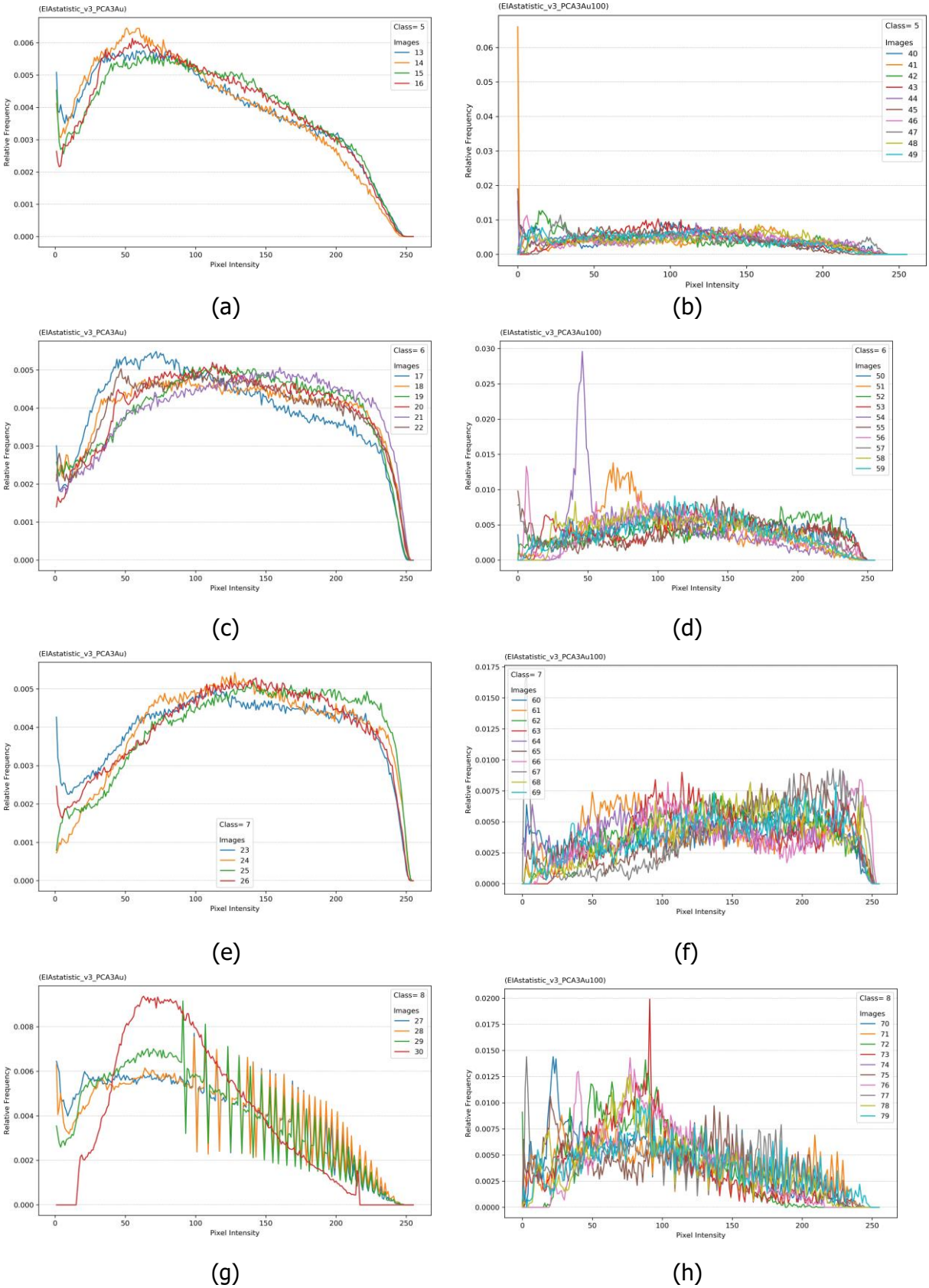


Figura A.7 – Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 100 pixels do conjunto de dados agrupados por classe (5 a 8).

Fonte: Elaborada pelo autor.

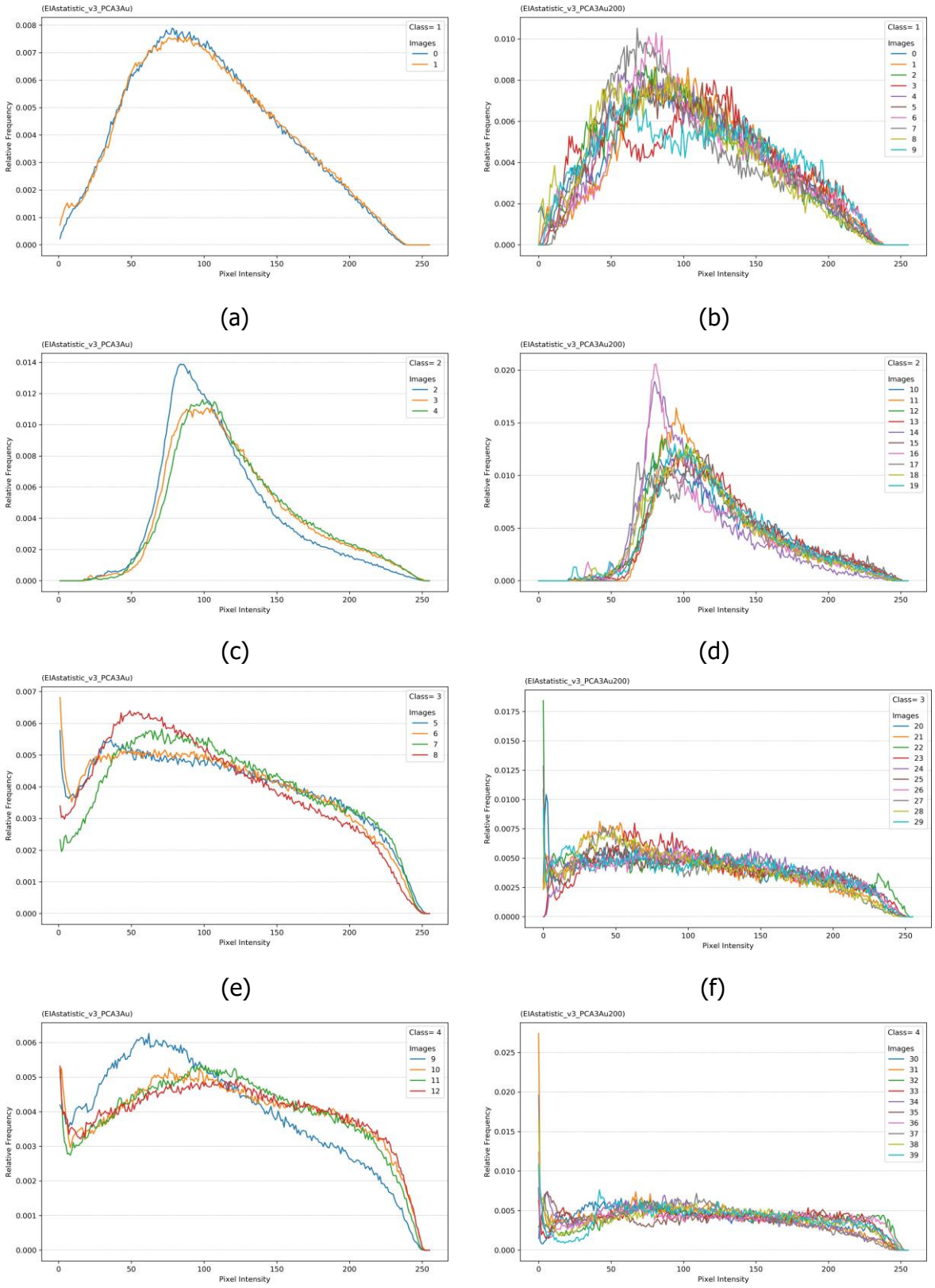


Figura A.8 – Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 200 pixels do conjunto de dados agrupados por classe (1 a 4).

Fonte: Elaborada pelo autor.

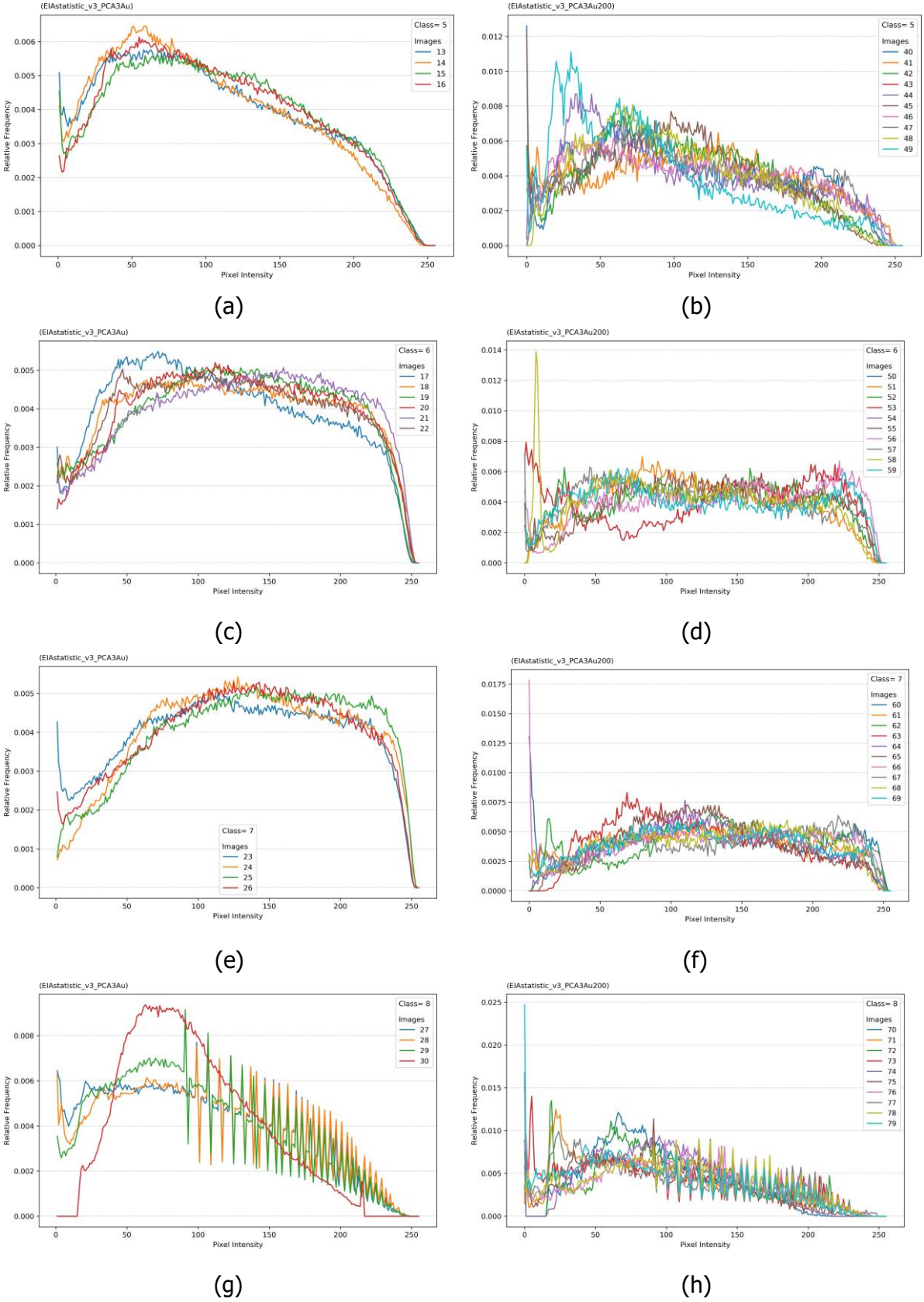


Figura A.9 – Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 200 pixels do conjunto de dados agrupados por classe (5 a 8).

Fonte: Elaborada pelo autor.

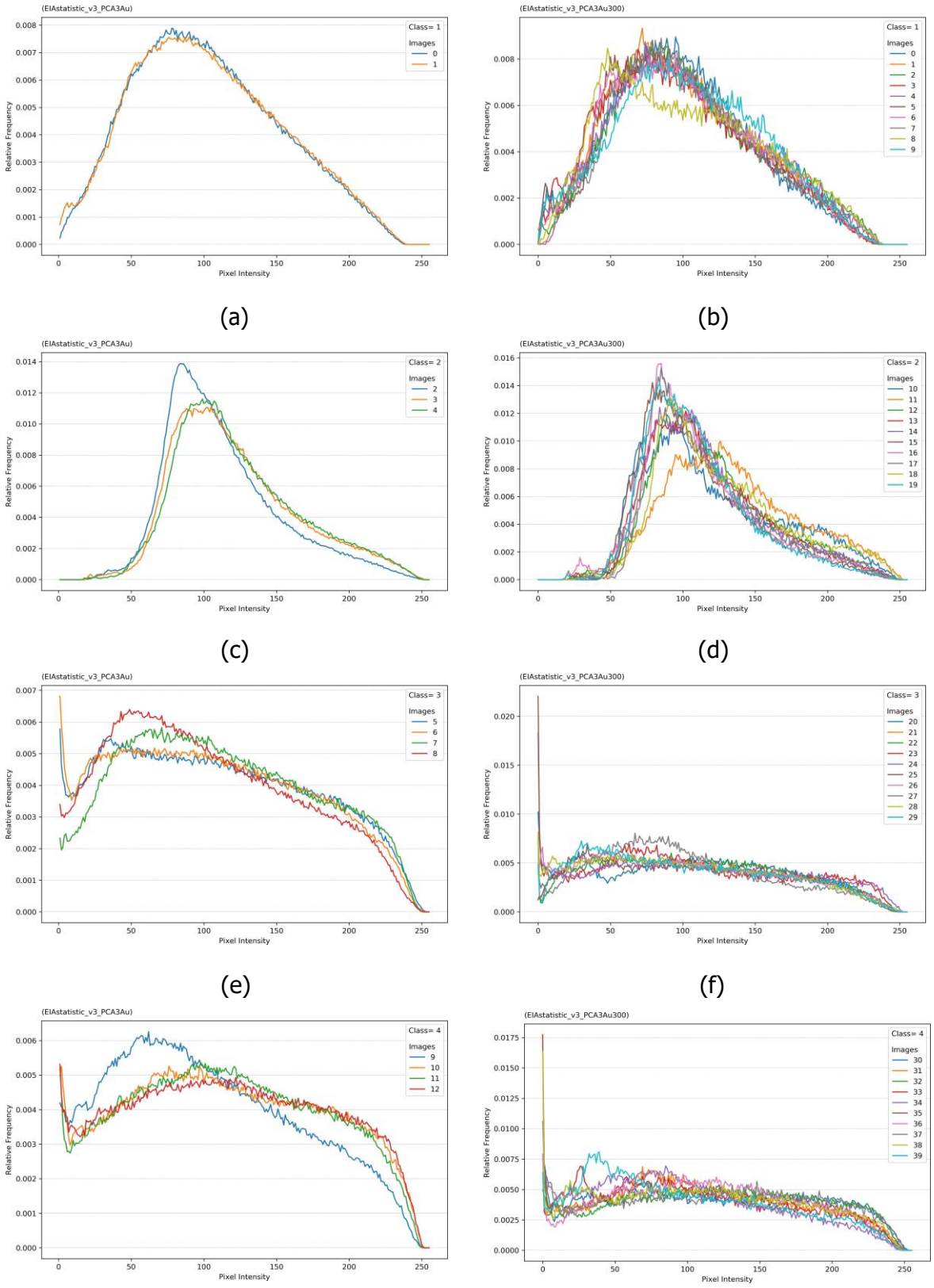


Figura A.10 – Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 300 pixels do conjunto de dados agrupados por classe (1 a 4).

Fonte: Elaborada pelo autor.

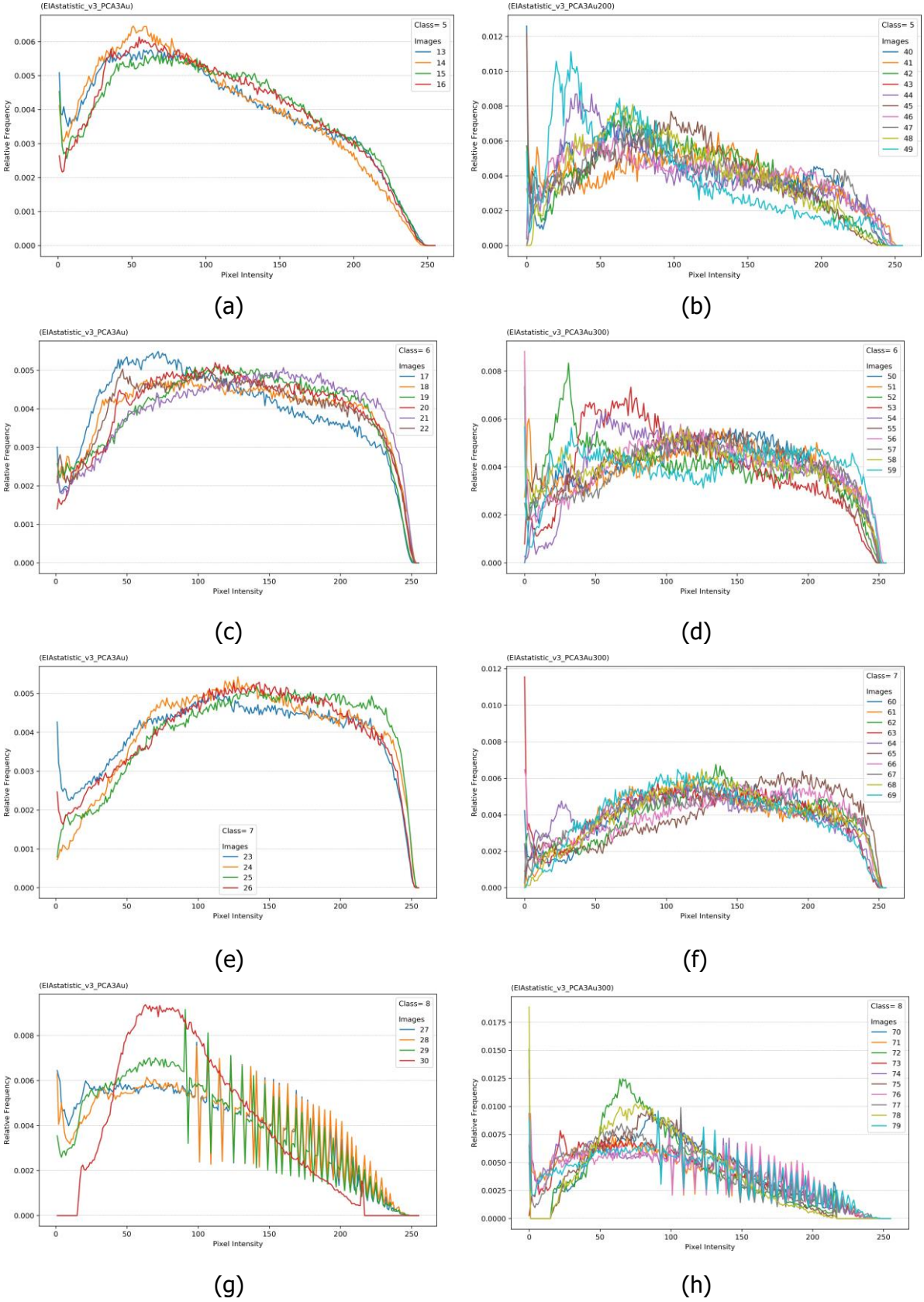


Figura A.11 – Histogramas das intensidades dos pixels das imagens (a, c, e, g) e das respectivas janelas (b, d, f, h) de 300 pixels do conjunto de dados agrupados por classe (5 a 8).

Fonte: Elaborada pelo autor.

Tabela A.10 – Acurácia média (desvio padrão) em (%) da análise de classificação computada com os algoritmos LDA, SVM e 1-NN para os casos binário e multiclasse para os conjuntos de atributos de textura extraídos das janelas com dimensões 100x100 pixels.

Conjuntos de atributos de textura	Binário			Multiclasse		
	LDA	SVM	1-NN	LDA	SVM	1-NN
GLDM	98.6	98.7	95.3	77.6	78.7	46.5
	(0.4)	(0.5)	(0.9)	(1.1)	(1.1)	(1.5)
Fourier	96.7	97.2	97.8	67.3	69.2	57.9
	(0.6)	(0.5)	(0.7)	(1.1)	(1.1)	(1.7)
CNTD	98.3	98.2	89.8	73.3	75.0	43.7
	(0.4)	(0.4)	(1.4)	(1.2)	(1.3)	(1.5)
Fractal	95.5	93.6	91.4	65.0	52.9	40.2
	(0.7)	(0.8)	(1.3)	(1.2)	(1.2)	(1.6)
AHP	98.5	97.9	95.6	71.4	73.6	56.8
	(0.5)	(0.5)	(0.8)	(1.2)	(1.3)	(1.5)
LBP	91.1	92.7	85.9	64.0	65.0	49.7
	(1.4)	(1.3)	(1.3)	(1.4)	(1.4)	(1.6)
CNRNN	99.7	98.2	97.0	81.2	75.0	54.0
	(0.2)	(0.5)	(0.7)	(1.3)	(1.3)	(1.3)
LCFNN	99.0	98.9	93.2	81.5	80.1	54.5
	(0.4)	(0.4)	(0.8)	(1.1)	(1.4)	(1.6)

Fonte: Elaborada pelo autor.

Tabela A.11 – Acurácia média (desvio padrão) em (%) da análise de classificação computada com os algoritmos LDA, SVM e 1-NN para os casos binário e multiclasse para os conjuntos de atributos de textura extraídos das janelas com dimensões 200x200 pixels.

Conjuntos de atributos de textura	Binário			Multiclasse		
	LDA	SVM	1-NN	LDA	SVM	1-NN
GLDM	99.7	99.3	98.2	83.4	83.5	55.1
	(0.5)	(0.6)	(0.8)	(2.4)	(2.3)	(3.0)
Fourier	90.2	99.0	98.7	65.0	74.7	65.3
	(3.2)	(0.7)	(0.7)	(3.2)	(2.4)	(2.6)
CNTD	97.1	97.8	95.9	76.2	80.1	54.1
	(1.8)	(1.0)	(1.8)	(2.3)	(2.2)	(3.2)
Fractal	98.2	97.2	96.2	71.3	61.6	48.0
	(1.2)	(1.1)	(1.9)	(3.0)	(2.7)	(3.0)
AHP	95.9	97.9	97.4	75.0	80.1	67.5
	(2.0)	(0.5)	(1.5)	(2.8)	(2.6)	(3.1)
LBP	95.1	98.8	90.9	51.9	73.2	62.8
	(2.6)	(0.9)	(2.0)	(3.7)	(2.6)	(3.4)
CNRNN	95.4	99.5	99.0	71.0	80.4	65.0
	(2.6)	(0.5)	(0.7)	(3.2)	(2.1)	(2.8)
LCFNN	99.3	99.7	99.2	72.7	86.9	72.1
	(0.8)	(0.4)	(0.7)	(3.4)	(2.2)	(2.7)

Fonte: Elaborada pelo autor.

APÊNDICE B – LÍNGUA ELETRÔNICA E ESPECTROSCOPIA DE IMPEDÂNCIA PARA DIAGNÓSTICO DE CÂNCER DE BOCA: INFORMAÇÕESCOMPLEMENTARES

Esta seção contem os resultados dos experimentos exploratórios para conhecimento, avaliação e comparação de atributos básicos dos espectros de impedância, fornecendo subsídios para geração de modelos de AM. Os resultados encontram-se organizados na pasta do repositório desta aplicação e podem ser acessados em **https://drive.google.com/drive/folders/1t37DmFlZecBIMZxxdY5zWbI-LGFLv3l6?usp=sharing**.

APÊNDICE C – IMUNOSSENSOR E VOLTAMETRIA PARA DIAGNÓSTICO DE CÂNCER: INFORMAÇÕES COMPLEMENTARES

Esta seção contem os resultados dos experimentos exploratórios para conhecimento, avaliação e comparação de atributos básicos dos voltamogramas, fornecendo subsídios para geração de modelos de AM. Os resultados encontram-se organizados na pasta do repositório desta aplicação e podem ser acessados em **https://drive.google.com/drive/folders/1_PRQ_XHiDarLDlmLHLIFMjcrPIHCbjNQ?usp=sharing**.

APÊNDICE D – IMUNOSSENSOR E SERS PARA DIAGNÓSTICO DE COVID-19: INFORMAÇÕES COMPLEMENTARES

Esta seção contem os resultados dos experimentos exploratórios para conhecimento, avaliação e comparação de atributos básicos dos mapas hiperspectrais de SERS, fornecendo subsídios para geração de modelos de AM. Os resultados encontram-se organizados na pasta do repositório desta aplicação e podem ser acessados em **https://drive.google.com/drive/folders/1tww8mpud0AES4wWArw5yHdU-_DVqFU3a?usp=sharing**.

APÊNDICE E – AM E REDES COMPLEXAS APLICADAS AO DIAGNÓSTICO DE OSTEOARTRITE

E.1 Introdução

A Osteoartrite (OA) é uma doença que acomete a região das articulações (ombros, joelhos, etc), sendo caracterizada pela presença de alterações na estrutura das articulações tais como: degradação da cartilagem articular, espessamento do osso subcondral (osso abaixo da cartilagem), formação dos osteofitos, graus variáveis de inflamação da sinóvia, degeneração dos ligamentos e, no joelho, meniscos, e hipertrofia da cápsula articular. Essas modificações resultam em dor, deformidades, limitações aos movimentos e até em perda da função das articulações.(149)

A prevalência da OA é significativa. Segundo a Organização Mundial de Saúde (OMS) a OA afeta cerca de 9,6% dos homens e 18% das mulheres com idade superior a 60 anos. Espera-se que aumentos na expectativa de vida e no envelhecimento da população tornem a osteoartrite a quarta principal causa de incapacidade até o ano 2020.(150)

Em termos de exame de imagem, a modalidade radiografia é o padrão-ouro para diagnóstico da OA(151,152), definido pela presença de algumas características na imagem(153), e no caso de jelhos, principalmente, pela presença de osteofitos nas margens da articulação(154). O sistema atualmente aceito para classificação do nível de severidade da OA foi proposto por Kellgren-Lawrence (151, 154) (KL) possuindo 5 níveis de classificação: 0-normal, 1-duvidoso, 2-mínimo, 3-moderado e 4-severo.

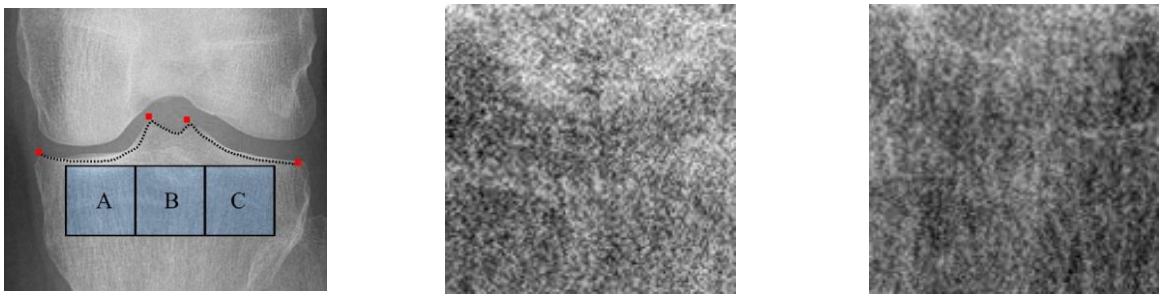
Uma maneira de observar as modificações na estrutura óssea do joelho devido à OA é através das modificações no padrão de textura das imagens da região do osso subcondral da tíbia. Na literatura há vários trabalhos propondo métodos computacionais para análise da textura dessa região da imagem para auxílio ao diagnostico da OA.(155–157)

Em específico, o trabalho de RIAD *et al.*(158) preparou e utilizou o mesmo conjunto de dados analisado nesta pesquisa de doutorado. Nesse trabalho é proposto um novo método baseado em wavelets para realizar a análise de textura de radiografias do joelho, obtendo uma acurácia de 80.38% na classificação de imagens com níveis de OA $KL = 0$ e $KL = 2$.

Neste contexto, e a partir da experiência do Grupo de Computação Científica do IFSC/USP, liderado pelo Prof. Odemir Martinez Bruno (orientador deste trabalho de aplicação), em pesquisas envolvendo redes complexas e análise de texturas em imagens(159–161), o objetivo deste trabalho é propor novas abordagens computacionais para auxiliar o diagnóstico da OA em joelhos através do emprego dos conceitos e métricas de redes complexas e aplicação de Aprendizado de Máquina para a análise de texturas de radiografias.

E.2 Conjunto de dados

Tendo sido preparado(158), utilizado e cedido pelo Prof. Dr. Rachid Jennane da *University of Orléans* (França), o conjunto de dados utilizado neste estudo é composto da região B da extremidade superior do osso da tíbia (Figura E.1a) em 688 imagens radiográficas do joelho direito de diferentes pacientes. As imagens originais estão disponíveis na base de dados de acesso aberto da *Osteoarthritis Iniative (OAI)**. Foram utilizadas imagens com níveis de OA (ou classes) KL=0 (ausência) (Figura E.1b) e KL=2 (mínimo) (Figura E.1c). As imagens possuem 128 x 128 pixels de tamanho correspondendo à região central da extremidade superior do osso da tíbia. As intensidades dos pixels são quantizadas em 8 bits (escala de cinza), variando de 0-255.



(a) (b) (c)

Figura E.1 – Imagens da região de interesse (a) e pertencentes ao conjunto de dados de OA com níveis KL=0 (ausência)(b) e KL=2 (mínimo)(c).

Fonte – RIAD *et al.* (158)

E.3 Metodologia de análise

A abordagem proposta está seguindo as etapas do processo CRISP-DM, conforme apresentado na Capítulo 2. A etapa 1 (Entendimento do negócio/pesquisa) foi realizada com a revisão da literatura sobre métodos computacionais para auxílio ao diagnóstico de OA em joelhos através de imagens, especificamente de radiografias. Com o auxílio da revisão bibliográfica inicial, foram definidos os objetivos e a metodologia deste trabalho. Os resultados dessa etapa foram apresentados na seção E.1.

A etapa 2 (Compreensão dos dados) foi realizada com a coleta e análise estatística descritiva do conjunto de dados (imagens). As etapas 3 (Preparação dos dados), 4 (Modelagem dos dados) e 5 (Validação do modelo) subsequentes, e suas respectivas tarefas, tiveram como ponto de partida a reprodução e avaliação da abordagem proposta por BACKES, CASANOVA e BRUNO.(85).

A preparação dos dados (etapa 3) envolveu a tarefa de criação de novos conjuntos de dados através do mapeamento das imagens em redes e da extração de novos atributos

* https://nda.nih.gov/oai/

(Capítulo 2). No mapeamento imagem-rede realizado (85), os nós e ligações da rede são definidos da seguinte forma: os pixels da imagem são os nós e as ligações (e) entre nós são efetuadas conforme a regra estabelecida na Equação E.1. Por essa regra, dois nós são ligados se a distância entre eles for menor ou igual a uma distância (parâmetro r) definida pelo analista. As ligações estabelecidas são não-direcionadas e ponderadas, isto é, possuem um peso ($w(e)$) dado pela Equação E.2, cujos valores estão na faixa $0.0 \leq w(e) \leq 1.0$. Possuem maior peso as ligações entre nós mais distantes e com maior diferença entre as intensidades dos pixels correspondentes da imagem.

$$E = \{ e = (v_{xy}, v_{x'y'}) \in I \times I \mid \sqrt{(x - x')^2 + (y - y')^2} \leq r \} \tag{E.1}$$

$$w(e) = \frac{[(x - x')^2 + (y - y')^2] + r^2 * \frac{|I(x,y) - I(x',y')|}{L}}{r^2 + r^2}, \forall e = (v_{xy}, v_{x'y'}) \in E \tag{E.2}$$

Os novos conjuntos de dados obtidos após a aplicação da modelagem imagem-rede dependem de quatro parâmetros: distância entre nós (r), threshold (ou limiar) mínimo (*tmin*), threshold incremental (*tinc*) e threshold máximo (*tmax*). O parâmetro r determina a estrutura (ou topologia) da rede, e os parâmetros *tmin, tinc e tmax* são utilizados para criar uma dinâmica evolutiva específica dessa estrutura, a qual possibilita a construção do conjunto de dados. A tarefa de extração de atributos de cada rede (85) para formação de um conjunto de dados foi realizada através das seguintes ações:

1. $threshold = tmax$
2. corte (ou exclusão) das ligações da rede com $w(e) > threshold$
3. medição da densidade de probabilidade(47) ($p(i)$) (Equação E.3) do histograma de graus da nova rede
4. cálculo dos atributos energia (E) (Equação E.4) e/ou constraste (C) (Equação E.5) e/ou entropia (H) (Equação E.6)
5. $threshold = threshold - tinc$
6. repita 2 a 5 enquanto $threshold \geq tmin$

$$p(i) = \frac{h(i)}{\sum_{i=0}^{k} h(i)} \tag{E.3}$$

Onde i representa os graus presentes na rede, variando de 0 ao grau máximo k.

$$E = \sum_{i=0}^{k} p(i)^2 \tag{E.4}$$

$$C = \sum_{i=0}^{L-1} [p(i)]\, i^2 \tag{E.5}$$

$$H = -\sum_{i=0}^{L-1} [p(i)]\, [\log_2 p(i)] \tag{E.6}$$

Dessa forma, um conjunto de dados terá o seguinte tamanho: objetos (em linhas) = número de redes, atributos (em colunas) = número de atributos escolhidos x número de evoluções de threshold + 1 (etiqueta da classe a que pertence a rede).

Para determinação dos melhores parâmetros *tmin, tinc e tmax*, fixa-se o parâmetro *r*, e emprega-se a análise de classificação (etapa 4) utilizando o algoritmo *Linear Discriminant Analysis (LDA)* (41) com esquema de validação *Leave-one-out Cross-Validation (LOOCV)* (30) sobre somente os atributos energia de todos os objetos do conjunto de dados gerado, avaliando-se a acurácia do classificador. Após terem sido determinados, fixam-se esses parâmetros e faz-se o mesma análise de classificação variando-se o parâmetro *r*.

Uma vez fixados os quatro parâmetros de ajuste da modelagem imagem-rede, passou-se à tarefa de seleção de atributos. Essa foi realizada através da abordagem *Wrapper* (Capítulo 2) empregando-se a análise de classificação utilizando o algoritmo LDA com esquema de validação LOOCV sobre todas as possíveis combinações dos atributos (energia, contraste e entropia) de todos os objetos do conjunto de dados gerados, e avaliando-se a acurácia do classificador.

Em todos os experimentos realizados, foi utilizado um PC com processador Intel Core i5 (2,2 GHz), 6,00 GB de memória RAM e Windows 10. Os códigos para execução desses experimentos foram desenvolvidos através das seguintes ferramentas de programação (Capítulo 2): linguagem Python e bibliotecas Numpy, Imageio, Scikit-Learn, entre outras.

E.4 Resultados

Compreensão dos dados

A análise estatística descritiva do conjunto de dados (imagens) foi realizada através do cálculos de atributos (grandezas estatísticas) sobre as intensidades dos pixels de cada imagem. Foram calculadas a Média, Valor Minimo, Mediana, Valor Máximo. Para efeito de comparação entre as classes de imagens, foi realizada uma segunda análise descritiva estatística sobre esses atributos, agrupando-os nas suas respectivas classes, como pode ser visto nas Tabela E.1 e Tabela E.2. Como exemplo de comparação, obteve-se 110 para a média do grandeza "Média"de cada uma das 344 imagens da classe KL=0. Esse valor foi de 112 para a classe KL=2.

Preparação dos dados

A determinação dos parâmetros *tmin, tinc e tmax* da modelagem imagem-rede foi realizada fixando-se $tinc = 0.05$ e $r = 2$. A estrutura inicial da rede possui: 16384 nós,

Tabela E.1 – Análise estatística descritiva dos atributos das 344 imagens da classe KL=0.

	Atributos das imagens (KL=0)			
Estatística	Média	Minimo	Mediana	Máximo
Média	110	0	108	255
Variância	187	0	214	0
Desvio padrão	14	0	15	0
Mínimo	58	0	57	255
25%	102	0	99	255
Mediana	111	0	109	255
75%	119	0	119	255
Máximo	151	0	152	255
Intervalo	93	0	95	0

Fonte – Elaborada pelo autor.

Tabela E.2 – Análise estatística descritiva dos atributos das 344 imagens da classe KL=2.

	Atributos das imagens (KL=2)			
Estatística	Média	Minimo	Mediana	Máximo
Média	112	0	110	255
Variância	185	0	237	0
Desvio padrão	14	0	15	0
Mínimo	75	0	67	255
25%	103	0	100	255
Mediana	112	0	110	255
75%	122	0	121	255
Máximo	151	0	153	255
Intervalo	77	0	86	0

Fonte – Elaborada pelo autor.

97026 conexões, $k_{min} = 5$, $< k >= 11.8$, $k_{max} = 12$. Como pode ser visto na Tabela E.3, os resultados dos experimentos 1 a 7 mostram que as faixas 0.2 a 0.31 (exp. 2) e 0.5 a 0.61 (exp. 5) forneceram as maiores acurácias. Executou-se então o experimento 8, que amplia a faixa de variação, e a acurácia aumentou. Com isso, definiram-se *tmin* e *tmax*.

A determinação do parâmetro *tinc* foi realizada fixando-se os parâmetros $r = 2$ e *tmin* e *tmax* conforme definidos anteriormente. Como pode ser visto na Tabela E.4, os resultados

Tabela E.3 – Experimentos para ajuste dos parâmetros *tmin e tmax*.

Experimento	Raio	tmin	tmax	tinc	Acurácia (%)
1	2	0.1	0.21	0.05	67.587
2	2	0.2	0.31	0.05	70.203
3	2	0.3	0.41	0.05	67.296
4	2	0.4	0.51	0.05	74.564
5	2	0.5	0.61	0.05	75.145
6	2	0.6	0.71	0.05	52.761
7	2	0.7	0.81	0.05	47.674
8	2	0.2	0.61	0.05	77.907

Fonte – Elaborada pelo autor.

dos experimentos 8 a 13 mostram que a maior acurácia foi obtida para $tinc = 0.01$ (exp. 9). Com isso, definiram-se *tmin, tmax* e *tinc*.

Tabela E.4 – Experimentos para ajuste do parâmetro *tinc*.

Experimento	Raio	tmin	tmax	tinc	Acurácia (%)
8	2	0.2	0.61	0.05	77.907
9	2	0.2	0.61	0.01	78.343
10	2	0.2	0.61	0.02	76.163
11	2	0.2	0.61	0.03	77.325
12	2	0.2	0.61	0.04	75.436
13	2	0.2	0.61	0.005	76.017

Fonte – Elaborada pelo autor.

Uma vez determinados os parâmetros *tmin, tmax* e *tinc*, realizou-se um experimento para verificar se o aumento do parâmetro *r* para 3 melhoraria o resultado. A estrutura inicial da rede possui: 16384 nós, 224786 conexões, $k_{min} = 10$, $< k >= 27.4$, $k_{max} = 28$. A acurácia obtida foi de 75.581%, inferior à obtida para $r = 2$ (78.343%). Logo, definiu-se o parâmetro *r*.

Seleção dos atributos

Como pode ser visto na Tabela E.5, as maiores acurácias foram obtidas para os experimentos 17, 19 e 20. Entretanto, pelo número menor de atributos necessários, os experimentos 17 e 19 apresentaram os melhores resultados.

Tabela E.5 – Seleção dos Atributos.

Experimentos	Atributos	Acurácia (%)	Qtde. Atributos
14	Energia	78.343	40
15	Contraste	76.163	40
16	Entropia	77.616	40
17	Energia, Contraste	79.651	80
18	Energia, Entropia	77.616	80
19	Contraste, Entropia	79.651	80
20	Energia, Contraste, Entropia	79.651	120

Fonte – Elaborada pelo autor.

Os resultados preliminares obtidos até o presente momento foram apresentados São Paulo School of Advanced Science on Learning from Data (São Paulo/SP, 2019). Eles evidenciam o grande potencial da abordagem empregada, pois a melhor acurácia obtida (79.65%) está bem próxima da obtida por RIAD (158) (80.38%) sobre o mesmo conjunto de dados. As perspectivas de superar esse valor são grandes, o que poderá resultar em publicações. As atividades previstas para a sequência do trabalho são:

- Extrair novas métricas (atributos) da rede e melhorar os resultados da classificação sobre os modelos imagem-rede já executados.
- Empregar outros algoritmos de AM encontrados na literatura.
- Empregar outras métricas de avaliação de desempenho dos classificadores.
- Propor novas formas de modelagem imagem-rede.
- Comparar os resultados dos diversos modelos e métricas utilizadas
- Comparar os resultados obtidos com os de outras abordagens da literatura.

APÊNDICE F – GUIA PARA TREINAMENTO DE ALUNOS DE IC SOBRE AM APLICADO A BIOSSENSORES

Esse guia foi proposto visando elencar os tópicos básicos de estudo e fontes bibliográficas, e propor exercícios para treinamento de alunos de Iniciação Científica (IC) sobre AM aplicado a dados de biossensores. Os alunos de graduação em Ciência da Computação da Unesp (Presidente Prudente/SP) Lorrany Vitoria Gomes de Barros, Lucas Vinícius Voltera e Lucas Bernardo de Souza fizeram uso e contribuíram para o aperfeiçoamento desse guia.

F.1 Aprendizado de máquina

Tópicos para estudo

Programação em Python. Estatística e Probabilidade. Exploração de dados. Visualização de dados. Pré-processamento de dados. Engenharia de atributos. Teoria de Aprendizagem de Máquina. Algoritmos básicos de AM. Construção de modelos de AM. Técnicas de validação de modelos de AM. Análise e avaliação de modelos de AM. Ambientes, ferramentas e técnicas de programação para AM.

Exercício 1: análise de dados em saúde

Objetivo: Compreender e fixar os conceitos, métodos e ferramentas de AM.

Conjunto de dados: Escolher um conjunto de dados de acesso livre sobre algum tema em saúde, que seja supervisionado (para classificação) e que os dados estejam organizados em pelos menos 2 classes.

Atividades (Notebook Python):

- Empregar métodos de pré-processamento para tratar os dados (se necessário).
- Explorar os dados com métodos de estatística descritiva e de visualização de dados.
- Realizar seleção de atributos (pelos menos 1 método).
- Aplicar AM para agrupamento dos dados (pelos menos 1 algoritmo).
- Aplicar AM para classificação dos dados (pelos menos 1 algoritmo).

Referências

Lista de referências básicas e de abordagem prática em Python:

- IGUAL, L. Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications, 2017.

- GRUS, J. Data Science from Scratch: First Principles with Python, 2015.
- HAN, J. Data Mining: Concepts and Techniques, 2011.
- MYATT, G. J. Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining, 2014.
- BROWNLEE, J. Machine Learning Mastery with Python, 2016.
- MULLER, A. C. Introduction to Machine Learning with Python A Guide for Data Scientists, 2016.
- DUDA, R. O. Pattern classification, 2001.
- BISHOP, C. M. Pattern Recognition and Machine Learning, 2006.
- BRUCE, P. Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python, 2020.
- LI, R. Essential Statistics for Non-STEM Data Analysts, 2020.
- MORETTIN, P. A. Estatística Básica, 2017.
- PERKOVIC, L. Introduction to Computing Using Python: An Application Development Focus, 2015.
- DOWNEY, A. B. Think Python: How to Think Like a Computer Scientist, 2016.
- https://drivendata.github.io/cookiecutter-data-science/
- https://www.kdnuggets.com/tutorials/index.html
- https://www.analyticsvidhya.com/blog/

F.2 Biossensores

Tópicos para estudo

Definição. Princípios básicos. Componentes. Classificação. Biossensores e Técnicas Analíticas.

Exercício 2: análise de dados de biossensor

Objetivo: Utilizar algoritmos de AM para verificar qual modelo tem melhor desempenho nas tarefas de aglomeração e classificação de espectros de impedância de sensor para deteccao dos fármacos Triclosan e Ibuprofeno considerando-se todas as configurações experimentais utilizadas, isto é, para cada sensor, solução, concentração.

Configurações experimentais:

- Sensores (6): Cu, MWNCTS, Ni, PEDOT, PPy, PSS
- Soluções (4): Tric, Ibu+Tric (Ibu constante), Tric+Ibu (Tric constante), Ibu.
- Concentrações (6): Em cada solução, um dos solutos teve sua concentração variada 0 a $10^{-12} umol/L$. Para cada concentração, foram adquiridos 3 espectros.
- Solução TricWpss (Tric, sensor sem filme PSS) não precisa considerar.
- Medidas: espectros de impedância vs. sensor/solução/concentração

Conjunto de Dados: TRIC.xls

Procedimentos:

- Compreender o problema e os dados fornecidos.
- Organizar os dados em formato adequado para as análises.
- Empregar métodos de pré-processamento para tratar os dados (se necessário).
- Empregar métodos de estatística descritiva e de visualização de dados para apresentar os dados.
- Empregar métodos para seleção de atributos para melhoria de desempenho.
- Empregar métodos para extração de atributos para melhoria de desempenho.
- Empregar métodos não-supervisionados para aglomeração. Apresentar resultados com visualizações e medidas de desempenho.
- Empregar métodos supervisionados para classificação. Apresentar resultados com visualizações e medidas de desempenho.

Referências

Biossensores

KARUNAKARAN, Biosensors and Bioelectronics, 2015.

PANDEY, Biosensors - fundamentals and applications, 2019.

ALTINTAS, Biosensors and nanotechnology -applications in health care diagnostics, 2018.

ZHANG, Electrochemical Sensors, Biosensors and their Biomedical Applications, 2007.

YOON, Introduction to Biosensors -From Electric Circuits to Immunosensors, 2016.

SHIMIZU, Electronic Tongues - Fundamentals and recent advances, 2021.

TURNER, Biosensors - Sense and Sensilility. Chem. Soc. Rev. 2013, 42 (8), 3184.

MOHANTY, Biosensors - A Tutorial Review. IEEE Potentals 2006, 25 (2), 35–40.

Quimiometria e Quimioinformática

GRESSLING, Data Science in Chemistry - Artificial Intelligence, Big Data, Chemometrics and Quantum Computing with Jupyter, 2020.

BEEBE, Chemometrics - A Practical Guide, 1998.

BRERETON, Chemometrics - Data driven extraction for Science, 2018.

BRERETON, Chemometrics for Pattern Recognition, 2009.

GASTEIGER, Chemoinformatics, 2003.

F.3 Metodologia e escrita científica

Tópicos para estudo

Tipos de pesquisa. Planejamento da pesquisa. Revisão bibliográfica (revisão sistemática). Leitura, compreensão e extração de informações de trabalhos científicos. Organização e desenvolvimento do trabalho.

Exercício 3: Escrita de projeto de IC para pedido de bolsa (c.f. FAPESP)

Objetivo: Aplicar os conhecimentos e habilidades na preparação de um projeto de IC para pedido de bolsa à FAPESP.

Atividades:

- Cadastro Curriculum Lattes (se não possuir cadastro)
- Preparação dos documentos pessoais exigidos
- Preparação da Súmula Curricular (c.f. FAPESP)
- Cadastro Orcid, etc
- Estudo e planejamento do projeto junto ao orientador
- Escrita do projeto junto ao orientador (c.f. FAPESP)

Referências

WAZLAWICK, Metodologia de Pesquisa Para Ciência da Computação, 2014.

BOTELHO, Metodologia científica, 2013.

MARCONI, Fundamentos de Metodologia Científica, 2017.

MEDEIROS, Redação Científica, 2006.

BLOOMBERG, Completing Your Qualitative Dissertation - A Road Map from Beginning to End, 2018.

https://lattes.cnpq.br/

https://fapesp.br/bolsas/ic