Classificadores bayesianos gaussianos aplicados em sinais de PPG para identificação de cardiomiopatias

Patrícia Tavares Leitão
Prog. de Pós-Graduação em Eng. de Teleinformática
Universidade Federal do Ceará
CE, Brasil
patricia.tavares@alu.ufc.br

Igor Rocha de Sousa
Prog. de Pós-Graduação em Eng. de Teleinformática
Universidade Federal do Ceará
CE, Brasil
igor.sousa@alu.ufc.br

Jermana Lopes de Moraes
Engenharia da Computação
Universidade Federal do Ceará
CE, Brasil
jermana@ufc.br

Auzuir Ripardo de Alexandria
Prog. de Pós-Graduação em Eng. de Telecomunicações
Inst. Fed. de Educação, Ciência e Tecnologia do Ceará
CE, Brasil
auzuir@ifce.edu.br

***Resumo*—Neste trabalho, é realizada a classificação das cardiomiopatias idiopáticas, chagásica e isquêmica utilizando os classificadores bayesanos gaussianos LDA (*Linear Discriminant Analysis*), QDA (*Quadratic Discriminant Analysis*) e *Naive* Bayes. São aplicadas técnicas de aprimoramento no banco de dados a fim de torná-lo mais adequado para os classificadores utilizados, como a transformação de Box-Cox. É realizado um aumento artificial de dados nas classes com pouquíssimas amostras a fim de melhorar a qualidade dos classificadores. O classificador LDA obteve o melhor resultado, com 96,5%, 100% e 92,5% na classificação entre pessoas saudáveis e as cardiomiopatias idiopáticas, chagásica e isquêmica, respectivamente. É realizada uma comparação destes resultados com os encontrados na literatura, em que o classificador LDA, de baixa complexidade, obteve resultados similares aos alcançados por classificadores não lineares como MLP, SOM, K-means e KNN.**

***Index Terms*—Cardiomiopatias, reconhecimento de padrões, classificadores bayesianos guassianos.**

I. Introdução

As doenças cardiovasculares são as principais causas de morte no mundo. Por ano, morrem cerca de 17,9 milhões de pessoas. Essas doenças são distúrbios do coração e dos vasos sanguíneos, incluindo doenças cardíacas, cerebrovasculares e outras condições [1].

Essa tendência global se observa também no Brasil, em que as doenças cardiovasculares são a primeira causa de óbito entre os brasileiros, apresentando 27% no total de mortes. Esses óbitos decorrem principalmente de doenças coronárias, insuficiência cardíaca e acidente vascular cerebral [2]. Com destaque para as doenças cardíacas, essas doenças impõem limitações a qualidade de vida relacionadas a aspectos físicos, sociais, financeiros e de saúde [3]. Porém, apesar da elevada mortalidade, ela é altamente prevenível [4].

O monitoramento cardíaco, realizado em ambientes clínicos, é fundamental para proporcionar um diagnóstico precoce dessas doenças [5]. Esse monitoramento pode ser feito pelo método de análise de alterações nas atividades elétricas do coração, o eletrocardiograma (ECG), ou pelo método de análise de fluxo sanguíneo, a fotopletismografia (*photoplethysmography* - PPG) [6].

O PPG é uma técnica que detecta a variação da absorção óptica da pele causada pela mudança do volume sanguíneo durante o ciclo cardíaco [7]. Em ambientes clínicos, a aquisição do PPG é realizada em dispositivos de oximetria de pulso, utilizados para medir a saturação de oxigênio [6]. Os dados obtidos através do PPG possibilitam a extração de parâmetros fisiológicos tais como a frequência cardíaca (FC) e a variabilidade da frequência cardíaca (VFC), que podem ser indicativos de desordens em razão de inconstâncias da pressão sanguínea.

A aplicação de técnicas de inteligência computacional é de grande importância na categorização de doenças cardiovasculares [8]. Essas técnicas tem sido cada vez mais empregadas para auxílio ao diagnóstico médico, detectando com boa taxa de acerto, a ocorrência ou não dessas doenças [9].

Como exemplo de trabalhos sobre classificação de doenças cardiovasculares, pode-se citar Bemando *et al.* [10], que utilizam o classificador *Naive* Bayes (NB), Li *et al.* [11] que utilizaram o classificador *Linear Discriminant Analysis* (LDA) e Shariatnia *et al.* [12] que utilizaram o classificador *Quadratic Discriminant Analysis* (QDA).

O NB, assim como o LDA e o QDA, são classificadores bayesanos que, assumindo a hipótese gaussiana sobre a distribuição dos dados, são chamamos de classificadores bayesianos gaussianos. Possuem fácil implementação e alcançam altas taxas de acurácia com baixo esforço computacional devido a sua simplicidade [13]. São considerados classificadores ótimos para dados com distribuição gaussiana, podendo superar métodos de classificação mais sofisticados [14].

Neste trabalho, é realizada a classificação de cardiomiopatias utilizando classificadores bayesianos gaussianos, tais como o LDA, o QDA e o NB. O banco de dados utilizado

O presente trabalho foi realizado com apoio do CNPq (# 305359/2021-5) e da FUNCAP (BMD-0008-00514.01.22/23).

foi coletado e tratado pelo grupo de pesquisa o qual este trabalho faz parte no Hospital Dr. Carlos Studart Gomes, também conhecido por Hospital do Coração de Messejana, em Fortaleza, Ceará, Brasil, e é descrito em [15].

A organização deste trabalho segue a seguinte ordem. Na Seção II, são apresentadas as cardiomiopatias presentes no banco de dados, assim como métodos de extração de características. Na Seção III, são apresentados os classificadores bayesianos gaussianos. Já na Seção IV, são apresentadas técnicas de aprimoramento do banco de dados. A metodologia é apresentada na Seção V, enquanto que os resultados e conclusões estão presentes nas Seções VI e VII, respectivamente.

II. Cardiomiopatias

Cardiomiopatias são um conjunto de doenças cardíacas que consistem em alterações miocárdicas decorrentes de diversas causas [16]. As principais miocardiopatias são: hipertrófica, dilatada idiopática, chagásica e displasia arritmogênica do ventrículo direito [17]. O banco de dados utilizado neste trabalho possui três tipos de cardiomiopatias dilatadas: idiopática, chagásica e isquêmica.

A. Cardiomiopatia idiopática

A cardiomiopatia dilatada idiopática é um distúrbio crônico que apresenta dilatação nas camadas cardíacas, fazendo com que a musculatura fique atrofiada ou hipertrofiada. Essa condição resulta na insuficiência cardíaca, tendo como alguns dos principais sinais clínicos a dificuldade respiratória, edema pulmonar, aumento da silhueta cardíaca no exame radiográfico e perda de peso. Dentro das cardiomiopatias, a dilatada é a mais comum e é chamada de idiopática quando uma determinada causa não é identificada [18].

B. Cardiomiopatia chagásica

A doença de Chagas é uma antropozoonose, transmitida pelo protozoário *Trypanosoma cruzi* [19]. Essa doença apresenta duas fases de evolução clínica: aguda e crônica. Na fase aguda pode apresentar sintomas como: febre, anorexia e taquicardia. Apresenta duração de 4 a 8 semanas, porém se não tratada, pode evoluir para a cronificação da enfermidade [20]. A fase crônica está associada à miocardite fibrosante focal de baixa intensidade e incessante que surge por meio da infecção causada pelo *Trypanosoma cruzi* e se apresenta de forma indeterminada e determinada [19]. A forma indeterminada não possui acometimento clínico ou sintomas, enquanto que na forma determinada, apresenta sintomas cardíacos, digestivos, ou cardiodigestivos [21].

C. Cardiomiopatia isquêmica

O estreitamento das artérias coronárias, causado pela formação de placas de ateroma, dificultam a passagem de sangue e reduzem seu fluxo para o coração, causando isquemia miocárdica e consequentemente a insuficiência cardíaca [22]. Contudo, a obstrução coronária é apenas um elemento de um complexo processo fisiopatológico que leva à isquemia. Outras condições como hipertensão arterial sistêmica, diabetes melito, consumo de cigarro, obesidade e hiperlipidemia são fatores de risco para essa cardiomiopatia [23]. Seu tratamento é feito através de medicamentos que reduzem os batimentos cardíacos, controlam os níveis da pressão arterial, dilatam os vasos do coração e diminuem as placas de gordura e a formação de coágulos sanguíneos [22].

D. Extração de características

As cardiomiopatias podem ser diagnosticadas por meio da análise de algumas características cardíacas identificadas pela aquisição do sinal PPG, que detecta a variação da absorção óptica da pele, causada pela mudança do volume sanguíneo durante o ciclo cardíaco [7]. A saída do sensor PPG é um sinal elétrico de uma dimensão, do qual alguns parâmetros podem ser extraídos, tais como:

• Média das durações dos batimentos cardíacos, em um intervalo de tempo:

$$M_dNN = \frac{1}{Q_{NN}} \sum_{k\in[T_i,T_f]} NN[k]\,, \tag{1}$$

em que T_i e T_f representam o tempo inicial e o tempo final, Q_{NN} representa a quantidade de amostras de intervalos NN no intervalo de $[T_i,T_f]$.

• Desvio-padrão de todos os intervalos de tempo entre os batimentos, em um intervalo de tempo:

$$SDNN = \sqrt{\frac{1}{Q_{NN}} \sum_{k\in[T_i,T_f]} (NN[k] - M_dNN)^2} \tag{2}$$

• Desvio-padrão das médias dos intervalos de tempo entre os batimentos, a cada 5 minutos em um intervalo de tempo:

$$SDANN = \sqrt{\frac{1}{M} \sum_{j=1}^{M} (M_{d_j} - M_dNN_M)^2}\,, \tag{3}$$

em que M representa a quantidade de segmentos de 5 minutos, M_{d_j} é a média do segmento de 5 minutos e M_dNN_M é a média de todos os segmentos de 5 minutos.

• Média do desvio-padrão dos intervalos de tempo entre os batimentos, a cada 5 minutos:

$$SDNN_{index} = \frac{1}{M} \sum_{j=1}^{M} SDNN_j \tag{4}$$

• Raiz quadrada da média do quadrado das diferenças entre intervalos de tempo entre batimentos consecutivos, em um intervalo de tempo:

$$RMSSD = \sqrt{\frac{1}{Q_{NN}-1} \sum_{k\in[T_i,T_f]} (NN_d[k])^2} \tag{5}$$

• Desvio-padrão de todos os intervalos de tempo entre os batimentos consecutivos, em um intervalo de tempo:

$$SDSD = \sqrt{\frac{1}{Q_{NN}-1} \sum_{k\in[T_i,T_f]} (NN_d[k] - M_dNN_d)^2} \tag{6}$$

• Variabilidade da frequência cardíaca (VFC): A VFC pode ser calculada através de índices geométricos, desde que haja a eliminação dos batimentos ectópicos e a retirada de ruídos do sinal PPG adquiridos [15].

III. Classificadores Bayesianos Gaussianos

Classificadores bayesianos são classificadores estatísticos que utilizam probabilidade para classificar uma amostra pertencente a uma determinada classe. Para classificar uma amostra $\mathbf{x} \in \mathbb{R}^p$ em uma classe c_i, em que p é o número de atributos, é considerada a maior probabilidade a *posteriori* $P(c_i/\mathbf{x})$, dentre todas as possíveis classes, ou seja, a amostra $\mathbf{x}$ pertence à classe c_i sempre que

$$P(c_i/\mathbf{x}) > P(c_j/\mathbf{x}) \tag{7}$$

é satisfeita [24]. Tais classificadores seguem os princípios do teorema de Bayes que é descrito como

$$P(c_i/\mathbf{x}) = \frac{P(\mathbf{x}/c_i)P(c_i)}{P(\mathbf{x})}\,, \tag{8}$$

em que $P(\mathbf{x}/c_i)$ é a função densidade de probabilidade condicional, $P(c_i)$ é a probabilidade a *priori* e $P(\mathbf{x})$ é a probabilidade do elemento [25]. Os classificadores bayesianos gaussianos consideram que $P(\mathbf{x}/c)$ é gaussiana, ou seja,

$$P(c_i/\mathbf{x}) = \frac{P(c_i)}{P(\mathbf{x})}\left\{\frac{1}{|\mathbf{\Sigma}_i|^{\frac{1}{2}}(2\pi)^{\frac{D}{2}}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\mathbf{m}_i)\right]\right\}, \tag{9}$$

em que $\mathbf{m}_i \in \mathbb{R}^p$ é o vetor média e $\mathbf{\Sigma}_i \in \mathbb{S}^p$ é a matriz de covariância da classe i. A partir de (9), são desenvolvidos os classificadores utilizados neste trabalho.

A. Quadratic Discriminant Analysis (QDA)

Este é o classificador bayesiano gaussiano mais abrangente, pois considera que cada classe possui sua própria matriz de covariância $\mathbf{\Sigma}_i$. Desenvolvendo (9), chega-se ao discriminante quadrático em $\mathbf{x}$,

$$\begin{aligned} g_i(\mathbf{x}) = &\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_i^{-1}\mathbf{x} + \mathbf{m}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{x} + \\ &+ \left[\ln(P(c_i)) - \frac{1}{2}\ln(|\mathbf{\Sigma}_i|) - \frac{1}{2}\mathbf{m}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{m}_i\right]\,. \end{aligned} \tag{10}$$

Em alguns casos, as funções lineares podem não criar a melhor separação das classes e o uso de funções discriminantes quadráticas pode ser mais apropriado [12].

B. Linear Discriminant Analysis (LDA)

Neste classificador, assume-se que a matriz de covariância $\mathbf{\Sigma}_i$ é a mesma para todas as classes, ou seja, $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \ldots, \mathbf{\Sigma}_n = \mathbf{\Sigma}$. Aplicando essas suposições em (10), chega-se ao discriminante linear em $\mathbf{x}$,

$$g_i(\mathbf{x}) = \mathbf{m}_i^T\mathbf{\Sigma}^{-1}\mathbf{x} + \left[\ln(P(c_i)) - \frac{1}{2}\mathbf{m}_i^T\mathbf{\Sigma}^{-1}\mathbf{m}_i\right]\,. \tag{11}$$

Em situações específicas, o classificador LDA pode apresentar resultados superiores aos do QDA devido à questões relacionadas à matriz de convariância $\mathbf{\Sigma}_i$, como em [12].

C. Naive Bayes (NB)

Neste clasificador, além de assumir que a matriz de covariância é a mesma para todas as classes como no LDA, assume-se que todos os atributos são estatisticamente independentes [26]. Assim, todas as classes possuem a mesma matriz de covariância $\mathbf{\Sigma}_{NB}$ e o discriminante linear em $\mathbf{x}$ é definido como

$$g_i(\mathbf{x}) = \mathbf{m}_i^T\mathbf{\Sigma}_{NB}^{-1}\mathbf{x} + \left[\ln(P(c_i)) - \frac{1}{2}\mathbf{m}_i^T\mathbf{\Sigma}_{NB}^{-1}\mathbf{m}_i\right]\,. \tag{12}$$

Este é um importante classificador, sendo extensivamente utilizado em diagnóstico de doenças cardíacas devido sua simplicidade e não necessidade de métodos complexos para inversão de matrizes [10].

IV. Aprimoramento do banco de dados

Entre as diversas técnicas existentes para aprimoramento de banco de dados, pode-se citar o aumento artificial de dados, a transformação de Box-Cox e a regularização de Tikhonov.

A. Aumento artificial de dados

O aumento de dados é um método usado para ampliar o número de amostras sem a necessidade de coletar novos dados. Isto pode ser útil quando se possui poucas amostras, o que pode infuenciar negativamente na construção de um bom classificador. Uma das formas de aumento de dados artificais é criando cópias de algumas amostras originais e adicionando um ruído gaussiano de baixa variância. Entretanto, esse dados aumentados não devem ser muito diferentes dos dados originais, pois podem distorcer o conjunto de dados [27].

B. Transformação de Box-Cox

A transformação de Box-Cox [28] é uma operação que faz com que variáveis não gaussianas tenham uma distribuição com uma forma mais simétrica, mais próxima da normal. Este tipo de operação pode ser útil, visto que normalidade é uma importante suposição para muitas técnicas e métodos estatísticos, como os classificadores bayesianos gaussianos, e muitas variáveis não seguem uma distribuição normal. A transformação é dada como

$$x^*(\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & ,\lambda \neq 0 \\ \ln(x) & ,\lambda = 0, \end{cases} \tag{13}$$

em que x^* representa o valor transformado, x representa o valor a ser transformado e λ é o parâmetro da transformação de Box-Cox [29].

C. Regularização de Tikhonov

O desempenho dos classificadores apresentados depende da invertibilidade da matriz de covariância $\mathbf{\Sigma}_i$, que está intimamente associada ao seu posto ser completo. Quando se tem poucos dados, $\mathbf{\Sigma}_i$ pode apresentar problemas de inversão. Uma maneira de tornar a matriz de covariância inversível, com posto completo, é aplicar a regularização de Tikhonov sobre a matriz de covariância $\mathbf{\Sigma}_i$, como

$$\mathbf{\Sigma}_i(\alpha) = \mathbf{\Sigma}_i + \alpha\mathbf{I}_p\,, \quad 0 \leq \alpha \ll 1\,, \tag{14}$$

em que α é o coeficiente de regularização. O efeito prático da regularização de Tikhonov é adicionar um pequeno valor α aos elementos da diagonal principal de $\mathbf{\Sigma}_i$, tornando-a diagonalmente dominante. Consequentemente, suas linhas tornam-se linearmente independentes e seu posto completo. A regularização em (14) equivale à adição de um ruído branco gaussiano aos dados originais, em que o parâmetro α pode ser interpretado como a variância deste ruído branco [30].

V. Metodologia

A sequência de passos utilizada na metodologia deste trabalho é apresentada na Fig. 1. Primeiramente, é realizado o aumento de dados das classes que contenham poucas amostras. Em seguida, os dados passam pela transformação de Box-Cox, na tentativa de tornar suas funções de densidade de probabilidade mais próximas da normal. O conjunto de dados resultante são então utilizados para treinamento e teste dos classificadores bayesianos gaussianos.

A. Aumento de dados

O banco de dados, criado pelo grupo de pesquisa deste trabalho e apresentado em [15], foi construído a partir da aquisição do sinal PPG em 35 pacientes cardiopatas do Hospital do Coração de Messejana em Fortaleza, Ceará e em dez pessoas saudáveis. O banco de dados possui quatro classes e sete atributos: $MdNN$, $SDNN$, $SDANN$, $SDNN_{index}$, $RMSSD$, $SDSD$, VFC, descritos na Subseção II-D. Os dados são divididos em:

- Classe 0: Miocardiopatia Idiopática - 29 amostras;
- Classe 1: Miocardiopatia Chagásica - 3 amostras;
- Classe 2: Miocardiopatia Isquêmica - 3 amostras;
- Classe 3: Pessoas Saudáveis - 10 amostras.

As classes de miocardiopatia chagásica e de miocardiopatia isquêmica possuem apenas três amostras, o que dificulta a construção de um bom classificador. Assim, decidiu-se realizar o aumento do número de amostras destas classes a fim de melhorar a qualidade dos classificadores.

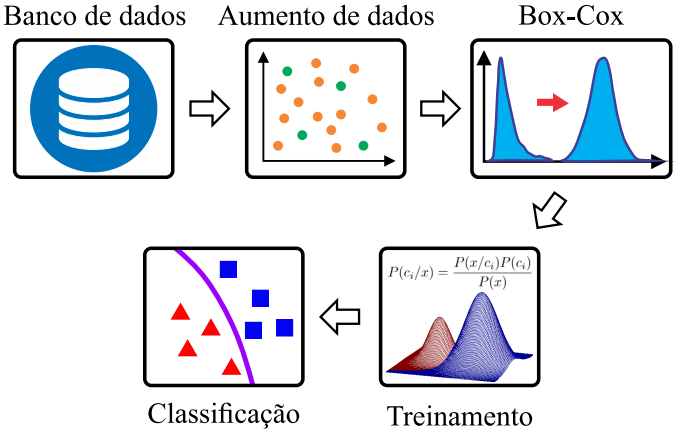


Fig. 1: Passos utilizados para a classificação.

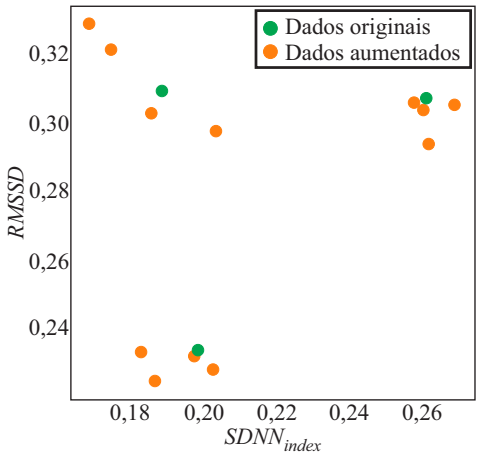
Para a realização do aumento das amostras, é adicionado um ruído branco com variância de 10^{-4} em cópias das amostras originais, gerando assim novas amostras. Dessa forma, as classes de miocardiopatia chagásica e miocardiopatia isquêmica, que antes possuíam apenas três amostras, agora possuem 15, como pode ser observado na Fig. 2.

Para que os efeitos da transformação de Box-Cox nos dados sejam avaliados, é realizado um treinamento dos classificadores bayesianos gaussianos com o banco de dados aumentado antes da transformação de Box-Cox. Para isso, os banco de dados aumentado é normalizado entre -1 e 1.

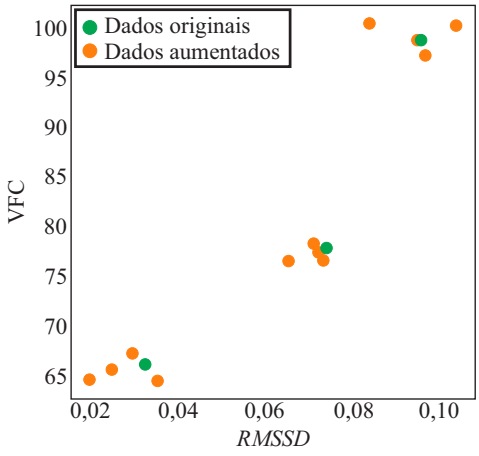
B. Aplicação da transformação de Box-Cox

Posteriormente o passo do aumento de dados, é aplicada a transformação de Box-Cox para fazer com que os dados aumentados fiquem mais próximos da distribuição normal, ou seja, mais próximo da Gaussiana. Isto porque os classificadores utilizados são bayesianos gaussianos, ou seja, assume-se que a função densidade de probabilidade é gaussiana. Para isso, os dados foram normalizado entre 0,1 e 1.

Os valores de λ utilizados na transformação dos atributos $MdNN$, $SDNN$, $SDANN$, $SDNN_{index}$, $RMSSD$, $SDSD$ e



(a) Miocardiopatia Chagásica.



(b) Miocardiopatia Isquêmica.

Fig. 2: Amostras aumentadas.

VFC são 0,8508, 0,1247, -0,6107, 0,1189, 0,1221, 0,2320 e 0,2808, respectivamente. Estes valores foram encontrados através da busca pela maximização do logaritmo natural da verossimilhança.

A comparação das distribuições de dois atributos do banco de dados aumentado pode ser observada na Fig. 3. As distribuições destes atributos após a aplicação da técnica são visivelmente não gaussianas. Porém, é importante ressaltar que para esses dados aplicados à transformação de Box-Cox, essas são as distribuições que mais se aproximam da gaussiana.

C. Treinamento dos classificadores

São avaliados os classificadores QDA, LDA e NB. Os dados são divididos em quatro grupos:

- Grupo 1: Miocardiopatia Idiopática + Pessoas Saudáveis;
- Grupo 2: Miocardiopatia Chagásica + Pessoas Saudáveis;
- Grupo 3: Miocardiopatia Isquêmica + Pessoas Saudáveis;
- Grupo 4: Cardiomiopatias + Pessoas Saudáveis.

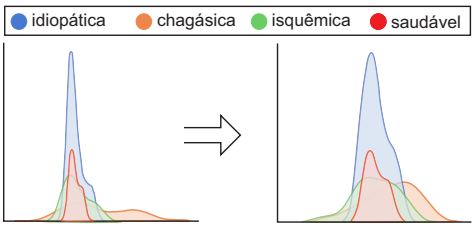
Além disso, os dados são separados de forma aleatória em 70% dos dados para treinamento dos classificadores e 30% para teste. No total, são executadas 100 realizações deste procedimento, em que são avaliados a acurácia média, seu desvio-padrão, a sensibilidade e precisão média,

$$\text{acurácia} = \frac{VP + VN}{N} \tag{15}$$

$$\text{precisão} = \frac{VP}{VP + FP} \tag{16}$$

$$\text{sensibilidade} = \frac{VP}{VP + FN}, \tag{17}$$

em que VP denota a quantidade de verdadeiros positivos, VN os verdadeiros negativos, FP os falso positivos, FN os falso negativos e N a quantidade total de dados.



(a) atributo *SDANN*.

(b) atributo VFC.

Fig. 3: Comparação das distribuições antes e depois da aplicação da transformada de Box-Cox.

VI. Resultados

Os resultados das 100 realizações são apresentados na Fig. 4, em que são ilustrados os *boxplots* das acurácias dos classificadores nos dados de teste. Observa-se que a classificação com a transformação de Box-Cox nos dados (Fig. 4b) obteve melhor desempenho para todos os classificadores e em todos os grupos em relação à classificação sem a transformação de Box-Cox nos dados (Fig. 4a), com melhores acurácias e menores dispersões.

Os valores de acurácia, seu desvio-padrão (d.p.), precisão e sensibilidade das classificações do banco de dados aumentado com transformação de Box-Cox são exibidos nas Tabelas I, II e III. Avaliando estes resultados juntamente com a Fig. 4b, constata-se que o classificador LDA obteve o melhor desempenho em todos as métricas e em todos os grupos.

O classificador *Naive* Bayes obteve um bom resultado, atingindo 100% de acerto no grupo 2 e 81,42% no grupo 4. Porém, de forma geral, seu desempenho é inferior ao do LDA. Pode-se supor que as covariâncias entre os atributos não são desprezíveis, como assume o classificador NB.

O classificador QDA apresentou problemas relacionados à inversão das matrizes de covariância das classes durante o treinamento. Por este motivo, e apenas na utilização deste classificador, decidiu-se adicionar um ruído gaussiano de baixa variância ao banco de dados aumentado com transformação de Box-Cox, o que equivale à regularização de Tikhonov nas matrizes de covariância (vide Seção IV-C). Após testes prelimirares, decidiu-se usar um ruído com variância $\sigma^2 = 10^{-5}$, possibilitando prosseguir com a utilização deste classificador.

Mesmo com a regularização das matrizes de covariância, o classificador QDA obteve o pior resultado entre os classi-

TABELA I: Resultados do Classificador LDA por grupos.

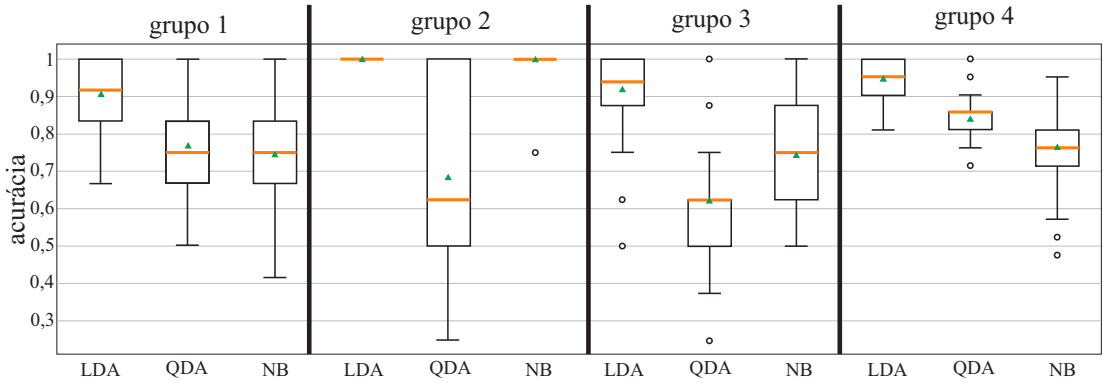
Grupo	*acurácia e d.p.*	*precisão*	*sensibilidade*
1	$96{,}50 \pm 4\%$	$96{,}62\%$	$94{,}43\%$
2	$\mathbf{100 \pm 0\%}$	$\mathbf{100\%}$	$\mathbf{100\%}$
3	$92{,}50 \pm 11\%$	$93{,}00\%$	$94{,}00\%$
4	$96{,}95 \pm 3\%$	$93{,}77\%$	$93{,}00\%$

TABELA II: Resultados do Classificador QDA por grupos.

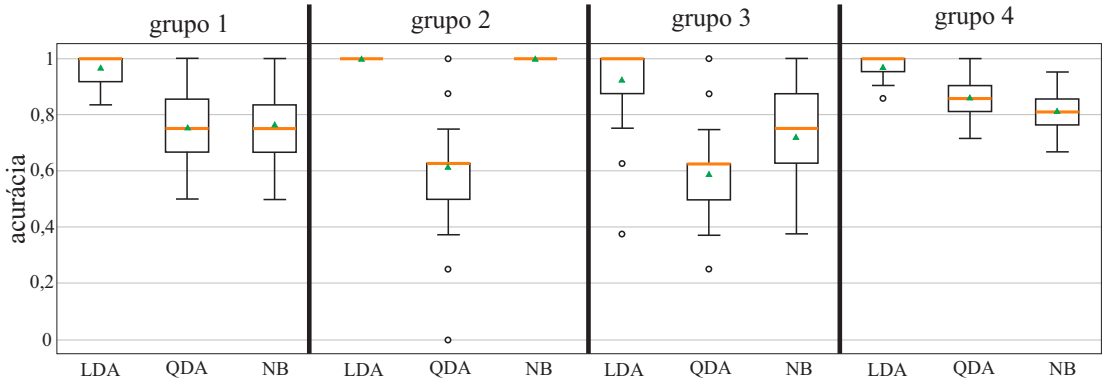
Grupo	*acurácia e d.p.*	*precisão*	*sensibilidade*
1	$75{,}41 \pm 13\%$	$\mathbf{48{,}98\%}$	$58{,}65\%$
2	$61{,}50 \pm 21\%$	$41{,}31\%$	$\mathbf{59{,}00\%}$
3	$58{,}87 \pm 17\%$	$37{,}91\%$	$56{,}98\%$
4	$\mathbf{85{,}90 \pm 7\%}$	$48{,}44\%$	$54{,}75\%$

TABELA III: Resultados do Classificador NB por grupos.

Grupo	*acurácia e d.p.*	*precisão*	*sensibilidade*
1	$76{,}50 \pm 11\%$	$70{,}25\%$	$73{,}36\%$
2	$\mathbf{100 \pm 0\%}$	$\mathbf{100\%}$	$\mathbf{100\%}$
3	$72{,}00 \pm 16\%$	$73{,}45\%$	$72{,}24\%$
4	$81{,}42 \pm 7\%$	$67{,}93\%$	$76{,}65\%$



(a) Classificação do banco de dados aumentado.



(b) Classificação do banco de dados aumentado com a transformação de *Box-Cox*.

Fig. 4: *Boxplots* das acurácias dos classificadores para dados de teste.

ficadores utilizados, apresentando baixos valores de acurácia, precisão, sensibilidade e altos valores de desvio-padrão. Obteve acurácia similar ao classificador NB no grupo 1 e melhor no grupo 4.

É interessante notar que para os grupos 1, 2 e 3, o classificador QDA apresentou baixos índices de acurácia. Porém, para o grupo 4, que agrupa todas as cardiomiopatias em uma mesma classe, seu resultado é superior ao do classificador NB. De fato, esta foi a melhor acurácia do classificador QDA entre todos grupos. Pode-se supor que, como o grupo 4 possui mais dados na classe "cardiomiopatias", sua matriz de covariância possui melhor qualidade. Consequentemente, o classificador obtém um melhor desempenho.

Este problema não ocorre com o classificador LDA, que usa todos as amostras de um grupo para estimar a matriz de covariância única. De fato, o LDA utiliza 39 amostras para estimar a matriz de covariância única do grupo 1, 25 para a do grupo 2, 25 para a do grupo 3 e 69 para a do grupo 4. Já o QDA utiliza 29 amostras para estimar a matriz de covariância de uma classe e 10 para a da outra classe do grupo 1, 15 e 10 amostras para as duas classes do grupo 2, 15 e 10 amostras para as duas classes do grupo 3 e 59 e 10 amostras para as duas classes do grupo 4.

Pode-se supor que o LDA obteve melhores resultados que o QDA devido a baixa quantidade de amostras das classes. Em uma futura expansão do banco de dados realizada por este grupo de pesquisa, aconselha-se a aplicação do classificador QDA para avaliação de seu desempenho com uma quantidade maior de dados.

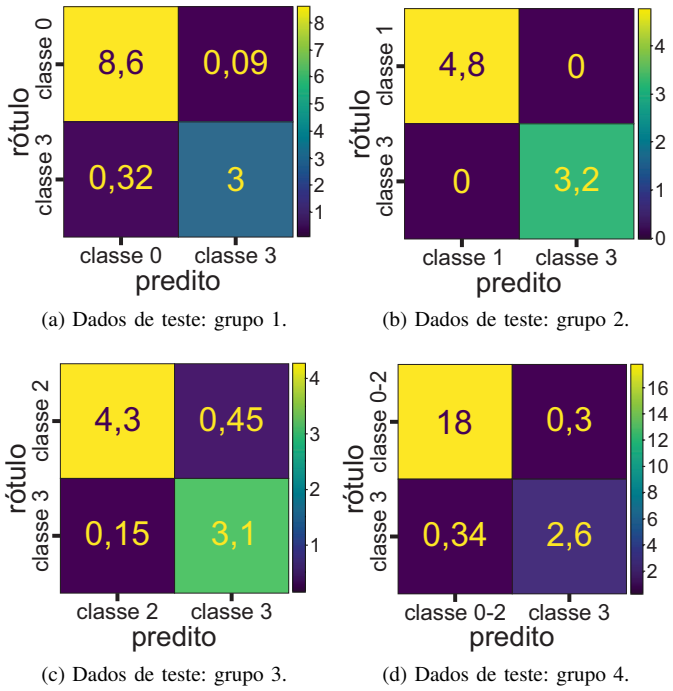
As matrizes de confusão média do classificador LDA para os dados de teste são ilustradas na Fig. 5. Observa-se que de maneira geral, existem poucos falsos positivos e falsos negativos, com destaque para o grupo 2, que em que não há predições errôneas.

Classificações de cardiomiopatias com este banco de dados, sem as técincas de aprimoramento de dados utilizadas neste trabalho, foram realizadas em [15]. Foram utilizados classificadores não lineares como a rede neural MLP (*multilayer perceptron*), rede SOM (*self-organizing map*), *K-means* e KNN (*k nearest neighbor*). Comparações entre os resultados obtidos em [15] e os alcançados neste trabalho são exibida nas Tabelas IV e V.

Observa-se que o classificador LDA utilizado neste trabalho obteve resultados similares aos classificadores não lineares apresentados em [15]. No grupo 1, o LDA obteve acurácia média superior aos classificadores MLP, *K-means* e KNN, com $96{,}5\%$. Já no grupo 2, obteve o mesmo valor que os classificadores MLP, SOM e *Kmeans*, com 100%. No grupo 3, obteve acurácia superior aos classificadores SOM e KNN, e apenas $0{,}83\%$ abaixo da MLP. Por fim, no grupo 4, obteve resultado melhor que todos os classificadores não lineares utilizados em [15], com $96{,}95\%$.

TABELA IV: Comparativo da acurácia média e desvio-padrão.

Banco de Dados	Classificador	*Grupo 1*	*Grupo 2*	*Grupo 3*	*Grupo 4*
apresentado em [15]	MLP	88,57 ± 0,05%	**100 ± 0%**	93,33 ± 0,08%	90,00 ± 0,04%
	SOM	**100 ± 0,08%**	**100 ± 0%**	87,50 ± 0,21%	87,50 ± 0,07%
	K-means	92,85 ± 0,34%	**100 ± 0%**	**100 ± 0%**	91,85 ± 0,34%
	KNN	92,60 ± 0,03%	100 ± 0,06%	90,00 ± 0,10%	**93,00 ± 0,05%**
Dados aumentados com transformação de Box-Cox	LDA	**96,50 ± 4%**	**100 ± 0%**	**92,50 ± 11%**	**96,95 ± 3%**
	QDA	75,41 ± 13%	61,50 ± 21%	58,87 ± 17%	85,90 ± 7%
	NB	76,50 ± 11%	**100 ± 0%**	72,00 ± 16%	81,42 ± 7%



(a) Dados de teste: grupo 1. (b) Dados de teste: grupo 2.

(c) Dados de teste: grupo 3. (d) Dados de teste: grupo 4.

Fig. 5: Matrizes de confusão média do classificador LDA.

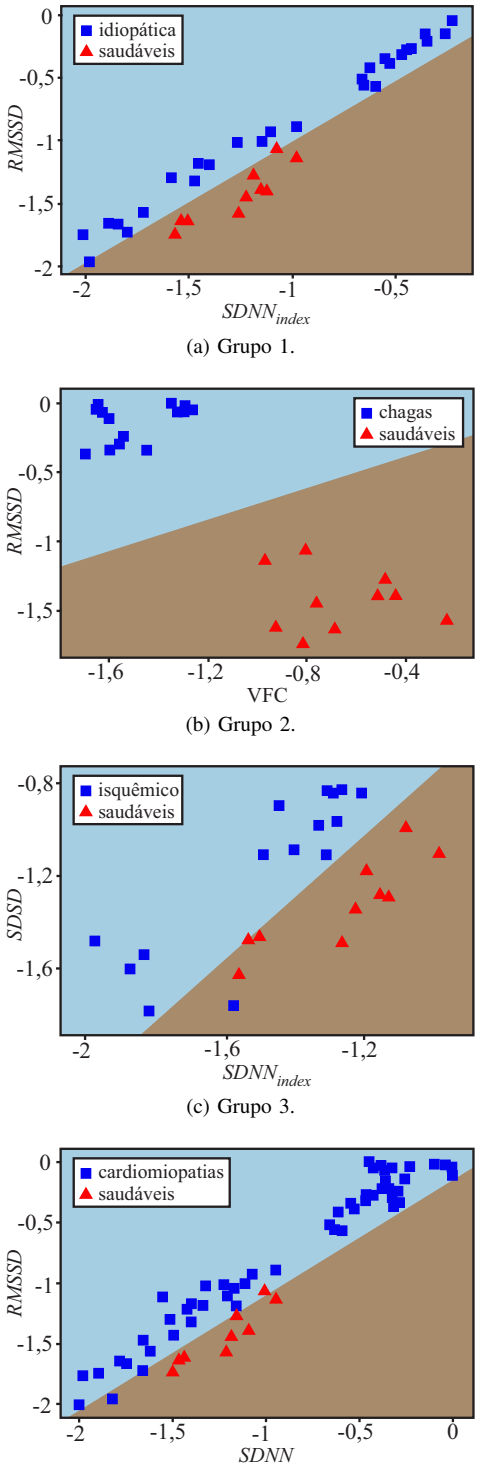
TABELA V: Comparativo da sensibilidade.

Dados	Classif.	*Grupo 1*	*Grupo 2*	*Grupo 3*	*Grupo 4*
apresentado em [15]	MLP	87,5%	**100%**	96,9%	88,5%
	SOM	**100%**	**100%**	90,0%	90,0%
	K-means	90,6%	**100%**	**100%**	**92,8%**
	KNN	89,5%	**100%**	90,0%	89,5%
dados aum. com transf. de Box-Cox	LDA	**94,4%**	**100%**	**94,0%**	**93,0%**
	QDA	58,6%	59,0%	56,9%	54,7%
	NB	73,36%	**100%**	72,24%	76,65%

A sensibilidade do LDA segue os mesmos padrões que a acurácia em relação aos classificadores não lineares. Na Fig. 6, são apresentadas algumas representações visuais das classificações utilizando o LDA. Nota-se que a classificação do grupo 2 é um problema linearmente separável razoavelmente simples, conforme já demosntrado na Tabela I e Fig. 4b, não necessitando da utilização de classificadores mais sofisticados.

VII. Conclusão

Este trabalho abordou o problema de classificação das cardiomiopatias dilatadas idiopática, chagásica e isquêmica utilizando classificadores bayesanos gaussianos: LDA, QDA e *Naive* Bayes. Foram realizadas melhorias no banco de dados



(a) Grupo 1.

(b) Grupo 2.

(c) Grupo 3.

(d) Grupo 4.

Fig. 6: Representações visuais das classificações.

original (coletado e tratado pelo grupo de pesquisa deste trabalho) voltadas para este tipo de classificador, como o aumento de dados artificiais em classes com pouquíssimas amostras, a aplicação da transformação de Box-Cox e a adição de ruído branco nos dados, o que equivale à regularização de Tikhonov das matrizes de covariância na classificação com o QDA.

A aplicação da transformação de *Box-cox* mostrou-se satisfatória, conforme ilustrado na Fig. 4. O classificador LDA obteve melhor desempenho entre os utilizados, com $96{,}5\%$ de acurácia média no grupo 1 (idiopática $\times$ saudáveis), 100% no grupo 2 (chagásica $\times$ saudáveis), $92{,}5\%$ no grupo 3 (isquêmica $\times$ saudáveis) e $96{,}95\%$ no grupo 4 (cardiomiopatias $\times$ saudáveis). O classificador QDA apresentou baixo desempenho devido ao pequeno número de amostras por classe e não deve ser descartado em futuros trabalhos de classificação, quando este grupo de pesquisa expandir a quantidade de amostras no banco de dados.

A classificação do grupo 2 mostrou-se como um problema linearmente separável, não havendo a necessidade de classificadores mais complexos. O classificador LDA obteve, no geral, resultados similares aos alcançados por classificadores não lineares em [15], como a MLP, SOM, *K-means* e KNN.

Com estes resultados, constata-se que o classificador LDA é uma poderosa ferramenta para classificação de dados de cardiomiopatias, principalmente em cenários onde se deseja utilizar um classificador em sistemas embarcados de menor poder computacional, comum em dispositivos *wearable*.

Referências

[1] WHO, "Cardiovascular disease," *world health organization*, 2022. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] L. R. Costa, E. V. Passos, and O. M. Silvestre, "O redescobrimento do Brasil cardiovascular: Como prevenimos e tratamos a doença cardiovascular em nosso país," pp. 117–118, 2021.

[3] B. Stevens, L. Pezzullo, L. Verdian, J. Tomlinson, A. George, and F. Bacal, "Os custos das doenças cardíacas no Brasil," *Arquivos Brasileiros de Cardiologia*, vol. 111, pp. 29–36, 2018.

[4] I. M. Bensenor, "Prevalencia de fatores de risco cardiovascular no mundo e no Brasil," *Rev. Soc. Cardiol. Estado de São Paulo*, pp. 18–24, 2019.

[5] R. A. Pinto *et al.*, "Rede neural convolucional u-net para inferência do sinal eletrocardiograma a partir do sinal fotopletismograma," Master's thesis, Universidade Federal do Amazonas, 2022.

[6] H. S. Oliveira, "Estimativa dos pontos de sistole e diástole para identificação de hipertensão a partir de sinais de fotopletismografia," Master's thesis, Universidade Federal do Amazonas, 2022.

[7] P. H. d. B. Souza, "Método para estimação da frequência cardíaca e variabilidade cardíaca com base em fotopletismografia por vídeo," Master's thesis, Universidade de Brasilia, 2019.

[8] J. B. Azzi *et al.*, "Utilização de técnicas de inteligência computacional na caracterização de pacientes com doenças cardiovasculares," 2018.

[9] V. Tragante, "seleção de atributos por meio de algotirimos genéticos para diagnóstico auxiliado por computador em cardiopatia isquêmica," Master's thesis, Universidade de São Paulo, 2006.

[10] C. Bemando, E. Miranda, and M. Aryuni, "Machine-learning-based prediction models of coronary heart disease using naïve Bayes and random forest algorithms," in *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*. IEEE, 2021, pp. 232–237.

[11] B. Li, H. Ding, Z. Wang, Z. Liu, X. Cai, and H. Yang, "Research on the difference between patients with coronary heart disease and healthy controls by surface enhanced Raman spectroscopy," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 272, p. 120997, 2022.

[12] S. Shariatnia, M. Ziaratban, A. Rajabi, A. Salehi, K. Abdi Zarrini, and M. Vakili, "Modeling the diagnosis of coronary artery disease by discriminant analysis and logistic regression: a cross-sectional study," *BMC medical informatics and decision making*, vol. 22, no. 1, p. 85, 2022.

[13] S. Kharya, S. Agrawal, and S. Soni, "Naive Bayes classifiers: a probabilistic detection model for breast cancer," *Int. J. Comput. Appl*, vol. 92, no. 10, pp. 26–31, 2014.

[14] S. S. Bafjaish, "Comparative analysis of naive bayesian techniques in health-related for classification task," *Journal of Soft Computing and Data Mining*, vol. 1, no. 2, pp. 1–10, 2020.

[15] J. L. de Moraes, "Desenvolvimento de um sistema para estratificação de cardiopatias utilizando a fotopletismografia (PPG)," Master's thesis, Instituto Federal de educaçãa, ciência e tecnologia do ceara, 2017.

[16] L. C. P. Ribeiro, R. M. S. Felippe, M. R. de Sousa, E. T. M. da Fonseca, V. da Silva Baptista, and E. C. de Siqueira, "Uma análise sobre as cardiomiopatias: hipertrófica e dilatada," *Revista Eletrônica Acervo Saúde*, vol. 15, no. 8, pp. e10 740–e10 740, 2022.

[17] E. C. L. Santos, F. C. R. Figurinha, A. G. S. Lima, B. B. Henares, and F. Mastrocola, *Manual de cardiologia cardiopapers*. Editora Atheneu, 2015.

[18] L. M. M. Azevedo, "Cardiomiopatia dilatada idiopática," Master's thesis, FMUP Faculdade de Medicina da Universidade do Porto, 2014.

[19] J. P. de Sousa Neto, M. d. S. R. Mariano, and E. de Andrade Aoyama, "Principais alterações cardiovasculares decorrentes da doença de Chagas com ênfase à cardiopatia chagásica," *Revista Brasileira Interdisciplinar de Saúde*, 2020.

[20] C. M. B. Tameirão, L. J. d. Miranda, M. E. F. Gomes, M. G. E. d'AssumpÇão, and M. H. G. Junior, "A doença de Chagas e a cardiopatia chagásica croônica: revisão de literatura," *Brazilian Journal of Development*, 2021.

[21] MS, "Protocolo clínico e diretrizes terapêuticas para doença de Chagas: relatório de recomendação," *Ministerio da Saude*, 2023. [Online]. Available: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/doenca-de-chagas/protocolo-clinico-e-diretrizes-terapeuticas-para-doenca-de-chagas-_-relatorio-de-recomendacao.pdf/view

[22] R. M. de Oliveira Araújo and R. O. S. Junior, "Os exercícios físicos na prevenção e tratamento da cardiopatia isquêmica," *Research, Society and Development*, vol. 12, no. 2, 2023.

[23] L. Wannmacher and A. F. Costa, "Uso racional de estatinas na prevenção de cardiopatia isquêmica," *Brasília: Ministério da Saúde*, 2010.

[24] B. S. Santana, G. F. Gaiardo, L. P. Lucca, M. M. Donato, T. N. Tavares, V. F. Garcia, and Maria, "Utilização de classificadores bayesianos para predição de afinidade entre personagens literários," in *XIV Simposio de Informatica*, Santa Maria, Brasil, 2017, pp. 1–7.

[25] R. P. d. V. Sousa, "Análise dos componentes principais supervisionada: uma abordagem não-paramétrica," B.S. thesis, Brasil, 2019.

[26] H. Zhang, "The optimality of naive Bayes," *Proceedings of the 17th International FLAIRS conference*, p. 6, 2004.

[27] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 968–988. [Online]. Available: https://aclanthology.org/2021.findings-acl.84

[28] "Uma análise das transformações," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26.

[29] G. F. Bueno, L. S. Costa, and E. A. Costa, "Transformação Box-Cox e modelagem dendrométrica de árvores isoladas no bioma cerrado em minas gerais," *Caderno de Ciências Agrárias*, vol. 13, pp. 1–9, 2021.

[30] C. M. Bishop, "Training with noise is equivalent to tikhonov regularization," *Neural computation*, vol. 7, no. 1, pp. 108–116, 1995.