JHI JOURNAL OF HEALTH INFORMATICS

CBIS'20 7 a 11 de Dezembro Foz do Iguaçú - Paraná Evento Virtual

Método de Aprendizagem de Máquina para Classificação da intensidade do desvio vocal utilizando *Random Forest*

Machine Learning Method for Classifying Vocal Deviation Intensity Using Random Forest

Método de aprendizaje automático para clasificar la intensidad de la desviación vocal mediante Random Forest

Danilo Rangel Arruda Leite[1], Ronei Marcos de Moraes[1], Leonardo Wanderley Lopes[1]

Descritores: Aprendizagem de máquina; Distúrbios da voz; Espectrografia do som

RESUMO

Objetivo: Utilizar imagens espectrográficas da voz para classificar a intensidade do desvio vocal, avaliar e comparar a eficiência do modelo de classificação *Random Forest* (RF) com o *Naive Bayes* (NB) e *Support Vector Machine* (SVM). **Método**: Foram selecionadas, aleatoriamente, 198 amostras de indivíduos com desvio vocal classificados com intensidade entre leve e moderada. A vogal /ɛ/ sustentada foi selecionada para este estudo, pois é a vogal mais comumente utilizada para avaliação da qualidade vocal na realidade brasileira. **Resultado**: O RF obteve o melhor resultado, com acurácia de 78% e *Kappa* 0,41. Os resultados deste trabalho foram considerados satisfatórios. **Conclusão**: O modelo de classificação RF obteve resultados satisfatórios. Foi utilizado o *Short-Time Fourier Transform* para gerar os espectrogramas do sinal de voz. As intensidades do desvio vocal utilizadas nesse trabalho, foram as do tipo leve e moderado. A metotodologia de classificação utilizada mostrou-se relevante para o processo de classificação da intensidade do desvio.

Keywords: Machine learning; Voice disorders; Sound spectrography

ABSTRACT

Objective: Using spectrogram images of the voice signal to classify the intensity of the vocal deviation in pathologically voices, to evaluate and to compare the efficiency of the Random Forest (RF) classification model with the Naive Bayes (NB) and Support Vector Machine (SVM). **Method**: 198 samples were chosen randomly according to voice disorder classified by intensity from soft to moderate. A sustained vowel /ɛ/ was specified for this study, as it is a vowel most commonly used for the assessment of vocal quality in Brazilian reality. **Result**: Random Forest obtained the best result, with an accuracy of 78% and Kappa 0.41. The results in this work were considered satisfactory. **Conclusion**: the RF classification model obtained satisfactory results. Short-Time Fourier Transform was used to generate the spectrograms of the voice signal. Intensities of the vocal deviation used in this work, were mild and moderate.

Descriptores: Automático; Trastornos de la Voz; Espectrografia do som

RESUMEN

Objetivo: Utilizar imágenes de espectrograma de la señal de voz para clasificar la intensidad de la desviación vocal en voces patológicamente, y evaluar y comparar la eficiencia del modelo de clasificación Random Forest (RF) con Naive Bayes (NB) y Support Vector Machine (SVM). **Método**: 198 amuestras fueron elegidas aleatoriamente de acuerdo con el trastorno de la voz clasificado con intensidad entre ligero y moderado. Se especificó una vocal sostenida /ɛ / para este estudio, por ser una vocal más comúnmente utilizada para la evaluación de la calidad vocal en la realidad brasileña. **Resultado**: Random Forest obtuvo el mejor resultado, con una precisión del 78% y Kappa 0.41. Los resultados de este trabajo se consideraron satisfactorios. **Conclusión**: el modelo de clasificación de RF obtuvo resultados satisfactorios. Se utilizó Short-Time Fourier Transform para generar los espectrogramas de la señal de voz. Las intensidades de la desviación vocal utilizadas en este trabajo fueron leves y moderadas.

[1] *Departamento de Estatística. Programa de Pós-graduação em Modelos de Decisão em Saúde. Universidade Federal da Paraíba - UFPB, João Pessoa (PB) Brasil.*

Autor Coorespondente: **Danilo Rangel Arruda Leite**
e-mail: **danilorangel@buscapb.com.br**

INTRODUÇÃO

A voz é um dos principais meios de comunicação do ser humano; é um fenômeno que envolve grande variações anatomofuncionais e comportamentais e a sua emissão deve ser agradável, sem esforços e conforme aos interesses profissionais, sociais e pessoais do interlocutor. Qualquer alteração na sua emissão pode ser classificada como distúrbio de voz ou desvio vocal[1].

A presença de desvio na voz pode causar mudanças significativas nos padrões vibratórios, afetando assim, a qualidade da produção normal da voz, podendo influenciar negativamente na qualidade de vida de um indivíduo, limitando a comunicação no seu trabalho, entre outros[1].

Nesse sentido, alterações vocais devem ser diagnosticadas e tratadas o mais precocemente possível. A análise da voz possibilita alcançar resultados que mostram a condição de um desvio vocal com mais eficiência[1-2]. Nesse cenário, a análise acústica é um procedimento não invasivo que utiliza técnicas de processamento digital de sinal de voz, podendo contribuir com medidas acústicas para a construção de ferramentas de classificação do desvio de voz, bem como sua intensidade[1,3].

A análise acústica pode compreender a análise descritiva de padrões visuais, como o espectrograma de faixa larga e faixa estreita, o diagrama de desvio fonatório e o espectro de longo termo. Dentre as possibilidades de análise acústica, a espectrografia é um recurso de grande relevância, a partir dele podem ser visualizadas informações como presença de ruído em média e altas frequências, intensidade, instabilidade dos harmônicos, quebras de sonoridade, entre outras[1,4].

Nesse contexto, as imagens espectrográficas da voz podem proporcionar a construção de ferramentas para classificar o desvio vocal e sua intensidade, de forma não invasiva e com menor custo. Em virtude das possibilidades de informações fornecidas através de espectrogramas, gerou-se interesse em sistemas de classificação que possam auxiliar os especialistas em voz, utilizando imagens espectrográficas para identificar os tipos de sinais de voz e classificar a intensidade do desvio vocal[4].

Sendo assim, esse artigo tem como objetivo utilizar os recursos discriminativos dos sinais de voz para classificar a intensidade do desvio vocal, utilizando imagens espectrograficas, além de avaliar e comparar a eficiência do modelo de classificação RF com o NB e SVM.

Análise espectrográfica da voz

Um espectrograma é uma ferramenta que pode proporcionar a análise de diversos atributos e características do sinal da voz. Ele pode ser formado a partir do cálculo da transformada de *Fourier* de curta duração (*Short-Time Fourier Transform* – STFT) aplicado ao sinal digital do som, conforme Equação 1.

$$X[n,\lambda] = \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-j\lambda m} \quad (1)$$

onde $w[n]$ é uma sequência de janela que determina a parte do sinal a ser analisado. $X[n, \lambda]$ é uma função de duas variáveis, o índice de tempo discreto n e a frequência contínua λ, $x[n=m]$ deslocada no tempo, conforme visto através da janela $w[m]$ e X representa a energia do sinal, em uma função de n e de λ .

O eixo no STFT mostra tempo e frequência, e a escala de cores da imagem do espectrograma representa a amplitude da frequência (figura 1). A base para a representação STFT é conhecida como uma soma de uma série de sinusóides (senos e cossenos)[5].

O espectrograma é um dos principais recursos utilizados na análise acústica e pode ser definido como um gráfico que mostra a intensidade por meio do escurecimento ou coloração do traçado, as faixas de frequência no eixo vertical e o tempo no eixo horizontal. Sua representação mostra estrias horizontais, denominadas harmônicos. O espectrograma demonstra visualmente as características acústicas da emissão, contudo, essas informações exigem interpretação por parte do avaliador[1,6].

Espectrogramas são mais utilizados para classificar sinais de voz de diferentes tipos, como proposto em Yanagihara[7] que classifica os sinais nos tipos I, II, III e IV e Titze[8] que classifica os sinais nos tipos I, II e III. A classificação do sinal de voz em três tipos, de acordo com seu grau de periodicidade, é uma tarefa conhecida como digitação de sinal, ou seja, é uma etapa relevante de pré-processamento antes de calcular qualquer medida de perturbação[1].

De acordo com Titze[8] e Sprecher et al[9], a classificação dos tipos de sinais, são:

- **Sinal Tipo I:** quase periódico; sem modulações ou sub-harmônicos; série rica de harmônicos definidos até 4 KHz; regularidade no traçado; ausência ou presença de ruído de baixa amplitude, em relação à estrutura harmônica;
- **Sinal Tipo II:** oscilam entre quase periódicos e aperiódicos; com presença de modulações e bifurcações; presença de sub-harmônicos claramente definidos, intermitentes (com duração d" 1s) e com intensidade próxima à intensidade de f0;
- **Sinal Tipo III**: aperiódicos, caóticos e com dimensão

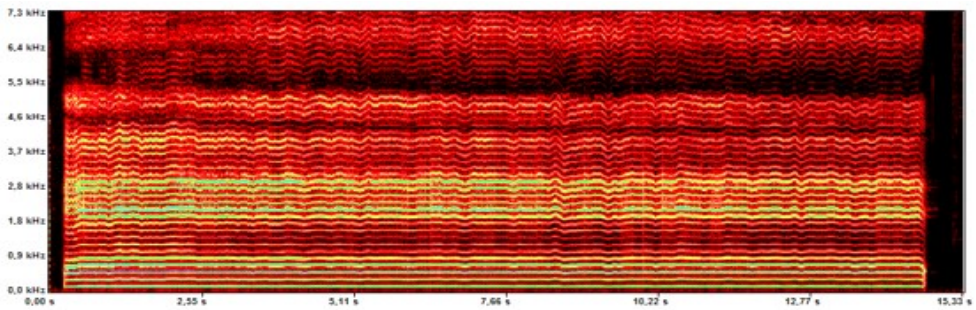

Figura 1 – Espectrograma extraído de um paciente com distúrbio de voz

finita; presença de energia de ruído entre os harmônicos em baixas frequências; maior definição nos primeiros harmônicos de baixa frequência, em detrimento da substituição dos harmônicos de alta frequência por ruído difuso; concentração de energia abaixo de 1,5 KHz;

· **Sinal Tipo IV:** aperiódicos e caóticos; com dimensão infinita e predomínio de ruído em relação à estrutura harmônica em toda a faixa de frequência.

A classificação realizada por Yanagihara[7], é mais utilizada no contexto clínico para caracterização do desvio observado no sinal e sua relação com a intensidade do desvio vocal, no plano perceptivo-auditivo, classifica os sinais em tipo I, II, III e IV, utilizando como critérios: a regularidade dos harmônicos, a presença de ruído em diferentes faixas de frequência e a relação entre a estrutura harmônica e o ruído presente no traçado.

A classificação proposta por Titze[8], é mais utilizada em procedimentos de pesquisa para determinação do tipo de análise a ser realizada (linear vs. não linear, análise de padrão visual vs. extração de medidas). É baseada no modelo de dinâmica não linear da produção vocal e caracteriza a mudança quantitativa no comportamento do sinal, advinda de mudança no padrão vibratório das pregas vocais.

Dessa forma, observa-se que, na literatura científica, existe a necessidade de construção de ferramentas computacionais que utilizem modelos de aprendizado de máquina para classificar a intensidade do desvio vocal através de imagens espectrográficas da voz[10-11]. A construção dessa ferramanta poderá auxiliar o clínico nos procedimentos de avaliação e monitoramento dos desvios da voz, assim como contribuir para o treinamento de novos profissionais, sejam eles acadêmicos ou profissionais.

Diante o exposto, o presente estudo analisou o RF e comparou seus resultados com os classificadores NB e SVM, na classificação da intensidade do desvio vocal em pacientes com disturbio de voz[1,12]. Para este estudo, foram utilizados os tipos II (leve) e III (moderado)[10].

Naive Bayes (NB)

A NB é um classificador probabilístico que possui uma abordagem de aprendizado de máquina supervisionado. Baseado no teorema de Bayes, é utilizado para determinar os resultados da classificação e assume que não há dependência entre as variáveis de um sistema. Mesmo que essa hipótese seja irrealista em alguns casos, o classificador NB é capaz de fornecer resultados satisfatórios[13]. Uma das vantagens da abordagem do NB é sua capacidade de avaliar dados para os quais ele nunca foi treinado. É amplamente utilizado para mineração de dados, classificação de dados e imagens, tomada de decisões, entre outros[14].

O classificador de Naive Bayes é comumente utilizado para solucionar tarefas de classificação em que S = $[s_1,...,s_p]$denota um vetor de características ou atributos e $w = \{w_1, w_2,...,w_k\}$refere-se a um conjunto com k problemas da voz ou classes. Esse classificador assume: (i) independência dos atributos dada a classe; (ii) não ocorrência de atributos escondidos ou latentes. Por (i) temos que:

$$P(s_1,\ldots,s_p) \approx \prod_{k=1}^{p} P(S_k|w_i) \quad (2)$$

Conceitualmente, a regra de Bayes fornece um mecanismo capaz de obter as probabilidades condicionais, $P(w_i/S_k)$, a partir de evidências fornecidas pelos atributos dada a classe, $P(S_k/w_i)$. O método NB é então expresso como:

$$P(w_i|S) = P(w_i) * (P(s_1|w_i) *\ldots* P(s_p|w_i)) = P(w_i) \prod_{k=1}^{p} P(S_k|w_i) \quad (3)$$

onde $P(w_i)$ é a probabilidade da classe do desvio da voz w_i $P(S_k/w_i)$, é a probabilidade de observar o padrão S dado que pertence à classe w_i. O NB assume distribuição multinomial, mas pode ser utilizado para variáveis contínuas, após discretização das informações. No entanto, esse procedimento pode levar à perda de informações[13-14]. A regra de decisão atribui s a w_i se $P(w_i/s) > P(w_j/s)$. Para todo j ε (1,...,k) com j diferente de i. Ou seja, a classe com maior valor é a mais provável.

Support Vector Machines (SVM)

O SVM é um algoritimo de aprendizado de máquina supervisionado que analisa os dados e reconhece padrões e pode ser aplicado para classificação e regressão. Tem boa capacidade de generalização, fundamenta-se no conceito de planos de decisão que definem os limites de decisão, construindo hiperplanos em um espaço multidimensional. Os hiperplanos separam os dados da amostra em diferentes categorias[15].

No espaço bidimensional, esse hiperplano é utilizado para classificar as amostras em duas categorias, tornando-o um classificador linear binário não probabilístico[15]. O hiperplano é construído de forma que aloca a maioria dos pontos da mesma categoria no mesmo lado do hiperplano, enquanto tenta maximizar a distância das amostras de dados de ambas as categorias a este hiperplano. O subconjunto de amostras de dados mais próximo do hiperplano de separação é denotado como os vetores de suporte.

Dado um conjunto separável de exemplos $S = \{(s_1, y_1),..., (s_n, y_n)\}$ de n elementos, onde s_i E R^p corresponde ao vetor de entrada de p-dimensão p e y_i a qual classe pertence y_i E{+1, -1}, i=1, ..., n, pode-se definir um hiperplano de separação como uma função linear, que é capaz de separar este conjunto sem erro[15].

$$S = \{(s_1, y_1, \ldots, s_n, y_n) \mid s_i \in \mathbb{R}^p, y_i \in \{+1, -1\}\}_{i=1}^{n} \quad (4)$$

O resultado fornecido pelo SVM para cada amostra do sinal de voz pode ser interpretado como a probabilidade da amostra pertencer a uma classe específica, considera um conjunto de dados de entrada e prevê para cada dado de entrada, a qual das possíveis classes o dado pertence[15].

Random Forest (RF)

O classificador RF gera uma série de árvores de decisão e votos nas classes classificadas por cada árvore para determinar a classe final. É um algoritmo típico de *Bootstrap Aggregation*, também conhecido como *Bagging*[16].

RF é um algoritimo de aprendizado de máquina supervisionado, utilizado para resolver problemas de classificação e regressão. Algoritmo do tipo *ensemble learning*, agrupa os resultados ou previsões de várias árvores de decisão[17], treinadas individualmente, na tentativa de produzir um melhor modelo preditivo para resolver o mesmo problema, diminuindo a variância e o viés. Tem a vantagem de ser eficiente para grandes conjuntos de dados[15].

O algoritmo usa as folhas, ou decisões finais, de cada nó para chegar a uma conclusão própria. Isso aumenta a precisão do modelo, uma vez que está observando os resultados de muitas árvores de decisão diferentes e encontrando uma média.

Para utilizar o RF a fim de resolver problemas de classificação, o RF utiliza o coeficiente de *Gini* para determinar qual das ramificações é mais provável de ocorrer, medindo o grau de heterogeneidade dos dados. Logo, pode ser utilizado para medir a impureza de um nó. Este índice num determinado nó é dado por:

$$Gini(S) = 1 - \sum_{i=1}^{c} (p_i)^2 \qquad (5)$$

onde o representa a frequência relativa da classe que está sendo observada no conjunto de dados e c representa o número de classe.

A entropia nos diz quão impuro ou não homogêneo é o conjunto de dados. Por exemplo, a entropia é zero quando o conjunto de dados for completamente homogêneo, ou um quando for igualmente dividida[18]. Dado um conjunto de dados (S) que pode ter *c* classes distintas, a entropia de S será dada por:

$$Entropia(S) = \sum_{i=1}^{c} - p_i \, log_2(p_i) \qquad (6)$$

Outra característica relacionada à entropia é o ganho de informação. O ganho de informação é baseado na redução da entropia depois que um conjunto de dados é dividido em um atributo. Construir uma árvore de decisão envolve encontrar o atributo que retorna o maior ganho de informação, ou seja, os ramos mais homogêneos.

Trabalhos correlatos

Alguns estudos descrevem a utilização de classificadores de aprendizado de máquina no desenvolvimento de modelos de detecção e classificação do desvio vocal. Em 2016, utilizou-se o classificador SVM com a técnica de *mel-frequency cepstral coefficients* (MFCC), obtendo-se um resultado de 95% de precisão[12].

Em 2020, Chen et al[16] proporam um RF difusa para reconhecimento de emoções de fala, os resultados experimentais mostraram que as precisões de reconhecimento da proposta são 87,34% RF, maiores do que as da rede neural de retropropagação que obteve 74,50%. Nos últimos anos, RF vem sendo utilizado para reconhecimento de linguagem natural, como descrito em[19].

Wu et al[20] proporam um sistema para detecção de voz disfônica utilizando a CNN como arquitetura básica, alcançando uma precisão geral de 71%. Em 2018, os métodos SVN, RF, *K-Nearest Neighbor* (K-NN), *Gradient Boosting* e *Ensemble Learning* (EL) com uma amostra de 50 vozes normais e 150 com distúrbios de voz[21]. Os resultados mostraram que o EL obteve maior pontuação de precisão, com 68,48%, seguido por *Gradient Boosting* com 67,35%. Segundo o estudo, os desvios padrão de todos os métodos de classificação foram razoavelmente baixos.

Trinh e O'Brien[22] propuseram uma estrutura da CNN e utilizando o *Keras*[23] para testar a abordagem, obtiveram uma precisão de classificação acima de 95%. Nesse trabalho, os espectrogramas são extraídos usando a biblioteca *Librosa*[24], a imagem gerada é a entrada da CNN para classificação utilizando o banco de dados *The Saarbrucken Voice Database*.

MÉTODOS

O conjunto de dados para análise é composto por 198 amostras selecionadas aleatoriamente de indivíduos com desvio, classificados com intensidade de desvio vocal entre leve (118 amostras) e moderada (78 amostras), proveniente do Laboratório Integrado de Estudos da Voz (LIEV) de uma universidade pública.

Os indivíduos foram selecionados conforme os seguintes critérios de elegibilidade: presença de queixa vocal; ter realizado gravação da vogal /ε/ sustentada com duração mínima de seis segundos; apresentar laudo otorrinolaringológico referente ao exame visual laríngeo para confirmação diagnóstica de desvio de voz; não apresentar comprometimento cognitivo ou neurológico que impedisse a gravação da voz; não ter realizado terapia vocal ou tratamento cirúrgico na laringe previamente. Posteriormente, foi extraído o espectrograma utilizando a transformada de *Fourier* de curta duração, a partir de amostras de 3 segundos da vogal sustentada /ε/, com taxa de amostragem de 44.100 Hz.

A vogal /ε/ foi selecionada para este estudo, pois é uma vogal oral, aberta, não arredondada e é considerada a vogal com a posição mais média no Português Brasileiro, o que permite uma posição mais neutra e intermediária do trato vocal. Além disso, é a vogal mais comumente utilizada para avaliação da qualidade vocal na realidade brasileira[1].

O modelo criado inclui quatro diferentes etapas, como coleta de dados, extração das imagens espectrográficas, construção dos modelos e avaliação de desempenho. A metodologia proposta é baseada em um esquema em cascata em que, primeiro foram classificadas as vozes com e sem desvio vocal e, em seguida, foi identificada a intensidade do desvio.

As etapas que descrevem a construção do modelo proposto, são listadas a seguir: **i**. No pré-processamento, foram utilizados arquivos de voz com no mínimo 6 segundos, foram extraídos 3 segundos do meio do arquivo de áudio da vogal sustentada /ε/, utilizando uma taxa de amostragem de 44.100 Hz; **ii.** Foi utilizada a biblioteca *Librosa*[23] para carregar os dados do arquivo em um vetor e retornar uma matriz de áudio e taxa de

amostragem. Foram extraídos espectrogramas do sinal da voz através da STFT, com janela de 64 amostras; **iii**. Três diferentes modelos de aprendizado de máquina supervisionado foram investigadas e analisados seus resultados, são eles: RF, NB e SVM (tabela 1). O *dataset* foi dividido em setenta porcento para treinamento e o restante para teste, utilizando uma abordagem de validação cruzada de 4 vezes na partição de treinamento 1 para validar; **iv.** Foi analisado o desempenho dos modelos para verificar qual deles obteve o melhor resultado na classificação dos dados. Comparamos o desempenho do RF com os modelos NB e SVM. A decisão final de detecção é então obtida de acordo com a classe que fornece a maior probabilidade de acerto entre os 3 classificadores.

RESULTADOS E DISCUSSÃO

A análise acústica é uma técnica não invasiva que pode possibilitar a geração de dados quantitativos e compará-los entre si, e com valores normativos relacionados a diferentes condições laríngeas e a diferentes tipos de desvio da qualidade vocal. O espectrograma é uma das principais ferramentas utilizadas na análise acústica, pois o exame visual do traçado espectrográfico de banda estreita é um dos procedimentos mais utilizados clinicamente, dado que possibilita a avaliação qualitativa do sinal, independente do grau de periodicidade e ruído presente na emissão[4,9].

Dessa forma, a inspeção acústica do traçado espectrográfico pode fornecer dados qualitativos e quantitativos do sinal, integrando-os às informações perceptivo auditivas e laríngeas[4,11].

Sendo assim, nesta pesquisa foi investigado o algoritimo RF para identificar os tipos de sinais de voz e classificar a intensidade do desvio vocal utilizando imagens espectrográficas geradas a partir da STFT e comparados os resultados obtidos com o SVM e NB.

Para avaliar a precisão dos resultados obtidos através dos classificadores, foram utilizados três medidas normalmente utilizados no meio científico, são elas[11,25]: acurácia; coeficiente de *Kappa*[26]; área sobre a curva. Essas medidas estão relacionadas com os resultados de classificação e diagnóstico verdadeiro.

O teste é considerado positivo (desvio) ou negativo (saudável), e o desvio presente ou ausente. O teste está correto quando ele é positivo na presença do desvio (Verdadeiro Positivo-VP) ou negativo na ausência do desvio (Verdadeiros Negativo-VN). Além disso, o teste está errado quando ele é positivo na ausência do desvio (Falso Positivo-FP), ou negativo quando o desvio está presente (Falso Negativo-FN)

Os resultados do experimento realizado para detecção e classificação da intensidade do desvio vocal, demosntraram que, em termos de acurácia, o modelo RF obteve melhor resultado, com acurácia de 0,78 de valores de classificação correta, conforme apresentado na tabela 1, o coeficiente de *Kappa* e sua respectiva interpretação (grau de concordância), por sua vez, foi classificado com um grau de concordância moderada, de 0,41.

O *Kappa* é um método estatístico para avaliar o nível de concordância ou reprodutibilidade entre dois conjuntos de dados. O coeficiente *Kappa* é calculado por:

$$kappa = \frac{P(O) - P(E)}{1 - P(E)} \quad (7)$$

em que:

P(O): proporção observada de concordâncias (soma das respostas concordantes dividida pelo total);

P(E): proporção esperada de concordâncias (soma dos valores esperados das respostas concordantes dividida pelo total).

Acurácia (ACC) mede a capacidade do teste de identificar corretamente quando há e quando não há presença do desvio. É definida como a relação entre o número de casos corretamente classificados e todos os casos expostos ao classificador:

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \quad (8)$$

A área média sob a curva (AUC) foi de 0.70. Segundo Cohen[26], quando a AUC estiver entre 0.7 e 0.8, o modelo tem poder discriminatório aceitável. AUC representa o grau ou medida de separabilidade. Discrimina o quanto o modelo é capaz de distinguir entre classes:

- Taxa de positivo verdadeiro (TPR) é um sinônimo de *recall* e, portanto, é definido por:

$$TPR = \frac{VP}{VP + FN} \quad (9)$$

- Taxa de falso positivo (FPR) é definida por:

$$FPR = \frac{FP}{FP + VN} \quad (10)$$

A arquitetura do modelo de classificação RF utilizou a entropia como medida, com número de estimadores de 20. Número de estimadores indica a quantidade de árvores construídas pelo algoritmo antes de tomar uma votação ou fazer uma média de predições[18].

Os parâmetros utilizados para cada algorítimo são apresentados a seguir:

- SVM – *Kernel*= Polinomial, *Gamma*= 0.1, *degree*=3;
- RF – Estimadores= 20, critério = entropia;
- NB – configuração padrão.

Alguns estudos descreveram a utilização do RF, bem como SVM, KNN, CNN, entre outros, também com resultados considerados bons[11,19-20], evidenciando que RF é um modelo que pode ser utilizado para classificação do desvio de voz, bem como sua intensidade.

Tabela 1 - Acurácia entre os modelos

Modelo	Kappa	Acurácia	AUC
RF	0,41	0.78	0.70
NB	0,38	0.71	0.59
SVM	0,30	0.65	0.50

O classificador de RF tende a superar a maioria dos outros métodos de classificação em termos de precisão, evitando problemas de sobreajuste. É um algoritmo do tipo *ensemble learning*, que agrupa os resultados ou previsões de várias árvores de decisão[16], treinadas individualmente, na tentativa de produzir um melhor modelo preditivo para resolver o mesmo problema, diminuindo a variância e o viés.

Embora Hossain e Mohammad[12], Wu et al[20], e Pham, Lin e Zhang[21] não tenham utilizados os modelos para classificar a intensidade do desvio vocal, o presente estudo pode fornecer informações relevantes para trabalhos futuros.

CONCLUSÃO

Este estudo analisou os modelos de aprendizado de máquina RF, SVM e NB, para classificar a intensidade do desvio vocal, utilizando o STFT para gerar os espectrogramas do sinal de voz. As intensidades do desvio vocal utilizadas neste trabalho, foram as do tipo leve e o moderado. Foi observado que o modelo de classificação RF obteve resultados satisfatórios, ficando com a acurácia de 78% e com coeficiente de *Kappa*[25] com um grau de concordância moderado de 0,41.

Sendo assim, a contribuição apresentada neste artigo mostra que a metodologia de classificação escolhida é relevante para o processo de classificação da intensidade do desvio vocal. Contudo, destacamos a necessidade de avaliar e comparar outros classificadores, individualmente ou em conjunto, bem como testar outros parâmetros de configurações.

REFERÊNCIAS

1. Lopes LW, Alves JN, Evangelista DS, França FP, Vieira VJD, Lima-Silva MFB et al. Accuracy of traditional and formant acoustic measurements in the evaluation of vocal quality. CoDAS. 2018; 30(5): e20170282. doi: 10.1590/2317-1782/20182017282
2. Guan H, Lerch A. Learning Strategies for Voice Disorder Detection. IEEE 13th Int Conf Semant Comput (ICSC). 2019. pp. 295-301. doi: 10.1109/ICOSC.2019.8665504
3. Pishgar M, Karim F, Majumdar S, Darabi H. Pathological Voice Classification Using Mel-Cepstrum Vectors and Support Vector Machine. Electrical Eng Syst Sci. 2018 [cerca de 5 p.]. Disponível em: https://arxiv.org/abs/1812.07729
4. Lopes LW, Silva ACF, Silva IM, Paiva MAA, Silva SIN, Almeida LNA et al. Evidence of Internal Consistency in the Spectrographic Analysis Protocol. J Voice. 2020. doi: 10.1016/j.jvoice.2020.07.013
5. Sakashita Y, Aono M. Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions. Detection and Classification of Acoustic Scenes and Events 2018. 2018 [cerca de 5 p.]. Disponível em: http://dcase.community/documents/challenge2018/technical_reports/DCASE2018_Sakashita_15.pdf
6. Ali SM, Karule PT. Spectral Analysis of Pathological & Normal Speech Signal. Int J Sci Eng Res. 2013; 4(2): [cerca de 3 p.]. Disponível em: https://www.ijser.org/paper/Spectral-Analysis-of-Pathological-Normal-Speech-Signal.html
7. Yanagihara N. Significance of harmonic changes and noise components in hoarseness. J Speech Hear Res. 1967; 10(3): 431-541. doi: 10.1044/jshr.1003.531
8. Titze IR. Workshop on Acoustic Voice Analysis: Summary Statement. National Center for Voice and Speech; 1995.
9. Sprecher A, Olszewski A, Jiang JJ. Updating signal typing in voice: addition of type 4 signals. J Acoust Soc Am. 2010;127:3710–3716.
10. Lopes LW, Silva IM, Sousa ESS, Silva ACF, Paiva MAA, Diniz EGR et al. Classificação espectrográfica do sinal vocal: relação com o diagnóstico laríngeo e a análise perceptivo-auditiva. Audiol Commun Res. 2020; 25: e2194 [cerca de 9 p.]. doi: 10.1590/2317-6431-2019-2194
11. Miramont JM, Restrepo JF, Codino J, Jackson-Menaldi C, Schlotthauer G. Voice Signal Typing Using a Pattern Recognition Approach. J Voice. 2020. doi: 10.1016/j.jvoice.2020.03.006
12. Hossain MS, Mohammad G. Healthcare Big Data Voice Pathology Assessment Framework. IEEE Access. 2016; 4: 7806-7815. doi: 10.1109/ACCESS.2016.2626316
13. Moraes RM, Machado LS. Gaussian Naive Bayes for Online Training Assessment in Virtual Reality-Based Simulator. Mathware Soft Comput. 2009. pp. 123-132. doi: 10.1142/9789812799470_0188
14. Moraes RM, Silva ILA, Machado LS. Online Skills Assessment in Training Based on Virtual Reality Using a Novel Fuzzy Triangular Naive Bayes Network. WSPC. 2020. doi: 10.1142/9789811223334_0054
15. Olson DL, Delen D. Advanced Data Mining Techniques. Alemanha: Springer; 2008.
16. Chen L, Su W, Feng Y, Wu M, She J, Hirota K. Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. Inf Sci. 2020; 509: 150-163. doi: 10.1016/j.ins.2019.09.005
17. Moraes RM, Martínez L. Computational Intelligence Applications for Data Science. Knowl Based Syst. 2015; 87: 1-2. doi: 10.1016/j.knosys.2015.07.038
18. Breiman L. Random Forests. Mach Learn. 2001; 45: 5-32. doi: 10.1023/A:1010933404324
19. Iliou T, Anagnostopoulos CN. Comparison of different classifiers for emotion recognition. IEEE Access. 2009. pp. 102-106. doi: 10.1109/PCI.2009.7
20. Wu H, Soraghan J, Lowit A, Di-Caterina G. A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks. Interspeech. 2018. pp. 446-450. doi: 10.21437/Interspeech.2018-1351
21. Pham M, Lin J, Zhang Y. Diagnosing Voice Disorder with Machine Learning. IEEE International Conference on Big Data. 2018. pp. 5263-5266. doi: 10.1109/BigData.2018.8622250
22. Trinh NH, O'Brien D. Pathological speech classification using a convolutional neural network. Irish Mach Vis Image Process Tech. 2019; Agosto 28-30. doi: 10.21427/9dnc-n002
23. Keras. Keras Documentation. [Online].; 2019. Available from: https://keras.io/.
24. McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E et al. librosa: Audio and music signal analysis in python. Proc 14th Python Sci Conf. 2015. pp. 18–25. Disponível em: http://conference.scipy.org/proceedings/scipy2015/pdfs/brian_mcfee.pdf
25. Mythili J, Vijaya M. Pathology Voice Detection and Classification Using Ensemble Learning. Int J Eng Sci Inv (IJESI). 2018; 7(8): 1-8. Disponível em: http://www.ijesi.org/papers/Vol(7)i8/Version-1/A0708010108.pdf
26. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Meas. 1960; 20(1): 37-46. doi: 10.1177/001316446002000104