

UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

Dissertação de Mestrado

Moluscos bivalves como biossensores de poluição aquática

Bruna de Vargas Guterres

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande - FURG, como requisito parcial para a obtenção do grau de Mestra em Engenharia de Computação

Orientador: Prof^ª. Dr^ª. Silvia Silva da Costa Botelho
Co-orientador: Prof^ª. Dr^ª. Juliana Zomer Sandrini

Rio Grande, 2021

Ficha Catalográfica

G983m Guterres, Bruna de Vargas.
Moluscos bivalves como biossensores de poluição aquática /
Bruna de Vargas Guterres. – 2021.
129 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande –
FURG, Programa de Pós-Graduação em Computação, Rio
Grande/RS, 2021.

Orientadora: Dra. Silvia Silva da Costa Botelho.

Coorientadora: Dra. Juliana Zomer Sandrini.

1. Biossensores 2. Análise Comportamental 3. Valvometria
4. Poluição Aquática I. Botelho, Silvia Silva da Costa II. Sandrini,
Juliana Zomer III. Título.

CDU 351.777.6

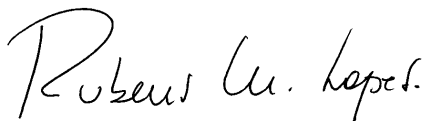
Catálogo na Fonte: Bibliotecário José Paulo dos Santos CRB 10/2344

DISSERTAÇÃO DE MESTRADO

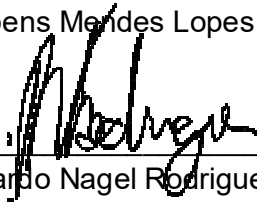
**MOLUSCOS BIVALVES COMO BIOSSENSORES DE
POLUIÇÃO AQUÁTICA**

Bruna de Vargas Guterres

Banca examinadora:



Prof. Dr. Rubens Mendes Lopes



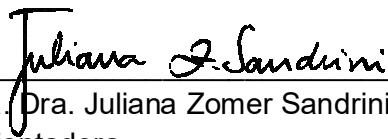
Prof. Dr. Ricardo Nagel Rodrigues



Prof. Dr. Paulo Lilles Jorge Drews Junior



Profa. Dra. Silvia Silva da Costa Botelho
Orientadora



Profa. Dra. Juliana Zomer Sandrini
Coorientadora

Dedico a presente dissertação de mestrado aos meus pais, Maria e Soni, que são os pilares de tudo me tornei e conquistei. Em especial, dedico este trabalho ao meu irmão, Denilson Guterres, que sempre foi meu exemplo e motivo de orgulho.

AGRADECIMENTOS

Gostaria de agradecer principalmente a minha família por todo apoio ao longo do mestrado e da vida, por me ensinarem os valores que carrego e a importância da educação como ferramenta de transformação. Agradeço aos meus amigos e colegas que se fizeram presentes ao longo desta caminhada e aos diversos professores que auxiliaram na construção dos conhecimentos adquiridos.

Agradeço ainda à minha orientadora Prof^a. Dr^a. Silvia Botelho pela confiança em mim depositada, pelo apoio e por todo auxílio ao longo da execução da presente dissertação de mestrado. Gostaria de agradecer a minha co-orientadora Prof^a. Dr^a Juliana Sandrini por aceitar auxiliar na construção do biossensor proposto e por toda ajuda fornecida.

Da mesma forma, agradeço a Amanda Guerreiro pelo auxílio nos testes laboratoriais executados e na escrita de artigos e a Jhordan Silveira, Je Junior e Eduardo Leivas pela parceria e disponibilidade a qualquer tempo. Por fim, agradeço a CAPES pelo fomento financeiro.

*Tenho a impressão de ter sido uma criança brincando à beira-mar,
divertindo-me em descobrir uma pedrinha mais lisa
ou uma concha mais bonita que as outras, enquanto o imenso oceano
da verdade continua misterioso diante de meus olhos.*

— ISAAC NEWTON

RESUMO

GUTERRES, Bruna de Vargas. **Moluscos bivalves como biossensores de poluição aquática**. 2021. 128 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande - FURG, Rio Grande.

Além dos poluentes aquáticos conhecidos, poluentes emergentes são gerados através de fontes antropogênicas diversas. Como se proteção de um ecossistema complexo, pode-se fazer uso de organismos presentes nesse meio na biomonitoração de ambientes aquáticos. Trata-se da associação de componentes biológicos e detectores físico-químicos como dispositivos analíticos conhecidos como biossensores que possibilitam o monitoramento em tempo real de todo o espectro de substâncias tóxicas e efeitos sinérgicos. Respostas comportamentais de moluscos bivalves podem ser utilizadas na biomonitoração automática provendo uma maior sensibilidade e relevância ecológica que testes de toxicidade padrões. O presente trabalho propõe e valida em laboratório modelos computacionais de visão e inteligência artificial na construção de biossensores de poluição aquática baseados na análise comportamental de moluscos bivalves tendo em vista uma arquitetura modular. Um modelo baseado em visão computacional é apresentado como técnica de valvometria não invasiva capaz de eliminar a necessidade de fixação de componentes às conchas de bivalves. Modelos de classificação baseados em grupo de bivalves e na combinação de classificadores são apresentados como ferramentas capazes de inferir a presença de contaminantes a partir de parâmetros comportamentais atingindo acurácias superiores à 90%. O emprego de redes neurais artificiais auto-regressivas não lineares na modelagem comportamental de moluscos individuais é apresentado com o intuito de considerar sua natureza. Os resíduos entre os comportamentos real e modelado são empregados em um gráfico de controle capaz de inferir a presença de contaminantes. Os modelos propostos demonstraram vantagens em relação ao emprego de métodos estatísticos tradicionais na detecção de poluentes.

Palavras-chave: Biossensores, análise comportamental, valvometria, poluição aquática.

LISTA DE FIGURAS

Figura 1	Diagrama em blocos acerca da estrutura de um biossensor de monitoramento aquático baseado na análise comportamental de moluscos bivalves.	24
Figura 2	Modelo de um modelo artificial (HAYKIN, 2007).	25
Figura 3	Estrutura de uma NARNET (BENMOUIZA; CHEKNANE, 2013, Adaptado) para estimar o valor atual ($\hat{y}_t$) de uma série temporal com base em seus últimos p valores. Cada entrada da rede encontra-se atrasada em relação à outra em uma unidade de tempo (z^{-1})	26
Figura 4	Espaços de características linearmente separáveis (A) e linearmente inseparáveis (B) tendo em vista as classes vermelha e azul. A linha escura esboça a fronteira ideal de decisão no espaço original (SANTOS, 2002, Adaptado).	27
Figura 5	Mapeamento de um espaço de entrada X não linearmente separável e um espaço de características linearmente separável Z através de uma função <i>kernel</i> ϵ . (SANTOS, 2002).	28
Figura 6	Estrutura exemplo de uma árvore de decisão binária em que nós terminais são representados através de quadrados e nós não terminais são representados por meio de círculos (OSHIRO, 2013).	29
Figura 7	Diagrama em blocos sobre o funcionamento dos métodos <i>Bagging</i> (OSHIRO, 2013).	30
Figura 8	Matriz de Confusão gerada a partir de um conjunto de instâncias (SILVA, 2006, Adaptado).	31
Figura 9	Matriz de confusão expandida na qual métricas de avaliação como acurácia e precisão são facilmente visualizadas.	32
Figura 10	Exemplo de curva com ROC com área sob a curva (AUC) igual à 0,78. Linha em vermelho ilustra a curva de um classificador com saídas totalmente aleatórias (CORTES; MOHRI, 2004, Adaptado). . .	32
Figura 11	Exemplo de curva de Precisão-Revocação através de uma linha azul. A região cinza destaca a área sob a curva PR (AUC-PR) (CHICCO, 2017, Adaptado).	33
Figura 12	Ponto $Q=(X,Y,Z)$ projetado na imagem através de um raio que passa pelo centro de projeção resultando em um ponto $q = (z, y, f)$ projetado na imagem através de uma câmera com distância focal f . (BRADSKI; KAEHLER, 2008)	34

Figura 13	Exemplo de operação morfológica de erosão na qual a imagem original (A) passa por um processo de convolução com um elemento estruturante retangular (B) resultando na imagem a direita (C) (BARELLI, 2018).	37
Figura 14	Exemplo de operação morfológica de dilatação na qual a imagem original (A) passa por um processo de convolução com um elemento estruturante em forma de cruz (B) resultando na imagem a direita (C) (BARELLI, 2018).	38
Figura 15	Exemplo da aplicação do operador morfológico de abertura à uma imagem em escala de cinza (A). O resultado (B) quando subtraído da imagem original fornece uma imagem com iluminação mais uniforme (C) que pode ser melhorada através do ajuste de seu contraste (D) (BARELLI, 2018).	38
Figura 16	O espaço RGB é representado por um cubo (A) através do qual é possível visualizar todas as cores passíveis de obtenção a partir da combinação das componentes vermelha, verde e azul variando de 0 à 1 deste espectro. É comum visualizar este cubo ao longo de seu eixo neutro (eixo da origem ao ponto 1,1,1). A seção transversal do formato resultante corresponde ao cone que representa o espaço HSV (B). Fonte: Corrêa ¹ et al. (2014) (adaptado).	39
Figura 17	Modelo de regressão linear baseado em pontos amostrais de duas variáveis dependente (y) e independente (x). A reta modelada não estima corretamente todos valores de y tendo como base os valores de x havendo, dessa forma, erros definidos como a diferença entre os valores real e estimado da variável dependente (BLACK, 2009)(adaptado)	41
Figura 18	Estimativa normal de densidade por <i>kernel</i> (linha sólida) e contribuições do <i>kernel</i> (linhas pontilhadas) centrado em cada X_i da amostra $x_1, x_2, \dots, x_4$. A EDK em qualquer x é a soma das contribuições do <i>kernel</i> centrado em cada X_i (GIVENS; HOETING, 2012)(adaptado).	43
Figura 19	Exemplo de gráfico de controle EWMA para uma série temporal. As linhas de limite central (CL), superior (UCL) e inferior (LCL) encontram-se ilustradas. Os círculos indicam os valores resultantes da média ponderada e as cruzes indicam os valores da série analisada a cada intervalo de tempo (SIREGAR et al., 2013, Adaptado).	44
Figura 20	Contribuições propostas	72
Figura 21	Modelo de uma técnica de valvometria baseada em segmentação de cores. Moluscos bivalves com suas conchas superior e inferior pintadas em cores distintas, como por exemplo laranja e verde, são fixados de forma a fornecer vista frontal à uma câmera. Um sistema embarcado é responsável por capturar as imagens, processá-las e inferir a amplitude de abertura de cada concha.	73
Figura 22	Diagrama em blocos ilustrando o pré-processamento e extração de ROIs. As conchas dos moluscos bivalves são pintados em cores distintas como, por exemplo, verde e laranja para posterior análise.	74

Figura 23	Diagrama de processamento de cada ROI. Ele pode ser dividido em segmentação de cores, redução de ruído através de operadores morfológicos de abertura e determinação do pixel mais próximo da extremidade da concha para cada cor de interesse.	74
Figura 24	Modelo computacional proposto para a construção de um módulo ferramenta de análise comportamental baseado na utilização de um classificador para, a partir dos parâmetros comportamentais horários de um grupo de moluscos bivalves, inferir o <i>status</i> de contaminação ambiente (contaminado ou não).	76
Figura 25	Modelo computacional proposto para a construção de um módulo ferramenta de análise comportamental baseado na combinação de classificadores. Quatro classificadores (<i>C1</i> à <i>C4</i>) são empregados para inferir a presença de contaminantes (classificação binária) a partir de parâmetros comportamentais horários. A saída destes classificadores (votos intermediários) são processadas através de um classificador de votação responsável por determinar o <i>status</i> final de contaminação, isto é, se há ou não contaminantes no meio aquático.	77
Figura 26	Modelo para a construção de um módulo ferramenta de análise comportamental baseado no treinamento de NARNETs para prever o comportamento individual mexilhões não expostos a contaminantes. Os erros entre os valores previstos e reais de diversos indivíduos são empregados em um gráfico de controle EWMA a fim de identificar o <i>status</i> de contaminação ambiente.	78
Figura 27	<i>Layout</i> do experimento com mexilhões <i>P. perna</i> . Uma câmera é posicionada de forma externa ao aquário de forma a obter uma vista frontal dos mexilhões instrumentados com ímãs e sensores pintados em verde e laranja. A câmera é conectada a um Odroid XU4 com o intuito de capturar e armazenar imagens de acordo com um <i>trigger</i> externo. Um Arduino Mega em conjunto com uma placa de aquisição de dados é responsável por armazenar as saídas dos sensores e enviar um <i>trigger</i> à câmera.	80
Figura 28	<i>Layout</i> do experimento realizado em triplicatas no qual mexilhões <i>P. perna</i> foram instrumentados com sensores de efeito Hall e ímãs. As saídas dos sensores foram conectadas a uma placa de aquisição de dados que, em conjunto com um Arduino Mega, salvou as respostas comportamentais em um arquivo CSV. Após um período não toxicológico de três dias, os mexilhões foram expostos a 0%, 5% e 20% de óleo diesel WAF	82
Figura 29	Divisão do experimento em período de exposição não tóxica e período de exposição toxicológica. Ambos períodos tem duração de 24 horas A partir das primeira 24 horas, os animais foram expostos à 5% e 20% de WAF de diesel S-500	84

Figura 30	Esquema do classificador de votação no qual quatro classificadores foram treinados com base em um conjunto de dados de treinamento contendo a amplitude média de abertura, percentual de tempo aberto, frequência de transição e quantidade de micromovimentos. As classificações (P_1, P_2, P_3 e P_4) adquiridas por meio de um novo conjunto de dados foram usadas para obter uma previsão final (P_f) através de um classificador de votação.	87
Figura 31	Arquitetura da NARNET para prever $Y(t + 1)$ com base nos dados comportamentais das últimas seis horas (últimas sessenta AOA) de $Y(t)$. A rede possui sessenta atrasos de entrada, uma camada oculta com quinze neurônios cuja função de ativação é uma tangente hiperbólica. A camada de saída abrange um neurônio com uma função de ativação linear.	87
Figura 32	Imagem antes (a) e após (b) melhoria de contraste e a respectiva enumeração dos animais.	90
Figura 33	Região de interesse (a), regiões de coloração laranja (b) e verde (c) detectadas a partir dos valores HSV médios. Os pontos verde e vermelhos ilustrados em (c) correspondem aos pontos máximo e mínimo das figuras (b) e (c), respectivamente. A distância d destacada em amarelo corresponde ao valor de amplitude detectado para o caso apresentado.	90
Figura 34	Comportamento de abertura e fechamento de mexilhões <i>Perna perna</i> tendo em vista o emprego dos sensores de efeito Hall e algoritmo de visão computacional	91
Figura 35	Erro entre sensores de efeito Hall e algoritmo de visão computacional para cada animal ao longo de duas horas de experimento.	92
Figura 36	Gráficos de dispersão do comportamento de mexilhões <i>Perna perna</i> de acordo com sensores de efeito Hall e algoritmo de visão computacional. A linha de regressão linear ($y = ax + b$) minimiza a soma residual dos quadrados entre os dados obtidos pelos sensores de efeito hall (y) e o algoritmo de visão computacional (x) de acordo com os coeficientes a e b . Os coeficientes de correlação de Pearson encontram-se ilustrados para cada bivalve.	93
Figura 38	Estimativa de densidade por kernel (KDE) de histogramas de erro entre sensores de efeito hall e algoritmo de visão computacional. As funções densidade de probabilidade e os histogramas de erro são demonstradas como linhas e barras azuis, respectivamente.	94
Figura 37	Estimativa de densidade por kernel (KDE) de histogramas de erro entre sensores de efeito hall e algoritmo de visão computacional para cada mexilhão <i>Perna perna</i> . As funções densidade de probabilidade e os histogramas de erro são demonstradas como linhas e barras azuis, respectivamente.	96
Figura 39	Exemplos de respostas comportamentais de mexilhões <i>Perna perna</i> expostos a 5% e 2 % de WAF de diesel. Sensores de efeito Hall foram empregados para adquirir dados comportamentais em um experimento que compreendeu 24 horas de exposição não toxicológica e toxicológica, como demonstrado pela linha tracejada preto.	97

Figura 40	Matriz de confusão, curva ROC e AUC referentes ao processo de treinamento de um classificador <i>ensemble bagged</i> para inferir o <i>status</i> de contaminação ambiente a partir de parâmetros comportamentais horários de um grupo de cinco moluscos bivalves.	98
Figura 41	Matrizes de confusão referentes ao processo de teste do classificador <i>bagged trees</i> (A) e ao emprego de métodos estatísticos na identificação de diferenças significativas entre os grupos de controle e de teste ($p < 0,05$) (B)	99
Figura 42	Matrizes de confusão, curvas ROC e AUC dos classificadores <i>bagged trees</i> (A), SVM cúbica (B), SVM quadrática (C) e SVM gaussiana (D) obtidas durante a primeira etapa de treinamento do modelo baseado na combinação de classificadores para detecção de compostos petroquímicos.	100
Figura 43	Matriz de Confusão, curva ROC e valor AUC obtidos durante o treinamento do classificador de votação para detecção de compostos petroquímicos.	100
Figura 44	Matrizes de Confusão, curvas ROC e valores AUC obtidos durante o teste do classificador de votação tendo em vista 5% (A) e 20% (B) de WAF de diesel.	101
Figura 45	Curvas Precisão-Revocação (PR) para 5% (A) e 20% (B) de WAF de diesel.	101
Figura 46	Exemplo de resposta obtida através uma NARNET durante o treinamento, validação e teste (A) e o respectivo erro ao longo do tempo (B).	103
Figura 47	Exemplo de desempenho de uma NARNET ao longo das épocas de treinamento (A) e respectivo gráfico de dispersão entre sua saída e a amplitude média esperada (B). Neste exemplo é possível verificar uma correlação muito forte em um intervalo de confiança de 95% de acordo com o coeficiente de correlação de Pearson (r).	103
Figura 48	Exemplo de funções de autocorrelação (ACF) para a amplitude média de abertura (A) e para os resíduos decorrentes do uso da NARNET (B) a um nível de significância de 5%.	104
Figura 49	Gráfico de controle EWMA para mexilhões não expostos à WAF de diesel S-500. Dados de calibração foram empregados para estabelecer os limites do gráfico de controle. As cruze representam as amplitude média de abertura e pontos vermelhos e pretos indicam os valores da média móvel exponencialmente ponderada que ultrapassam ou não os limites de controle superior (UCL) e inferior (LDL) a cada intervalo de 0,1 hora, respectivamente.	106
Figura 51	Gráfico de controle EWMA para mexilhões expostos a 20% de WAF de diesel. Dados de calibração foram empregados para estabelecer seus limites. A linha tracejada azul indica o início da exposição toxicológica e as cruze representam as amplitudes médias de abertura. Pontos vermelhos e pretos indicam os valores EWMA que ultrapassam ou não os limites de controle superior (UCL) e inferior (LDL), respectivamente.	106

Figura 50	Gráfico de controle EWMA para mexilhões expostos a 5% de WAF de diesel. Dados de calibração foram empregados para estabelecer seus limites. A linha tracejada azul indica o início da exposição toxicológica e as cruzes representam as amplitudes médias de abertura. Pontos vermelhos e pretos indicam os valores EWMA que ultrapassaram ou não os limites de controle superior (UCL) e inferior (LDL), respectivamente.	107
Figura 52	Matrizes de confusão referentes às saídas do modelo baseado na associação de NARNETs e gráfico de controle EWMA para os grupos expostos à 5% (A) e 20% de contaminante.	108
Figura 53	Resposta da NARNET, valores almejados de amplitude de abertura e resíduos ao longo do tempo para os grupos de animais expostos a 5% (A) e 20% (B) de WAF de diesel. Linhas pontilhadas em cinza indicam o início do período de contaminação.	110
Figura 54	Matrizes de confusão referentes ao emprego dos testes estatísticos na comparação das amplitudes de abertura de mexilhões <i>Perna perna</i> expostos à 5% (A) e 20% (B) de WAF de diesel S-500 em relação ao grupo de controle a cada 0,1 hora.	110
Figura 55	Comportamento médio de moluscos bivalves expostos à 5% de WAF de diesel S-500 a cada intervalo de 0,1 hora. A linha tracejada em azul indica o início da exposição toxicológica na qual a WAF de diesel foi diluída nos aquários. Pontos em vermelho e preto identificam diferenças significativas erroneamente e corretamente identificadas a um nível de significância de 5%, respectivamente.	111
Figura 56	Comportamento médio de moluscos bivalves expostos à 20% de WAF de diesel S-500 a cada intervalo de 0,1 hora. A linha tracejada em azul indica o início da exposição toxicológica na qual a WAF de diesel foi diluída nos aquários. Pontos em vermelho e preto identificam diferenças significativas erroneamente e corretamente identificadas a um nível de significância de 5%, respectivamente.	111

LISTA DE TABELAS

Tabela 1	Tabela de Análise de Variância	49
Tabela 2	Unidade de análise, período de aclimação, período de monitoramento (PM) e frequência de aquisição empregados em trabalhos relacionados.	65
Tabela 3	Técnicas de valvometria e algoritmos empregados em trabalhos relacionados.	66
Tabela 4	Técnicas de valvometria e algoritmos empregados em trabalhos relacionados.	67
Tabela 5	Parâmetros comportamentais e variáveis de influência em trabalhos relacionados. Dentre os parâmetros comportamentais encontram-se a amplitude média de abertura (AA), período de tempo aberto (PA), período de tempo fechado (PF), frequência de transição (FT) e <i>avoidance behavior</i>	68
Tabela 6	Parâmetros comportamentais e variáveis de influência em trabalhos relacionados. Dentre os parâmetros comportamentais encontram-se a amplitude média de abertura (AA), período de tempo aberto (PA), período de tempo fechado (PF), frequência de transição (FT) e <i>avoidance behavior</i>	69
Tabela 7	Coefficiente (r) e valor p resultante da análise de correlação de Pearson para cada animal inspecionado. A reta de regressão linear ($y = ax + b$) que minimiza a soma residual dos quadrados entre os dados obtidos pelos sensores de efeito hall (y) e algoritmo de visão computacional (x) de acordo com os coeficientes a e b . Também são apresentados o coeficiente de determinação (r^2) e o erro quadrático médio (MSE) entre o sensor de efeito Hall e o algoritmo de visão computacional.	92
Tabela 8	Performance dos classificadores durante o teste tendo em vista a exposição à 5% e 20% de WAF de diesel.	102
Tabela 9	Resumo da performance resultante do emprego de NARNETs na previsão da amplitude média de abertura de mexilhões <i>P. perna</i> a cada intervalo de 0,1 hora. A performance foi avaliada com base nos valores de acurácia, coeficiente de correlação de Pearson (r) e erro quadrático médio (MSE) referentes aos períodos de exposição não toxicológica e toxicológica.	105

LISTA DE ABREVIATURAS E SIGLAS

ACF	Função de Autocorrelação
ANOVA	Análise de Variância
ANP	Agência Nacional de Petróleo, Gás Natural e Biocombustíveis
AOA	Amplitude Media de Abertura
AUC	Área sob a Curva
AUC-PR	Área sob a Curva PR
BEWS	Sistema Biológico de Aviso Prévio
CL	Limite Central
CONECO	Laboratório de Microcontaminantes Orgânicos e Ecotoxicologia Aquática
CROSERF	<i>Chemical Response to Oil Spills Ecological Effects Research Forum</i>
CSV	Valores Separados por Vírgula
EDK	Estimativa de Densidade por Kernel
EWMA	Média Móvel Exponencialmente Ponderada
FAO	<i>Food and Agriculture Organization of the United Nations</i>
FDP	Função Densidade de Probabilidade
FN	Falsos Negativos
FP	Falsos Positivos
FURG	Universidade Federal do Rio Grande
HSV	Matiz (<i>Hue</i>), Saturação (<i>Saturation</i>) e Brilho (<i>Value</i>)
ICB	Instituto de Ciências Biológicas
LCL	Limite de Controle Inferior
LED	Diodo Emissor de Luz
MANOVA	Análise de Variância Multivariada
MSE	Erro Quadrático Médio
NARNET	Rede Neural Artificial Auto-regressiva Não-linear

OpenCV	Biblioteca Aberta de Visão Computacional
PERMANOVA	Análise de Variância Multivariada com Permutações
PR	Precisão-Recuperação
RNA	Rede Neural Artificial
RGB	Vermelho (<i>Red</i>), Verde (<i>Green</i>) e Azul(<i>Blue</i>)
ROC	Características Operacionais do Receptor
ROI	Região de Interesse
SDK	<i>Kit</i> de Desenvolvimento de <i>Software</i>
SVM	Máquinas de Vetores de Suporte
TN	Verdadeiros Negativos
TP	Verdadeiros Positivos
UCL	Limite de Controle Superior
UNEP	Programa de Meio Ambiente das Nações Unidas
VBM	Medição Baseada em Visão
WAF	Fração Acomodada em Água

SUMÁRIO

1	INTRODUÇÃO	18
1.1	Objetivo geral	22
1.2	Objetivos específicos	22
2	REVISÃO BIBLIOGRÁFICA	23
2.1	Redes Neurais Artificiais	24
2.1.1	Rede Neural Artificial Auto-regressiva Não-linear (NARNET)	25
2.2	Modelos de Classificação	26
2.2.1	Máquinas de Vetores de Suporte (SVM)	27
2.2.2	Árvores de Decisão	28
2.2.3	Métricas de Avaliação	30
2.3	Ferramentas de Visão Computacional	34
2.3.1	Calibração de Câmeras	34
2.3.2	Operadores Morfológicos	37
2.3.3	Representação de cores em imagens	38
2.4	Ferramentas de análise estatística	40
2.4.1	Regressão linear	40
2.4.2	Correlação	41
2.4.3	Função densidade de probabilidade	42
2.4.4	Gráficos de Controle de Qualidade	43
2.4.5	Métodos ANOVA e alternativas não-paramétricas	45
2.5	Considerações Finais	51
3	TRABALHOS RELACIONADOS	53
4	MODELOS COMPUTACIONAIS	72
4.1	Módulo Técnica de Valvometria	72
4.1.1	Modelo com Visão Computacional	73
4.2	Módulo Ferramentas de Análise Comportamental	74
4.2.1	Modelo com Classificador Baseado em Grupo	75
4.2.2	Modelo com Combinação de Classificadores	75
4.2.3	Modelo com NARNET e Gráfico de Controle EWMA	77
5	VALIDAÇÃO	79
5.1	Setup dos Experimentos	79
5.1.1	Módulo Técnica de Valvometria	79
5.1.2	Módulo Ferramentas de Análise Comportamental	81
5.2	Ferramentas de Avaliação	83

5.2.1	Módulo Técnica de Valvometria	83
5.2.2	Módulo Ferramentas de Análise Comportamental	84
6	RESULTADOS E DISCUSSÃO	89
6.1	Módulo Técnica de Valvometria	89
6.2	Módulo Ferramentas de Análise Comportamental	97
6.2.1	Modelo com Classificador Baseado em Grupo	98
6.2.2	Modelo com Combinação de Classificadores	99
6.2.3	Modelo com NARNETs e Gráfico de Controle EWMA	102
7	CONCLUSÃO	113
8	PUBLICAÇÕES	115
	REFERÊNCIAS	116
	APÊNDICES	124
A	Moluscos Bivalves	125
A.1	Mexilhões <i>Perna perna</i>	125
B	Sensores de efeito Hall	127
C	Fração Acomodada em Água (WAF)	128

1 INTRODUÇÃO

A indústria petrolífera fornece bens de difícil substituição na matriz produtiva de qualquer país sendo base para o modo de operação, consumo e mesmo cultura da sociedade moderna (CANELAS, 2007). A importância da indústria petrolífera, dentro da cadeia produtiva do petróleo, não se resume apenas à produção final de derivados, ela também se dá na proteção da qualidade do meio ambiente. Isto porque suas atividades de exploração, transporte, distribuição e armazenamento podem gerar danos ambientais. Dentre os ambientes suscetíveis a estes riscos, o meio aquático é um dos mais preocupantes em virtude da possibilidade de ocorrência de vazamentos acidentais de óleo e efluentes de refinarias que podem comprometer recursos hídricos e conferir sérios riscos aos organismos expostos.

Além dos poluentes aquáticos decorrentes da indústria petrolífera, poluentes emergentes são gerados através de fontes antropogênicas diversas e se distribuem através das matrizes ambientais (GAVRILESCU et al., 2015). Embora tenham sido feitos grandes avanços na detecção e análise de traços de poluentes durante as últimas décadas, há grande variedade de contaminantes não detectados de preocupação ambiental emergente que necessitam identificação e quantificação em vários componentes ambientais e tecidos biológicos (GAVRILESCU et al., 2015). Além disso, as atividades humanas resultaram na contaminação dos recursos hídricos com micropoluentes biológicos como vírus e bactérias, que apresentam potencial patogenicidade.

Justifica-se, portanto, a necessidade de pesquisa e inovação em tecnologias para detecção dos mais variados contaminantes ambientais mesmo que emergentes. Nesse sentido torna-se necessário adotar medidas e procedimentos para monitorar e proteger ambientes aquáticos a fim de que atitudes corretivas aplicáveis sejam empregadas o mais rápido possível. Compostos químicos hidrofóbicos tendem a se acumular nos tecidos de organismos aquáticos por diferentes mecanismos como absorção direta da água, absorção de partículas em suspensão e consumo de alimentos contaminados (OOST; BEYER; VERMEULEN, 2003) dando origem à bioacumulação. Dessa forma, métodos de bioacumulação baseiam-se na análise de tecidos animais a fim de inferir a contaminação ambiente já que a concentração de compostos químicos nestes tecidos pode refletir a condição de seu meio (NEFF, 2002).

Nesse sentido, o monitoramento ambiental de poluentes baseado em análises toxicológicas de bioacumulação e mortalidade apresentam como desvantagens o fato do equilíbrio de concentração normalmente ser obtido após semanas de exposição. Estes métodos são, portanto, inapropriados para aplicações que almejam uma rápida detecção da presença de poluentes (KRAMER; JENNER; ZWART, 1989). Embora existam técnicas analíticas tradicionais altamente sensíveis e seletivas como, por exemplo, cromatografia e espectrometria de massa, estas são demoradas, caras, demandam pessoal altamente treinado e não permitem o monitoramento frequente e em tempo real de contaminantes (DORST et al., 2010). Ademais, estes métodos apresentam certas desvantagens e limitações pois baseiam-se apenas em princípios físico-químicos e não fornecem uma avaliação eficiente e econômica acerca da situação ecossistêmica geral (STARODUB et al., 2005).

Os sistemas de controle para ambientes aquáticos baseiam-se principalmente em métodos físico-químicos que são demorados, caros e capazes de detectar uma estreita gama de poluentes sem prover dados acerca de sua influência sobre a biota (GREKOV; KUZMIN; MISHUROV, 2019). Dessa forma, tais métodos não são apropriados para detectar contaminantes de forma imediata nem avaliar seu efeito ecossistêmico rapidamente. A incorporação de componentes biológicos aos sistemas de monitoramento, em contraste com métodos físico-químicos tradicionais, permite avaliar a característica toxicológica integral do meio aquático independente da origem e composição dos poluentes (GREKOV; KUZMIN; MISHUROV, 2019).

Tratam-se da associação de componentes biológicos e detectores físico-químicos como dispositivos analíticos conhecidos como biossensores. A utilização de biossensores permite a análise de alterações na resposta biológica de organismos (resposta molecular, celular, bioquímica, fisiológica ou comportamental) que podem estar relacionadas com a exposição e efeitos de elementos tóxicos (NEWTON; COPE, 2007). Neste contexto, a escolha do componente biológico utilizado como biossensor é importante uma vez que alguns animais apresentam características mais propícias que outros.

Moluscos bivalves são animais invertebrados de corpo não articulado protegido por duas conchas (valvas) unidas por um ligamento e apresentam características convenientes à utilização como biossensores uma vez que alterações comportamentais de abertura e fechamento de suas conchas representam uma resposta biológica à presença de contaminantes sendo uma forma relativamente fácil e barata de monitorar ambientes aquáticos (NEWTON; COPE, 2007). Respostas comportamentais podem ser utilizadas na biomonitoração automática de ambientes aquáticos provendo uma maior sensibilidade e relevância ecológica que testes de toxicidade padrões (LIAO et al., 2009). O emprego destes organismos aquáticos como biossensores fornece monitoramento em tempo real de todo o espectro de substâncias tóxicas e considera seus efeitos sinérgicos (LYUBIMTSEV; KHOLODKEVICH; DRUZHININ, 2019). Além disso, o uso da análise do com-

portamento de bivalves como biomarcadores¹ sensíveis e subletais a poluentes pode não se restringir à detecção de compostos químicos sendo adequado para monitorar os efeitos da poluição difusa, como sedimentos finos e seus efeitos nos serviços ecossistêmicos (LUMMER; AUERSWALD; GEIST, 2016).

A análise comportamental de moluscos bivalves requer o emprego de técnicas de valvometria que permitem adquirir dados referentes à amplitude de abertura de suas conchas. Dentre estas técnicas, sensores de efeito Hall apresentam uma série de vantagens como durabilidade, leveza, facilidade de fixação e oferecem um menor estresse ao animal facilitando a medição de seus movimentos (NAGAI et al., 2006). Entretanto, dados resultantes do monitoramento contínuo de organismos bivalves são bastante complexos devido ao comportamento não linear de seu movimento tornando assim sua análise estatisticamente desafiadora (HARTMANN et al., 2016; BAE; PARK, 2014).

Ademais, a resposta individual dos animais, principalmente sob exposição toxicológica, varia e reflete a natureza adaptativa individual que leva organismos a diferirem em sua capacidade de resposta às variações ambientais (DINGEMANSE et al., 2010). Tem-se, portanto, uma variabilidade inerente dos padrões comportamentais entre indivíduos. Além disso, suas respostas comportamentais encontram-se fortemente relacionadas não só à atividades vitais como respiração, alimentação e excreção, mas também à condições ambientais como a presença de contaminantes e predadores. Biossensores baseados no monitoramento comportamental ou fisiológico de animais sentinela² fornecem informações sobre o bem-estar animal e suas respostas às mudanças ambientais e às ações de manejo (ANDREWARTHA et al., 2015). Logo, o apropriado entendimento do movimento desses animais sésseis pode auxiliar no gerenciamento da qualidade da água em habitats naturais (HARTMANN et al., 2016).

Dentre os moluscos bivalves, o mexilhão *Perna perna* é um molusco da família *Mytilidae* de maior tamanho e mais abundante na costa brasileira apresentando grande importância sócio-econômica em virtude de sua utilização na mitilicultura (JR; BRASIL; SALOMÃO, 2007). A análise comportamental de moluscos bivalves através de sensores de efeito Hall em conjunto com ímãs foi originalmente proposta em Wilson, Reuter e Wahl (2005) com base em sua aplicação na análise comportamental de pinguins. A partir de então, sensores de efeito Hall em conjunto com ímãs consolidaram-se como técnica de valvometria sendo empregados para monitoramento contínuo em tempo real de águas superficiais, água do mar e efluentes através de soluções comerciais³ e na realização de experimentos em ambientes naturais (COMEAU et al., 2018; GREKOV; KUZMIN;

¹Biomarcadores ou marcadores biológicos são entidades que podem ser medidas experimentalmente e indicam a ocorrência de uma determinada função normal ou patológica de um organismo ou uma resposta a um agente farmacológico (LDA, 2020)

²animais localizados em um ambiente comum capaz de fornecer um aviso prévio de condições abaixo do ideal que podem afetar o bem-estar ou sobrevivência individual ou da população

³Aquadect (2020) e MolluSCAN (2020)

MISHUROV, 2019).

Algumas técnicas de valvometria propostas na literatura são categorizadas como não invasivas embora necessitem fixar dispositivos às conchas dos moluscos bivalves (CO-MEAU et al., 2018; BASTI et al., 2009; NAGAI et al., 2006). Isso pode comprometer sua mobilidade e exigir uma adequação dos animais em virtude dessa nova configuração de peso em suas valvas. Além disso, O emprego de técnicas de exploração funcional não invasiva, ao contrário das técnicas letais ou invasivas, permitem que os organismos individuais sejam monitorados pelo tempo que for necessário (MOULLAC et al., 2020). Logo, surge a possibilidade do estudo acerca de outras formas de monitoramento comportamental que não exijam a anexação de quaisquer materiais a esses animais como, por exemplo, através de sinais visuais.

Em virtude dos avanços rápidos e significativos das tecnologias em sistemas de câmeras e computadores, o acesso a câmeras e unidades de computação mais baratas, mais rápidas e de maior qualidade foi facilitado. Como resultado, há uma tendência crescente no uso de medição baseada em visão (*Visual Based Measure* - VBM) em uma ampla variedade de aplicações e cenários automatizados (SHIRMOHAMMADI; FERRERO, 2014).

Os rápidos avanços no campo da inteligência artificial têm profundas implicações para a economia e também para a sociedade em geral através de contribuições importantes em áreas como a agricultura, engenharia e medicina (COCKBURN; HENDERSON; STERN, 2018). A inteligência artificial melhorou muito a coleta, análise e previsão de informações biológicas e médicas. A exemplo disso, algoritmos de inteligência artificial (IA) foram capazes de solucionar um dos maiores desafios biológicos prevendo como proteínas de uma cadeia linear de aminoácidos se enrolam em formas 3D (SERVICE, 2020).

A integração de inteligência artificial (IA) e biossensores pode preencher a lacuna entre a aquisição de dados e análise com melhor precisão diagnóstica e terapêutica (JIN et al., 2020) em aplicações biomédicas. Biossensores vestíveis tem ganhado atenção devido ao seu tamanho mínimo, característica não invasiva e capacidade de monitoramento contínuo. A versatilidade de algoritmos de inteligência artificial tem revolucionado a construção de biossensores vestíveis através de novas possibilidades de desenvolvimento como, por exemplo, biossensores para monitoramento de saliva e suor baseados em biomarcadores como *proxy* para diagnóstico de doenças (JIN et al., 2020).

Além disso, fazendas de aquicultura inteligentes equipadas com biossensores, sensores para monitorar variáveis ambientais e sistemas de gerenciamento tem potencial para revolucionar todos os sensores da indústria (ANDREWARTHA et al., 2015). Dados fisiológicos e comportamentais em tempo real que fornecem *insights* sobre o bem-estar animal têm a capacidade de auxiliar práticas de manejo e otimizar os processos de produção. As informações sobre o ambiente são frequentemente necessárias para interpretar as respostas fisio-comportamentais e o animal sentinela torna-se integrado em uma série de

sensores ambientais (ANDREWARTHA et al., 2015).

Inteligência artificial na construção de biossensores de poluição aquática baseados no comportamento de moluscos bivalves pode oferecer vantagens em relação ao uso de métodos estatísticos tradicionais. Ademais, uma técnica de valvometria baseada em visão computacional pode simplificar o sistema com uma menor quantidade de cabos e ligações, eliminar a necessidade de fixação de componentes às conchas, facilitar o processo de instrumentação e reduzir erros devido à falhas de impermeabilização e fixação.

De forma multidisciplinar, este trabalho busca, portanto, averiguar o uso de técnicas de visão computacional na construção de uma técnica de valvometria não invasiva capaz de eliminar a necessidade de fixação de componentes às conchas de moluscos bivalves. Ademais, almeja-se investigar o emprego de métodos computacionais de inteligência artificial na biomonitoração aquática de poluentes baseada na análise comportamental de mexilhões *Perna perna*. Nesse sentido, as subseções 1.1 e 1.2 apresentam os objetivos geral e específicos do presente trabalho.

1.1 Objetivo geral

A presente dissertação de mestrado tem como objetivo geral propor modelos computacionais para construção de biossensores de poluição aquática baseados na análise comportamental de moluscos bivalves. De forma mais precisa, pretende-se desenvolver uma arquitetura modular com contribuições baseadas em visão computacional e inteligência artificial para a construção destes biossensores.

1.2 Objetivos específicos

Pode-se elencar os seguintes objetivos específicos:

- Realizar o estudo bibliográfico acerca do emprego de moluscos bivalves como biossensores de poluição aquática;
- Identificar os requisitos e módulos principais associados a um biossensor baseado na análise comportamental de mexilhões *Perna perna*;
- Propor métodos computacionais baseados em visão e inteligência artificial para os módulos e seus componentes;
- Validar a metodologia por meio de comparação com técnica de valvometria estado da arte.

2 REVISÃO BIBLIOGRÁFICA

O emprego de moluscos bivalves como biossensores aquáticos baseados em sua análise comportamental requer o emprego de técnicas de valvometria para adquirir dados acerca da amplitude de abertura de suas conchas. Dentre estas técnicas, destacam-se pares de indutores, eletrodos, células de carga e sensores de efeito Hall (HARTMANN et al., 2016; TRAN et al., 2010; LIAO et al., 2009). Além disso, empregam-se parâmetros de análise comportamental capazes de descrever esse comportamento no tempo como, por exemplo, amplitude média de abertura, frequência de transição e percentual de tempo em que as conchas encontram-se abertas (COMEAU et al., 2018; HARTMANN et al., 2016).

Por fim, torna-se necessário o emprego de ferramentas de análise comportamental capazes de inferir alterações nestes parâmetros decorrentes de alterações ambientais de interesse. Dentre estas ferramentas pode-se citar verificação de critérios de comportamento considerados atípicos a fim de emitir sinal de alerta por meio de um BEWS (AQUADECT, 2020) e métodos estatísticos como, por exemplo, ANOVA (*Analysis Of Variance*) (HARTMANN et al., 2016).

Dessa forma, biossensores baseados na análise comportamental de bivalves podem ser entendidos como ferramentas capazes de verificar a condição ambiental e emitir sinais de alerta associados à condições ambientais inóspitas sejam elas decorrentes da presença de contaminantes, algas nocivas ou condições físico químicas inadequadas. Nesse sentido, pode-se definir uma sequência de módulos necessários à construção destes biossensores tendo em vista as funções desempenhadas por cada qual na determinação sinais de alerta ambiental conforme ilustrado pelo diagrama em blocos da Figura 1.

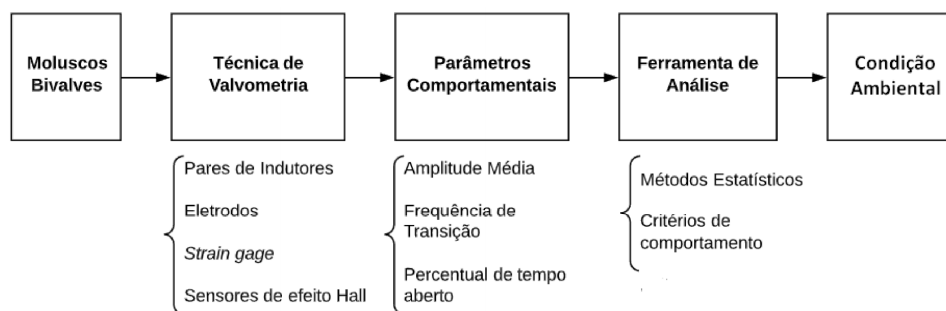


Figura 1: Diagrama em blocos acerca da estrutura de um biossensor de monitoramento aquático baseado na análise comportamental de moluscos bivalves.

A construção de biossensores de poluição aquática baseados na resposta comportamental de moluscos bivalves requer uma fundamentação teórica multidisciplinar abrangente sobre as principais características dos animais empregados, bem como a importância e propriedades dos meios de exposição selecionados para testes laboratoriais. Além disso, torna-se necessário apresentar os principais dispositivos e ferramentas empregados na aquisição de dados comportamentais acerca da amplitude de abertura de suas conchas. Nesse contexto, os apêndices A, B e C versam sobre os conceitos básicos como moluscos bivalves, sensores de efeito Hall e Fração Acomodada em Água (WAF) como meio de exposição padronizado, respectivamente.

Tendo os objetivos da presente dissertação de mestrado, propõe-se uma técnica de valvometria baseada em visão computacional capaz de eliminar a necessidade de fixar componentes às conchas de moluscos bivalves. Ademais, tendo em vista a necessidade de inferir a presença de contaminação ambiente a partir de parâmetros comportamentais, propõe-se o emprego de métodos inteligentes de classificação como ferramentas de análise. Por fim, almejando considerar a natureza adaptativa de bivalves individuais, propõe-se a modelagem de seu comportamento através de redes neurais auto-regressivas não lineares e posterior inferência do *status* de contaminação ambiente a partir do emprego de gráficos de controle de qualidade.

Dessa forma, as subseções 2.1, 2.2 e 2.3 versam sobre Redes Neurais Artificiais (RNAs), modelos de classificação e ferramentas de visão computacional empregados ao longo da presente dissertação. A validação dos modelos propostos é efetuada através de comparações com métodos estatísticos tradicionalmente empregados na construção de biossensores baseados em moluscos bivalves e apresentados através da subseção 2.4.

2.1 Redes Neurais Artificiais

Redes Neurais Artificiais (RNAs) são modelos matemáticos inspirados na estrutura neural de organismos inteligentes que aprendem por meio da experiência (HERCULANO-HOUZEL, 2009). Ela se assemelha ao cérebro humano por ser cons-

tituída por unidades básicas que possibilitam adquirir conhecimento através de um processo de aprendizado e mantê-lo com base em forças de conexão entre neurônios (pesos sinápticos) (HAYKIN, 2007).

Conjuntos de neurônios são organizados de modo a formar uma rede neural artificial. Estes neurônios são matematicamente modelados por meio de pesos sinápticos ($w_{k1}, w_{k2}, \dots, w_{km}$) associados à cada uma das entradas ($X_1, X_2, \dots, X_m$) da rede neural (Figura 2). O somatório dos sinais de entrada ponderados pelos respectivos pesos sinápticos e um *biases* (b_k) são entradas de uma função de ativação ($\varphi(\cdot)$) responsável por modelar influência sobre a saída (y_k) conforme equação 1.

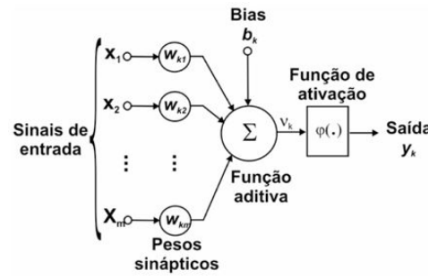


Figura 2: Modelo de um modelo artificial (HAYKIN, 2007).

$$y_k = \varphi\left(\sum_{i=1}^n w_{kn} \cdot x_n + b_k\right) \quad (1)$$

Em geral pode-se identificar duas classes de arquiteturas fundamentalmente diferentes: redes alimentadas em adiante (*feedforward*) e redes recorrentes (HAYKIN, 2007). Redes recorrentes são caracterizadas pelo fato de existir pelo menos um laço de realimentação da saída de neurônios como entradas de neurônios de camadas anteriores (HAYKIN, 2007). Isto impacta fortemente a capacidade de aprendizagem da rede e envolve o emprego de elementos de atraso unitário (z^{-1}) que resulta em um comportamento dinâmico não linear (HAYKIN, 2007).

Dentre as redes neurais recorrentes, encontram-se as Redse Neural Artificial Auto-regressiva Não-linear (NARNET) que possibilitam a modelagem de séries temporais não lineares conforme subseção a seguir.

2.1.1 Rede Neural Artificial Auto-regressiva Não-linear (NARNET)

Redes neurais artificiais auto-regressivas não-lineares (*Non-linear Auto regressive Neural networks- NARNETs*) podem ser utilizadas para modelar séries temporais uma vez que são capazes de reconhecer seus padrões e características não lineares (IBRAHIM et al., 2016). Por meio destes modelos é possível determinar a série temporal no instante t tendo em vista os valores da série nos últimos p instantes de tempo ($t - 1, t - 2, \dots$,

$t - p$) (BENMOUIZA; CHEKNANE, 2013) através de um sistema discreto não linear auto-regressivo conforme equação 2.

$$\hat{y}(t) = f(y(t-1), y(t-2), \dots, y(t-p)) \quad (2)$$

Em que $\hat{y}$ representa os valores estimados da série temporal e f uma função não linear de aproximação na qual os valores estimados dependem unicamente dos últimos p valores da série temporal. A Figura 3 ilustra a estrutura de uma NARNET com uma única camada oculta.

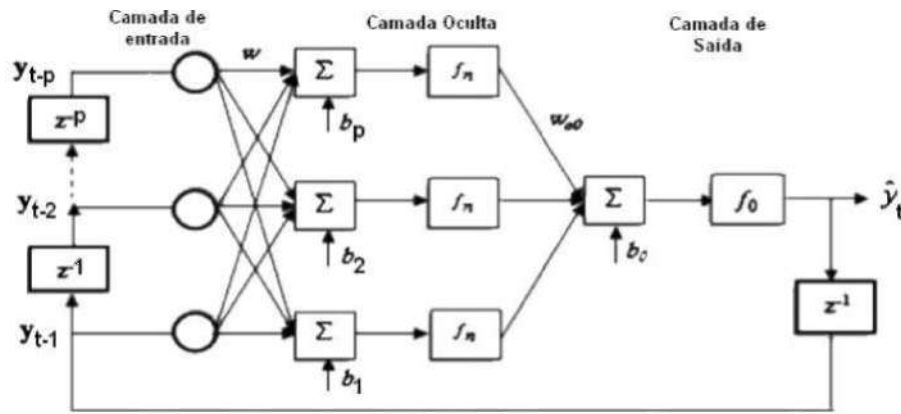


Figura 3: Estrutura de uma NARNET (BENMOUIZA; CHEKNANE, 2013, Adaptado) para estimar o valor atual ($\hat{y}_t$) de uma série temporal com base em seus últimos p valores. Cada entrada da rede encontra-se atrasada em relação à outra em uma unidade de tempo (z^{-1})

Conforme ilustrado através da Figura 3, NARNETs são redes neurais recorrentes baseadas em um modelo linear regressivo nas quais as p últimas saídas são realimentadas como entradas (BENMOUIZA; CHEKNANE, 2013).

2.2 Modelos de Classificação

A classificação consiste em, dado um conjunto de dados de treinamento, encontrar regras ou maneiras de classificar um novo exemplar tendo em vista suas características e atributos bem como um conjunto de classes previamente definido (PAULA, 2002). Embora existam metodologias não supervisionadas para a construção de classificadores, a subseções a seguir versam sobre as abordagens supervisionadas associadas à Máquinas de Vetores de Suporte (SVM) e árvores de decisão. Além disso, métricas de avaliação da *performance* de classificadores são apresentadas.

2.2.1 Máquinas de Vetores de Suporte (SVM)

Máquinas de Vetores de Suporte (*Support Vector Machines* - SVMs) são uma categoria de redes *feedforward* que podem ser usadas para classificação de padrões e regressão linear. A ideia base de SVMs é definir um hiperplano como superfície de decisão capaz de maximizar a separação entre duas classes de dados (HAYKIN, 2007). Nesse sentido, SVMs podem fornecer boa generalização em problemas de classificação de padrões.

O modelo mais simples de SVMs consiste em um Classificador de Margem Máxima que permite a separação entre dados linearmente separáveis. Embora esta característica limite a empregabilidade deste modelo, ele é fundamental ao entendimento de SVMs mais modernas. A Figura 4 ilustra espaços de características linearmente separáveis e inseparáveis nos quais uma linha escura representa o hiperplano de decisão (separação) ideal entre as classes.

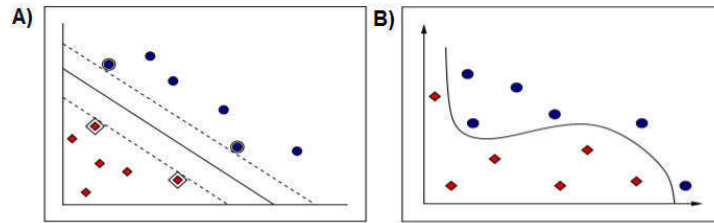


Figura 4: Espaços de características linearmente separáveis (A) e linearmente inseparáveis (B) tendo em vista as classes vermelha e azul. A linha escura esboça a fronteira ideal de decisão no espaço original (SANTOS, 2002, Adaptado).

Um hiperplano de Margem Máxima (ou de Separação Ótima) é aquele que separa um conjunto de vetores sem erro maximizando a distância entre os vetores (classes opostas) mais próximos ao hiperplano (SANTOS, 2002). Com o intuito de tolerar ruídos e *outliers*, as seguintes equações definem o modelo SVM.

$$x_i \cdot w + b \geq +1 - \xi_i \quad \text{para} \quad y_i = +1 \quad (3)$$

$$x_i \cdot w + b \leq -1 + \xi_i \quad \text{para} \quad y_i = -1 \quad (4)$$

$$\xi_i \geq 0 \quad \forall i \quad (5)$$

onde x_i , y_i e ξ_i definem as i -ésimas entrada, saída/classe alvo e diferença entre a saída alvo e o valor estipulado pela SVM. Da mesma forma, w e b representam os pesos e *biases* da rede. Com o intuito de separar conjuntos de dados não linearmente separáveis, SVMs realizam, antes da definição do hiperplano, um mapeamento não-linear dos vetores de entrada em um espaço de características com alta dimensão (SANTOS, 2002).

Trata-se de uma estratégia de pré-processamento dos dados através de funções *kernel* ($\epsilon(x)$) responsáveis por mapear o espaço de entrada X em um novo espaço $Z =$

$\{\phi(x)|x \in X\}$ denominado espaço de características. A Figura 5 ilustra um mapeamento de um espaço de entrada linearmente inseparável para um espaço de características de maior dimensão no qual os dados podem ser linearmente separados.

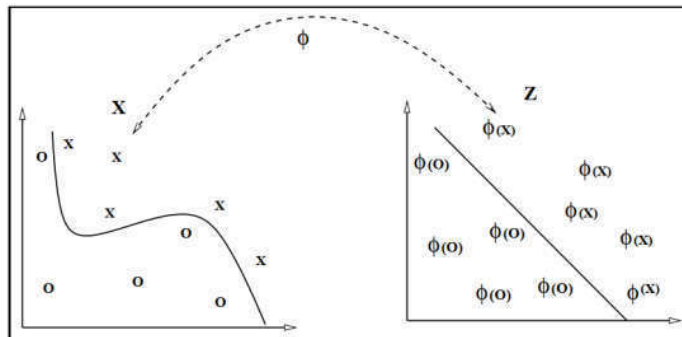


Figura 5: Mapeamento de um espaço de entrada X não linearmente separável e um espaço de características linearmente separável Z através de uma função *kernel* ϕ . (SANTOS, 2002).

Logo, a escolha da função *kernel* empregada em SVMs é importante (SANTOS, 2002) e define diferentes modelos SVM como, por exemplo, SVMs quadrático, cúbico e gaussiano.

2.2.2 Árvores de Decisão

Árvores de Decisão são algoritmos de aprendizagem de máquina que utilizam a estratégia de dividir para conquistar dividindo, portanto, um problema de classificação em subproblemas recursivamente (OSHIRO, 2013). Sua construção baseia-se na escolha de atributos para particionar as entradas em subconjuntos correspondentes. Dessa forma, uma árvore de decisão consiste em nodos (nós) e ramos. O nodo inicial é denominado nó raiz. A cada nó um atributo é verificado e, de acordo com o resultado, a árvore se ramifica até atingir os nodos terminais onde o resultado da classificação é designado (PAULA, 2002).

A Figura 6 ilustra a estrutura de uma árvore de decisão binária capaz de classificar dados em duas classes (C_1 e C_2). Independente de sua arquitetura, árvores de decisão apresentam caminhos determinados por uma sequência de verificações (efetuadas a cada nodo) que conduzem dados de entrada a uma classificação (PAULA, 2002).

O processo de aprendizagem da estrutura de uma árvore de decisão (ou de regras equivalentes) é conhecido como indução (PAULA, 2002). Uma árvore pode ser induzida pela seleção de algum atributo ou teste que servirá para particionar, a cada nodo, os exemplos em subconjuntos de acordo com o resultado do teste (OSHIRO, 2013). Dessa maneira, nodos tornam-se terminais quando todos exemplares restantes em um subconjunto pertencem a uma única classe ou quando o número de casos restantes for menor que um dado limiar (PAULA, 2002). Nesse caso, o nodo é associado à classe que apresenta maior

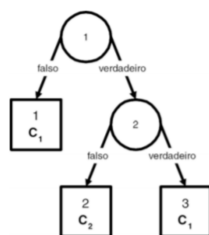


Figura 6: Estrutura exemplo de uma árvore de decisão binária em que nós terminais são representados através de quadrados e nós não terminais são representados por meio de círculos (OSHIRO, 2013).

frequência no grupo.

A seleção randômica de atributos resulta em árvores extensas e específicas para os dados de treinamento (*overfitting*) (OSHIRO, 2013). Nesse caso, o modo de seleção dos atributos (regras) consiste em avaliar sua qualidade ou representatividade de modo a aumentar a performance da árvore. A função de avaliação é, portanto, uma heurística que objetiva selecionar bons atributos para cada nodo tendo em vista informações incompletas. Após inferida, uma árvore de decisão pode ser empregada para classificar novas entradas (OSHIRO, 2013).

As classificações efetuadas por árvores de decisão podem ser combinadas na classificação de novas entradas (OSHIRO, 2013). Nesse sentido, métodos *ensemble* são algoritmos de aprendizado que constroem um conjunto de classificadores que combinam suas saídas (votos) na classificação de novos dados. Dentre os métodos *ensembles* populares, destacam-se os métodos *boosting*, *bagging* e *Random Forests*. Na subseção a seguir o método *ensemble bagging* é apresentado.

2.2.2.1 Bagging

O diagrama em blocos referente ao algoritmo *ensemble bagging* encontra-se ilustrado através da Figura 7.

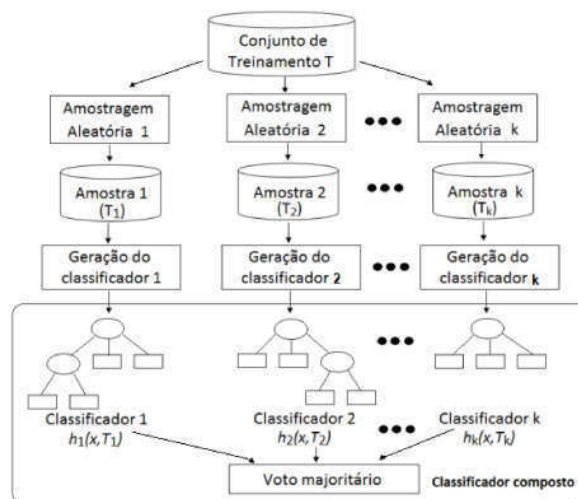


Figura 7: Diagrama em blocos sobre o funcionamento dos métodos *Bagging* (OSHIRO, 2013).

O método *bagging* é um algoritmo de aprendizado *ensemble* que emprega amostras *bootstrap* do conjunto de treinamento para treinar diferentes classificadores (OSHIRO, 2013). Na amostragem *bootstrap* E_0 , por exemplo, um conjunto de treinamento T com n exemplos é dividido em amostras (subconjuntos) T_k com reposição contendo n exemplos.

A seguir, cada T_k é empregado para treinar um classificador diferente (h_k). A estratégia de combinação dos classificadores consiste no voto majoritário (OSHIRO, 2013) conforme ilustrado através da Figura 7).

2.2.3 Métricas de Avaliação

Em aprendizado de máquina utilizam-se algumas ferramentas e métricas para avaliar o desempenho de classificadores. Dentre estas, pode-se citar matrizes de confusão que fornecem informações relevantes ao cálculo de diferentes parâmetros necessários à utilização de curvas de Características Operacionais do Receptor (ROC) e curvas de Precisão-Revocação (*Precision-Recall* - PR) (SILVA, 2006). A exemplo disso, pode-se citar a taxa de positivos verdadeiros ou revocação/sensibilidade, taxa de falsos positivos, acurácia e a métrica *F-score*. Dessa forma, as subseções a seguir abordam matrizes de confusão, curvas ROC e curvas PR bem como métricas associadas.

2.2.3.1 Matrizes de Confusão

Matrizes de confusão auxiliam na visualização do desempenho de classificadores. Tratam-se de matrizes quadradas cuja dimensão corresponde à quantidade de classes de um problema de classificação. Elas relaciona as saídas do classificados com as saídas almejadas de forma que em sua diagonal principal encontram-se a quantidade de acertos tendo em vista cada classe enquanto que os valores fora da diagonal principal representam

a quantidade de dados erroneamente classificados (OSHIRO, 2013).

Em um classificador binário é possível definir valores positivos verdadeiros (*True Positives* - TP) como a quantidade de exemplos rotulados corretamente como positivos, falsos positivos (*False Positives* - FP) como a quantidade de entradas negativas incorretamente classificadas como positivas, verdadeiros negativos (*True Negatives* - TN) correspondem a negativos rotulados corretamente como negativos e os falsos negativos (*False Negatives* - FN) compreendem exemplos positivos rotulados incorretamente como negativos (SILVA, 2006). Figura 8 exemplifica a matriz de confusão de um classificador binário.

As informações disponibilizadas através de uma matriz de confusão são utilizadas no cálculo de uma série de métricas como a taxa de positivos verdadeiros ou revocação/sensibilidade (equação 6) e a taxa de falsos positivos (*False Positive Rate* - equação 7) que servem como base para a construção das curvas ROC (SILVA, 2006). De forma similar, pode-se calcular a precisão como a fração de exemplos classificados como positivos que são verdadeiramente positivos (DAVIS; GOADRIC, 2006)(equação 8). Por fim, define-se a acurácia como a taxa de acerto geral do classificador, isto é, a razão entre a soma dos acertos e o número total de instâncias (equação 9).

		Classe Verdadeira	
		Positivo	Negativo
Classe Predita	Positivo	TP	FP
	Negativo	FN	TN

Figura 8: Matriz de Confusão gerada a partir de um conjunto de instâncias (SILVA, 2006, Adaptado).

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$FPR = \frac{FP}{FP + TN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Accuracy = \frac{TP + TN}{P + N} \quad (9)$$

onde $P = TP + FN$ e $N = FP + TN$ são as quantidades totais de dados pertencentes às classes positiva e negativa. Da mesma forma, *Recall*, *FPR*, *Precision* e *Accuracy* representam as métricas de revocação, taxa de falsos positivos, precisão e acurácia, respectivamente.

Podem ser utilizadas algumas variantes deste modelo de matriz de confusão capazes demonstrar mais facilmente métricas importantes como acurácia e precisão. Dessa forma,

na presente dissertação de mestrado uma matriz de confusão expandida será empregada conforme ilustrado pela Figura 9.

Classe Predita	0	TN	FN	B 1-B
	1	FP	TP	Precision 1-Precision
		A 1-A	Recall 1-Recall	Accuracy 1-Accuracy
		0	1	Classe Verdadeira

Figura 9: Matriz de confusão expandida na qual métricas de avaliação como acurácia e precisão são facilmente visualizadas.

Nota-se por meio da Figura 9 que a matriz de confusão expandida demonstra claramente as métricas acurácia (*accuracy*), precisão (*Precision*) e revocação ou sensibilidade (*Recall*) bem como quantidade de verdadeiros negativos (TN), verdadeiros positivos (TP), falsos negativos (FN) e falsos positivos (FP). Nota-se ainda o percentual de instâncias corretamente classificadas como negativas em relação à quantidade de classificados como negativos (B). Por fim, é possível visualizar o percentual de instâncias corretamente classificadas como negativas em relação ao total de instâncias que deveriam ter sido classificadas como negativas (A).

2.2.3.2 Curvas ROC

A curva ROC de um classificador binário discreto plota a sensibilidade (taxa de positivos verdadeiros) em função da taxa de falsos positivos conforme ilustrado na Figura 10.

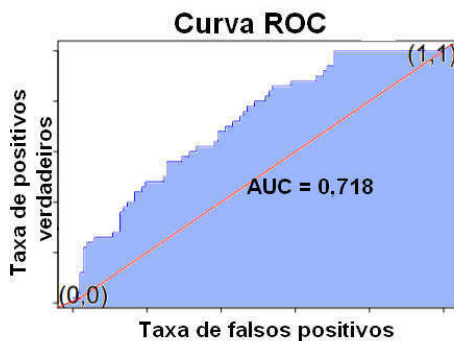


Figura 10: Exemplo de curva com ROC com área sob a curva (AUC) igual à 0,78. Linha em vermelho ilustra a curva de um classificador com saídas totalmente aleatórias (CORTES; MOHRI, 2004, Adaptado).

Um classificador com saídas totalmente aleatórias apresenta uma linha reta desde a origem até (1,1) em sua curva ROC. Curvas pelo menos um pouco acima dessa linha apresentam aprimoramento em relação à realização de classificações aleatórias (CORTES; MOHRI, 2004). Ademais, como curvas ROC correspondem à uma demonstração bidimensional da *performance* de um classificador, torna-se necessário reduzir essa curva à um valor escalar para comparar classificadores. Dessa forma, a área sob a curva ROC (*Area Under the Curve* - AUC) varia de 0 à 1 e resume estatisticamente a qualidade do ranking de uma solução de classificação (CORTES; MOHRI, 2004).

2.2.3.3 Curvas PR

O emprego de curvas ROC pode proporcionar resultados demasiadamente otimistas quanto à *performance* de um classificador quando há desbalanceamento considerável na distribuição entre as classes. Dessa forma, as curvas de Precisão-Revocação apresentam-se como uma alternativa à curvas ROC para tarefas que envolvam conjuntos de dados desbalanceados (DAVIS; GOADRIC, 2006). Nessas curvas, a precisão (equação 8) é plotada em função da revocação (equação 6). Em outras palavras, a fração de exemplos classificados como positivos que são verdadeiramente positivos é esboçada em função da fração de exemplos positivos que foram corretamente classificados conforme exemplificado na Figura 11.

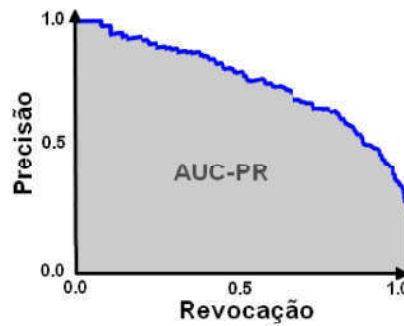


Figura 11: Exemplo de curva de Precisão-Revocação através de uma linha azul. A região cinza destaca a área sob a curva PR (AUC-PR) (CHICCO, 2017, Adaptado).

Assim como nas curvas ROC, a área sob as curvas PR (*Area Under the Curve Precision-Recall* - AUC-PR) é empregada como métrica na comparação de classificadores e quanto maior o seu valor, melhor o desempenho dos classificadores (DAVIS; GOADRIC, 2006). Além disso, como a precisão e a revocação descrevem aspectos complementares da correção de um sistema de classificação, pode-se combinar estes valores através da métrica *F-score* (SILVA, 2006) conforme equação 10.

$$F - score = \frac{2.Precision.Recall}{Precision + Recall} \quad (10)$$

2.3 Ferramentas de Visão Computacional

Com o intuito de desenvolver uma técnica de valvometria para a construção de biossensores baseados na análise comportamental de bivalves, propõe-se o emprego de câmeras em conjunto com algoritmos de visão computacional. Nesse sentido, as subseções que seguem versam sobre as ferramentas de visão computacional referentes à calibração de câmeras, operadores morfológicos e a representação de cores em imagens importantes ao desenvolvimento proposto.

2.3.1 Calibração de Câmeras

Calibração de câmera consiste no processo de determinar dados geométricos e ópticos da câmera, isto é, seus parâmetros intrínsecos e extrínsecos. Os parâmetros extrínsecos possibilitam relacionar o sistema de coordenadas da câmera com o sistema de coordenadas global (do “mundo”) enquanto que os parâmetros intrínsecos permitem modelar sua geometria e a distorção causada por suas lentes (BARELLI, 2018) conforme abordado nas subseções a seguir.

2.3.1.1 Parâmetros Intrínsecos

Estimar os parâmetros intrínsecos da câmera é essencial pois possibilita estimar a estrutura da cena no espaço Euclidiano e remover distorções causadas pelas lentes que comprometem a acurácia de algoritmos de visão computacional (BRADSKI; KAEHLER, 2008). Um modelo simples, mas útil da aquisição de imagens é o uso do modelo de câmera *pinhole* ilustrado na Figura 12. Neste modelo tanto o objeto quanto sua projeção apresentam a mesma orientação. Nesta imagem o local do orifício que permite passagem de luz corresponde ao centro de projeção e o ponto de intersecção entre o plano da imagem e o eixo óptico é referido como ponto principal.

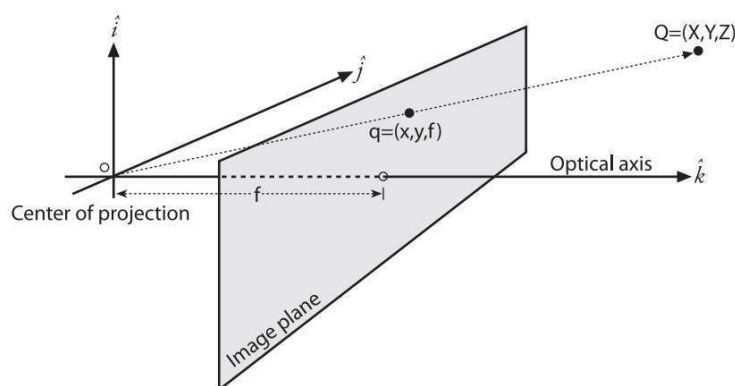


Figura 12: Ponto $Q=(X,Y,Z)$ projetado na imagem através de um raio que passa pelo centro de projeção resultando em um ponto $q = (z, y, f)$ projetado na imagem através de uma câmera com distância focal f . (BRADSKI; KAEHLER, 2008)

Nota-se através da Figura 12 que um ponto Q situado nas coordenadas X, Y, Z no mundo é representado no plano da imagem através de um ponto q de coordenadas x, y , f sendo f a distância focal da câmera. Dois parâmetros (c_x e c_y) são adicionados ao modelo com o intuito de considerar a diferença entre a posição do eixo óptico e o centro das coordenadas no plano de projeção. Consideram-se ainda duas distâncias focais f_x e f_y já que pixels individuais em câmeras de baixo custo são retangulares e não quadrados (BRADSKI; KAEHLER, 2008). As equações 11 e 12 descrevem o sistema em que um ponto $Q(X, Y, Z)$ no mundo é projetado no plano de projeção nas coordenadas x_{proj} e y_{proj} .

$$x_{proj} = f_x \cdot \frac{X}{Z} + c_x \quad (11)$$

$$y_{proj} = f_y \cdot \frac{Y}{Z} + c_y \quad (12)$$

A relação que descreve um ponto Q_i de coordenadas (X_i, Y_i, Z_i) no mundo em um ponto projetado com coordenadas (x_i, y_i) é chamada de transformação projetiva (BRADSKI; KAEHLER, 2008). É possível rearranjar os parâmetros que definem a câmera em uma única matriz 3×3 denominada matriz intrínseca da câmera (M) que relaciona a projeção de pontos do mundo (Q) no plano de imagem (q) conforme abaixo expresso:

$$q = MQ \quad (13)$$

Onde:

$$q = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad M = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \quad Q = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (14)$$

A matriz intrínseca da câmera não considera distorções porque um modelo pinhole não utiliza lentes (BRADSKI; KAEHLER, 2008). Existem dois tipos básicos de distorção causadas por lentes: distorções radiais e distorções tangenciais. A distorção radial ocorre quando os raios de luz dobram mais perto das bordas de uma lente do que no centro óptico causando o efeito "olho de peixe" (MATHWORKS, 2019). Nesse fenômeno, a distorção é nula no centro óptico e aumenta à medida que se aproxima das bordas sendo modelada com base na expansão em série de Taylor em torno de um raio nulo ($r = 0$) centrado no centro óptico (BRADSKI; KAEHLER, 2008) conforme segue.

$$x_{corrigido} = x (1 + k_1 \cdot r^2 + k_2 \cdot r^4 + k_3 \cdot r^6) \quad (15)$$

$$y_{corrigido} = y (1 + k_1 \cdot r^2 + k_2 \cdot r^4 + k_3 \cdot r^6) \quad (16)$$

onde (x, y) é a coordenada original distorcida e $(x_{corrigido}, y_{corrigido})$ corresponde à coordenada corrigida. A distorção tangencial decorre de defeitos de fabricação que fazem com que a lente não esteja posicionada perfeitamente paralela ao plano da imagem (BRADSKI; KAEHLER, 2008). Os coeficientes de distorção tangencial modelam esse tipo de distorção de acordo com a equação abaixo:

$$x_{distorcido} = x + [2 \cdot p_1 \cdot x \cdot y + p_2 \cdot (r^2 + 2 \cdot x^2)] \quad (17)$$

$$y_{distorcido} = y + [p_1 \cdot (r^2 + 2 \cdot y^2) + 2 \cdot p_2 \cdot x \cdot y] \quad (18)$$

onde (x, y) são a localização do pixel sem distorção, p_1 e p_2 são os coeficientes de distorção tangencial da lente e $(x_{distorcido}, y_{distorcido})$ são a localização dos pixel distorcidos.

2.3.1.2 Parâmetros extrínsecos

Para cada imagem capturada pela câmera de um objeto específico, pode-se descrever a pose do objeto em relação ao sistema de coordenadas da câmera em termos de rotação e translação (BRADSKI; KAEHLER, 2008). O vetor de translação (T) descreve uma mudança de um sistema de coordenadas para outro sistema cuja origem é deslocada (BRADSKI; KAEHLER, 2008).

Uma matriz de rotação em três dimensões pode ser decomposta em uma rotação bidimensional em torno de cada eixo, no qual as medições do eixo pivô permanecem constantes (BRADSKI; KAEHLER, 2008). Se girarmos em torno dos eixos x , y e z em sequência com os respectivos ângulos de rotação ψ, φ e θ , o resultado é uma matriz de rotação total R que é dada pelo produto das três matrizes $R_x(\psi), R_y(\varphi)$ e $R_z(\theta)$ (BRADSKI; KAEHLER, 2008), em que:

$$R_x(\psi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \psi & \sin \psi \\ 0 & -\sin \psi & \cos \psi \end{bmatrix}, R_y(\varphi) = \begin{bmatrix} \cos \varphi & 0 & -\sin \varphi \\ 0 & 1 & 0 \\ \sin \varphi & 0 & \cos \varphi \end{bmatrix} \quad (19)$$

$$R_z(\theta) = \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (20)$$

2.3.2 Operadores Morfológicos

Operações morfológicas são operações que modificam a estrutura dos objetos representados em uma imagem. Embora possam ser utilizadas em qualquer tipo de imagem, elas são normalmente empregadas em imagens binárias para amenizar ruídos ou realçar objetos de interesse (BARELLI, 2018). Estas operações são realizadas pixel a pixel com base na convolução de um elemento estruturante (normalmente representado por uma matriz quadrada) com a imagem a ser modificada.

As operações morfológicas de erosão e dilatação são as mais fundamentais uma vez que todas as demais podem ser descritas basicamente como uma determinada sequência de erosões e dilatações (BARELLI, 2018). A operação de erosão permite a corrosão de arestas do objeto de interesse e pode levar à eliminação de pequenos objetos presentes na imagem sendo empregada na eliminação de pequenos ruídos. Sua realização consiste em sobrepor cada pixel da imagem com o centro do elemento estruturante. O pixel analisado somente será representado na imagem resultante se todos os pixels demarcados no elemento estruturante coincidirem com pontos do objeto de interesse conforme exemplificado na Figura 13.

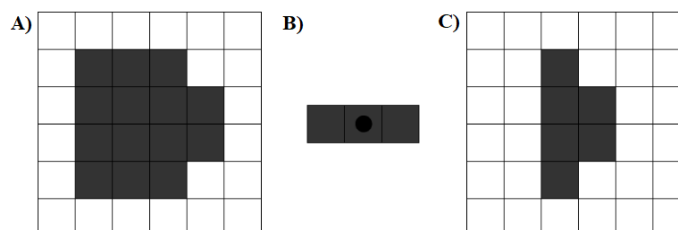


Figura 13: Exemplo de operação morfológica de erosão na qual a imagem original (A) passa por um processo de convolução com um elemento estruturante retangular (B) resultando na imagem a direita (C) (BARELLI, 2018).

A operação morfológica de dilatação possibilita alargar a área de objetos representados. Dessa forma, os artefatos presentes na imagem tendem a se tornar maiores (BARELLI, 2018) conforme exemplificado na Figura 14. No processo de convolução se pelo menos um pixel do elemento coincidir com um pixel representado na imagem, a posição do pixel central fará parte da imagem resultante.

Com o intuito de eliminar ruídos e ao mesmo tempo preservar as regiões de interesse pode-se fazer uso do operador morfológico de abertura que consiste em uma operação de erosão seguida de uma operação de dilatação (BARELLI, 2018). De forma análoga, o operador morfológico de fechamento é caracterizado pela utilização destes operadores em ordem inversa e permite preencher pontos de objetos que foram danificados durante o processo de binarização.

As operações de abertura e fechamento são utilizadas com o objetivo de uniformizar a iluminação dos objetos uma vez que tendem a suprimir pequenas regiões brilhantes e

escuras, respectivamente (BARELLI, 2018). A Figura 15 ilustra o operação de abertura a uma imagem em escala de cinza e sua subtração com a imagem original na correção da iluminação da imagem.

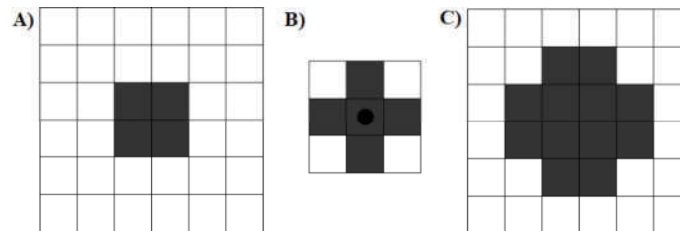


Figura 14: Exemplo de operação morfológica de dilatação na qual a imagem original (A) passa por um processo de convolução com um elemento estruturante em forma de cruz (B) resultando na imagem a direita (C) (BARELLI, 2018).

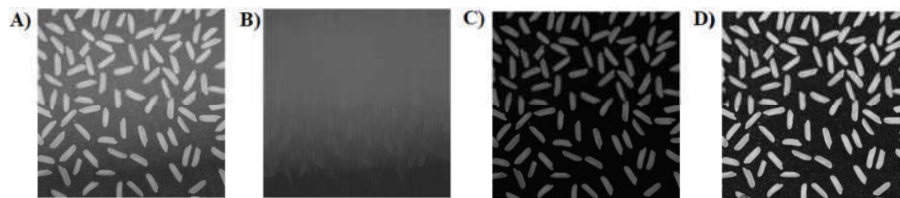


Figura 15: Exemplo da aplicação do operador morfológico de abertura à uma imagem em escala de cinza (A). O resultado (B) quando subtraído da imagem original fornece uma imagem com iluminação mais uniforme (C) que pode ser melhorada através do ajuste de seu contraste (D) (BARELLI, 2018).

2.3.3 Representação de cores em imagens

Cor é uma experiência rica e complexa geralmente causada pelo fato do sistema visual humano responder de forma alterada a diferentes comprimentos de onda da luz (FORSYTH; PONCE, 2003). A representação de imagens coloridas requer o emprego de múltiplos canais enquanto que imagens em escala de cinza (monocromáticas) possuem apenas um canal que representa a luminância (Y) em todos os pontos de uma dada cena (DAWSON-HOWE, 2014). As informações de cor podem ser representadas de diversas formas como, por exemplo, através dos espaços de cor RGB (*Red-Green-Blue*) e HSV (*Hue, Saturation and Value*) abordados nas subseções a seguir e ilustrados por meio da Figura 16.

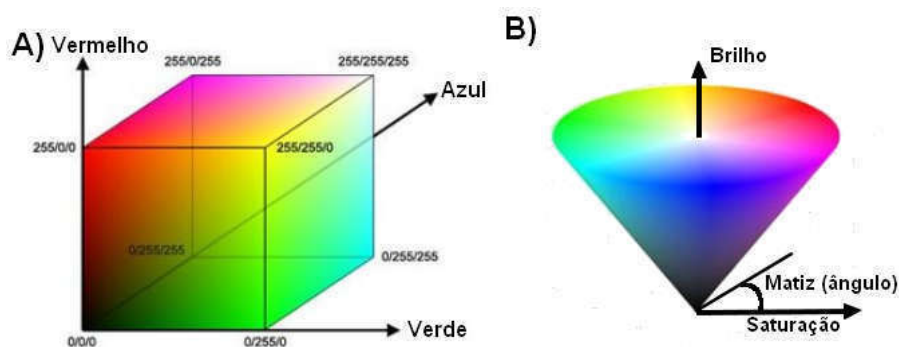


Figura 16: O espaço RGB é representado por um cubo (A) através do qual é possível visualizar todas as cores passíveis de obtenção a partir da combinação das componentes vermelha, verde e azul variando de 0 à 1 deste espectro. É comum visualizar este cubo ao longo de seu eixo neutro (eixo da origem ao ponto 1,1,1). A seção transversal do formato resultante corresponde ao cone que representa o espaço HSV (B). Fonte: Corrêa¹ et al. (2014) (adaptado).

2.3.3.1 Espaço RGB

De acordo com Dawson-Howe (2014), a representação de imagens coloridas mais comum consiste na utilização de três canais que representam, de forma resumida, os comprimentos de onda vermelho (*Red* - 700 nm), verde (*Green* - 546 nm) e azul (*Blue* - 435 nm). Trata-se do espectro de cores linear RGB no qual cada pixel é representado pela mistura de componentes vermelha, verde e azul cujos valores variam de 0 à 255 conforme ilustrado na Figura 16A.

Embora a utilização deste espaço possibilite representar aproximadamente 16,8 milhões ($256 \times 256 \times 256$) cores, segundo Dawson-Howe (2014) algumas cores não podem ser representadas através desse modelo.

2.3.3.2 Espaço HSV

Embora o espaço de cor RGB seja amplamente utilizado, coordenadas de cor em um espaço linear não capturam intuições humanas acerca da topologia de cores (FORSYTH; PONCE, 2003). Dessa forma, pode-se construir espaços de cor que refletem essas relações através da transformação não linear do espaço RGB. Dentre estes encontra-se o espaço HSV (*Hue, Saturation and Value*) cujas componentes encontram-se abaixo descritas conforme Forsyth e Ponce (2003). Ademais, este espaço é representado através de um cone (Figura 16B).

- Matiz (*hue*)- propriedade de uma cor que varia na passagem de vermelho para verde;
- Saturação (*saturation*) - propriedade de uma cor que varia na passagem de vermelho

para rosa;

- Brilho - também chamado de valor (*value*), corresponde a propriedade que varia na passagem de preto para branco;

2.4 Ferramentas de análise estatística

A presente dissertação de mestrado propõe uma técnica de valvometria baseada em visão computacional e sua validação através de um comparativo com o emprego de sensores Hall por meio de ferramentas estatísticas. Nesse sentido, analisam-se as curvas de regressão linear e a correlação entre os dados obtidos por estas abordagens bem como a Função Densidade de Probabilidade (FDP) entre os erros. Empregam-se ainda gráficos de controle de qualidade como ferramenta de análise a fim de identificar alterações decorrentes de exposição toxicológica.

Por fim, diferenças significativas entre grupos de animais expostos e não expostos a contaminantes são investigadas a fim de comparar a performance dos modelos propostos com o emprego de métodos estatísticos tradicionais. Dentre estes métodos, pode-se citar Análise de Variância (ANOVA - *Analysis of variance*) de um ou múltiplos fatores e o teste não paramétrico de Kruskal-Wallis. Dessa forma, as subseções a seguir versam sobre regressão linear, correlação, FDP, gráficos de controle de qualidade, métodos ANOVA e Kruskal-Wallis.

2.4.1 Regressão linear

A análise de regressão linear bivariada permite a obtenção de um modelo matemático capaz de estimar o valor de uma variável (y - variável dependente) a partir de outra (x - variável independente) (BLACK, 2009). Nesta análise tem-se, portanto, a modelagem da relação entre as variáveis dependente e independente através de uma linha reta. Em estatística, o modelo probabilístico que descreve a linha de regressão através de pontos amostrais pode ser definida conforme equação 21. Figura 17 ilustra o modelo regressão linear para uma dada amostra de variáveis dependentes e independentes.

$$\hat{y} = b_0 + b_1.x \quad (21)$$

onde $\hat{y}$ corresponde ao valor estimado para a variável dependente, x representa a variável independente, b_0 o valor onde a reta intercepta o eixo y e b_1 é a inclinação da curva.

Torna-se necessário, portanto, determinar os coeficientes b_0 e b_1 . Nesse sentido, pode-se empregar o método dos mínimos quadrados uma vez que ele possibilita determinar os coeficientes em questão de forma a produzir a mínima soma dos valores de erro ao quadrado (BLACK, 2009). Deve-se, portanto, definir a soma do quadrado dos erros (SQ)

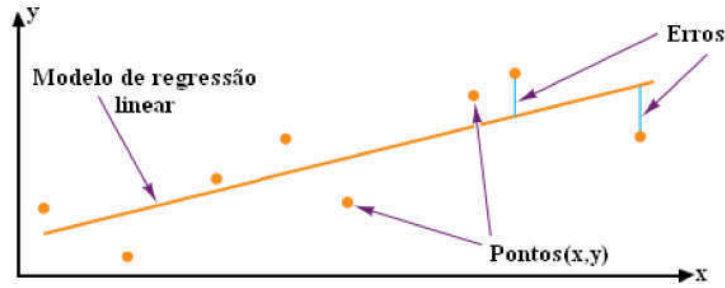


Figura 17: Modelo de regressão linear baseado em pontos amostrais de duas variáveis dependente (y) e independente (x). A reta modelada não estima corretamente todos valores de y tendo como base os valores de x havendo, dessa forma, erros definidos como a diferença entre os valores real e estimado da variável dependente (BLACK, 2009)(adaptado)

sendo estes definidos como as diferenças entre os valores estimados e reais da amostra da variável dependente ($e = \hat{y} - y$) de tamanho n conforme equação 22.

$$SQ = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 \cdot x_i))^2 \quad (22)$$

Com o intuito de encontrar os coeficiente que minimizam SQ , pode-se derivar a equação 22 em função de b_0 e b_1 e igualar a zero conforme as equações 23 e 24. Solucionando o sistema de equações resultante pode-se obter os coeficientes do modelo de regressão linear baseado no método de mínimos quadrados conforme equações 25 e 26 (BLACK, 2009).

$$n \cdot b_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \quad (23)$$

$$b_0 \sum_{i=1}^n x_i + b_1 \cdot \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i \cdot y_i \quad (24)$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} \quad (25)$$

$$b_1 = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2} \quad (26)$$

2.4.2 Correlação

Em muitas aplicações torna-se interessante analisar o relacionamento entre duas ou mais variáveis. Nesse sentido, os coeficientes de associação ou correlação permitem avaliar quantitativamente o grau de associação ou dependência linear entre duas variáveis (PUBLISHERS, 2010). A exemplo disso, pode-se citar o coeficiente de Pearson (r) que fornece valores entre -1 e +1 (equação 27). Valores iguais a -1, +1 e 0 indicam associações perfeitas negativas, perfeitas positivas e a não existência de associação li-

near entre as variáveis, respectivamente (PUBLISHERS, 2010). Além disso, a análise estatística r pode indicar diferentes níveis de correlação como correlações lineares muito fracas ($0 < |r| < 0.2$), fracas ($0.2 \leq |r| < 0.4$), moderadas ($0.4 \leq |r| < 0.6$), fortes ($0.6 \leq |r| < 0.8$) ou muito fortes ($0.8 \leq |r| < 1$) (CHRISTMANN; BADGETT, 2009).

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \quad (27)$$

onde r é o coeficiente de correlação de Pearson entre as variáveis x e y cujos valores médios são $\bar{x}$ e $\bar{y}$, respectivamente. Por fim, o quadrado deste coeficiente é conhecido como coeficiente de determinação (r^2) e possibilita avaliar a capacidade de um dado modelo de regressão linear de explicar um resultado (BLACK, 2009). Ele indica, portanto, a proporção da variância na variável dependente que é explicada pelo modelo de regressão linear e a variável independente.

2.4.3 Função densidade de probabilidade

A FDP é um conceito fundamental em estatística (SILVERMAN, 1986). Considerando a função densidade de probabilidade f de uma variável aleatória X , pode-se afirmar que f possibilita encontrar a probabilidade da variável X estar situada entre os valores a e b conforme descrito em 28.

$$P(a \leq X < b) = \int_a^b f(x)dx \quad \forall a < b \quad (28)$$

Nesse sentido, dado um conjunto de pontos amostrais de um dado processo ou fenômeno com uma função de densidade de probabilidade desconhecida torna-se interessante estimar $f(x)$ a partir dos dados observados. Uma abordagem paramétrica pode ser empregada com esta finalidade. Entretanto, nesta abordagem supõem-se que os dados extraídos pertencem a uma família paramétrica conhecida de distribuições (SILVERMAN, 1986). Muitas vezes, a utilização de abordagens não paramétricas torna-se mais adequada aos dados amostrais, visto que suposições menos rígidas são feitas acerca de sua distribuição.

Em estatística, a Estimativa de Densidade por *Kernel* (EDK) é uma forma não-paramétrica para estimar a FDP de uma variável aleatória (SILVERMAN, 1986). Sendo $x_1, x_2, \dots, x_n$ uma amostra de uma variável aleatória independente e identicamente distribuída com uma densidade desconhecida, pode-se estimar sua FDP ($\hat{f}$) através de uma EDK conforme 29.

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \quad (29)$$

onde $\hat{f}$ é a FDP estimada, K é a função *kernel* empregada, h é um parâmetro de

distância denominado de largura de banda que controla a influência de valores vizinhos em um *kernel*. Uma função *kernel* centrada em X_i atribui pesos às contribuições dadas por cada vizinho x à FDP estimada $\hat{f}$ de acordo com a proximidade de X_i a x . Normalmente, funções *kernel* são positivas em todos os lugares e simétricas em relação a zero. A Figura 18 exemplifica uma EDK utilizando uma função normal.

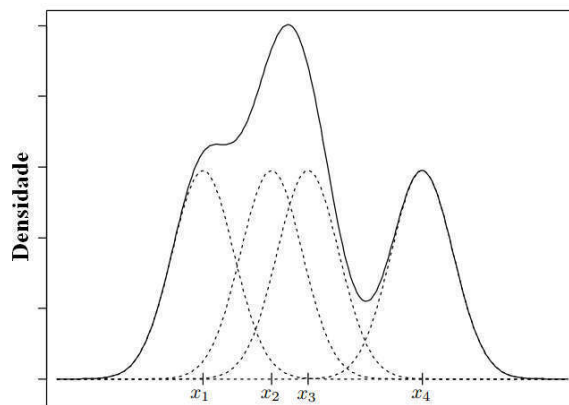


Figura 18: Estimativa normal de densidade por *kernel* (linha sólida) e contribuições do *kernel* (linhas pontilhadas) centrado em cada X_i da amostra $x_1, x_2, \dots, x_4$. A EDK em qualquer x é a soma das contribuições do *kernel* centrado em cada X_i (GIVENS; HOE-TING, 2012)(adaptado).

2.4.4 Gráficos de Controle de Qualidade

De acordo com Black (2008), qualidade é habilidade de um produto em fornecer características condizentes com suas especificações. Controle de qualidade pode ser entendido como um conjunto de estratégias, técnicas e ações utilizadas por uma organização com o intuito de assegurar a produção de produtos de qualidade.

Além disso, processo pode ser definido como uma série de ações, alterações ou funções que levam a um resultado. Nesse sentido, muitas ferramentas tem sido desenvolvidas com o intuito de auxiliar no controle de qualidade de processos. Dentre estas ferramentas, gráficos de controle são ferramentas estatísticas que englobam métodos gráficos para avaliar se um processo está ou não em um estado de controle estatístico (BLACK, 2008).

Cada gráfico de controle apresenta uma linha central ou limite de controle médio (*Central Line* - CL), e linhas de controle superior (*Upper Control Limit* - UCL) e inferior (*Lower Control Limit* - LCL). Estas ferramentas podem ser utilizados para verificar a previsibilidade de séries temporais fazendo uso de dados anteriores para determinar CL, UCL e LCL. A Figura 19 ilustra um exemplo de gráfico de controle EWMA.

Os limites estabelecidos no gráfico de controle (UCL, CL e LCL) podem ser empregados para concluir se a variação do processo é consistente (no controle) ou imprevisível (fora de controle) e, portanto, afetada por causas especiais de variação (TAGUE, 2005).

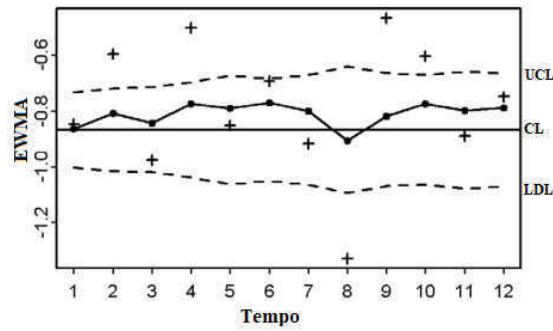


Figura 19: Exemplo de gráfico de controle EWMA para uma série temporal. As linhas de limite central (CL), superior (UCL) e inferior (LCL) encontram-se ilustradas. Os círculos indicam os valores resultantes da média ponderada e as cruzes indicam os valores da série analisada a cada intervalo de tempo (SIREGAR et al., 2013, Adaptado).

Dentre outros, o gráfico de controle baseado na média móvel exponencialmente ponderada (*Exponentially Weighted Moving Average* - EWMA) é um procedimento quase perfeitamente não paramétrico, muito eficaz contra pequenas variações no processo e sensível à autocorrelação dos dados (MONTGOMERY, 2009).

A EWMA é uma média ponderada das observações atuais e passadas. Os pesos são baseados no tempo de forma que as observações atuais apresentam peso exponencialmente maior que as mais antigas. Dessa forma, seu emprego baseia-se na configurações de dois parâmetros sendo eles o fator peso λ dado à observação mais recente e o fator k que determina o número de desvios padrões empregados para estabelecer os limites superior e inferior. Dessa forma, a estatística empregada nestes gráficos de controle pode ser calculada conforme equação 30.

$$z_i = \lambda \sum_{j=0}^{i-1} (1 - \lambda)^j x_{i-j} + (1 - \lambda)^i z_0 \quad (30)$$

onde $z_0 = \bar{x} = \mu_0$, z_i é o valor da média ponderada no i -ésimo intervalo e x representa os valores inspecionados através do mapa de controle EWMA. Nota-se através desta equação que os valores dos pesos decaem geometricamente. Além disso, UCL e LCL são estabelecidos com base na linha central (z_0) e na variância (σ) conforme as equações (31) e (32), respectivamente.

$$UCL = \mu_0 + k\sigma\sqrt{\frac{\lambda}{2 - \lambda}} \quad (31)$$

$$LCL = \mu_0 - k\sigma\sqrt{\frac{\lambda}{2 - \lambda}} \quad (32)$$

2.4.5 Métodos ANOVA e alternativas não-paramétricas

Segundo Publishers (2010), a análise da variância (*Analysis Of Variance* - ANOVA) pode ser pensada como um método para testar a hipótese de igualdade entre as médias de diferentes grupos por meio da análise das variâncias de suas amostras. Dessa forma, o método ANOVA pode ser empregado para determinar a existência de diferenças significativas entre grupos de acordo com um ou mais fatores.

O método ANOVA de um fator possibilita a comparação de médias univariadas de um ou mais grupos. Por sua vez, o método ANOVA multivariado (*Multivariate Analysis of Variance* - MANOVA) corresponde a uma extensão da análise de variância de um fator capaz de identificar diferenças significativas entre grupos com base em diversas variáveis respostas (fatores) simultaneamente (REIS, 2001).

As informações fornecidas pela modelagem univariada são contempladas pela extensão multivariada sem, no entanto, levar em consideração um nível de significância conjunto dos testes e um aproveitamento das correlações existentes entre variáveis (DEMÉTRIO, 1995). Ademais, existem diversas razões para adotar testes multivariados no lugar de testes univariados repetidos. Primeiramente, a utilização de testes estatísticos separados para cada variável provoca incremento demasiado de erros tipo I, ou seja, aumenta a probabilidade de rejeitar a hipótese nula quando ela é verdadeira (COUTO, 2006).

Além disso, embora não sejam verificadas diferenças significativas entre grupos na análise de cada variável individual, quando estas são inspecionadas em conjunto por meio de um método multivariado diferenças significativas poderão ser identificadas devido à combinação de pequenas diferenças entre variáveis individuais (COUTO, 2006). Nesses casos, testes multivariados tornam-se mais poderosos com maior probabilidade de rejeitar a hipótese nula quando ela é falsa (COUTO, 2006). Por fim, testes univariados não consideram informações importantes contidas nos dados como, por exemplo, a correlação entre variáveis que é incorporada na análise multivariada através de matrizes de variâncias e covariâncias.

O emprego do método MANOVA requer a verificação de alguns pressupostos derivados do método ANOVA de um fator sendo eles a normalidade multivariada e a igualdade de matrizes de variância e covariância (COUTO, 2006). Quando as suposições necessárias ao emprego dos métodos paramétricos não são verificados, alternativas não paramétricas devem ser empregadas. Dentre estas alternativas podem-se citar os métodos de Kruskal-Wallis para o caso univariado e a análise de variância multivariada permutacional (*Permutational Multivariate Analysis of Variance* - PERMANOVA) (FACHADA et al., 2016).

O método PERMANOVA foi proposto por McArdle e Anderson (2001) e, de acordo com Sun (2020), é um dos testes estatísticos não-paramétricos multivariados mais amplamente usados em ecologia microbiana. Ele é, portanto, um teste não paramétrico multivariado ANOVA com hipótese nula de não haver diferenças significativas entre os grupos

e baseia-se na análise e particionamento da soma dos quadrados das distâncias e processo de permutação.

Com o intuito de verificar os conceitos básicos referentes ao método ANOVA de um fator a partir do qual os métodos MANOVA e PERMANOVA são derivados bem como acerca de sua opção não paramétrica, as subseções a seguir apresentam os métodos ANOVA de um fator e Kruskal-Wallis.

2.4.5.1 ANOVA

Métodos ANOVA podem ser empregados para investigar a influência de um determinado fator sob um fenômeno observado. Cada valor observado pode, portanto, ser modelado conforme a equação 33 sendo composto por uma parte previsível (ou controlada) afetada pelo fator investigado e outra aleatória (não previsível).

$$y_{i,j} = \theta_i + e_{i,j} \quad (33)$$

onde $y_{i,j}$, θ_i e $e_{i,j}$ correspondem à j -ésima observação do grupo (nível) i , ao efeito comum a todas as observações do grupo (nível) i e ao efeito aleatório (não controlado) da j -ésima observação deste mesmo nível. O erro pode ser considerado como o efeito resultante de características não explicadas pelo modelo.

Considerando-se um fator de n níveis, deve-se estimar $\theta_1, \theta_2, \dots, \theta_n$ capazes de minimizar o somatório do quadrado dos ruídos e verificar se o conhecimento do fator escolhido influencia o fenômeno estudado. No entanto, esta abordagem requer o emprego de algumas restrições conforme expresso pelas equações 34 e 35 nas quais e_i corresponde ao erro não controlado das observações do nível (grupo) i .

$$E(e_i) = 0, \quad \forall i \in [1, n] \quad (34)$$

$$Var(e_i) = \sigma_i^2, \quad \forall i \in [1, n] \quad (35)$$

Considerando-se um fator de n níveis e que cada nível i possui n_i observações, o somatório do quadrado dos resíduos pode ser expresso pelas equações 36 e 37.

$$SQ(\theta_1, \dots, \theta_n) = \sum_{i=1}^n \sum_{j=1}^{n_i} e_{i,j}^2 \quad (36)$$

$$SQ(\theta_1, \dots, \theta_n) = \sum_{j=1}^{n_1} e_{1,j}^2 + \sum_{j=2}^{n_2} e_{2,j}^2 + \dots + \sum_{j=n}^{n_n} e_{n,j}^2 \quad (37)$$

Observa-se através da equação 37 que se as variâncias residuais $\theta_1^2, \theta_2^2, \dots, \theta_n^2$ não forem iguais, a soma dos quadrados dos resíduos será mais afetada pelo grupo com menor

variância. Para fins de simplificação considera-se variância constante dos erros, propriedade também conhecida como homoscedasticidade (equações 38 e 39). Admite-se, dessa forma, a mesma variabilidade residual entre os grupos.

$$Var(e_1) = Var(e_2) = \dots = Var(e_n) \quad (38)$$

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_e^2 \quad (39)$$

Com esse intuito de encontrar $\theta_1, \theta_2, \dots, \theta_n$ capazes de minimizar a soma dos quadrados dos resíduos dos diferentes níveis, efetua-se sua derivada função da parte previsível de cada grupo e iguala-se a zero conforme segue:

$$\frac{\partial SQ(\theta_1, \dots, \theta_n)}{\partial \theta_i} = 0 \quad (40)$$

$$-2 \sum_j^{n_i} (y_{i,j} - \hat{\theta}_i) = 0 \therefore \hat{\theta}_i = \frac{\sum_{j=1}^{n_i} y_{i,j}}{n_i} \quad (41)$$

Em que $i \in [1, n]$, n é a quantidade de níveis do fator considerado e n_i é a quantidade de observações de cada nível i . Analisando-se a expressão 41 nota-se que $\hat{\theta}_i = \bar{y}_i$. Pode-se afirmar, portanto, que a quantidade total de informação perdida pela adoção do modelo é dada por 42.

$$SQ(\theta_1, \dots, \theta_n) = \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2 \quad (42)$$

Analisando-se a variância residual de cada nível e a minimização proposta, pode-se afirmar que $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)$. Contudo, considerando a suposição de homoscedasticidade, pode-se afirmar que:

$$s_e^2 = s_i^2 \quad \forall i \in [1, n] \quad (43)$$

$$s_e^2 = \frac{\sum_{i=1}^n (n_i - 1) \cdot s_i^2}{\sum_{i=1}^n n_i - n} \quad (44)$$

$$s_e^2 = \frac{SQ(\theta_1, \dots, \theta_n)}{m - n} \quad (45)$$

Em que $m = \sum_{i=1}^n n_i$ e s_e^2 corresponde a quantidade média de informação perdida. A homoscedasticidade entre os níveis deve ser observada e torna-se necessário estabelecer suposições acerca do efeito residual e independência das observações tanto dentro quanto entre os níveis do fator considerado conforme segue.

- $e_{i,j} : N(0, \sigma_e^2) \quad \forall i \in [1, n] \text{ e } j \in [1, n_i]$

- $E(e_i.e_j) = 0 \quad \forall i \neq j$, indicando independência entre as observações dentro de cada população;

As suposições necessárias a adoção do modelo em questão requer, portanto, que os resíduos sigam uma distribuição normal com média 0 e variância σ_e^2 e que resíduos associados a pares de observações sejam independentes. Com base nessas suposições, pode-se afirmar que o conjunto de observações $y_1, y_2, \dots, y_n$ é uma amostra aleatória simples da população $N(\theta, \sigma_e^2)$.

As informações utilizadas para empregar a ANOVA podem ser resumidas através de uma tabela tendo em vista o modelo adotado. Conforme anteriormente exposto, a equação 42 corresponde a quantidade total de informações perdidas dentro dos grupos sendo, conforme a equação 46, referenciada pela tabela de análise de variância como SQ_{Den} . De forma análoga, a quantidade total de informações perdidas pela adoção do modelo em questão é dada pela equação 47.

$$SQ_{Den} = SQ(\theta_1, \dots, \theta_n) = \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2 \quad (46)$$

$$SQ_{Total} = \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{i,j} - \bar{y})^2 \quad (47)$$

em que $\bar{y}$ corresponde à média de todas observações conforme descrito pela equação 48 e m é a quantidade total de observações dada por $m = \sum_{i=1}^n n_i$.

$$\bar{y} = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} y_{i,j}}{m} \quad (48)$$

Subtraindo-se a quantidade total de informação perdida (SQ_{Total}) da soma dos quadrados dos resíduos dentro dos grupos (SQ_{Den}) tem-se a economia obtida ao se passar de um modelo para o outro (SQ_{Ent}) conforme descrito pela equação 49.

$$SQ_{Ent} = SQ_{Total} - SQ_{Den} \quad (49)$$

$$SQ_{Ent} = \sum_{i=1}^n (n_i(\bar{y}_i - \bar{y})) \quad (50)$$

Dessa forma, a expressão 50 mede uma certa distância entre a média de cada grupo e a média global dos dados analisados. De acordo com Bussab (1988), quanto maior a diferença entre as médias $\bar{y}_i$ mais eficaz é a consideração do fator na descrição do fenômeno analisado. Além disso, pode-se afirmar que a quantidade total média de informação perdida QM_{Total} é dada pela equação 51. De modo análogo, é possível afir-

mar que a quantidade média de informação perdida entre os níveis (QM_{Den}) pode ser expressa pela equação 52.

$$QM_{Total} = \frac{1}{n-1} \sum_i^n (y_i - \bar{y}) = \frac{1}{n-1} SQ_{Total} \quad (51)$$

$$QM_{Den} = \frac{1}{m-n} \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2 \quad (52)$$

$$QM_{Den} = \frac{1}{m-n} SQ_{Den} \quad (53)$$

em que $m = \sum_{i=1}^n n_i$ corresponde a quantidade total de observações. Estas informações são agrupadas através de uma tabela de análise de variância tal qual a Tabela 1.

Tabela 1: Tabela de Análise de Variância

Fonte de variação	Graus de Liberdade (gl)	Soma dos quadrados (SQ)	Quantidade média (QM)	F
Entre	$n-1$	SQ_{Ent}	QM_{Ent}	$\frac{QM_{Ent}}{QM_{Den}}$
Dentro	$m-n$	SQ_{Den}	QM_{Den}	
Total	$m-1$	SQ_{Total}		

Pode-se reescrever a equação 33 com base nas observações de todos os níveis do fator considerado tal qual a equação 54. Nota-se, todavia, que a influência desse fator segue compondo a parte aleatória do modelo.

$$y_{i,j} = \theta + e_{i,j} \quad (54)$$

Tem-se, por conseguinte, a adoção da hipótese nula $H_0 : \theta_1 = \theta_2 = \dots = \theta_n$ contra a hipótese alternativa de que pelo menos uma das igualdades não se verifica. Segundo Bussab (1988), rejeitar a hipótese nula significa afirmar que o fator é importante para explicar a variável de interesse ou fenômeno observado.

Considerando-se, para fins de simplificação, que todos os níveis contam com a mesma quantidade de observações, pode-se estabelecer a variância comum σ_e^2 através da estimativa ponderada s_e^2 abaixo reescrita.

$$s_e^2 = \frac{\sum_{i=1}^n (n_i - 1) s_i^2}{m - n} = \frac{SQ_{Den}}{m - n} = QM_{Den} \quad (55)$$

Tendo em vista, entretanto, o Teorema do Limite Central, pode-se afirmar que:

$$\bar{y}_i : N(\theta_i : \frac{\sigma_e^2}{n_n}) \quad \forall i \in [1, n] \quad (56)$$

Quando, por outro lado, não são encontradas evidências para rejeitar a hipótese nula, tem-se:

$$\bar{y}_i : N\left(\theta : \frac{\sigma_e^2}{n_n}\right) \quad \forall i \in [1, n] \quad (57)$$

Dessa maneira, $(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n)$ pode ser considerada uma amostra aleatória simples da população $N\left(\theta : \frac{\sigma_e^2}{n_n}\right)$ e, nesse cenário, as equações 58 e 59 descrevem estimadores não viesados de θ e $\frac{\sigma_e^2}{n_n}$, respectivamente.

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i \quad (58)$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 \quad (59)$$

Com base na equação 59, pode-se afirmar que $n_n s^2$ se trata de um estimador não viesado de σ_e^2 . Tendo em vista tais condições, pode-se afirmar que se a hipótese nula for verdadeira, a estatística F descrita pela equação 60 será aproximadamente uma unidade enquanto que, se forem encontradas evidências para rejeitar H_0 , ela adquirirá valores maiores que 1. Dessa forma, a estatística F permite testar $H_0 = \theta_1 = \theta_2 = \dots = \theta_n$. Ademais, observando-se as equações 58 e 59, pode-se aproximar a estatística F pela razão de dois qui-quadrados conforme expresso pela equação 60.

$$F = \frac{ms^2}{s_e^2} = \frac{QM_{Ent}}{QM_{Den}} : \frac{\frac{\chi^2(k-1)}{k-1}}{\frac{\chi^2(n-k)}{n-k}} \quad (60)$$

A estatística F corresponde, portanto, ao quociente de dois qui-quadrados seguindo, por conseguinte, uma distribuição F de Snedecor com graus de liberdade $n-1$ e $m-n$ (BUSSAB, 1988). Dessa forma pode-se afirmar por meio da tabela 1 as informações acerca tanto dos graus de liberdade quanto da estatística F são fornecidos.

2.4.5.2 Kruskal-Wallis

O teste não paramétrico de Wilcoxon permite verificar se duas populações podem ser consideradas semelhantes a partir de amostras independentes e é classificado como um teste não paramétrico. Esse teste baseia-se na classificação (posto) dos valores obtidos através da combinação das amostras de ambas populações para lidar com valores extremos que inviabilizam uma distribuição normal (OTT; LONGNECKER, 2015).

Considerando-se k o número de populações a serem comparadas e n_i a quantidade de observações da amostra da população $i \in [1, k]$, o método da soma dos postos das observações encontra-se descrito a seguir.

1. Ordena de forma crescente os dados resultantes da combinação das amostras populacionais;

2. Atribui valores de 1 à $n_T = \sum_{i=1}^k n_i$ a todas as observações sendo 1 associado ao menor valor e n_T ao maior;
3. Calcula T_i como sendo a soma das classes (postos) de todas as observações da população $i \in [1, k]$;

Com base na soma dos postos das observações, teste de Kruskal-Wallis adota a hipótese nula (H_0) de que todas as distribuições são idênticas. Conforme Ott e Longnecker (2015), a estatística utilizada nesse teste é descrita pela equação 61.

$$H = \frac{12}{n_T(n_T + 1)} \cdot \sum_{i=1}^k \frac{T_i^2}{n_i} - 3(n_T + 1) \quad (61)$$

2.5 Considerações Finais

Tendo em vista os objetivos da presente dissertação de mestrado, a seção de revisão bibliográfica apresentou os principais conceitos necessários ao entendimento dos modelos computacionais propostos bem como sua validação. A subseção 2.1 apresentou redes neurais artificiais de forma resumida bem como NARNETs como ferramentas propícias à modelagem de séries temporais que serão empregadas para modelar o comportamento de moluscos individuais com o intuito de considerar sua natureza adaptativa.

Ferramentas capazes de inferir a condição ambiente, como por exemplo a presença ou ausência de contaminantes, a partir da análise comportamental de bivalves são essenciais à construção dos modelos computacionais propostos. Dessa forma, a subseção 2.2 versou sobre os modelos de classificação e as métricas de avaliação que podem ser utilizadas na análise de sua *performance*.

A subseção 2.3 discorreu sobre ferramentas de visão computacional importantes à construção de uma técnica de valvometria totalmente não invasiva baseada em imagens e segmentação de cores. Por fim, a subseção 2.4 apresentou ferramentas de análise estatística empregadas não só na validação, mas também construção dos modelos computacionais propostos.

As ferramentas de regressão linear, correlação e função densidade de probabilidade são importantes na validação da técnica de valvometria proposta por meio de comparações com outra técnica de valvometria bem estabelecida. Gráficos de controle de qualidade permitem analisar a previsibilidade de séries temporais sendo utilizados, portanto, em conjunto com NARNETs na construção de biossensores de poluição aquática.

Métodos estatístico, como por exemplo ANOVA, são comumente empregados na construção de biossensores partir da identificação de diferenças significativas entre dados comportamentais de bivalves expostos e não expostos às condições ambientais de interesse. Dessa forma, estes métodos são empregados para determinar se os modelos

computacionais propostos apresentam vantagens para detecção de contraminantes apresentam vantagem em relação à estes métodos tradicionais, eles são empregados para fins de comparação na detecção de poluentes.