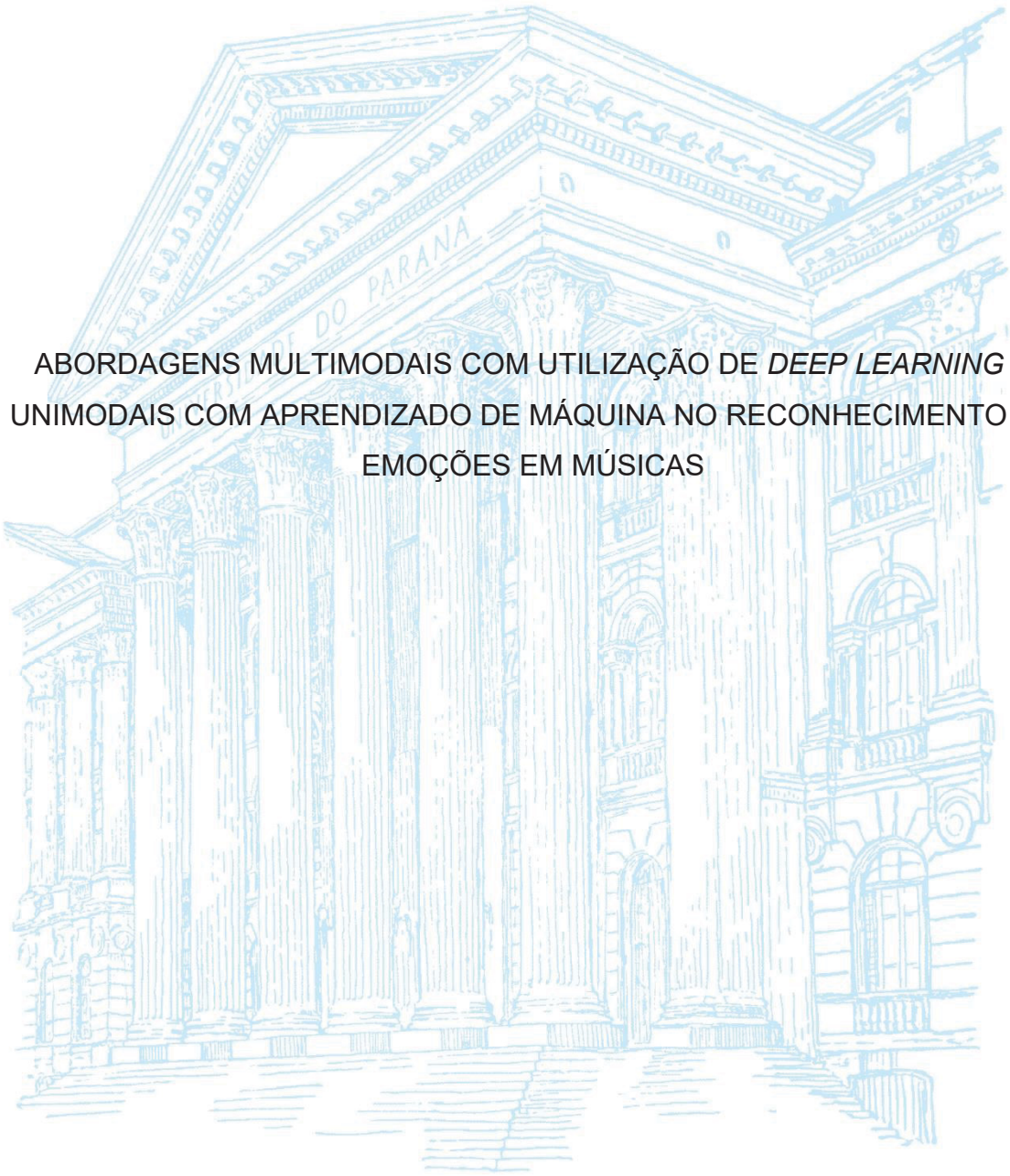
UNIVERSIDADE FEDERAL DO PARANÁ

PAULO SERGIO DA CONCEIÇÃO MOREIRA



ABORDAGENS MULTIMODAIS COM UTILIZAÇÃO DE *DEEP LEARNING* E UNIMODAIS COM APRENDIZADO DE MÁQUINA NO RECONHECIMENTO DE EMOÇÕES EM MÚSICAS

CURITIBA

2023

PAULO SERGIO DA CONCEIÇÃO MOREIRA

ABORDAGENS MULTIMODAIS COM UTILIZAÇÃO DE *DEEP LEARNING* E UNIMODAIS COM APRENDIZADO DE MÁQUINA NO RECONHECIMENTO DE EMOÇÕES EM MÚSICAS

Tese apresentada ao Programa de Pós-Graduação em Gestão da Informação (PPGGI), Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná (UFPR), como requisito parcial à obtenção do título de Doutor em Gestão da Informação.

Orientadora: Profª. Drª. Denise Fukumi Tsunoda
Coorientadora: Profª. Drª. Marília Nunes Silva

CURITIBA
2023

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS SOCIAIS APLICADAS

Moreira, Paulo Sergio da Conceição
Abordagens multimodais com utilização de *deep learning* e unimodais com aprendizado de máquina no reconhecimento de emoções em músicas / Paulo Sergio da Conceição Moreira. – Curitiba, 2023.
1 recurso on-line : PDF.

Tese (Doutorado) – Universidade Federal do Paraná, Setor de Ciências Sociais Aplicadas, Programa de Pós-Graduação em Gestão da Informação.
Orientadora: Profa. Dra. Denise Fukumi Tsunoda.
Coorientadora: Profa. Dra. Marília Nunes Silva.

1. Música. 2. Emoções. 3. Aprendizado do computador. 4. Abordagens multimodais. I. Tsunoda, Denise Fukumi. II. Silva, Marília Nunes. III. Universidade Federal do Paraná. Programa de Pós-Graduação em Gestão da Informação. IV. Título.

Bibliotecária: Maria Lidiane Herculano Graciosa CRB-9/2008

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO DA INFORMAÇÃO - 40001016058P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação GESTÃO DA INFORMAÇÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **PAULO SERGIO DA CONCEIÇÃO MOREIRA** intitulada: **ABORDAGENS MULTIMODAIS COM UTILIZAÇÃO DE DEEP LEARNING E UNIMODAIS COM APRENDIZADO DE MÁQUINA NO RECONHECIMENTO DE EMOÇÕES EM MÚSICAS**, sob orientação da Profa. Dra. DENISE FUKUMI TSUNODA, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 30 de Agosto de 2023.

Assinatura Eletrônica
05/09/2023 14:22:49.0
DENISE FUKUMI TSUNODA
Presidente da Banca Examinadora

Assinatura Eletrônica
05/09/2023 22:24:51.0
GLAUCIA MARIA BRESSAN
Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

Assinatura Eletrônica
06/09/2023 06:23:30.0
JOSÉ SIMÃO DE PAULA PINTO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
05/09/2023 16:58:30.0
MARILIA NUNES SILVA
Coorientador(a)

Assinatura Eletrônica
27/09/2023 02:48:16.0
CARLOS NASCIMENTO SILLA JUNIOR
Avaliador Externo (PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ)

Documento assinado eletronicamente de acordo com o disposto na legislação federal Decreto 8539 de 08 de outubro de 2015.
Gerado e autenticado pelo SIGA-UFPR, com a seguinte identificação única: 312562
Para autenticar este documento/assinatura, acesse https://siga.ufpr.br/siga/visitante/autenticacaoassinaturas.jsp e insira o codigo 312562

AGRADECIMENTOS

À Prof.ª Dr.ª Denise Fukumi Tsunoda, expresso minha profunda gratidão por sua orientação, paciência, confiança, generosidade e compreensão ao longo de toda a minha trajetória acadêmica.

Carinhosamente à Prof.ª Dr.ª Marília Nunes Silva, minha coorientadora, que, sem hesitar, aceitou o desafio de participar desta pesquisa, contribuindo de maneira imprescindível para a construção do estudo.

Aos familiares, amigos, colegas e conhecidos, especialmente à minha mãe, Sueli, ao meu irmão, Júlio, e à minha avó, Leonete.

À Thais, agradeço pelo companheirismo sem igual; pela presença nos momentos alegres e complicados; e pelo apoio e incentivo sempre que me vi desnorteado.

Aos membros da banca, representados pela Prof.ª Dr.ª Glaucia Maria Bressan, Prof.º Dr. Carlos Nascimento Silla Junior e Prof.º Dr. José Simão de Paula Pinto, agradeço por aceitarem o convite, pelas valiosas contribuições e estímulo.

À Simone, pela sempre prestativa ajuda.

Aos voluntários que gentilmente cederam parte do seu tempo participando dos experimentos de revisão e rotulação das músicas.

Ao corpo docente do Programa de Pós-Graduação em Gestão da Informação (PPGGI/UFPR).

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) pela bolsa concedida, que possibilitou a realização desta tese.

A todos que, de alguma forma, contribuíram para o desenvolvimento desta pesquisa.

RESUMO

Esta pesquisa foi realizada com base na compreensão da relevância da relação entre música e emoção na vida humana, abrangendo desde o lazer até estudos científicos. Embora a organização emocional da música seja intrínseca à natureza humana, o reconhecimento automático de emoções musicais enfrenta desafios, configurando-se como um tema complexo na recuperação de informações musicais. Nesse contexto, o propósito central desta tese foi investigar se a adoção de abordagens multimodais, envolvendo informações de diferentes fontes e arquiteturas de *deep learning*, pode superar o desempenho das abordagens unimodais baseadas em algoritmos de aprendizado de máquina. Essa indagação emergiu da carência de estratégias multimodais na área e da perspectiva de melhoria nos resultados de classificação reportados em pesquisas correlatas. Com cinco objetivos específicos, esta pesquisa abordou a identificação de um modelo cognitivo de emoções, definição de modalidades, construção de bases de dados multimodais, comparação de arquiteturas de *deep learning* e avaliação comparativa das abordagens multimodais com abordagens unimodais utilizando algoritmos tradicionais de aprendizado de máquina. A análise dos resultados demonstrou que as abordagens multimodais alcançaram desempenho superior em diversos cenários de classificação, comparadas às estratégias unimodais. Tais resultados contribuem positivamente para a compreensão da eficácia das abordagens multimodais e das arquiteturas de *deep learning* no reconhecimento de emoções em músicas. Adicionalmente, a pesquisa ressalta a necessidade de atenção aos modelos emocionais e metadados em plataformas *online*, visando evitar vieses e ruídos. Esta tese oferece contribuições relevantes na área de reconhecimento de emoções em músicas, particularmente no desenvolvimento de bases de dados multimodais, avaliação de arquiteturas de *deep learning* para problemas tabulares, protocolos de experimentos e abordagens voltadas à cognição musical. A comparação sistemática entre abordagens multimodais e unimodais evidencia as vantagens das primeiras, incentivando novas pesquisas nesse campo.

Palavras-chave: Música. Emoções. *Deep learning*. Aprendizado de máquina. Abordagens multimodais.

ABSTRACT

This research was conducted based on the understanding of the significance of the relationship between music and emotion in human life, spanning from leisure to scientific studies. Although the emotional organization of music is intrinsic to human nature, the automatic recognition of musical emotions faces challenges, manifesting as a complex theme in the retrieval of musical information. Within this context, the central purpose of this thesis was to investigate whether the adoption of multimodal approaches, involving information from different sources and deep learning architectures, can outperform unimodal approaches based on machine learning algorithms. This inquiry arose from the lack of multimodal strategies in the field and the prospect of improvement in classification results reported in related research. With five specific objectives, this research addressed the identification of a cognitive model of emotions, definition of modalities, construction of multimodal databases, comparison of deep learning architectures, and comparative evaluation of multimodal approaches with unimodal approaches using traditional machine learning algorithms. The analysis of results demonstrated that multimodal approaches achieved superior performance in various classification scenarios, compared to unimodal strategies. These findings positively contribute to the understanding of the effectiveness of multimodal approaches and deep learning architectures in the recognition of emotions in music. Additionally, the research emphasizes the need for attention to emotional models and metadata in online platforms, aiming to avoid biases and noise. This thesis offers relevant contributions to the field of music emotion recognition, particularly in the development of multimodal databases, evaluation of deep learning architectures for tabular problems, experimental protocols, and approaches focused on musical cognition. The systematic comparison between multimodal and unimodal approaches highlights the advantages of the former, encouraging new research in this field.

Keywords: Music. Emotions. Deep learning. Machine learning. Multimodal approaches.

LISTA DE FIGURAS

FIGURA 1 - ESTRUTURA DE UMA TRÍADE (TRÊS NOTAS).....42
FIGURA 2 - ESTRUTURA DE UM ACORDE DE SÉTIMA.....43
FIGURA 3 - MODELO CIRCUMPLEXO (RUSSELL, 1980) 52
FIGURA 4 - MODELO VETORIAL (BRADLEY *et al*., 1992).....53
FIGURA 5 - MODELO PANA (WATSON; TELLEGEN, 1985).....54
FIGURA 6 - RODA DAS EMOÇÕES DE PLUTCHIK 55
FIGURA 7 - MODELO BIDIMENSIONAL (THAYER 1989, 2012) 56
FIGURA 8 - MODELO TRIDIMENSIONAL (PLUTCHICK, 1970) 57
FIGURA 9 - MODELO TRIDIMENSIONAL (LÖVHEIM, 2012) 57
FIGURA 10 - PROCESSO BÁSICO DE RECONHECIMENTO DE EMOÇÕES EM MÚSICAS 60
FIGURA 11 - EXEMPLOS DE CARACTERÍSTICAS PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS 65
FIGURA 12 - MODELO DE HEVNER (1936).....68
FIGURA 13 - POSSIBILIDADES DE ABORDAGENS PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS 72
FIGURA 14 – ARQUITETURA MLP COM DUAS CAMADAS OCULTAS 78
FIGURA 15 - SINAIS (MLP) 79
FIGURA 16 - REPRESENTAÇÃO BÁSICA DE UM 80
FIGURA 17 - ESTRUTURA BÁSICA DE UMA CNN 82
FIGURA 18 - ESTRUTURA BÁSICA DE UMA DBN 84
FIGURA 19 - ESTRUTURA BÁSICA DE UMA RBM.....84
FIGURA 20 - REPRESENTAÇÃO BÁSICA (GANs).....86
FIGURA 21 - ESTRUTURA BÁSICA DE UMA RNN 88
FIGURA 22 - REPRESENTAÇÃO BÁSICA DE UMA LSTM 89
FIGURA 23 - REPRESENTAÇÃO BÁSICA DE UMA GRU.....90
FIGURA 24 - AQUITETURA *FT-TRANSFORMER* 92
FIGURA 25 - *FT-TRANSFORMER (FEATURE TOKENIZER)* 93
FIGURA 26 - *FT-TRANSFORMER* 94
FIGURA 27 - ARQUITETURA *NODE*.....95
FIGURA 28 – ARQUITETURA *NODE* (CAMADA *NODE* – EXEMPLO).....96
FIGURA 29 – MÓDULOS DA ARQUITETURA *TABNET* 97

FIGURA 30 - ARQUITETURA *TAB TRANSFORMER* 99
FIGURA 31 - REPRESENTAÇÃO DO MODELO DE EMOÇÕES PROPOSTO POR BIGAND *et al.* (2005) 112
FIGURA 32 - CROMAGRAMA (EXEMPLO).....128
FIGURA 33 - RESUMO VISUAL DOS ENCAMINHAMENTOS METODOLÓGICOS 147
FIGURA 34 - REPRESENTAÇÃO DE FIGURAS E PAUSAS.....223
FIGURA 35 - DIVISÃO PROPORCIONAL DOS VALORES.....223

LISTA DE QUADROS

QUADRO 1 - SÍNTESE DOS TRABALHOS IDENTIFICADOS COM ABORDAGEM MULTIMODAL PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS 30
QUADRO 2 - PROPRIEDADES DO SOM E SUAS REPRESENTAÇÕES NA MÚSICA 37
QUADRO 3 - NOMENCLATURA DAS NOTAS MUSICAIS.....38
QUADRO 4 - TIPOS DE INTERVALOS 39
QUADRO 5 - INTERVALOS DE ACORDO COM AS SUAS QUALIDADES 40
QUADRO 6 - INTERVALOS JUSTOS.....40
QUADRO 7 - INTERVALOS MAIORES E MENORES.....41
QUADRO 8 - FORMAÇÃO E REPRESENTAÇÃO DAS TRÍADES.....42
QUADRO 9 - FORMAÇÃO E REPRESENTAÇÃO DE ACORDES DE SÉTIMA.....44
QUADRO 10 - EXEMPLOS DE DEFINIÇÕES DE EMOÇÕES.....46
QUADRO 11 - EXEMPLOS DE EMOÇÕES BÁSICAS 49
QUADRO 12 - CARACTERÍSTICAS DAS EMOÇÕES BÁSICAS.....50
QUADRO 13 - SÍNTESE DOS MODELOS DIMENSIONAIS DE EMOÇÃO ANALISADOS 58
QUADRO 14 - EXEMPLOS DE CARACTERÍSTICAS EXTRAÍDAS DIRETAMENTE DO ÁUDIO.....65
QUADRO 15 - EXEMPLOS DE FERRAMENTAS UTILIZADAS PARA A EXTRAÇÃO DE CARACTERÍSTICAS RELACIONADAS AO ÁUDIO 66
QUADRO 16 - ATUALIZAÇÃO DO MODELO E HEVNER (1936) POR SCHUBERT (2003).....69
QUADRO 17 - ESCALA GEMS PROPOSTA POR ZENTNER, GRANDJEAN E SCHERER (2008) 70
QUADRO 18 - CLUSTERS EMOCIONAIS PROPOSTOS POR HU E DOWNIE (2007) 71
QUADRO 19 - EXEMPLOS DE ALGORITMOS EMPREGADOS (CLASSIFICAÇÃO *SINGLE-LABEL*).....73
QUADRO 20 - EXEMPLOS DE ALGORITMOS EMPREGADOS (CLASSIFICAÇÃO *MULTI-LABEL*) 74
QUADRO 21 - EXEMPLOS DE ALGORITMOS EMPREGADOS (CLASSIFICAÇÃO *FUZZY*).....75

QUADRO 22 - EXEMPLOS DE ALGORITMOS EMPREGADOS PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS 75
QUADRO 23 - PROCEDIMENTOS DE COLETA E ANÁLISE DOS DADOS 102
QUADRO 24 - PROCEDIMENTOS DOS EXPERIMENTOS PILOTO E AJUSTADO 106
QUADRO 25 - RELAÇÃO DOS ESTUDOS CONSULTADOS SOBRE EMOÇÕES 109
QUADRO 26 - PROTOCOLO DE AVALIAÇÃO DE PERCEPÇÕES SUBJETIVAS (PAPS) 113
QUADRO 27 - INDICATIVOS ACERCA DA DURAÇÃO DE SEGMENTOS PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS 115
QUADRO 28 - LISTA DE TERMOS EMPREGADOS NA PLATAFORMA *ALLMUSIC* 118
QUADRO 29 - CARACTERÍSTICAS EXTRAÍDAS PARA OS ACORDES 127
QUADRO 30 - SÍNTESE DAS CARACTERÍSTICAS RELACIONADAS AO ÁUDIO CONSIDERADAS 130
QUADRO 31 - EXEMPLOS DE TRANSCRIÇÕES OBTIDAS AUTOMATICAMENTE 131
QUADRO 32 - CARACTERÍSTICAS EXTRAÍDAS (LETRAS) 133
QUADRO 33 - CARACTERÍSTICAS SIMBÓLICAS CONSIDERADAS NESTA TESE 135
QUADRO 34 - CARACTERÍSTICAS RELACIONADAS À VOZ CONSIDERADAS NESTA TESE 138
QUADRO 35 - HIPERPARÂMETROS - ARQUITETURAS DE DEEP LEARNING . 140
QUADRO 36 - ALGORITMOS DE APRENDIZADO DE MÁQUINA UTILIZADOS NA TESE 141
QUADRO 37 - HIPERPARÂMETROS - ARQUITETURAS DE APRENDIZADO DE MÁQUINA 143
QUADRO 38 - PROTOCOLO DE EXPERIMENTOS 146
QUADRO 39 - SÍNTESE DAS PRINCIPAIS FERRAMENTAS EMPREGADAS NA TESE 148
QUADRO 40 - COMPARAÇÃO ENTRE PESQUISAS MULTIMODAIS 171
QUADRO 41 – REPRESENTAÇÃO DE PAUTA MUSICAL 220
QUADRO 42 - REPRESENTAÇÃO DAS CLAVES 221

QUADRO 43 - EXEMPLO DE POSICIONAMENTO DAS CLAVES NA PAUTA MUSICAL 222
QUADRO 44 - DESCRIÇÃO DOS ACIDENTES OU ALTERAÇÕES.....224
QUADRO 45 - CLASSIFICAÇÃO DAS ESCALAS MUSICAIS.....224
QUADRO 46 – GRAUS DA ESCALA DIATÔNICA 225
QUADRO 47 - ESCALAS (MAIOR E MENOR (NATURAL, HARMÔNICA E MELÓDICA) 226

LISTA DE TABELAS

TABELA 1 - ESTRATÉGIA DE BUSCA ADOTADA 27
TABELA 2 - DISTRIBUIÇÃO DAS MÚSICAS POR EMOÇÃO 120
TABELA 3 - TOTAL DE CARACTERÍSTICAS EXTRAÍDAS 138
TABELA 4 - CONCORDÂNCIA ENTRE VOLUNTÁRIOS E PLATAFORMA 149
TABELA 5 - REDISTRIBUIÇÃO DOS SEGMENTOS A PARTIR DA CONCORDÂNCIA 150
TABELA 6 - AVALIAÇÃO DA CONCORDÂNCIA ENTRE OS AVALIADORES DE ACORDO COM O COEFICIENTE KAPPA DE COHEN (κ) 151
TABELA 7 - PERCENTUAL DE REGISTROS POR EMOÇÕES (BASE UM) 152
TABELA 8 – RESULTADOS PARA A CONCORDÂNCIA ENTRE OS AVALIADORES 152
TABELA 9 - PERCENTUAL DE REGISTROS POR EMOÇÕES (BASE DOIS) 153
TABELA 10 - PERCENTUAIS DE MÚSICAS RETIDAS (MAIORIA OBSERVADA) 153
TABELA 11 - RESULTADOS PARA TODOS OS ATRIBUTOS E AJUSTE DE HIPERPARÂMETROS 154
TABELA 12 - RESULTADOS OBTIDOS COM O USO DE PCA, COM 90,00% DE VARIÂNCIA 156
TABELA 13 - RESULTADOS OBTIDOS COM O USO DE SELEÇÃO DE ATRIBUTOS 157
TABELA 14 - COMPONENTES PRINCIPAIS (BASES MAIORIA E CONCORDÂNCIA) 159
TABELA 15 - NÚMERO DE CARACTERÍSTICAS SELECIONADAS 160
TABELA 16 - RESULTADOS PARA A ESTRATÉGIA UNDERSAMPLING 161
TABELA 17 - RESULTADOS PARA A ESTRATÉGIA *OVERSAMPLING* 162
TABELA 18 – RESULTADOS: ARQUITETURAS DE DEEP LEARNING (DATA AUGMENTATION) 165
TABELA 19 – RESULTADOS: ALGORITMOS DE APRENDIZADO DE MÁQUINA (DATA AUGMENTATION) 166
TABELA 20 - MELHOR CENÁRIO DE DEEP LEARNING (BASE DOIS) 167
TABELA 21 - MELHOR CENÁRIO DE DEEP LEARNING (BASE UM) 168
TABELA 22 - DESEMPENHO POR EMOÇÃO - ALGORITMOS DE APRENDIZADO DE MÁQUINA (BASE DOIS) 169

TABELA 23 - DESEMPENHO POR EMOÇÃO - ALGORITMOS DE APRENDIZADO DE MÁQUINA (BASE UM) 170
TABELA 24 - MELHORES RESULTADOS OBTIDOS NESTA TESE 171

LISTA DE SIGLAS

[CLS]	-	*token* de classificação
5-HT	-	Serotonina/5-Hidroxitriptamina
AE	-	*Auto Encoder*
AEG	-	*Acoustic Emotion Gaussians*
BCRSN	-	*Bidirectional Convolutional Recurrent Sparse Network*
BDTD	-	Biblioteca Digital Brasileira de Teses e Dissertações
BN	-	*Bayesian Network*
BoostR	-	*AdaBoost.RT*
BOW	-	*Bag-Of-Words*
BPM	-	*Beats Per Minute*
BP-MLL	-	*Multi-label back-propagation*
BR	-	*Binary Relevance*
CCNF	-	*Continuous Conditional Neural Fields*
CLDNN	-	*Convolutional Long Short-Term Memory Deep Neural Network*
CLR	-	*Calibrated Label Ranking*
CNN	-	*Convolutional Neural Network*
DA	-	Dopamina
DAE	-	*Denoising Autoencoder*
DFA	-	*Detrended Fluctuation Analysis*
dB	-	Decibel
DBN	-	*Deep Belief Network*
DFT	-	*Discrete Fourier transform*
DGPs	-	*Deep Gaussian Process*
ECG	-	Eletrocardiograma
EEG	-	Eletroencefalograma
F_0	-	Frequência Fundamental
FC	-	Camada totalmente conectada
FFT	-	*Fast Fourier Transform*
fMRI	-	*Functional Magnetic Resonance Imaging*
FK-NN	-	*Fuzzy k-NN*
FNM	-	*Fuzzy Nearest-Mean*
FT-Transformer	-	*Feature Tokenizer + Transformer*

GANs - *Generative Adversarial Networks*
GEMS - *Geneva Emotional Music Scale*
GFCCs - *Gammatone Frequency Cepstral Coefficients*
GLP - *Gated Linear Unit*
GMD - *Greek Music Dataset*
GMM - *Gaussian Mixture Models*
GPU - *Graphics Processing Unit*
GRU - *Gated Recurrent Unit*
Hz - Hertz
HPCP - *Harmonic Pitch Class Profiles*
IDF - Frequência Inversa do Documento
kbps - *kilobits per second*
k-NN - *k-Nearest Neighbors*
LSTM - *Long Short-Term Memory*
LP - *Label Powerset*
MER - *Music Emotion Recognition*
MFCC - *Mel-Frequency Cepstral Coefficients*
MIDI - *Musical Instrument Digital Interface*
MIR - *Music Information Retrieval*
MIREX - *Music Information Retrieval Evaluation eXchange*
ML-k-NN - *Multi-label k-nearest neighbor*
MLP - *Multilayer Perceptron*
MLR - *Multiple Linear Regression*
NB - *Naïve Bayes*
NDLTD - *Networked Digital Library of Theses and Dissertations*
NE - Noradrenalina
NFNC - *Neuro-Fuzzy Networks Classification*
NLTK - *Natural Language Toolkit*
NNLS - *Non-Negative Least Squares*
NODE - *Neural Oblivious Decision Ensembles*
ODT - *Oblivious Decision Trees*
PAD - *Pleasure-Arousal-Valence*
PANA - *Positive Activation – Negative Activation*

PCA - *Principal Component Analysis*
PLP - *Predominant Local Pulse*
POS - *Part-Of-Speech*
PPGGI - Programa de Pós-Graduação em Gestão da Informação
PRISMA - *Preferred Reporting Items for Systematic Review*
RA*k*EL - *Random K-Labelsests*
RBM - *Restricted Boltzmann Machine*
R&B - *Rhythm and Blues*
ReLU - *Rectified Linear Unit*
REM - Reconhecimento de Emoções em Músicas
RF - *Random Forest*
RMS - *Root Mean Square*
RNN - *Recurrent Neural Network*
SAE - *Stacked Autoencoder*
SMOTE - *Synthetic Minority Oversampling Technique*
ST - Semitom
STFT - *Short-time Fourier transform*
SVM - *Support Vector Machine*
SVR - *Support Vector Regression*
T - Tom
TCLE - Termo de Consentimento Livre e Esclarecido
TF - *Term Frequency*
TF-IDF - *Term Frequency-Inverse Document Frequency*
URL - *Uniform Resource Locator*
ZCR - *Zero-crossing rate*

SUMÁRIO

1 INTRODUÇÃO 21
1.1 ABORDAGEM DO PROBLEMA 23
1.2 OBJETIVOS 25
1.2.1 Objetivo geral 26
1.2.2 Objetivos específicos 26
1.3 JUSTIFICATIVA 26
1.4 ORIGINALIDADE E NÃO TRIVIALIDADE DA TESE 32
1.5 DELIMITAÇÃO DO ESTUDO 33
1.6 ESTRUTURA DO DOCUMENTO 34
2 REFERENCIAL TEÓRICO 35
2.1 ABORDAGEM MULTIMODAL 35
2.2 FUNDAMENTOS DE TEORIA MUSICAL 37
2.2.1 Notação musical 38
2.2.2 Intervalos 39
2.2.3 Acordes 41
2.3 EMOÇÕES 45
2.3.1 Abordagem categórica das emoções 47
2.3.2 Abordagem dimensional das emoções 51
2.4 RECONHECIMENTO DE EMOÇÕES EM MÚSICAS 59
2.4.1 Obtenção de músicas para a construção de sistemas de reconhecimento de ... emoções em músicas 61
2.4.2 Extração de características para o reconhecimento de emoções em músicas 63
2.4.3 Abordagens emocionais consideradas para o reconhecimento de emoções em músicas 67
2.4.4 Algoritmos empregados para o reconhecimento de emoções em músicas 71
2.5 *DEEP LEARNING* 76
2.5.1 *Multilayer Perceptron* 78
2.5.2 *Autoencoders* 80
2.5.3 *Convolutional Neural Network* 81

2.5.4 *Deep Belief Network*....83
2.5.5 *Generative Adversarial Networks*....85
2.5.6 *Recurrent Neural Networks*....87
2.6 *DEEP LEARNING* PARA DADOS TABULARES....91
2.6.1 *FT-Transformer*....92
2.6.2 *Neural Oblivious Decision Ensembles*....95
2.6.3 *TabNet*....96
2.6.4 *TabTransformer*....98
3 ENCAMINHAMENTOS METODOLÓGICOS....101
3.1 CARACTERIZAÇÃO DA PESQUISA....101
3.2 COLETA E ANÁLISE DE DADOS....102
3.3 CONSTRUÇÃO DOS EXPERIMENTOS....104
3.4 DEFINIÇÃO DO MODELO DE EMOÇÕES....106
3.4.1 Modelo de emoções considerado para o experimento piloto....110
3.4.2 Modelo de emoções considerado para o experimento ajustado....111
3.5 ESTRUTURA E CONSTRUÇÃO DAS BASES DE DADOS....113
3.5.1 Duração e segmentação das músicas....115
3.5.2 Identificação e coleta das músicas para o experimento piloto....117
3.5.3 Revisão das músicas selecionadas para o experimento piloto....121
3.5.4 Identificação e coleta das músicas para o experimento ajustado....122
3.5.5 Avaliação das músicas selecionadas para o experimento ajustado....123
3.5.6 Extração das características....125
3.6 ESTRATÉGIAS DE CLASSIFICAÇÃO....139
3.6.1 Arquiteturas de *deep learning* empregadas....139
3.6.2 Algoritmos de aprendizado de máquina empregados....140
3.6.3 Protocolo de experimentos....143
3.7 SÍNTESE DA METODOLOGIA DESENVOLVIDA....147
4 ANÁLISE E DISCUSSÃO DOS RESULTADOS....149
4.1 RESULTADOS DA REVISÃO DAS MÚSICAS (EXPERIMENTO PILOTO)....149
4.2 RESULTADOS DA AVALIAÇÃO DAS MÚSICAS (EXPERIMENTO AJUSTADO)....151

4.3 RESULTADOS PARA O CENÁRIO DE CLASSIFICAÇÃO A 153
4.3.1 Resultados alcançados com as abordagens multimodais 154
4.3.2 Resultados obtidos com os algoritmos de aprendizado de máquina 155
4.4 RESULTADOS PARA O CENÁRIO DE CLASSIFICAÇÃO B 156
4.4.1 Resultados alcançados com as abordagens multimodais 156
4.4.2 Resultados obtidos com os algoritmos de aprendizado de máquina 159
4.5 RESULTADOS PARA O CENÁRIO DE CLASSIFICAÇÃO C 161
4.5.1 Resultados alcançados com as abordagens multimodais 161
4.5.2 Resultados obtidos com os algoritmos de aprendizado de máquina 162
4.6 RESULTADOS PARA O CENÁRIO DE CLASSIFICAÇÃO D 164
4.7 COMPARAÇÃO E SÍNTESE DOS RESULTADOS 166
4.8 PUBLICAÇÕES 173
5 CONSIDERAÇÕES FINAIS 175
5.1 CONTRIBUIÇÕES 177
5.2 TRABALHOS FUTUROS 179
REFERÊNCIAS 181
ANEXO A – MODELO QUADRIDIMENSIONAL (FONTAINE *et al.*, 2007) ... 218
ANEXO B – CONTATO COM A PLATAFORMA *ALLMUSIC* 219
APÊNDICE A – COMPLEMENTO PARA OS FUNDAMENTOS DE TEORIA MUSICAL 220
APÊNDICE B -TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (EXPERIMENTO PILOTO) 228
APÊNDICE C – INSTRUÇÕES E RECOMENDAÇÕES AOS VOLUNTÁRIOS (EXPERIMENTO PILOTO) 230
APÊNDICE D – ENDEREÇO DAS *PLAYLISTS* ANALISADAS 232
APÊNDICE E – *FOLDER* DE DIVUGALÇÃO DA PESQUISA 235
APÊNDICE F – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (EXPERIMENTO AJUSTADO) 236
APÊNDICE G – INSTRUÇÕES E RECOMENDAÇÕES AOS VOLUNTÁRIOS (EXPERIMENTO AJUSTADO) 237
APÊNDICE H – ESTRUTURA DAS ARQUITETURAS DE *DEEP LEARNING* 238

APÊNDICE I - ESTATÍSTICAS DESCRITIVAS (BASE UM) 239
APÊNDICE J - ESTATÍSTICAS DESCRITIVAS (BASE DOIS) 246
APÊNDICE K – HIPERPARÂMETROS DAS ARQUITETURAS DE *DEEP LEARNING* SELECIONADOS PARA AS MELHORES ACURÁCIAS NO CENÁRIO DE CLASSIFICAÇÃO A 253
APÊNDICE L - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (CENÁRIO A – TODOS OS ATRIBUTOS) 254
APÊNDICE M - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (CENÁRIO A – HIPERPARÂMETROS) 255
APÊNDICE N - HIPERPARÂMETROS DOS ALGORITMOS DE APRENDIZADO DE MÁQUINA SELECIONADOS 256
APÊNDICE O - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (PCA – CENÁRIO B) 258
APÊNDICE P - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (SELEÇÃO DE ATRIBUTOS – CENÁRIO B) 259
APÊNDICE Q - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (*UNDERSAMPLING* – CENÁRIO C) 260
APÊNDICE R - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (*OVERSAMPLING* – CENÁRIO C) 261
APÊNDICE S – COMPARAÇÃO DO PERCENTUAL DE ACERTO DAS ARQUITETURAS DE DEEP LEARNING (BASE DOIS) 262
APÊNDICE T – COMPARAÇÃO DO PERCENTUAL DE ACERTO DAS ARQUITETURAS DE DEEP LEARNING (BASE UM) 263
APÊNDICE U – MATRIZES DE CONFUSÃO PARA OS MELHORES RESULTADOS OBTIDOS COM ALGORITMOS DE APRENDIZADO DE MÁQUINA (BASE DOIS) 264

APÊNDICE V – MATRIZES DE CONFUSÃO PARA OS MELHORES RESULTADOS OBTIDOS COM ALGORITMOS DE APRENDIZADO DE MÁQUINA (BASE UM) 266

1 INTRODUÇÃO

A música é parte da vida cotidiana (Cohrdes; Grolig; Schroeder, 2018; Fuentes; Hagberg; Kjellberg, 2019; Lonsdale, 2019; Terry *et al*., 2020), e vai além das atividades de lazer (Yadati *et al*., 2017; Mélo; Andrade, 2020), apresentando-se como um complexo recurso para o Marketing (Stewart; Koh, 2017) e um fator de interesse para estudos científicos sobre transtornos mentais (Cho, 2018; Mcferran *et al*., 2018), por exemplo.

Pode-se dizer que a música ocorre em todas as culturas conhecidas (Juslin, 2019), e acompanha o ser humano desde o início da sua vida (Cohrdes; Grolig; Schroeder, 2018; Juslin, 2019); é parte importante em celebrações ao longo dos anos, como cerimônias de iniciação e casamentos (Terry *et al*., 2020); e permanece presente até o final da vida (Juslin, 2019), quando assume um papel importante no processo de luto, preenchendo lacunas psicossociais e espirituais (Viper; Thyrén; Horwitz, 2020).

Com o surgimento dos serviços de *streaming* como *Spotify*, *Pandora* ou *Apple Music* (Schedl *et al*. 2018; Joseph; Lekamge, 2019), observa-se uma mudança completa na forma de se ouvir música (Fabio *et al*., 2019), em que milhões de músicas podem ser acessadas (Schedl *et al*. 2018; Joseph; Lekamge, 2019), transformando a música em um elemento onipresente (Joseph; Lekamge, 2019; Rajesh; Nalini, 2020). Nesse cenário, a audição de canções é realizada em diferentes meios, como dispositivos móveis, estações de rádio e sistemas de recomendação (Cheng *et al*., 2017; Rajesh; Nalini, 2020).

No entanto, o aumento na disponibilidade de música dificulta a escolha do que o usuário irá ouvir (Joseph; Lekamge, 2019; Yu *et al*., 2020; Mélo; Andrade, 2020) e, nesse contexto, mecanismos destinados à recuperação de informações musicais (Crayencour; Cella, 2019; Russo *et al*., 2020) passaram a utilizar modalidades como emoção, gênero musical e letras nas estratégias de pesquisa (Sarkar *et al.*, 2019; Liu; Than, 2020).

Todavia, se por um lado há esforços consideráveis no quanto às estratégias baseadas no gênero musical (Sarkar *et al.*, 2019; Yu *et al*., 2020), por exemplo; por outro, em características como a emoção, é necessária uma investigação mais profunda (Fabio *et al*., 2019; Sarkar *et al.*, 2019), pois, embora existam aplicações baseadas em emoções, estas tendem a ser elaboradas manualmente (Hult *et al*.,

2020), ou seja, o processo para novos recursos é dependente do trabalho de um ser humano.

A elaboração manual, porém, não é produtiva, seja por conta da quantidade de músicas disponível digitalmente (Sarkar *et al.*, 2019; Hult *et al*., 2020; Xu *et al*., 2020), do tempo empregado ou dos custos envolvidos (Russo *et al*., 2020). Somado a essa questão, o desenvolvimento de estratégias para a recuperação de informações musicais baseadas em emoções justifica-se pela capacidade que a música tem em recuperar memórias vividas e emocionais de distintos períodos e eventos da vida (Jakubowski; Ghosh, 2019).

Ademais, o impacto emocional desempenha um papel importante na música (Yang; Dong; Li, 2017; Zhang *et al.*, 2018a), sendo uma das principais razões para o ser humano ouvir música (Juslin, 2019; Rajesh; Nalini, 2020). Quando comparada a outros materiais, a música pode induzir uma experiência emocional mais profunda e duradoura (Bo *et al.*, 2020), capaz de despertar diferentes tipos de emoções (Balasubramanian *et al.*, 2018; Rajesh; Nalini, 2020).

Para mais, visualiza-se que a música é capaz de expressar e refletir a emoção do ouvinte (Deng, J. J. *et al*., 2015), permitindo, por exemplo, que este se expresse por meio de sinais como sorrisos, risos, choros e o franzir das sobrancelhas (Juslin, 2019). Para além, as emoções evocadas pela música provocam mudanças nos principais componentes da emoção, como sentimentos subjetivos, alterações autonômicas e endócrinas, expressão motora de emoção e tendências de ação (dançar, cantar, bater os pés, tocar um instrumento etc.) (Koelsch, 2014).

As aplicações desenvolvidas a partir da relação “música-emoção” apresentam contribuições para diversos campos: da musicoterapia à criação de trilhas sonoras; da saúde ao marketing; além dos jogos eletrônicos (Juslin, 2019), por exemplo. No âmbito profissional, músicos, críticos musicais, professores, produtores, enfermeiras, engenheiros de som e psicoterapeutas estão entre aqueles que podem se beneficiar (Juslin, 2019).

Com relação às possíveis aplicações baseadas no aspecto emocional de canções, é possível citar: a) sistemas de recomendação de músicas baseados em comunidades de usuários, críticos e músicos; b) gestão de coleções pessoais de música; c) geração de *playlists* ou listas automáticas por meio de inteligência artificial; e d) categorização musical para fins de avaliação, consumo e crítica musical (Huq; Bello; Rowe, 2010; Villar; Luján, 2020).

1.1 ABORDAGEM DO PROBLEMA

A organização de músicas de acordo com o seu impacto emocional, apesar de natural para os seres humanos, é uma tarefa difícil para um computador (Yang; Dong; LI, 2017), devido aos desafios relacionados à variabilidade do conteúdo musical entre os gêneros, ao contexto cultural dos ouvintes e à confiabilidade dos dados utilizados (Fan *et al*., 2017).

Somado a esses fatores, o reconhecimento de emoções em músicas é uma temática interessante e desafiadora dentro da comunidade de recuperação de informação musical (*Music Information Retrieval* (MIR)) (Fan *et al*., 2017; Rajesh; Nalini, 2020), em razão da influência de fatores como, por exemplo, eventos relacionados à experiência de vida do ouvinte (Deng, J. J. *et al*., 2015) e à falta de total clareza a respeito de quais elementos da música despertam emoções específicas nos ouvintes (Cho *et al*., 2016; Rajesh; Nalini, 2020).

Nesse sentido, as estratégias para reconhecer emoções em músicas tendem a considerar apenas características acústicas (Bhaskar; Sruthi; Nedungadi, 2015; Aljanaki; Yang; Soleymani, 2017; Panwar *et al*., 2018). Muitas vezes, essas características assemelham-se às empregadas para classificação de gênero musical (Murthy; Koolagudi, 2019), representando propriedades intrínsecas do áudio musical como o ritmo, o timbre e a melodia (Cornelis *et al*., 2010; Yang; Chen, 2012).

Todavia, elencar modalidades para representar o conteúdo emocional de uma música é uma etapa “complicada”, pois a percepção emocional de uma canção não se limita ao áudio, uma vez que, letras (Shi; Feng, 2018; Simonetta; Ntalampiras; Avanzini, 2019), acordes (CHO *et al*., 2016; GREER *et al*. 2019), partituras (Simonetta; Ntalampiras; Avanzini, 2019), e informações extramusicais (Wang *et al*., 2010), como a biografia do artista, resenhas e capas de músicas e álbuns (Kim *et al*., 2010; Simonetta; Ntalampiras; Avanzini, 2019) também podem influenciar a percepção do ouvinte.

Tendo em vista essa dificuldade, a multimodalidade[1] apresenta-se como alternativa (Thammasan; Fukui; Numao, 2017), posto que se trata de uma abordagem em que se trabalha com diferentes atributos de entrada, representando uma mesma

[1] Neste estudo, ressalta-se que o conceito empregado para multimodalidade consiste na combinação de múltiplas fontes de informação, denominadas aqui como modalidades. Maiores detalhes são apresentados na seção do Referencial Teórico.

peça musical (Simonetta; Ntalampiras; Avanzini, 2019). Ao mesmo tempo, a combinação de modalidades tende a melhorar os resultados em comparação ao uso de apenas uma modalidade (Wang *et al*., 2010; Yang; Chen, 2012; Grekow, 2017).

Do mesmo modo, a combinação de fontes se justifica por conta de a percepção da emoção musical ser multidimensional (Yang; Chen, 2012), além da própria natureza multimodal da música (Kim *et al*., 2019) enquanto forma de arte (Li *et al*., 2019). Entretanto, apesar desses indicativos, considera-se um problema na área de reconhecimento de emoções em músicas a falta de estratégias e conjuntos de dados destinados ao estudo da multimodalidade da relação “música-emoção” (Pesek *et al*., 2017).

Referente aos métodos empregados na construção de modelos de reconhecimento de emoções em músicas, costuma-se recorrer aos algoritmos de aprendizado de máquina (*machine learning*), como *Support Vector Machine* (SVM), k-*Nearest Neighbors* (k-NN), *Bayesian Network* (BN), *Gaussian Mixture Models* (GMM), *Naïve Bayes* (NB), modelos baseados em árvores de decisão e redes neurais (Kaminskas; Ricci, 2012; Yang; Chen, 2012; Yang; Dong; Li, 2017; Joseph; Lekamge, 2019).

Contudo, a partir do desenvolvimento dos métodos de aprendizado profundo (*deep learning*), uma melhora nos resultados das classificações pôde ser observada (Yang; Dong; Li, 2017). Além disso, esses métodos passaram a ser empregados em áreas correlatas, como processamento de voz e reconhecimento de fala (Yang; Dong; Li, 2017; Hossain; Muhammad, 2019; Sarkar *et al*., 2019), classificação de gêneros musicais (PANWAR *et al*., 2017), sistemas de recomendação de músicas (Schedl, 2019), geração de músicas (Briot; Pachet, 2018) e reconhecimento de emoções em videoclipes (Pandeya; Lee, 2020).

No campo do reconhecimento de emoções em músicas, estudos como os de Schmidt e Kim (2011), Li, J. *et al*. (2015) e Thammasan, Fukui e Numao (2016) foram construídos a partir do uso de *deep learning*. Nestes estudos, os autores empregaram uma arquitetura denominada *Deep Belief Network* (DBN) para o reconhecimento de emoções em músicas por meio de características acústicas (Schmidt; Kim, 2011), letras (LI, J. *et al*., 2015) e de eletroencefalografia (EEG) (Thammasan; Fukui; Numao, 2016).

Além dessa, arquiteturas como *Deep Gaussian Process* (DGPs) (Chen, S. H. *et al*., 2015), *Bidirectional Convolutional Recurrent Sparse Network* (BCRSN) (Dong

et al., 2019), *Convolutional Neural Network* (CNN) (Malik *et al*., 2017; Sarkar *et al*., 2019; Xie; Kim; Park, 2020), *Recurrent Neural Networks* (RNN) (Malik *et al*., 2017) e estratégias não supervisionadas como *Auto Enconder* (AE) e *Restricted Boltzmann Machine* (RBM) também foram consideradas para o problema em questão (Zhou; Chen; Yang, 2019).

No entanto, mesmo que os métodos de *deep learning* tenham despertado a atenção dos pesquisadores e provocado avanços em campos como processamento de imagens (Bayle; Robine; Hanna, 2018) e de linguagem (Cheah *et al*., 2019), no tocante à música, o uso desses métodos pode ser considerado em fase inicial (Bayle; Robine; Hanna, 2018; Briot; Pachet, 2018), devido, por exemplo, à falta de um grande conjunto de dados musicais que possibilite a aplicação desses (Bayle; Robine; Hanna, 2018) e à indefinição de quais arquiteturas de *deep learning* empregar (Purwins *et al*., 2019).

Assim sendo, visualiza-se como uma oportunidade de pesquisa na área de reconhecimento de emoções em músicas, a construção de estudos a partir de diferentes arquiteturas de *deep learning*. Outrossim, a falta de estudos considerando o uso de múltiplas fontes de informação (multimodalidade) também se manifesta como uma questão pouco explorada. Desse modo, a partir da união desses tópicos, a pergunta norteadora deste estudo é: **para o reconhecimento de emoções em músicas, abordagens multimodais, somadas ao uso de arquiteturas de *deep learning*, são capazes de superar o desempenho de classificação de abordagens unimodais e construídas a partir de algoritmos de aprendizado de máquina[2]?**

1.2 OBJETIVOS

Para responder à pergunta norteadora desta proposta, definiu-se um objetivo geral e cinco objetivos específicos, conforme elencado na sequência.

[2] Para este estudo, a separação entre algoritmos de aprendizado de máquina e arquiteturas de *deep learning* justifica-se pela diferença fundamental em suas abordagens: os algoritmos de aprendizado de máquina “tradicionais” – também denominados como rasos - dependem de recursos cuidadosamente projetados, enquanto as arquiteturas de *deep learning* têm a capacidade de aprender representações de alto nível diretamente dos dados brutos, tornando-as mais adequadas para tarefas de reconhecimento de padrões complexos em larga escala.

1.2.1 Objetivo geral

Investigar se a adoção de abordagens multimodais e arquiteturas de *deep learning* é capaz de superar o desempenho de classificação obtido por abordagens unimodais, construídas com o emprego de algoritmos de aprendizado de máquina.

1.2.2 Objetivos específicos

Os seguintes objetivos específicos foram estabelecidos para esta pesquisa:

a) identificar o modelo de emoções a ser utilizado;
b) identificar quais modalidades empregar para o reconhecimento de emoções em músicas;
c) construir uma base multimodal composta por características relacionadas às modalidades identificadas;
d) comparar diferentes arquiteturas de *deep learning* para o reconhecimento de emoções na base elaborada;
e) comparar o desempenho de classificação obtido com as abordagens multimodais desenvolvidas com o desempenho proporcionado por abordagens unimodais, construídas a partir de algoritmos de aprendizado de máquina.

1.3 JUSTIFICATIVA

Diante da necessidade de se encontrar trabalhos com estratégia multimodal para o reconhecimento de emoções em músicas, realizou-se uma revisão integrativa da literatura (Aldrighi *et al*., 2016; Shishehgar *et al*., 2017), considerando a recomendação *Preferred Reporting Items for Systematic Review* (PRISMA) (Moher *et al*., 2009). Para isto, utilizou-se a estratégia de busca listada na TABELA 1.

TABELA 1 - ESTRATÉGIA DE BUSCA ADOTADA

Expressão	Base	Campos	Quantidade
((“*multimodal*” *OR* “*multi-modal*” *OR* "*multifactor*" *OR* "*multi-factor*") *AND* (“*music emotion recognition*” *OR* “*music emotion classification*” *OR* “*music mood recognition*” *OR* “*music mood classification*”))	*Dimensions*	Título e resumo	17
	IEEE Xplore	-	6
	Science Direct	-	10
	Scopus	Título, Resumo, Palavras-chave	28
	Web of Science	Título, Resumo, Palavras-chave do autor e *Keywords Plus*	13
Total			**74**

FONTE: O autor (2020).

A busca apresentada na TABELA 1 foi conduzida no dia 30 de novembro de 2020, e possibilitou a recuperação de 74 documentos. Quanto à exclusão de registros, a partir da leitura dos títulos, resumos e palavras-chave desses registros, os seguintes critérios de exclusão foram adotados: a) registros duplicados (18); b) registros de eventos inteiros, sem trabalhos relacionados ao tema de pesquisa (cinco); c) trabalhos que não abordavam o reconhecimento de emoções em músicas (30); d) estratégias com uma fonte de informação (áudio, letras, imagens etc.) (um); e e) estratégias bimodais (áudio e letras; áudio e imagens; letras e acordes etc.) (17). Sendo assim, estes registros foram eliminados, restando três trabalhos para análise completa do texto.

Para complementar a revisão conduzida, cinco trabalhos foram acrescentados após pesquisa no *Google Scholar*, na Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) e na *Networked Digital Library of Theses and Dissertations* (NDLTD), totalizando oito estudos multimodais para análise completa.

O primeiro estudo analisado foi desenvolvido por Lu *et. al* (2010), no qual os autores utilizaram características provenientes do áudio, das letras e de arquivos do tipo *Musical Instrument Digital Interface* (MIDI). Neste estudo, 500 canções de música *pop* chinesa foram consideradas, com a associação destas às emoções conduzida por oito participantes, de acordo com o modelo de valência e excitação proposto por Thayer (1989). Para a classificação, este modelo foi dividido em quatro classes: a)

“feliz”; b) “triste”; c) “raiva” e d) “relaxante”, e os algoritmos *AdaBoost* e SVM foram empregados.

O segundo trabalho recuperado foi elaborado por Zhao, Yang e Chen (2010). Neste estudo, os autores utilizaram três modalidades (áudio, letras e MIDI) relacionadas a 981 músicas chinesas, divididas em uma variação do modelo dimensional proposto por Russell (1980), com as seguintes classes: a) “alegria”; b) “contentamento”; c) “raiva”; e d) “depressão”. Como métodos de classificação, os autores utilizaram os algoritmos SVM, NB e J48.

O terceiro trabalho correspondeu ao estudo de Schuller, Weninger e Dorfner (2011), cuja abordagem baseou-se no uso de quatro modalidades (áudio, letras, acordes e metadados). Neste estudo, 2.648 músicas de diferentes estilos foram utilizadas, e o algoritmo *REPtree* empregado para estimar os valores de valência e excitação (RUSSELL, 1980) destas músicas.

Na busca direta no *Google Scholar*, dois trabalhos foram recuperados: a) Schuller *et al*., (2010); e b) Schuller, Dorfner e Rigoll (2010). No estudo de Schuller *et al*. (2010), 195 canções de vários estilos musicais foram representadas por quatro modalidades (áudio, letras, acordes e gênero musical) e o algoritmo SVM foi empregado para a tarefa de classificação.

Entretanto, os autores não conseguiram encontrar a totalidade de acordes para 30,00% dos registros. Desta forma, esta modalidade foi descartada e considerada como incremento para a obtenção de recursos tonais (áudio). Como modelo emocional, Schuller *et al*. (2010) optaram por dois modelos: a) valência e excitação (Thayer, 1989); e b) lista de adjetivos proposta por Hevner (1936).

O estudo de Schuller, Dorfner e Rigoll (2010) consistiu em um experimento multimodal (áudio, acordes, letras e metadados) com 2.648 canções de diferentes estilos, distribuídas de acordo com as suas posições no modelo de valência e excitação, considerando duas estratégias: a) cinco (-2, -1, 0, 1, 2); e b) três classes (-1, 0, 1). Como algoritmos, os seguintes foram utilizados pelos autores: a) SVM; b) *Random Forest* (RF); c) BN; e d) k-NN.

Em termos de teses e dissertações, encontrou-se uma dissertação na BDTD, cujo objetivo foi classificar automaticamente emoções em músicas latinas, considerando três fontes de informação (áudio, letras e cifras) (Przybysz, 2016). Neste estudo, o autor construiu uma base multimodal composta por músicas latinas, a partir de uma base já existente (Santos; Silla JR., 2015), cuja associação das músicas às

emoções foi elaborada por um músico profissional brasileiro, com os rótulos “alegria”, “amor”, “decepção”, “entusiasmo”, “paixão” e “tristeza”. Além destes rótulos, uma classificação com três classes (negativa; neutra; positiva) foi realizada.

Ademais, em sua dissertação, Przybysz (2016) elencou trabalhos correlatos ao tema, em que constatou a predominância de estudos com uma modalidade (áudio; letras; informações culturais etc.) e abordagens bimodais (áudio e imagens; áudio e letras). Dos trabalhos analisados, apenas Panda *et al*. (2013) conduziram uma abordagem multimodal (áudio, letras e MIDI).

No estudo conduzido por Panda *et al*. (2013), os autores utilizaram uma base de dados com a taxonomia de emoções formulada pela *Music Information Retrieval Evaluation eXchange* (MIREX), que consiste em cinco *clusters* com diversos rótulos emocionais. Sobre os métodos de classificação, os seguintes foram adotados: a) SVM; b) k-NN; c) C4.5; e d) NB.

Para a pesquisa na NDLTD, dois registros baseados em multimodalidade foram recuperados: a) a dissertação elaborada por Przybysz (2016); e b) a tese desenvolvida por Imbrasaite (2015). No tocante a esta tese, trata-se de um estudo realizado na Universidade de Cambridge que utilizou características relacionadas à voz, ao áudio e às letras.

O conjunto de dados usado neste estudo continha 186 músicas, considerando os valores de valência e excitação. Quanto aos métodos empregados pela pesquisadora, a pesquisa foi conduzida a partir dos seguintes algoritmos de regressão: a) *Support Vector Regression* (SVR); e b) *Continuous Conditional Neural Fields* (CCNF) (Imbrasaite, 2015).

Aprofundando a discussão dos estudos multimodais, verifica-se que os resultados para a busca de registros nas bases de dados e nas demais fontes apresentaram a predominância de estratégias com uma única fonte - geralmente o áudio - e estratégias bimodais, o que vai ao encontro de resultados obtidos em outros estudos, como os trabalhos de Yang, Dong e Li (2017), Panda (2019) e Joseph e Lekamge (2019).

Yang, Dong e Li (2017) conduziram uma revisão de literatura sobre reconhecimento de emoções em músicas com trabalhos publicados entre 2003 e 2017, em que constataram que, dos 27 registros estudados, apenas o elaborado por Hu e Yang (2014) utilizava mais de uma modalidade (áudio e vídeo) (Yang; Dong; Li, 2017). Joseph e Lekamge (2019), em uma revisão sistemática da literatura,

analisaram 14 trabalhos relacionados ao reconhecimento de emoções em músicas e constataram que apenas características acústicas como intensidade, ritmo, harmonia, melodia e timbre foram consideradas nestes estudos.

Panda (2019), em sua tese, compilou 64 estudos sobre reconhecimento de emoções em músicas elaborados entre 1998 e 2018, dos quais apenas o estudo de Panda *et al*. (2013) - o mesmo presente em Przybysz (2016) - foi elaborado a partir de uma abordagem multimodal.

Finalmente, a síntese dos trabalhos que adotaram múltiplas fontes é apresentada no QUADRO 1.

QUADRO 1 - SÍNTESE DOS TRABALHOS IDENTIFICADOS COM ABORDAGEM MULTIMODAL PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS

(continua)

Autores	**País**	**Modalidades**	**Registros**	**Emoções**	**Métodos**
Imbrasaite (2015)	Inglaterra.	Voz, áudio e letras.	186 músicas.	Valência e excitação.	SVR e CCNF.
Lu *et al*. (2010)	China.	Áudio, letras e MIDI.	500 canções de música pop chinesa.	Feliz, triste, raiva e relaxante (THAYER, 1989).	*AdaBoost* e SVM.
Panda *et al.* (2013)	Portugal.	Áudio, letras e MIDI.	193 músicas.	Cinco *clusters* (MIREX)	SVM, k-NN, C4.5 e NB.
Przybysz (2016)	Brasil.	Áudio, letras e cifras.	784 músicas latinas.	Seis emoções: alegria; amor; decepção; entusiasmo; paixão e tristeza. Três emoções: positiva, neutra e negativa.	SVM, k-NN e J48.
Schuller, Dorfner e Rigoll (2010)	Alemanha.	Áudio; letras; acordes e metadados.	2.648 canções de diferentes gêneros musicais.	Valência e excitação, com cinco (-2, -1, 0, 1, 2) e três classes (-1, 0, 1)	SVM, RF, BN e k-NN.

FONTE: O autor (2020).

QUADRO 1 - SÍNTESE DOS TRABALHOS IDENTIFICADOS COM ABORDAGEM MULTIMODAL PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS

(conclusão)

Autores	País	Modalidades	Registros	Emoções	Métodos
Schuller *et al*. (2010)	Alemanha.	Áudio, letras, acordes e gênero musical.	195 de diferentes gêneros musicais.	Valência e excitação (THAYER, 1989). Hevner (1936)	SVM.
Schuller, Weninger e Dorfner (2011)	Alemanha.	Áudio; letras; acordes e metadados.	2.648 canções de diferentes gêneros musicais.	Valência e excitação (RUSSELL, 1980).	*REPTree*.
Zhao, Yang e Chen (2010)	China.	Áudio, letras e MIDI.	981 canções chinesas	Alegria, contentamento, raiva e depressão (RUSSELL, 1980).	SVM; NB e J48.

FONTE: O autor (2020).

Visualiza-se no QUADRO 1 que o maior número de modalidades empregadas simultaneamente foi igual a quatro (Schuller *et al*. 2010; Schuller; Dorfner; Rigoll, 2010; Schuller; Weninger; Dofner, 2011). Entretanto, sete diferentes modalidades foram apresentadas (acordes, áudio, gênero musical, letras, metadados, MIDI e voz). Ademais, apenas um trabalho considerou separar a voz da parte instrumental para a extração de características (Imbrasaite, 2015).

No âmbito social, este estudo pode contribuir, por exemplo, com a Musicoterapia no tocante ao tratamento de possíveis distúrbios emocionais e à melhoria de desempenho e bem-estar de indivíduos saudáveis (Huq; Bello; Rowe, 2010). Além disso, pesquisadores que se esforçam no uso da relação “música-emoção” para estudar distúrbios como demência (Baird; Samson, 2015), depressão (Garrido; Schubert, 2015), Alzheimer (Ratovohery *et al*., 2019) e síndrome de Williams (Andrade, 2017) podem se beneficiar desta pesquisa, principalmente caso necessitem de recursos tecnológicos para a coleta, análise e visualização de dados; além de sistemas de recomendação e classificação automática de músicas.

Para além de aplicações relacionadas à saúde, contribuições sociais derivadas desta tese podem associar-se, ainda: a) ao auxílio de sistemas de composição de músicas baseadas em emoções (Ferreira, Whitehead, 2019; Grekow, 2023); b) à otimização da experiência de busca por músicas, resultando na redução do tempo despendido nessa tarefa para os usuários (Deng, S. *et al*., 2015; Sana *et*

al., 2022); e c) à adaptação e ao bem-estar de estudantes em processo de intercâmbio (Vidas, Nelson, Dingle, 2023).

Em termos mercadológicos, verifica-se potencial de aplicação nesta área, posto que o reconhecimento de emoções em músicas pode servir como instrumento para a construção de ações relacionadas às decisões de vendas de produtos, bem como ser uma ferramenta capaz de detectar a emoção do consumidor e, consequentemente, auxiliar na recomendação de produtos e serviços específicos (Bhatti *et al*., 2016; Panwar *et al*., 2018).

Com relação à aderência ao Programa de Pós-Graduação em Gestão da Informação (PPGGI), o presente estudo é a continuação da dissertação desenvolvida pelo autor neste mesmo programa (Moreira, 2019). Além disso, conteúdos presentes no escopo das disciplinas como semiótica, recuperação de informação e inteligência artificial, compõem o estudo. Por fim, verifica-se aderência à linha de pesquisa "Informação e Tecnologia", do PPGGI, cujo objetivo é o "desenvolvimento de métodos, técnicas e ferramentas com vistas à transformação de dados e informações como insumo para o desenvolvimento e aperfeiçoamento dos processos e produtos tecnológicos" (Programa de Pós-Graduação em Gestão da Informação, 2020, p. 6).

Ademais, a temática é uma abordagem interdisciplinar, formada por conhecimentos de áreas como Psicologia, Neurociência, Ciência da Computação, Ciência Cognitiva, Inteligência Artificial (Li *et al*., 2018) e Música, sendo, portanto, aderente à proposta interdisciplinar do PPGGI.

1.4 ORIGINALIDADE E NÃO TRIVIALIDADE DA TESE

Esta tese aborda dois tópicos que são tendências de pesquisa no campo de reconhecimento de emoções em músicas: a) a adoção de uma abordagem multimodal em vez de um processo com uma fonte (ex: áudio); e b) a comparação de resultados classificação obtidos com arquiteturas de *deep learning* frente aos resultados alcançados com algoritmos de aprendizado de máquina (*machine learning*) (Han *et al*., 2022).

Conforme relatado na seção de justificativa, observou-se, por exemplo, a falta de estratégias multimodais para o reconhecimento de emoções em músicas que realizassem a separação do arquivo musical em diferentes modalidades (voz, letras, instrumentos, acordes etc.), possibilitando, assim, a extração de características

específicas para cada modalidade, evitando a consulta em diferentes bases de dados, a redução do conjunto de dados analisado (Imbrasaite, 2015) e a eliminação de modalidades (Schuller *et al*., 2010).

Sendo assim, esta pesquisa preenche esta lacuna, buscando, a partir de um único arquivo, separar e extrair o maior número possível de modalidades relacionadas à música, contribuindo com o aumento de modalidades empregadas simultaneamente, bem como apresenta novos procedimentos de pré-processamento, extração e análise. Entende-se que a separação de fontes é um processo importante, uma vez que a presença de várias fontes em um sinal pode afetar negativamente as tarefas de processamento (Purwins *et al*., 2019).

Por fim, em relação aos métodos empregados, não se constatou o uso de arquiteturas de *deep learning* para a classificação das emoções, sendo o algoritmo SVM o método mais popular, com aplicação em seis dos oito estudos analisados.

1.5 DELIMITAÇÃO DO ESTUDO

Apesar do objetivo de combinar diferentes modalidades para o reconhecimento de emoções em músicas, ressalta-se que este estudo não contempla características pertencentes aos usuários, como indicadores relacionados à frequência cardíaca, à pressão arterial, à frequência respiratória e às expressões faciais, tendo em vista a natureza da pesquisa e do Programa ao qual ela está inserida, bem como a dificuldade da coleta desses elementos, mediante o uso de equipamentos específicos, além do contexto de pandemia, cujas recomendações dos órgãos de saúde recomendavam, no início da pesquisa, a adoção de distanciamento social.

Para mais, neste estudo, ainda que exista a coleta de dados demográficos dos usuários, tais como sexo e idade, esses são considerados apenas para fins descritivos dos avaliadores, não sendo, portanto, trabalhados como modalidades ou elementos de análise quanto a possíveis divergências entre percepções emocionais. Por fim, a presente pesquisa não visa exaurir todos os aspectos psicológicos e neurobiológicos associados às emoções e aos diferentes modelos de emoções relatados na literatura.

Tendo em mente que as plataformas selecionadas para a coleta das músicas deste estudo são compostas, em sua grande maioria, por músicas escritas em inglês,

apenas músicas neste idioma são consideradas. Por fim, apesar deste trabalho analisar diferentes obras musicais, de diferentes gêneros e artistas, ressalta-se que obras de música de concerto - também referenciadas como música clássica ou música erudita -, como, por exemplo, óperas, cantatas e oratórios, não são consideradas.

1.6 ESTRUTURA DO DOCUMENTO

O restante desta tese está organizado da seguinte maneira: a seção 2 apresenta o referencial teórico correspondente aos conceitos básicos para a compreensão desta pesquisa, sendo estes: abordagem multimodal, fundamentos de teoria musical, emoções, reconhecimento de emoções em músicas, *deep learning* e *deep learning* para dados tabulares.

A seção 3 diz respeito aos encaminhamentos metodológicos definidos, demonstrando os procedimentos de coleta e análise de dados, a caracterização da pesquisa, e o percurso metodológico estabelecido para a construção da base de dados. Nesta seção, apresentam-se os dois experimentos conduzidos para a rotulação das bases de dados desta pesquisa, assim como ilustra-se o protocolo de desenvolvido para a elaboração dos experimentos de classificação, considerando tanto algoritmos de aprendizado de máquina, quanto arquiteturas de *deep learning*.

Na seção 4 são relatados e discutidos os resultados obtidos e, por fim, a seção 5 é destinada às considerações finais do estudo, contemplando as contribuições, destacando e apresentando as possibilidades para a continuidade da pesquisa.

2 REFERENCIAL TEÓRICO

Nesta seção são explorados os principais conceitos relacionados ao tema e à problemática desta tese. Apresentam-se: abordagem multimodal; fundamentos de teoria musical; emoções; reconhecimento de emoções em músicas; e *deep learning*. Em relação à seleção desses tópicos, julga-se que todos contribuem para a compreensão dos diferentes conhecimentos envolvidos na construção da pesquisa, e servem para instrumentalizar o autor ao longo do trajeto de desenvolvimento da pesquisa.

Neste sentido, pesquisou-se sobre abordagem multimodal para apresentar o que pode ser definido como uma abordagem ou problema multimodal, buscando, principalmente, compreender qual seria o papel do uso de múltiplas fontes na resolução de um problema. Quanto aos fundamentos de teoria musical, em virtude de o autor desta pesquisa não ser músico, entendeu-se ser necessário compreender os principais elementos deste campo, tendo em vista que as características extraídas para a construção da base de dados representarem conceitos como melodia, harmonia, ritmo, entre outros.

No que diz respeito ao estudo das emoções, buscou-se sistematizar as principais abordagens teóricas do campo – categórica e dimensional -, apresentando algumas definições e os principais modelos emocionais, uma vez que essas abordagens relacionam diretamente com o tema reconhecimento de emoções em músicas, cujas principais etapas para a construção de um sistema para essa temática foram abordadas. Para a subseção sobre *deep learning*, o principal objetivo consistiu em apresentar um conjunto de arquiteturas e o básico do funcionamento destas, posto que a abordagem proposta neste estudo é desenvolvida pensando no uso de diferentes arquiteturas de *deep learning*.

Finalmente, de maneira complementar à subseção das arquiteturas de *deep learning*, arquiteturas de *deep learning*, projetadas para cenários que envolvem dados tabulares, são detalhadas.

2.1 ABORDAGEM MULTIMODAL

Considera-se uma abordagem multimodal ou problema como multimodal a partir do momento em que se trabalha com informações de diferentes fontes

(modalidades) (Mansoorizadeh; Charkari, 2010; Ngiam *et al*., 2011; Baltrušaitis; Ahuja; Morency, 2018). Do mesmo modo, interpreta-se como multimodalidade as etapas de extração e combinação de informações inter-relacionadas de várias modalidades (Mansoorizadeh; Charkari, 2010).

Uma abordagem multimodal demonstra eficácia no estudo e modelagem das interações humano-computador (Worsley; Blikstein, 2015), uma vez que se pode resolver ambiguidades presentes em uma única modalidade (Huang *et al*., 1999). Além disso, é possível que uma modalidade seja insuficiente para refletir todos os aspectos de um fenômeno (Li, C. *et al*., 2015). Sendo assim, empregar recursos multimodais pode fornecer uma compreensão aprofundada dos fenômenos (Baltrušaitis; Ahuja; Morency, 2018).

Em relação à presença de elementos multimodais na sociedade, entende-se que a própria maneira como os seres humanos expressam suas emoções é multimodal (Poria *et al*., 2015, 2016). No dia a dia, modalidades textuais, auditivas e visuais são exploradas de forma simultânea para permitir a extração eficaz das informações semânticas e afetivas transmitidas durante a comunicação (Poria *et al*., 2015, 2016). Por meio de expressões faciais, posturas corporais, entonações e escolha de palavras, por exemplo, as pessoas são extremamente eficientes em expressar sentimentos diferentes (Morency; Mihalcea; Doshi, 2011).

Percebe-se que a abordagem multimodal é um campo multidisciplinar de crescente importância (Poria *et al*., 2015; Soleymani *et al*., 2017; Baltrušaitis; Ahuja; Morency, 2018), com potencial de aplicação em campos como a inteligência artificial (Baltrušaitis; Ahuja; Morency, 2018) e a análise de sentimentos (Poria *et al*., 2015; Soleymani *et al*., 2017).

Estratégias com múltiplas fontes de recursos foram beneficiadas pelos avanços em áreas como o processamento de sinais e a inteligência artificial (Poria *et al*., 2015), além da proliferação das mídias sociais (Soleymani *et al*., 2017). Estes fatores colaboraram para a criação de sistemas inteligentes capazes de detectar e processar informações afetivas contidas em fontes multimodais, como o reconhecimento de emoções e sentimentos em vídeos, por meio de informações visuais, textuais, áudio, modulações vocais e expressões faciais (Poria *et al*., 2015; Soleymani *et al*., 2017).

Finalizada a subseção sobre abordagem multimodal, a próxima subseção corresponde aos fundamentos de teoria musical.

2.2 FUNDAMENTOS DE TEORIA MUSICAL

De maneira sintética, pode-se entender a música como a arte de utilizar os sons para fins expressivos (Solomon, 2019). Outra possível definição consiste em compreendê-la como a combinação de sons de forma simultânea e sucessiva (Med, 2017), mediante “ordem, equilíbrio e proporção dentro do tempo” (Med, 2017, p.11). Por fim, ainda é possível entender a música como uma sucessão de tons dispostos em um ritmo específico (Miller, 2016). Todavia, sabe-se que a matéria-prima da música é o som (Med, 2017), cujas propriedades também se manifestam na música, conforme ilustra-se no QUADRO 2.

QUADRO 2 - PROPRIEDADES DO SOM E SUAS REPRESENTAÇÕES NA MÚSICA

Propriedade	**O que representa no som?**	**O que representa na música?**	**Como identificar na música?**
Altura	Frequência das vibrações. Tem como unidade de medida o Hertz (Hz).	**Melodia:** a partir da alternância de notas com alturas diferentes; **Harmonia:** a partir da simultaneidade de sons com alturas diferentes.	Posição da nota no pentagrama e pela clave.
Duração	Tempo de emissão das vibrações (extensão do som).	**Ritmo:** mediante a alternância de diferentes durações.	Figura da nota e do andamento.
Intensidade	Amplitude das vibrações. Tem como unidade de medida o decibel (dB).	**Dinâmica:** a partir da alternância de intensidades.	Sinais de dinâmica.
Timbre	“Combinação de vibrações determinadas pela espécie do agente que a produz” (Med, 2017, p. 12).	**Instrumentação:** por meio da combinação e da alternância de timbres.	“Pela indicação da voz ou do instrumento que deve executar a música” (Med, 2017, p. 12).

FONTE: Adaptado de Med (2017, p. 12).

Embora existam diferentes definições para a música (Miller, 2016), além de vários atributos perceptivos que a caracterizam (Deng, J. J. *et al*, 2015), é possível estabelecer que existem elementos fundamentais ou primários para a sua compreensão, como a melodia, a harmonia e o ritmo (Meyer, 1996; Priolli, 2006; Med, 2017).

Dentre esses elementos, a melodia caracteriza-se como o “conjunto de sons dispostos em ordem sucessiva” (Med, 2017, p.11), e manifesta-se como a parte mais simples a ser assimilada pelo ser humano ao escutar uma canção (Pilhofer; day,

2019). Para mais, consiste na linha principal de uma música, servindo como base para a construção da harmonia (Pilhofer; Day, 2019).

A harmonia consiste em um “conjunto de sons dispostos em ordem simultânea” (Med, 2017, p.11), e relaciona-se diretamente com a construção dos acordes (Pilhofer; Day, 2019). Por fim, o ritmo é definido como “ordem e proporção em que estão dispostos os sons que constituem a melodia e a harmonia” (Med, 2017, p.11). É por meio do ritmo, por exemplo, que se diferenciam os diversos gêneros musicais (Pilhofer; Day, 2019).

Tendo conhecimentos dos elementos básicos para a compreensão da música, na sequência são explorados os elementos relacionados à notação musical.

2.2.1 Notação musical

Os sons empregados na construção da música são representados a partir de sinais gráficos denominados notas (Zamacois, 2007; Med, 2017), cujo papel é determinar a altura e a duração desses sons (Zamacois, 2007; Med, 2017). A nomenclatura das sete notas musicais existentes é apresentada no QUADRO 3.

QUADRO 3 - NOMENCLATURA DAS NOTAS MUSICAIS

Idioma	**Notas**						
Português	Dó	Ré	Mi	Fá	Sol	Lá	Si
Inglês	C	D	E	F	G	A	B
Alemão	C	D	E	F	G	A	H

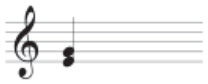
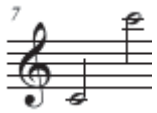
FONTE: Adaptado de Med (2017, p. 272).

A notação das notas musicais (QUADRO 3) é realizada em uma representação denominada pauta musical (Pilhofer; Day, 2019; Solomon, 2019) - ou pentagrama (Zamacois, 2007; Med, 2017) -, que consiste em um conjunto de cinco linhas paralelas horizontais, com quatro espaços entre elas (Pilhofer; Day, 2019).

2.2.2 Intervalos

Um intervalo é a diferença de altura entre dois sons (Pilhofer; Day, 2019; Solomon, 2019). No que diz respeito aos tipos de intervalos, estes são denominados de acordo com os critérios apresentados no QUADRO 4.

QUADRO 4 - TIPOS DE INTERVALOS

Critério	Caraterística	Classificação	Exemplo	Fonte
Notas	Notas tocadas simultaneamente	Harmônico		Pilhofer e Day (2019, p. 104)
	Notas tocadas sucessivamente	Melódico		
Oitava	Dentro de uma oitava	Simples		Pilhofer e Day (2019, p. 120-121)
	Ultrapassa uma oitava	Composto		

FONTE: O autor (2021).

No caso dos intervalos melódicos, estes ainda são divididos em: a) ascendentes: quando “a primeira nota é mais grave do que a segunda” (Med, 2017, p. 60); e b) descendente: quando “a primeira nota é mais aguda do que a segunda” (Med, 2017, p. 60). Com relação à classificação dos intervalos, estes são medidos de forma numérica (quantidade) e qualitativa (qualidade) (Med, 2017; Pilhofer; Day, 2019).

Para se classificar um intervalo de forma numérica, verifica-se a distância entre as notas na pauta musical (Med, 2017; Pilhofer; Day, 2019). Sendo assim, um intervalo simples que contém duas notas, é denominado segunda; com três notas, recebe o nome de terça; e assim até chegar a uma oitava (oito notas). A partir de nove notas (nona), tem-se um intervalo composto, que apresenta a mesma lógica de nomenclatura do intervalo simples: dez notas (décima); 11 notas (décima primeira) etc. (Med, 2017).

Para complementar a classificação numérica, adota-se uma classificação qualitativa, por meio do número de tons e semitons presentes no intervalo (Med, 2017; Pilhofer; Day, 2019). Nessa classificação, as seguintes qualidades são empregadas:

a) justo; b) maior; c) menor; d) aumentado; e e) diminuto (Miller, 2016; Pilhofer; Day, 2019; Solomon, 2019). As características de cada qualidade são apresentadas no QUADRO 5.

QUADRO 5 - INTERVALOS DE ACORDO COM AS SUAS QUALIDADES

Qualidade	**Intervalos**
Justo/Perfeito	Primeira, quarta, quinta e oitava
Maior	Segunda, terça, sexta e sétima
Menor	
Aumentado	Todos
Diminuto	Todos, com exceção à primeira

FONTE: Adaptado de Pilhofer e Day (2019, p.107).

Quanto às características dos intervalos justos, estas são listadas no QUADRO 6.

QUADRO 6 - INTERVALOS JUSTOS

Intervalo	**Característica**
Primeira	Dois sons com o mesmo nome e altura
Quarta	Dois tons e um semitom
Quinta	Três tons e um semitom
Oitava	Cinco tons e dois semitons

FONTE: Adaptado de Med (2017, p. 62-63).

Em relação às características dos intervalos maiores e menores, estas são listadas no QUADRO 7.

QUADRO 7 - INTERVALOS MAIORES E MENORES

Intervalo (maior)	**Característica**	**Intervalo (menor)**	**Característica**
Segunda	Um tom	Segunda	Um semitom
Terça	Dois tons	Terça	Um tom e um semitom
Sexta	Quatro tons e dois semitons	Sexta	Três tons e dois semitons
Sétima	Cinco tons e um semitom	Sétima	Quatro tons e dois semitons

FONTE: Adaptado de Med (2017, p. 64-67).

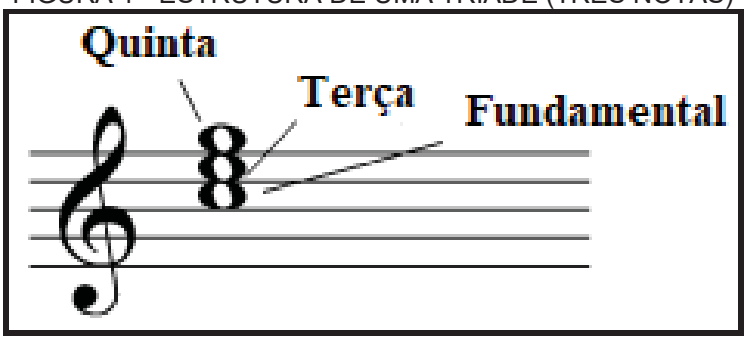
Finalmente, quanto aos intervalos aumentados e diminutos, é possível diferenciá-los da seguinte forma: a) intervalo aumentado: é aquele com um semitom cromático a mais do que os intervalos justos ou maiores (Med, 2017); e b) intervalo diminuto: possui um semitom cromático a menos do que os intervalos justos ou menores (Med, 2017).

2.2.3 Acordes

Um acorde consiste no conjunto de três ou mais notas tocadas simultaneamente (Miller, 2016; Pilhofer; Day, 2019; Solomon, 2019). No que tange aos tipos de acordes, é possível classificá-los de acordo com o "intervalo existente entre as notas extremas" (Med, 2017, p.275). Todavia, apesar de variações, os acordes mais comuns são os formados a partir de três notas - tríade ou acordes de quinta - ou quatro notas (acordes de sétima) (Solomon, 2019).

Quanto à estrutura, uma tríade (FIGURA 1) é construída a partir de três notas, que recebem os seguintes nomes: a) fundamental (*root* - em inglês): nota mais baixa da tríade e a responsável por nomear o acorde (Miller, 2016; Pilhofer; Day, 2019; Solomon, 2019); b) terça (*third* - em inglês): nota intermediária, situando-se uma terça acima da nota fundamental (Miller, 2016; Pilhofer; Day, 2019; Solomon, 2019); e c) quinta (*fifth* - em inglês): nota superior, situada uma quinta acima da fundamental (Miller, 2016; Pilhofer; Day, 2019; Solomon, 2019).

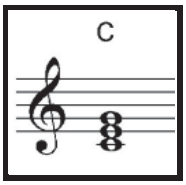
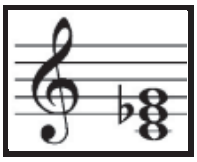
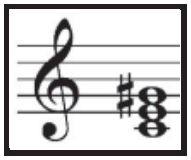
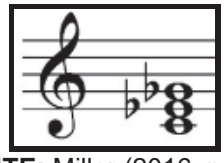
FIGURA 1 - ESTRUTURA DE UMA TRÍADE (TRÊS NOTAS)



FONTE: Adaptado de Miller (2016, p.149).

No que tange à tipologia, as tríades são divididas em quatro grupos, respeitando as seguintes qualidades: a) maior; b) menor; c) aumentada; e d) diminuta (Miller, 2016; Pilhofer; Day, 2019; Solomon, 2019). A formação e a representação destes grupos são apresentadas no QUADRO 8.

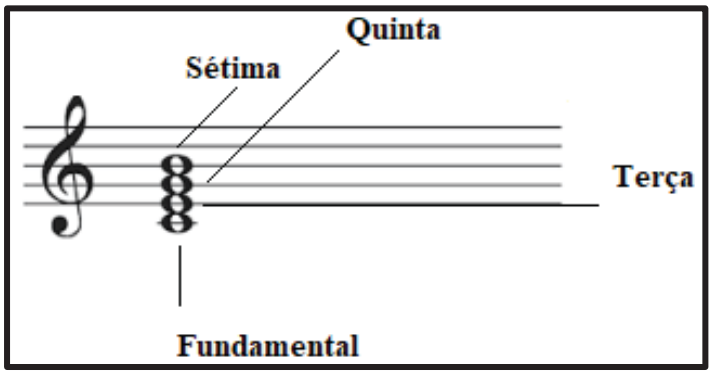
QUADRO 8 - FORMAÇÃO E REPRESENTAÇÃO DAS TRÍADES

Tríade	Formação	Exemplo
Maior	Uma terça maior + uma terça menor	 **FONTE:** Miller (2016, p.151)
Menor	Uma terça menor + uma terça maior	 **FONTE:** Miller (2016, p.152)
Aumentada	Duas terças maiores	**FONTE:** Miller (2016, p.154)
Diminuta	Duas terças menores	**FONTE:** Miller (2016, p.153)

FONTE: Adaptado de Med (2017, p.277-278).

Além do mais, as tríades maiores e menores são acordes consonantes (Med, 2017; Solomon, 2019), ou seja, transmitem uma sensação de “estabilidade, relaxamento e tranquilidade” (Med, 2017, p.277). Em contrapartida, as tríades aumentadas e diminutas são acordes dissonantes (Med, 2017; Solomon, 2019), uma vez que “criam a impressão de intranquilidade e tensão” (Med, 2017, p.277).

Além das tríades, o outro tipo de acorde comumente empregado consiste nos acordes de sétima - ou tétrade (Med, 2017) -, cuja estrutura (FIGURA 2) é semelhante à tríade, porém, apresentando uma nota acima da quinta - sétima (*seventh* - em inglês) -, por estar situado no intervalo de uma sétima acima da nota fundamental (Pilhofer; Day, 2019; Solomon, 2019).

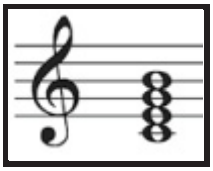
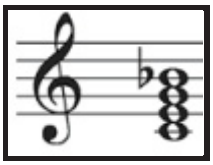
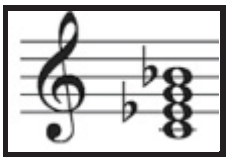
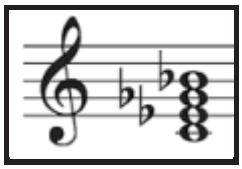
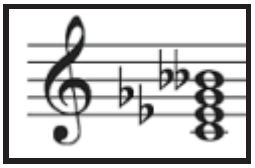
FIGURA 2 - ESTRUTURA DE UM ACORDE DE SÉTIMA



FONTE: Adaptado de Pilhofer e Day (2019, p.138).

Referente aos tipos de acordes de sétima, o QUADRO 9 contempla alguns exemplos de formação e representação destes acordes.

QUADRO 9 - FORMAÇÃO E REPRESENTAÇÃO DE ACORDES DE SÉTIMA

Acorde	Formação	Exemplo
Sétima maior	Tríade maior + sétima maior	 **FONTE:** Pilhofer e Day (2019, p.138)
Sétima menor	Tríade menor + sétima menor	**FONTE:** Pilhofer e Day (2019, p.138)
Sétima dominante	Tríade maior + sétima menor	 **FONTE:** Pilhofer e Day (2019, p.139)
Acorde de sétima da sensível	Quinta diminuta + sétima menor	**FONTE:** Pilhofer e Day (2019, p.140)
Sétima diminuta	Quinta diminuta + sétima diminuta	**FONTE:** Pilhofer e Day (2019, p.140)

FONTE: Adaptado de Med (2017, p.347-348).

Ademais, além das tríades e dos acordes de sétima, quando necessário, é possível trabalhar com acordes mais complexos, como os acordes de nona, acordes com décima primeira e acordes com décima terceira (Miller, 2016; Med, 2017; Pilhofer; Day, 2019).

Finalizada a subseção acerca dos fundamentos de teoria musical, a seguir discute-se os principais elementos relacionados às emoções. Complementos aos conceitos de teoria musical apresentados nesta subseção, como valores, acidentes e escalas, são fornecidos e detalhados no APÊNDICE A.

2.3 EMOÇÕES

Emoções desempenham um papel central na vida humana (Adolphs; Anderson, 2018; Shackman; Wager, 2019), apresentando-se como um dos aspectos mais fascinantes e, ao mesmo tempo, misteriosos da função cerebral (Zych; Gogolla, 2021). A busca por compreendê-las motiva, por exemplo, discussões filosóficas desde a Grécia Antiga (Scarantino, 2016), por meio de filósofos como Platão e Aristóteles (Lyons, 1999; Dias; Cruz; Fonseca, 2008; Rosenwein; Cristiani, 2018). Posteriormente, filósofos como René Descartes e Baruch Spinoza também discorreram acerca do tema (Lyons, 1999; Plamper, 2015; Rosenwein; Cristiani, 2018).

Além dos filósofos, a necessidade de se entender as emoções desperta a atenção de cientistas, médicos, bem como do público em geral (Shackman; Wager, 2019). Todavia, uma definição precisa do que é uma emoção é uma tarefa complexa (Schmeichel; Inzlicht, 2013), tendo este conceito diferentes definições (Kleinginna; Kleinginna, 1981; Mulligan; Scherer, 2012; Celeghin *et al*., 2017), as quais variam muito entre as abordagens teóricas (Lange *et al*., 2020).

Como exemplo da dificuldade em se obter um consenso sobre o conceito de emoção, pode-se recorrer à pesquisa de Kleinginna e Kleinginna (1981), na qual os autores identificaram 92 definições diferentes para emoções. Para mais, desde a publicação do ensaio "*What is an emotion*" (James, 1884), diversos autores buscam uma definição para o tema (Dias; Cruz; Fonseca, 2008; Lange *et al*., 2020), mesmo este sendo um assunto complexo para se explicar cientificamente (Adolphs; Anderson, 2018).

Entretanto, mesmo com divergências quanto a uma definição, é possível ilustrar, minimamente, o que alguns autores entendem por emoções, ao ter contato com algumas definições (QUADRO 10).

QUADRO 10 - EXEMPLOS DE DEFINIÇÕES DE EMOÇÕES

Definição	Fonte
Estados psicológicos formados por elementos como experiência pessoal, reações fisiológicas, comportamentais e comunicativas.	Akçay e Oğuz (2020, p. 58, tradução nossa)
Reação afetiva intensa e com curta duração. É composta, geralmente, por subcomponentes - subjetividade, excitação fisiológica, expressão, tendência de ação e regulação - com um certo grau de "sincronização".	Juslin (2013a, p. 236, tradução nossa)
Estados cognitivos ou fisiológicos que acontecem para antecipar ou responder a informações afetivas, que refletem o valor e / ou a intensidade dos estímulos eliciadores.	Kryklywy *et al*. (2020, p. 120, tradução nossa)
Processos psiconeurais que são especialmente influentes no controle do vigor e da padronização das ações.	Panksepp (1998, p. 44, tradução nossa)
Respostas multifacetadas que envolvem mudanças na experiência consciente, no comportamento e na fisiologia.	Schmeichel e Inzlicht (2013, p. 281, tradução nossa)
Conjuntos generalizados e, teoricamente, coordenados de processos comportamentais, fisiológicos, cognitivos e afetivos, selecionados para promover respostas automáticas e adaptativas a eventos ambientais recorrentes que apresentam desafios de aptidão.	Shariff e Tracy (2011, p.396, tradução nossa)
Conceito “guarda-chuva” que se caracteriza por mudanças afetivas, cognitivas, comportamentais, expressivas e fisiológicas.	Tyng *et al*. (2017, p. 3, tradução nossa)

FONTE: O autor (2021).

Embora as definições elencadas no QUADRO 10 apresentem particularidades, é possível visualizar elementos em comum para os autores, como a presença de dimensões psicofisiológicas, processos perceptuais-cognitivos, percepção subjetiva e motivação para ação.

No campo científico, Charles Darwin, com a publicação do livro “*The expression of the emotions in man and animals*” (Darwin, 1872), pode ser considerado o primeiro cientista moderno das emoções (Moors *et al*., 2013; Rosenwein; Cristiani, 2018; Gross; Preston, 2020). Ao escrever esse livro, Darwin argumentava que as emoções eram aspectos herdados e conservados evolutivamente (Barrett, 2017; Rosenwein; Cristiani, 2018; Adolphs; Anderson, 2018; Gross, Preston, 2020).

Darwin visualizava, ainda, as emoções - bem como as suas expressões - como parte de uma natureza humana universal (Barrett, 2017), cuja evolução destinava-se a comunicar os estados internos dos animais e, ao mesmo tempo, refletir processos fisiológicos subjacentes e adaptativos (Gross; Preston, 2020). Essa visão serviu de base para, por exemplo, os estudos sobre expressão emocional, conduzidos

por pesquisadores como Paul Ekman e Wallace Friesen (Rosenwein; Cristiani, 2018; Zych; Gogolla, 2021). A hipótese central desses estudos consistia em entender que algumas expressões - principalmente as faciais - eram adaptações evoluídas biologicamente e básicas quanto à sua forma e à sua função (Barrett, 2011).

Por fim, algumas das ideias de Darwin ainda são relevantes para o estudo atual das emoções, como: a) a função comunicativa das emoções; b) a homologia entre as espécies; c) a importância no comportamento social; d) o enraizamento fisiológico; e e) as funções adaptativas das emoções (Gross; Preston, 2020, p.179).

Além de abordagens inspiradas por Darwin, o estudo das emoções apresenta períodos em que uma abordagem é dominante, conforme defendem Moors *et al*. (2013). Para os autores, é possível citar, por exemplo, o domínio do pensamento conhecido como *Appraisal Theory*[3] durante a década de 1960, quando pesquisadores como Magda Arnold e Richard Lazarus, ao sistematizarem os pensamentos de filósofos como Aristóteles e Baruch Spinoza (Moors *et al*., 2013), argumentaram sobre a existência de um processo de "avaliação" de estímulos - uma espécie de "etapa adicional" entre o ser humano e mundo - no qual o organismo é informado se o estímulo observado é elegível para despertar uma emoção (Moors, 2010; Barrett, 2017).

Todavia, ao longo dos anos, duas abordagens se destacaram, concentrando os estudos acerca das emoções, sendo essas: a) a abordagem categórica ou discreta; e b) a abordagem dimensional (Scherer, 2005; Dias, Cruz; Fonseca, 2008; Fanselow, 2018).

2.3.1 Abordagem categórica das emoções

A abordagem categórica - ou discreta - das emoções relaciona-se diretamente com o conceito das emoções básicas[4] (Scherer, 2005; Eerola; Vuoskoski, 2013), cuja principal característica é representar as emoções em um grupo finito de termos (categorias) (Eerola; Vuoskoski, 2013; Saarimäki *et al*., 2016; Celeghin *et al*., 2017). Ademais, trata-se de uma abordagem fundamentada nas ideias evolucionistas de Charles Darwin (Scherer, 2005; Crivelli; Fridlund, 2019; Keltner *et al*., 2019), na qual

[3] Optou-se por manter a expressão no idioma original. Em português, por exemplo, encontra-se como possível tradução a expressão "Teoria da Avaliação".
[4] Também denominadas como emoções primárias ou fundamentais (Ortony; Turner, 1990).

as emoções são vistas como uma série de estados separados e discretos (Ekman, 1992; Gu *et al*., 2019), responsáveis por diferentes respostas e padrões fisiológicos, comportamentais e neurais (Barrett, 2006; Ekman; Cordaro, 2011; Colombetti, 2014; Saarimäki *et al*., 2016; Hutto; Robertson; Kirchhoff, 2018).

Já em relação à quantidade exata de categorias para representar as emoções nessa abordagem, constata-se a falta de consenso entre os pesquisadores acerca desse ponto (Gu *et al*., 2019). Entretanto, com o passar dos anos, seis emoções se consolidaram como "universais", sendo essas: a) raiva; b) medo; c) nojo; d) felicidade; e) tristeza; e f) surpresa (Saarimäki *et al*., 2016; An *et al*., 2017; Wang *et al*., 2019; Wingenbach *et al*., 2020).

A popularidade deste conjunto é resultado das conclusões apresentadas em pesquisas sobre expressões faciais (An *et al*., 2017; Digirolamo; Russell, 2017; Wingenbach *et al*., 2020). Esses estudos, geralmente, consistem em apresentar uma série de fotos de diferentes expressões faciais a uma determinada amostra, solicitando que, para cada imagem, seja determinada a emoção mais adequada para representar a foto visualizada (Colombetti, 2014; Digirolamo; Russell, 2017).

Dos estudos com o objetivo de associar emoções a expressões faciais, o mais famoso e influente foi conduzido por Ekman e Friesen (1971) (Colombetti, 2014; Keltner *et al*., 2019). Neste estudo, os autores apresentaram imagens com diferentes expressões faciais relacionadas às emoções raiva, medo, felicidade, nojo, tristeza e surpresa para membros de uma tribo na Nova Guiné que, após ouvirem pequenas histórias acerca do que cada emoção significava, indicaram a imagem mais pertinente para representar a emoção descrita na história (Ekman; Friesen, 1971; Colombetti, 2014).

De todo modo, mesmo com indícios para sustentar a hipótese de existência de seis emoções básicas, diversos autores propuseram ao longo dos anos diferentes conjuntos de emoções (QUADRO 11), envolvendo grupos compostos por apenas duas emoções (Mowrer, 1960) até alternativas maiores, com dez emoções (Arnold, 1960).

QUADRO 11 - EXEMPLOS DE EMOÇÕES BÁSICAS

Autores	Emoções
Arnold (1960)	Raiva (*Anger*), Aversão (*Aversion*), Coragem (*Courage*), Desânimo (*Dejection*), Desejo (*Desire*), Desespero (*Despair*), Medo (*Fear*), Ódio (*Hate*), Esperança (*Hope*), Amor (*Love*), Tristeza (*Sadness*)
Ekman, Friesen e Ellsworth (1982)	Raiva (*Anger*), Nojo (*Disgust*), Medo (*Fear*), Alegria (*Joy*), Tristeza (*Sadness*), Surpresa (*Surprise*)
Gray (1982)	Raiva e Terror (*Rage and Terror*), Ansiedade (*Anxiety*), Alegria (*Joy*)
Izard (1971)	Raiva (*Anger*), Desprezo (*Contempt*), Nojo (*Disgust*), Angústia (*Distress*), Medo (*Fear*), Culpa (*Guilt*), Interesse (*Interest*), Alegria (*Joy*), Vergonha (*Shame*), Surpresa (*Surprise*)
James (1884)	Medo (*Fear*), Pesar (*Grief*), Amor (*Love*), Raiva (*Rage*)
McDougall (1926)	Raiva (*Anger*), Nojo (*Disgust*), Euforia (*Elation*), Medo (*Fear*), Sujeição (*Subjection*), Ternura (*Tender-emotion*), Admiração (*Wonder*)
Mowrer (1960)	Dor (*Pain*) e Prazer (*Pleasure*)
Oatley e Johnson-Laird (1987)	Raiva (*Anger*), Nojo (*Disgust*), Ansiedade (*Anxiety*), Felicidade (*Happiness*), Tristeza (*Sadness*)
Panksepp (1982)	Expectativa (*Expectancy*), Medo (*Fear*), Raiva (*Rage*), Pânico (*Panic*)
Plutchik (1980)	Aceitação (*Acceptance*), Raiva (*Anger*), Antecipação (*Anticipation*), Nojo (*Disgust*), Alegria (*Joy*), Medo (*Fear*), Tristeza (*Sadness*), Surpresa (*Surprise*)
Tomkins (1984)	Raiva (*Anger*), Interesse (*Interest*), Desprezo (*Contempt*), Nojo (*Disgust*), Angústia (*Distress*), Medo (*Fear*), Alegria (*Joy*), Vergonha (*Shame*), Surpresa (*Surprise*)
Watson (1930)	Medo (*Fear*), Amor (*Love*), Raiva (*Rage*)
Weiner e Graham (1984)	Medo (*Fear*), Amor (Love), Raiva (Rage), Felicidade (*Happiness*), Tristeza (*Sadness*)

FONTE: Adaptado de Ortony e Turner (1990, p. 316).

O uso do adjetivo “básico” empregado nesta abordagem corresponde às suposições de que as emoções: a) são “universais” e estão presentes em todas as culturas conhecidas (Ortony; Turner, 1990); b) estão associadas universalmente a determinadas expressões faciais específicas (Ekman; Friesen, 1971; Ortony; Turner, 1990; Wingenbach *et al*., 2020); c) relacionam-se com funções biológicas ligadas às necessidades básicas de sobrevivência (Ortony; Turner, 1990; Gu *et a*l., 2019); e d) evoluíram de maneira a se adaptarem a determinadas situações humanas, como perdas, frustrações, sucessos e alegrias (Ekman; Cordaro, 2011).

Quanto às características que diferenciam as emoções básicas de outros estados afetivos, como o humor, por exemplo, Ekman e Cordaro (2011) apresentam

que quase todas as emoções básicas apresentam até 13 elementos representativos, conforme elenca-se no QUADRO 12.

QUADRO 12 - CARACTERÍSTICAS DAS EMOÇÕES BÁSICAS DE ACORDO COM EKMAN E CORDARO (2011)

#	**Característica**
1	Sinais universais distintos
2	Fisiologia distinta
3	Avaliação automática
4	Eventos antecedentes universais distintos
5	Presença em outros primatas
6	Possibilidade de início rápido
7	Pode ser de curta duração
8	Ocorrência inesperada
9	Pensamentos, memórias e imagens distintas
10	Experiência subjetiva distinta
11	O período refratário filtra as informações disponíveis a fim de ajustar a emoção sentida
12	Sem restrições quanto ao alvo da emoção
13	A emoção pode ser representada de forma construtiva ou destrutiva

FONTE: Adaptado de Ekman e Cordaro (2011, p. 365).

Contudo, é preciso atentar-se que, apesar da teoria das emoções básicas ter colaborado com a Psicologia em áreas como o estudo de expressões faciais, Neuroimagem e Psicologia Evolutiva, essa abordagem sofre com críticas no que se refere aos seus procedimentos metodológicos e às suas bases teóricas (Hutto; Robertson; Kirchhoff, 2018). Neste sentido, por exemplo, Scherer (2005) visualiza a dificuldade em se encontrar um número ideal para se representar todas as emoções existentes.

Fanselow (2018) defende que esta visão pode levar a uma proliferação de emoções individuais caso não haja regras para limitar o que se pode entender como uma emoção. Por seu turno, Colombetti (2014) questiona nessa abordagem o fato de que, se existe um pequeno número de emoções consideradas básicas, então como ocorre a relação entre essas e aquelas que não são consideradas básicas?

Por fim, em relação às expressões faciais como “prova” da existência de determinadas emoções, Keltner *et al*. (2019) concentram a sua crítica, principalmente, ao trabalho de Ekman e Friesen (1971), argumentando que as pessoas expressam suas emoções de diferentes maneiras - como pistas vocais, movimentos corporais, gestos, postura etc. - e não apenas por meio de movimentos dos músculos faciais disparados com palavras ou situações.

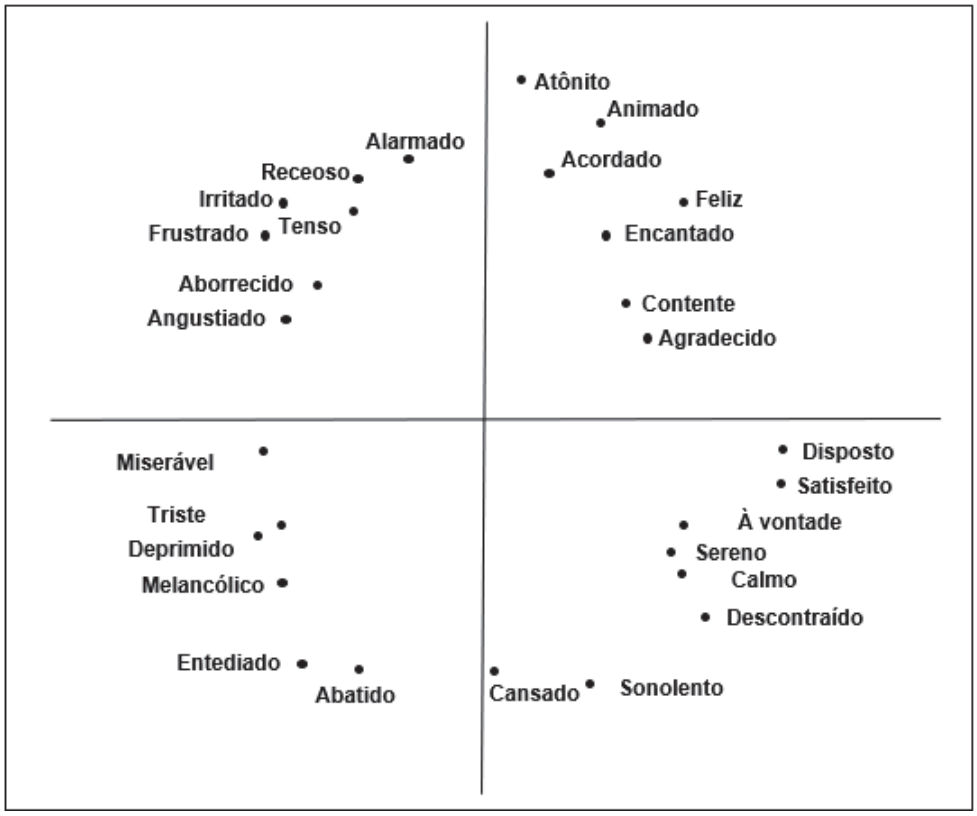
2.3.2 Abordagem dimensional das emoções

A abordagem dimensional das emoções consiste em um conjunto de modelos parcimoniosos (Bradley *et al*., 1992; Fanselow, 2018) que propõem a representação das emoções em um pequeno número de dimensões em vez de assumir a existência de estados emocionais específicos e independentes como os adotados na abordagem categórica (Bradley *et al*., 1992; Colombetti, 2014). Trata-se de uma abordagem cujo marco inicial consiste nos estudos do psicólogo alemão Wilhelm Wundt (Scherer, 2000, 2005), que apresentou um modelo tridimensional formado pelas seguintes dimensões: a) agradável-desagradável; b) tenso-relaxado; e c) animado-calmo (Wundt, 1922; Scherer, 2005; An *et al*., 2017; Diriwächter, 2021).

A partir desse modelo, pode-se visualizar diferentes modelos dimensionais, dos quais três apresentam-se como dominantes nos estudos sobre emoções (Rubin; Talarico, 2009; An *et al*., 2017), sendo esses: a) o modelo circumplexo, desenvolvido por Russell (1980); b) o modelo vetorial, elaborado por Bradley *et al*. (1992); e c) o modelo denominado *Positive Activation – Negative Activation* (PANA), proposto por (Watson; Tellegen, 1985).

No que diz respeito ao modelo circumplexo (FIGURA 3), este modelo foi proposto pelo psicólogo norte-americano James A. Russell (Russell, 1980), e tem como principal característica a representação de 28 termos emocionais como um único ponto em um espaço bidimensional composto pelas dimensões valência (*valence*), situada no eixo horizontal; e excitação (*arousal*), localizada no eixo vertical (Russell, 1980; Barrett, 2017). Ademais, a dimensão da valência varia entre agradável e desagradável, enquanto os graus de excitação variam de calmante a excitante (Hinojosa *et al*. 2016).

FIGURA 3 - MODELO CIRCUMPLEXO (Russell, 1980)

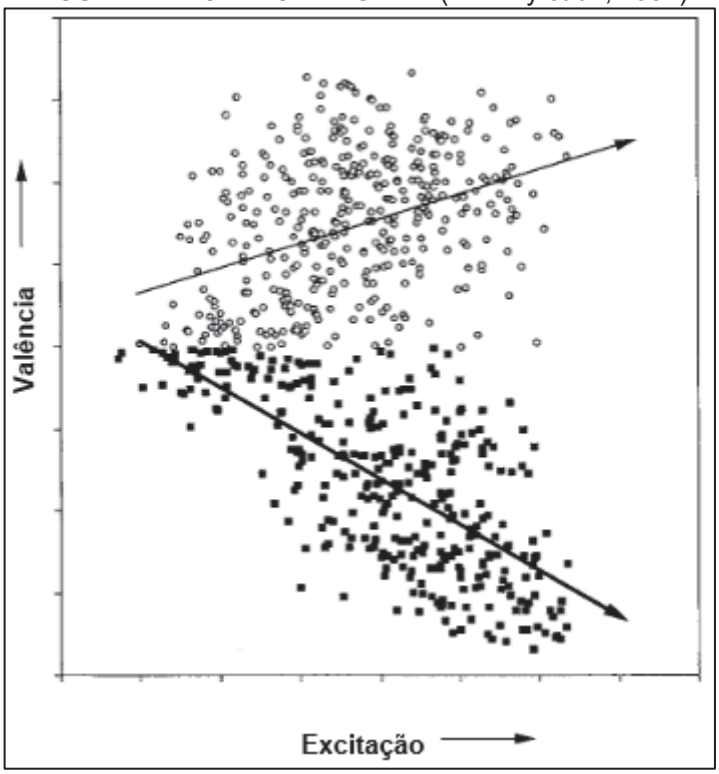


FONTE: Adaptado de Russell (1980, p. 1169).

Com base na FIGURA 3, pode-se analisar o modelo circumplexo da seguinte maneira: a) o primeiro quadrante é formado por estados com valência positiva (+) e excitação positiva (+) (ex: encantado e feliz); b) o segundo quadrante contempla os estados com valência positiva (+) e excitação negativa (-) (ex: satisfeito e calmo); c) no terceiro quadrante encontram-se os termos com valência negativa (-) e excitação negativa (-) (ex: entediado e deprimido); e d) finalmente, no quarto quadrante estão os estados com valência negativa (-) e excitação positiva (+) (ex: tenso e irritado).

Com relação ao modelo vetorial, trata-se de uma abordagem bidimensional elaborada por Bradley *et al*. (1992). Consiste em um modelo composto pelas dimensões de valência (eixo vertical) e excitação (eixo horizontal), no qual há dois vetores com início nos pontos relacionados à excitação zero e à valência neutra, com um dos vetores prosseguindo para a valência positiva, enquanto o outro direciona-se à valência negativa (Rubin; Talarico, 2009) (FIGURA 4).

FIGURA 4 - MODELO VETORIAL (Bradley *et al*., 1992)

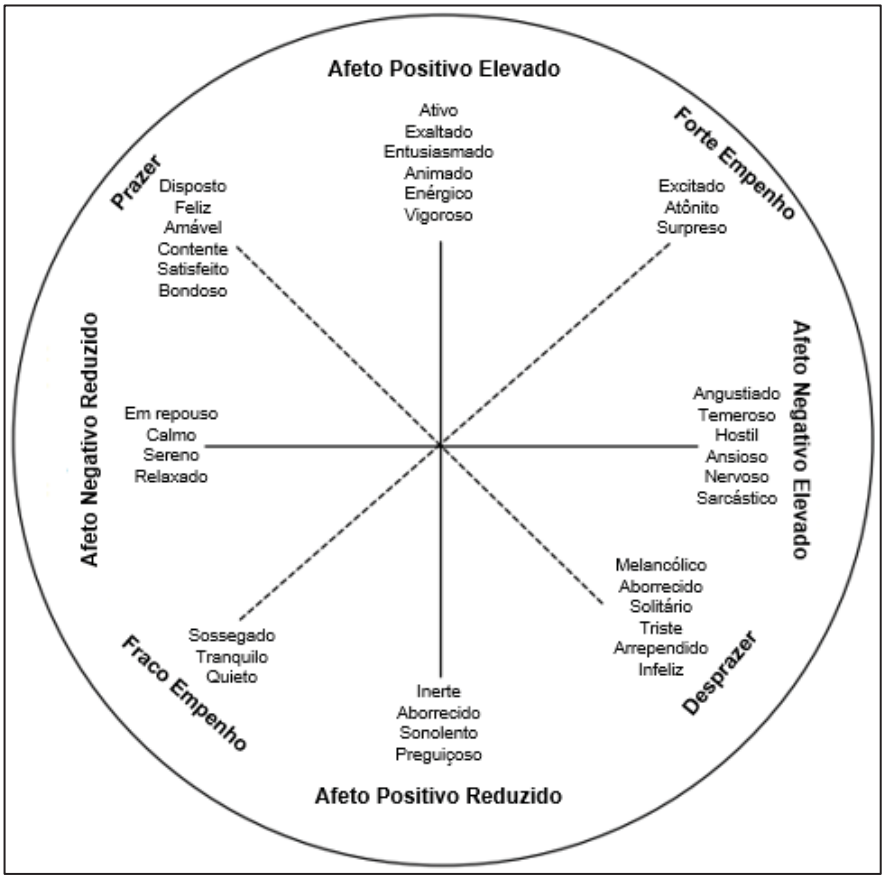


FONTE: Adaptado de Bradley *et al*. (2001, p. 277).

Como resultado, é possível observar uma forma de “bumerangue” (Bradley *et al*., 2001; Talarico; Labar; Rubin, 2004) que representa o direcionamento aos pontos de maior excitação, e que refletem a base motivacional dos julgamentos afetivos (Bradley *et al*., 2001).

A respeito do modelo PANA (Watson; Tellegen, 1985), este esquema reflete a experiência emocional em duas dimensões formadas pelos fatores principais intitulados afeto positivo (eixo vertical) e afeto negativo (eixo horizontal) (FIGURA 5). Para mais, há duas linhas tracejadas neste modelo denominadas como prazer/desprazer e forte empenho/fraco empenho, que consistem em fatores secundários, obtidos posteriormente aos dois fatores principais do modelo (Watson; Tellegen, 1985).

FIGURA 5 - MODELO PANA (Watson; Tellegen, 1985)



FONTE: Adaptado de Watson e Tellegen (1985, p. 221).

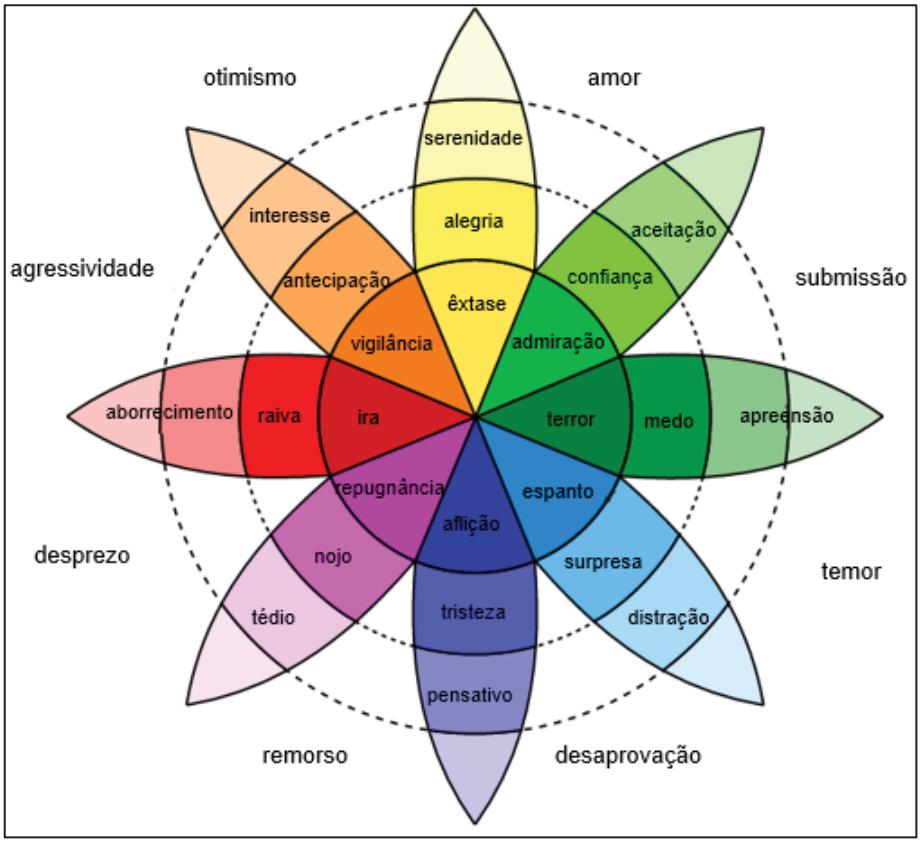
Neste modelo, os termos que se encontram no mesmo espaço apresentam uma alta correlação positiva, ao passo que aqueles em divisões adjacentes demonstram uma correlação positiva moderada. Ademais, palavras localizadas a 90° não apresentam correlação com as demais enquanto aquelas situadas a 180° são opostas em significado, com uma alta correlação negativa (Watson; Tellegen, 1985).

Além dos três modelos já apresentados, a abordagem dimensional das emoções apresenta outros modelos dimensionais, quer sejam bidimensionais (Plutchik, 1980, Thayer, 1989, 2012), tridimensionais (Schlosberg, 1954; Plutchik, 1970; Russell; Mehrabian, 1977; Lövheim, 2012) e até modelos quadridimensionais (Fontaine *et al*., 2007).

Em relação ao modelo bidimensional apresentado por Plutchik (1980) (FIGURA 6), este é popularmente conhecido como “Roda das Emoções de Plutchik” (Acheampong; Wenyu Nunoo-Mensah, 2020; Shiraishi; Reilly, 2020) e compreende oito emoções básicas (alegria, tristeza, confiança, nojo, medo, raiva, surpresa e

antecipação) (Plutchik, 1980; Acheampong; Wenyu Nunoo-Mensah, 2020; Shiraishi; Reilly, 2020).

FIGURA 6 - RODA DAS EMOÇÕES DE PLUTCHIK

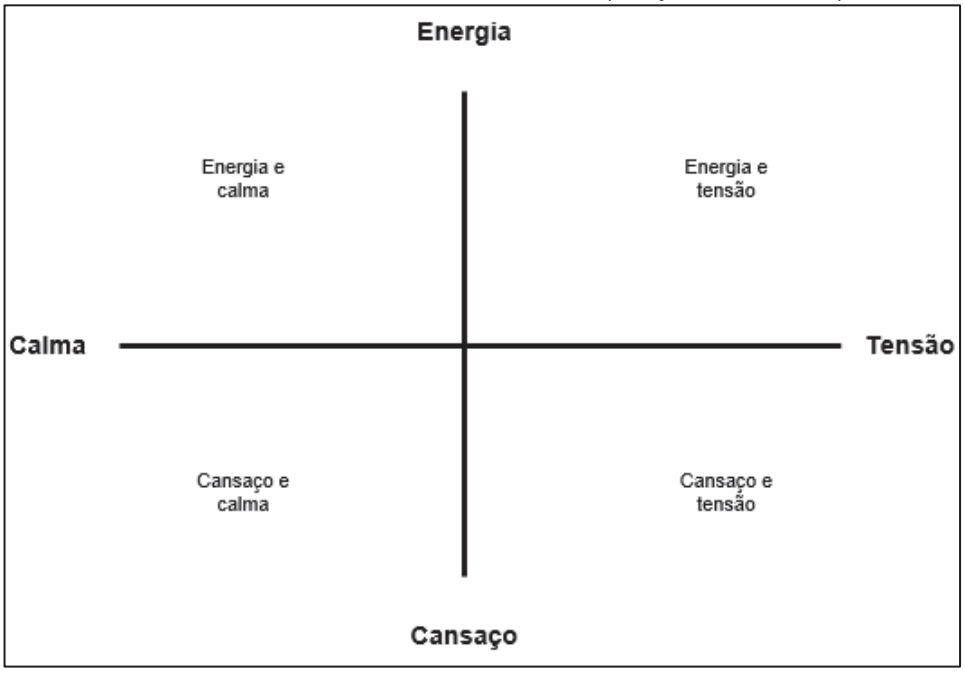


FONTE: Adaptado de Wang *et al*. (2020, p. 13:5).

O modelo ilustrado na FIGURA 6 apresenta as seguintes características: a) emoções internas são mais intensas, como é o caso da alegria em relação à serenidade, ou admiração no tocante à confiança, por exemplo; b) a combinação das emoções básicas produz novas emoções (ex: alegria + confiança = amor; tristeza + nojo = remorso etc.); e c) a presença de polos opostos entre as emoções (alegria x tristeza; confiança x nojo; medo x raiva; e antecipação x surpresa) (Cohen-Kalaf *et al*., 2021).

Outro modelo bidimensional a ser destacado é o proposto por Thayer (1989, 2012) (FIGURA 7). Esse modelo apresenta duas dimensões bipolares nomeadas como energia/cansaço e tensão/calma. A partir dessas dimensões, o autor argumenta que é possível obter quatro estados de humor conhecidos como: a) energia e calma; b) energia e tensão; c) cansaço e calma; e d) cansaço e tensão (Thayer, 1989, 2012).

FIGURA 7 - MODELO BIDIMENSIONAL (Thayer 1989, 2012)



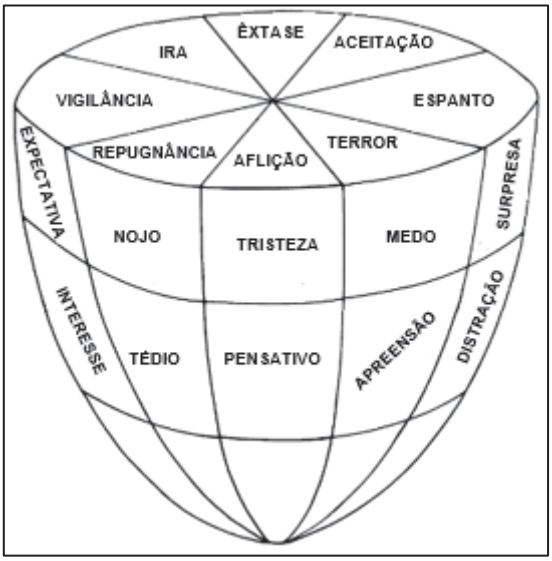
FONTE: Adaptado de Thayer (2012, p. 410).

No âmbito dos modelos tridimensionais, Schlosberg (1954) refinou em seu estudo o modelo proposto por Wundt (1922), apresentando as dimensões prazer-desprazer, atenção-rejeição e ativação (Schlosberg, 1954, An *et al*., 2017), com dimensão ativação apresentando dois níveis denominados como descanso e tensão (Schlosberg, 1954).

Russell e Mehrabian (1977) propuseram a distribuição dos estados afetivos em três dimensões denominadas prazer, valência e dominância, elaborando, assim, um modelo conhecido como PAD (*Pleasure-Arousal-Dominance*). Neste modelo, a dimensão correspondente ao prazer refere-se a sentir-se bem, feliz, satisfeito ou alegre; excitação diz respeito ao quanto um indivíduo se sente estimulado, excitado, atento ou ativo; por fim, dominância representa o grau em que um indivíduo se sente influente, no controle ou importante (Andersson *et al*., 2012).

No que concerne ao modelo tridimensional proposto por Plutchik (1970), sua representação consiste em um cone invertido (FIGURA 8), e abarca três características: a) intensidade, responsável por distinguir emoções como o medo e o pânico, por exemplo; b) graus de similaridade, tornando emoções como vergonha e culpa semelhantes; e c) polaridade, indicando que emoções como alegria e tristeza como opostas (Plutchik, 1970).

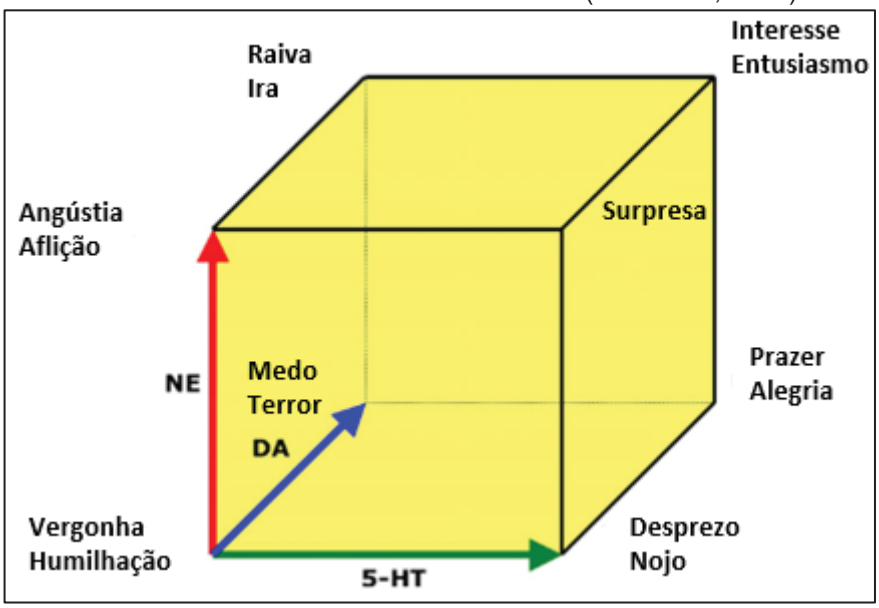
FIGURA 8 - MODELO TRIDIMENSIONAL
(Plutchick, 1970)



FONTE: Adaptado de Plutchik (1970, p. 10).

Em seu modelo tridimensional, Lövheim (2012) representa oito emoções básicas (interesse/excitação, prazer/alegria, surpresa, angústia, medo/terror, vergonha/humilhação, desprezo/repulsa e raiva) em um sistema de coordenadas ortogonais correspondentes aos três eixos monoaminérgicos principais: a) Serotonina/5-hidroxitriptamina (5-HT); b) Dopamina (DA); e c) Noradrenalina (NE) (FIGURA 9).

FIGURA 9 - MODELO TRIDIMENSIONAL (LÖVHEIM, 2012)



FONTE: Adaptado de Lövheim (2012, p. 342).

Por fim, o modelo quadridimensional de Fontaine *et al*. (2007) (ANEXO A) considera as dimensões avaliação-agradabilidade, controle-potência, ativação-excitação e imprevisibilidade, englobando 24 termos afetivos distribuídos em três painéis: a) valência x potência; b) valência x excitação; e c) valência x imprevisibilidade. Neste modelo, a dimensão avaliação-agradabilidade é composta pela dinâmica de oposição entre as emoções agradáveis e desagradáveis.

A segunda dimensão, controle-potência, caracteriza-se por avaliações de controle, englobando sentimentos de poder ou fraqueza, por exemplo. A terceira dimensão, ativação-excitação, tem como característica, principalmente, elementos como batimento cardíaco acelerado e prontidão para a ação. A quarta e última dimensão, imprevisibilidade, apresenta avaliações de novidade, representadas, por exemplo, em ações como o levantar de sobrancelhas e exclamações espontâneas (Fontaine *et al*., 2007).

Uma síntese dos modelos dimensionais de emoção apresentados neste estudo é demonstrada no QUADRO 13.

QUADRO 13 - SÍNTESE DOS MODELOS DIMENSIONAIS DE EMOÇÃO ANALISADOS

Autores	**Número de Dimensões**	**Características**
Bradley *et al*. (1992)	Duas	Valência e Excitação
Fontaine *et al*. (2007)	Quatro	Valência, Potência, Excitação e Imprevisibilidade
Lövheim (2012)	Três	Serotonina (5-HT), 5-hidroxitriptamina), Dopamina (DA) e Noradrenalina (NE)
Plutchik (1970)	Três	Intensidade, Graus de Similaridade e Polaridade
Plutchik (1980)	Duas	Oito emoções básicas
Russell (1980)	Duas	Valência (Prazer/Desprazer) e Excitação (Ativação/Desativação)
Russell e Mehrabian (1977)	Três	Prazer, Excitação e Dominância
Schlosberg (1954)	Três	Prazer/Desprazer, Atenção/Rejeição e Ativação
Thayer (1989, 2012)	Duas	Energia/Cansaço e Tensão/Calma
Watson e Tellegen (1985)	Duas	Afeto Positivo e Afeto Negativo
Wundt (1922)	Três	Agradável/Desagradável, Tenso/Relaxado e Animado/Calmo

FONTE: O autor (2021).

Por último, cabe ressaltar que, assim como acontece com a abordagem categórica, os modelos dimensionais de emoções não estão imunes às críticas. Neste sentido, por exemplo, Colombetti (2014) considera essas propostas pouco convincentes, tendo em vista que os aspectos biológicos presentes nas emoções são radicalmente rejeitados.

Akçay e Oğuz (2020) apontam para a dificuldade em se diferenciar determinadas emoções como medo e raiva. Além disso, os autores atentam para a dificuldade de representar algumas emoções, citando o caso da surpresa, que pode apresentar valência positiva ou negativa de acordo com o contexto da sua ativação (Akçay; Oğuz, 2020). Finalmente, quanto aos modelos bidimensionais, Adolphs e Aderson (2018) argumentam que esses esquemas são incapazes de capturar toda a variação existente entre as emoções, sendo necessária a adoção de um número maior de dimensões (Adolphs; Anderson, 2018).

Com os principais elementos condizentes com as emoções expostos nesta subseção, o próximo tema apresentado e discutido consiste no reconhecimento de emoções em músicas

2.4 RECONHECIMENTO DE EMOÇÕES EM MÚSICAS

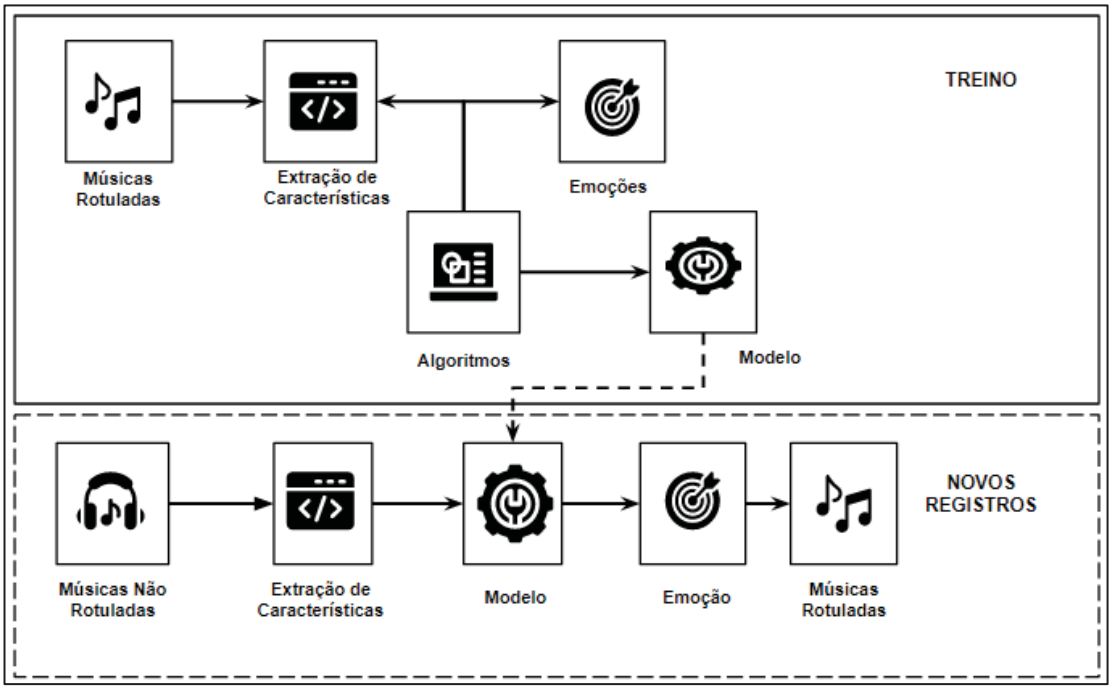
Antes de versar acerca do reconhecimento de emoções em músicas, é necessário esclarecer que há outras abordagens na qual essa temática pode ser empregada e que diferem, portanto, da abordagem adotada neste trabalho. Em um primeiro momento, por exemplo, é possível associar o reconhecimento de emoções em músicas ao estudo de “pistas” relacionadas aos correlatos neurais, conforme aponta Koelsch (2014).

Em outra perspectiva, Juslin e Laukka (2004) também apresentam uma abordagem na qual visualiza-se o emprego do tema reconhecimento de emoções em músicas, que consiste em utilizar três tipos de evidências para inferir acerca de emoções em músicas, sendo essas: a) recorrer a um autorrelato, quer seja verbal, como questionários ou listas de verificação de adjetivo; ou não verbais, como ações envolvendo controles deslizantes e desenhos; b) medir o comportamento, como expressões faciais, vocalizações e linguagem corporal, além de medidas como a velocidade de escrita; e c) reações fisiológicas, como frequência cardíaca, respiração, condutância da pele e sobressalto resposta, tensão muscular (Juslin; Laukka, 2004).

Neste sentido, nesta pesquisa, ressalta-se que o conceito de reconhecimento de emoções em músicas (REM) (do inglês *Music Emotion Recognition* (MER)) refere-se exclusivamente ao subcampo da recuperação de informação musical, cujo objetivo é identificar o conteúdo afetivo - emoções, neste caso - associado a um segmento musical (Kim *et al*., 2010; Dong *et al*., 2019; Hizlisoy; Yildirim; Tufekci, 2021), utilizando, por exemplo, técnicas de aprendizado de máquina e processamento de sinais (Hizlisoy; Yildirim; Tufekci, 2021).

Dessa forma, em termos de etapas, um sistema de REM contempla um processo básico (FIGURA 10) que, de maneira geral, é composto pelos seguintes passos: a) extração e análise de características musicais relacionadas a um conjunto de músicas rotuladas, que representam um conjunto ou valores de emoções (emoções básicas, valência, excitação etc.); b) estabelecimento de relações entre essas características e as emoções selecionadas; e c) utilização de algoritmos para treinar modelos destinados a reconhecer emoções de músicas ainda não rotuladas (Yang; Dong; Li, 2017; Panwar *et al*., 2018; Xu *et al*. 2020; Hizlisoy; Yildirim; Tufekci, 2021).

FIGURA 10 - PROCESSO BÁSICO DE RECONHECIMENTO DE EMOÇÕES EM MÚSICAS



FONTE: Adaptado de Yang, Dong e Li (2017, p. 366).

A seguir são detalhados cada um dos passos relacionados à criação de um sistema de REM.

2.4.1 Obtenção de músicas para a construção de sistemas de reconhecimento de emoções em músicas

Obter músicas devidamente associadas a um conjunto de emoções (rotuladas) é um desafio para o desenvolvimento de sistemas de REM, uma vez que, devido às restrições de direitos autorais, os dados usados raramente são divulgados e reutilizados em outros estudos (Aljanaki; Yang; Soleymani, 2017; ChaturvedI *et al*., 2021).

Para contornar esse desafio, visualiza-se a adoção de duas estratégias para a criação de bases de dados próprias: a) a coleta de dados em plataformas digitais como a *Last.FM* e a *AllMusic* por meio de mineração de *tags* sociais (Song; Dixon; Pierce, 2012; Aljanaki; Wiering; Veltkamp, 2016; Aljanaki; Yang; Soleymani, 2017; Chaturvedi *et al*., 2021); e b) o desenvolvimento de um processo de anotação conduzido junto a uma amostra (Chen, Y. H. *et al.*, 2015; Aljanaki; Wiering; Veltkamp, 2016; Zhang *et al*., 2017; Zhang, K. *et al*., 2018; Hizlisoy; Yildirim; Tufekci, 2021).

O uso de plataformas digitais como fonte de coleta de músicas rotuladas apresenta-se como um mecanismo que proporciona a coleta de uma grande quantidade de dados por meio de mineração de *tags* sociais (Aljanaki; Wiering; VEltkamp, 2016), cuja execução pode ser realizada mediante *scripts* de busca direto na *Uniform Resource Locator* (URL) da plataforma selecionada (Chaturvedi *et al*., 2021).

Como exemplos de pesquisas elaboradas a partir da mineração de *tags* sociais, citam-se os estudos de Lin, Yang e Chen (2011); Song, Dixon e Pierce (2012); e Çano e Morisio (2017). No estudo conduzido por Lin, Yang e Chen (2011), os autores desenvolveram uma base composta por 7.922 músicas associadas a 183 emoções listadas na plataforma *AllMusic*[5]. Para mais, as músicas selecionadas representam seis gêneros musicais populares: a) *blues*; b) *country*; c) *jazz*; d) *rhythm and blues* (R&B); e) *rap*; e f) *rock*.

Song, Dixon e Pierce (2012) utilizaram a *Last.FM* no desenvolvimento de uma base de dados com 2.904 músicas associadas a quatro estados afetivos: a) feliz (*happy*); b) irritado (*angry*); c) triste (*sad*); e d) relaxado (*relaxed*). Por fim, Çano e Morisio (2017) adotam a mesma estratégia de Song, Dixon e Pierce (2012) no que diz

[5] Disponível em: https://www.allmusic.com/moods. Acesso em: 06 maio 2021.

respeito à plataforma e às emoções para elaborarem duas bases de dados destinadas ao reconhecimento de emoções em músicas considerando as letras como principal recurso.

No entanto, é preciso ressaltar que as informações disponíveis nas plataformas digitais podem estar seriamente comprometidas (Chaturvedi *et al*., 2021), principalmente por não apresentarem a homogeneidade e o controle que anotações oriundas de um ambiente experimental controlado são capazes de fornecerem (Aljanaki; Wiering; Veltkamp, 2016).

Em relação à coleta de músicas rotuladas junto a uma amostra, essa estratégia costuma ser conduzida por meio da consulta a voluntários, quer sejam especialistas (músicos, psicólogos etc.) ou não. Nessa abordagem, é possível citar como exemplo os trabalhos realizados por Chen, Y. H. *et al*. (2015); Zhang *et al*. (2017); Zhang, K. *et al*. (2018); Hizlisoy, Yildirim e Tufekci (2021).

Chen, Y. H. *et al*. (2015) recorreram a 665 voluntários para desenvolverem a base de dados nomeada como AMG1608, cujo total de registro é de 1.608 músicas com 30 segundos de duração, apresentando valores correspondentes à valência e à excitação. Em sua pesquisa, Zhang *et al*. (2017) desenvolveram uma base de dados com 171 músicas populares chinesas para representar as emoções exuberância, ansiedade, depressão e contentamento.

O processo de rotulagem foi conduzido por 10 voluntários. Por sua vez, para a construção da base denominada *PMEmo*, Zhang, K. *et al*. (2018) recorreram a 457 voluntários para a anotação de 794 músicas *pop*. Em termos representativos, os 457 voluntários do estudo correspondem a: a) 366 universitários chineses sem formação musical; b) 44 estudantes chineses cursando Música; e c) 47 falantes de língua inglesa.

Hizlisoy, Yildirim e Tufekci (2021) construíram uma base composta por 124 canções tradicionais turcas e que representam valores de valência e excitação em uma escala de -5 a 5 para cada uma das duas dimensões. Neste estudo, a anotação foi realizada por 21 estudantes voluntários com grande conhecimento acerca da música turca e com habilidades para tocar instrumentos musicais.

Do ponto de vista da “qualidade” dos dados, o uso de anotações fornecidas por seres humanos possibilita, de certa forma, a verificação da consistência das informações fornecidas por meio, por exemplo, de testes estatísticos como o Alfa de Cronbach (Aljanaki; Yang; Soleymani, 2017), cujo resultado permite estimar a

confiabilidade de escalas (Field; Males; Field, 2012). Além disso, a coleta de dados de forma controlada pode proporcionar o enriquecimento desse processo, possibilitando a coleta de indicadores fisiológicos e a exclusão de fatores externos que possam influenciar o resultado (Aljanaki; Wiering; Veltkamp, 2016).

Entretanto, conduzir um processo de anotação por amostra é trabalhoso e demorado (Aljanaki; Wiering; Veltkamp, 2016; Aljanaki; Yang; Soleymani, 2017; Zhang, K. *et al*., 2018), tendo em vista a subjetividade inerente à tarefa e a necessidade de muitas anotações para cada música (Aljanaki; Yang; Soleymani, 2017), especialmente se este processo for realizado em um ambiente experimental controlado (Zhang, K. *et al*., 2018).

2.4.2 Extração de características para o reconhecimento de emoções em músicas

Uma das etapas mais importantes para a construção de sistemas de REM consiste na extração de características relacionadas ao conteúdo musical (Zhang, K. *et al*., 2018; Russo *et al*., 2020). O objetivo nessa etapa é transformar o conjunto de músicas selecionado em características que possam descrevê-lo totalmente como, por exemplo, componentes acústicos e musicais das canções (Yang; Dong; Li, 2017) relacionados ao timbre, ao ritmo e à harmonia (Yang; Chen, 2011), assim como as letras das canções (Yang; Dong; Li, 2017).

Todavia, extrair características exige uma análise profunda do conteúdo e das características musicais (Medina; Beltrán; Baldassarri, 2020) e, antes dessa etapa, é necessário adotar dois procedimentos referentes ao pré-processamento: a) padronização dos dados; e b) definição da duração dos registros (Yang; Dong; Li, 2017). Em relação à padronização dos dados, este procedimento consiste em transformar os dados brutos (arquivos de músicas) em um formato mais “adequado” para a análise (Chaturvedi *et al*., 2021).

Nesse sentido, uma prática comum é adotar um formato padrão para todos os arquivos como, por exemplo, estabelecendo uma frequência de amostragem de 22.0050 Hz, com precisão de 16 bits e canal mono (Yang; Chen, 2011; Zhang *et al*., 2017; Chaturvedi *et al*., 2021). Além dessa ação, outro procedimento relacionado à padronização dos dados consiste em definir a duração das canções selecionadas.

Nessa questão, Yang, Dong e Li (2017) visualizaram em seu estudo de revisão, que a duração média das amostras adotadas pelos pesquisadores correspondia a

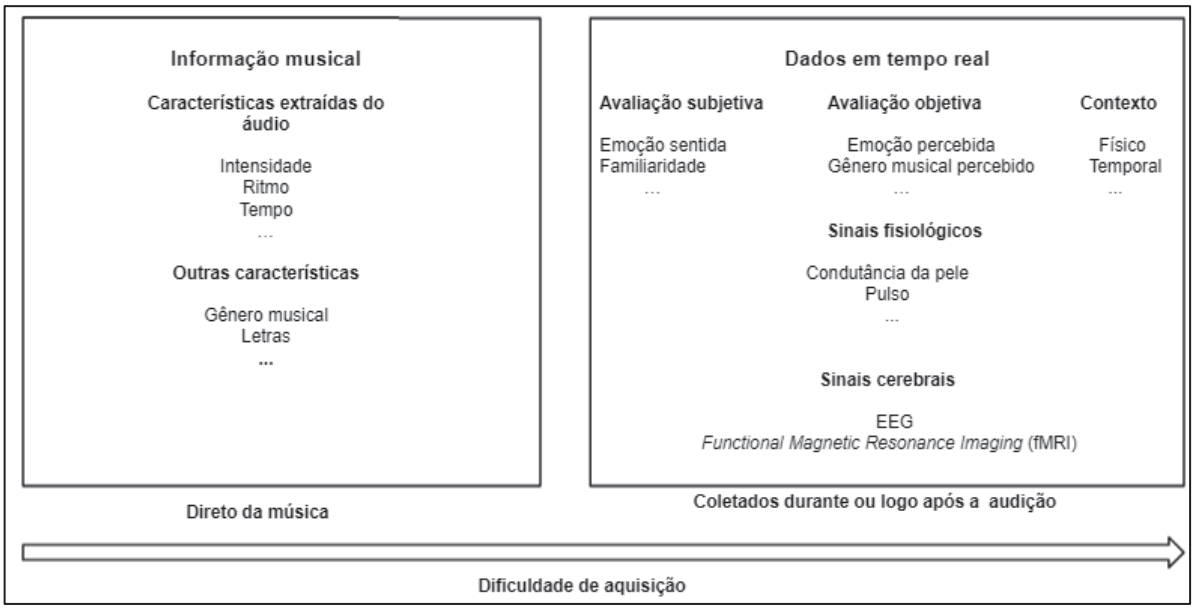
segmentos de 25 a 30 segundos. Essa duração corresponde à visão de Yang e Chen (2011), que consideram que amostras de 20 a 30 segundos são suficientemente representativas.

Chaturvedi *et al*. (2021) afirmam que é possível trabalhar com amostras representativas considerando segmentos com 20 a 60 segundos de duração, enquanto Grekov (2018) argumenta que os seis segundos iniciais são a menor duração para que especialistas consigam detectar o conteúdo afetivo para uma determinada amostra. Contudo, para realizar a segmentação das canções, é prudente evitar segmentos não representativos, como momentos de silêncio ou interferências como aplausos (Dufour; Tzanetakis, 2018).

Após os procedimentos de pré-processamento, é necessário selecionar características para se construir um sistema de REM. Nesse sentido, essas podem ser divididas em dois grandes grupos compostos por: a) recursos de áudio; e b) recursos fisiológicos induzidos pela música (Zhang *et al*., 2019). Esses grupos assemelham-se aos apresentados no trabalho de Xu *et al*. (2020) (FIGURA 11).

Embora Xu *et al*. (2020) considerem a existência de três grupos, visualiza-se que os dados demográficos (sexo, idade etc.) e questões acerca dos hábitos e habilidades musicais dos usuários (Yang *et al*., 2007; Xu *et al*., 2020) não exigem um processo de extração, sendo a coleta conduzida sem a necessidade de o usuário estar escutando músicas, por exemplo (Xu *et al*., 2020).

FIGURA 11 - EXEMPLOS DE CARACTERÍSTICAS PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS



FONTE: Adaptado de Xu *et al*. (2020, p.3).

Conforme observa-se na FIGURA 11, o primeiro grupo de características é obtido diretamente do áudio (músicas), sem considerar os dados dos usuários (Zhang *et al*., 2019; Xu *et al*., 2020). Trata-se do grupo mais empregado em estudos de REM (Xu *et al*., 2020), contemplando características ligadas aos elementos musicais como ritmo, intensidade, melodia e timbre (Yang; Chen, 2011; Xu *et al*., 2020), além de outras possibilidades, como as letras das canções (Xu *et al*., 2020).

Em relação às características extraídas diretamente do áudio, uma breve lista com algumas possibilidades é apresentada no QUADRO 14.

QUADRO 14 - EXEMPLOS DE CARACTERÍSTICAS EXTRAÍDAS DIRETAMENTE DO ÁUDIO

Dinâmica	**Harmonia**	**Melodia**	**Ritmo**	**Timbre**
Low Energy Rate; Root Mean Square (RMS); *Volume*	*Chromagram*; *Key Strength*; *Modality Estimation*	*Pitch Estimation; Pitch Strength*	*Beat Spectrum; Beats Loudness Onset*; *Tempo Estimation*	*Mel-Frequency Cepstral Coefficients* (MFCCs); *Spectral Centroid*; *Spectral Flatness; Spectral Flux*; *Spectral Rolloff; Zero-crossing rate* (ZCR)

FONTE: Adaptado de Panda, Malheiro e Paiva (2020a, p. 616).

Para a extração das características associadas diretamente ao áudio, é comum recorrer a um conjunto de técnicas relacionadas à Transformada de Fourier (Mckay, 2010; Müller, 2015), sendo essas: a) a transformada discreta de Fourier (do inglês *Discrete Fourier Transform* (DFT)) (Mckay, 2010; Müller, 2015); b) a

transformada rápida de Fourier (do inglês *fast Fourier transform* (FFT)) (Mckay, 2010; Müller, 2015); e c) a transformada de Fourier de tempo curto (do inglês *Short-Time Fourier Transform* (STFT)) (Elbier *et al*., 2018; Mueller; Mcfee; Kinnaird, 2021).

Para mais, destaca-se que o processo de extração de características diretamente do áudio pode ser conduzido por meio de diferentes *softwares* e bibliotecas disponíveis (Panda; Rocha; Paiva, 2015; Medina; Beltrán; Baldassarri, 2020), conforme apresenta-se no QUADRO 15.

QUADRO 15 - EXEMPLOS DE FERRAMENTAS UTILIZADAS PARA A EXTRAÇÃO DE CARACTERÍSTICAS RELACIONADAS AO ÁUDIO

Ferramenta	**Endereço**	**Referência**
Essentia	https://essentia.upf.edu/	Bogdanov *et al*. (2013)
JAudio	https://sourceforge.net/projects/jaudio/	McKay (2010)
Librosa	https://librosa.org/	McFee *et al*. (2015)
Marsyas	http://marsyas.info/index.html	Tzanetakis e Cook (2000)
MIRToolbox	https://bityli.com/3f5UJ	Lartillot, Toiviainen e Eerola (2008)
openSMILE	https://www.audeering.com/opensmile/	Eyben, Wöllmer e Schuller (2010)
PsySound3	http://www.psysound.org/	Cabrera, S. Ferguson and E. Schubert (2007)

FONTE: Elaborado a partir de Aljanaki, Yang e Soleymani (2017, p. 15); Dufour e Tzanetakis (2018, p.5).

Além das características ligadas ao sinal de áudio, é possível incorporar outros elementos extraídos diretamente das músicas, como os associados às letras, por exemplo. Para essas características, o emprego de procedimentos de mineração de texto (Malheiro *et al*., 2018; Wang; Su; Wongchaisuwat, 2021) permite a obtenção de características como *Bag-of-Words* (BOW), *Part-of-Speech* (POS), *Term Frequency-Inverse Document Frequency* (TF-IDF) e N-*Grams* (Malheiro *et al*., 2018; Patra; Das; Bandyopadhyay, 2018; Chatarin *et al*., 2020; Wang; Su; Wongchaisuwat, 2021).

Por fim, em relação às características fisiológicas obtidas durante a audição (tempo real) (Zhang *et al*., 2019; Xu *et al*., 2020), a coleta dessas características ocorre em virtude das seguintes hipóteses: a) sinais fisiológicos possibilitam a extração de informações ricas relacionadas às emoções induzidas; b) é possível capturar variações fisiológicas induzidas pela audição de música; e c) pode-se conseguir

melhores resultados ao se combinar dados fisiológicos aos recursos provenientes do áudio (Zhang *et al*., 2019).

Como exemplos de características inseridas neste grupo, encontram-se aquelas baseadas em indicadores coletados com o auxílio de instrumentos específicos, como EEG, eletrocardiograma (ECG), reconhecimento de voz, condução pela pele, frequência cardíaca e respiratória, temperatura dos dedos, gestos, movimento do músculo facial, entre outros (Hsu *et al*., 2018).

2.4.3 Abordagens emocionais consideradas para o reconhecimento de emoções em músicas

No estudos de REM, visualiza-se a busca pelo reconhecimento de uma das seguintes categorias de emoções: a) emoções percebidas, que representam a percepção emocional do ouvinte ao ouvir uma determinada música; b) emoções induzidas pela música, que contemplam as emoções "sentidas" pelo ouvinte, e que podem ser diferentes das emoções percebidas; e c) emoções transmitidas, quando deseja-se reconhecer a emoção que o compositor ou intérprete almejou transmitir com a sua obra (Yang; Chen, 2011; Panda; Malheiro; Paiva, 2020a).

Em termos de taxonomia, as pesquisas de REM seguem as abordagens tradicionais dominantes dos estudos sobre emoções (abordagem categórica ou dimensional), permitindo a construção, por exemplo, de sistemas destinados ao reconhecimento de emoções básicas ou à identificação de valores correspondentes à valência e à excitação fornecidos por modelos dimensionais, como os desenvolvidos por Russell (1980) e Thayer (1989) (Kaminskas; Ricci, 2012; Dufour; Tzanetakis, 2018; Hsu *et al*., 2018; Chaturvedi *et al*., 2021; Hizlisoy; Yildirim; Tufekci, 2021).

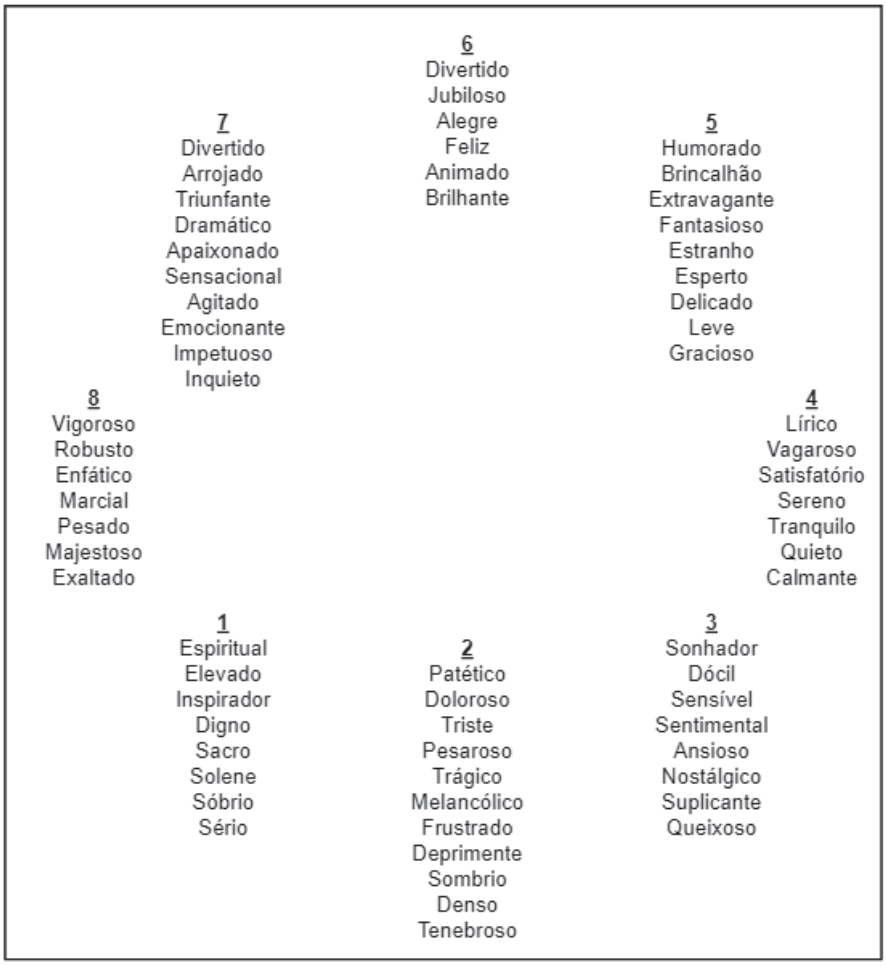
Contudo, além das abordagens construídas a partir de taxonomias tradicionais das emoções, é possível visualizar o uso de taxonomias mais alinhadas ao estudo sobre emoções musicais, como as propostas de Hevner (1936) (Yang; Dong; Li, 2017; Dufour; Tzanetakis, 2018) e Zentner, Grandjean e Scherer (2008) (Aljanaki; Wiering; Veltkamp, 2016; Aljanaki; Yang; SoleymanI, 2017), além de alternativas customizadas, como os *clusters* emocionais elaborados por Hu e Downie (2007) (Aljanaki; Yang; Soleymani, 2017; Dufour; Tzanetakis, 2018).

A taxonomia proposta por Hevner (1936) consiste em um *checklist* composto por 66 adjetivos emocionais divididos em oito *clusters* homogêneos. Os *clusters*

propostos formam uma escala mais ou menos contínua, que já considera possíveis erros relacionados às divergências entre os anotadores acerca das definições precisas dos adjetivos, além de possíveis interferências provocadas pelas condições fisiológicas e o efeito de humores momentâneos no momento da audição (Hevner, 1936; Dufour; Tzanetakis, 2018).

Uma adaptação da proposta de Hevner (1936) é apresentada na FIGURA 12.

FIGURA 12 - MODELO DE HEVNER (1936)



FONTE: Adaptado de Hevner (1936, p. 249).

Posteriormente, a taxonomia de Hevner (1936) foi atualizada por Schubert (2003). O autor conduziu um experimento com 133 indivíduos com pelo menos um ano de estudo universitário em Música, no qual o número de adjetivos foi reduzido para 46, e o número de *clusters* aumentado para nove (Schubert; 2003; Dufour; Tzanetakis, 2018), conforme observa-se no QUADRO 16.

QUADRO 16 - ATUALIZAÇÃO DO MODELO E HEVNER (1936) POR SCHUBERT (2003)

Cluster	**Adjetivos**
A	Alegre, Animado, Feliz, Jubiloso
B	Humorístico, Leve, Lírico, Divertido, Brincalhão
C	Calmo, Delicado, Gracioso, Quieto, Relaxado, Sereno, Tranquilo, Sensível, Sossegado
D	Sonhador, Sentimental
E	Trágico, Nostálgico
F	Tenebroso, Depressivo, Sombrio, Melancólico, Pesaroso, Triste, Solene
G	Pesado, Majestoso, Sacro, Sério, Espiritual, Vigoroso
H	Dramático, Emocionante, Divertido, Apaixonado, Sensacional, Elevado, Triunfante
I	Agitado, Nervoso, Inquieto, Tenso

FONTE: Adaptado de Schubert (2003, p.1121).

Outra taxonomia relacionada às emoções musicais consiste na *Geneva Emotional Music Scale* (GEMS) desenvolvida por Zentner, Grandjean e Scherer (2008). Trata-se de uma escala específica para descrever emoções induzidas pela música (Aljanaki; Wiering; Veltkamp, 2016).

A escala original é composta por nove fatores de primeira ordem validados a partir de análises fatoriais exploratórias e confirmatórias (Zentner; Grandjean; Scherer, 2008; Chełkowska-Zacharewicz; Janowski, 2020) e mais três fatores de segunda ordem (Zentner; Grandjean; Scherer, 2008; Aljanaki; Wiering; Veltkamp, 2016), conforme apresenta-se no QUADRO 17.

QUADRO 17 - ESCALA GEMS PROPOSTA POR ZENTNER, GRANDJEAN E SCHERER (2008)

Fatores de Segunda Ordem	**Fatores de Primeira Ordem**	**Explicação**
Sublimidade (*Sublimity*)	**Espanto**/Admiração (***Amazement***/*Wonder*)	Diz respeito aos sentimentos de admiração e felicidade
	Solenidade/Transcendência (***Solemnity***/*Transcendence*)	Sensação de transcendência e inspiração
	Vigor (*Power*)	Sentir-se forte, heroico, triunfante e enérgico
	Ternura (*Tenderness*)	Relacionado à sensualidade, ao afeto e ao amor
	Nostalgia (*Nostalgia*)	Sentir-se sonhador e melancólico
Vitalidade (*Vitality*)	Calma/ Paz (***Calmness***/*Peacefulness*)	Referente aos sentimentos de relaxamento, serenidade e meditação
	Alegria (*Joyful Activation*)	Relacionado à vontade de dançar; sentir-se animado e saltitante; divertir-se
Inquietação (*Unease*)	Tensão (*Tension*)	Sentir-se nervoso, impaciente e irritado
	Tristeza (*Sadness*)	Sentir-se deprimido e triste

FONTE: Adaptado de Zentner, Grandjean e Scherer (2008, p. 507) e Aljanaki, Wiering e Veltkamp (2016, p. 118).

É preciso ressaltar que os termos em negrito - *amazement*, *solemnity* e *calmness* - são sinônimos propostos por Aljanaki, Wiering e Veltkamp (2016) em virtude de os autores visualizarem, em outros estudos, indícios de dificuldades na compreensão dos termos originais *wonder*, *transcendence* e *peacefulness*.

Por fim, uma alternativa para o reconhecimento de emoções em músicas baseada no uso de plataformas digitais consiste nos *clusters* propostos por Hu e Downie (2007) (QUADRO 18).

QUADRO 18 - *CLUSTERS* EMOCIONAIS PROPOSTOS POR HU E DOWNIE (2007)

***Cluster* 1**	***Cluster* 2**	***Cluster* 3**	***Cluster* 4**	***Cluster* 5**
Turbulento	Amigável/Boa índole	Letrado	Espirituoso	Volátil
Vibrante	Doce	Melancólico	Engraçado	Impetuoso
Confiante	Divertido	Agridoce	Excêntrico	Visceral
Violento	Brincalhão	Outonal	Irônico	Agressivo
Apaixonado	Alegre	Taciturno	Exagerado	Tenso/Ansioso
		Pungente	Peculiar	Intenso
			Bobo	

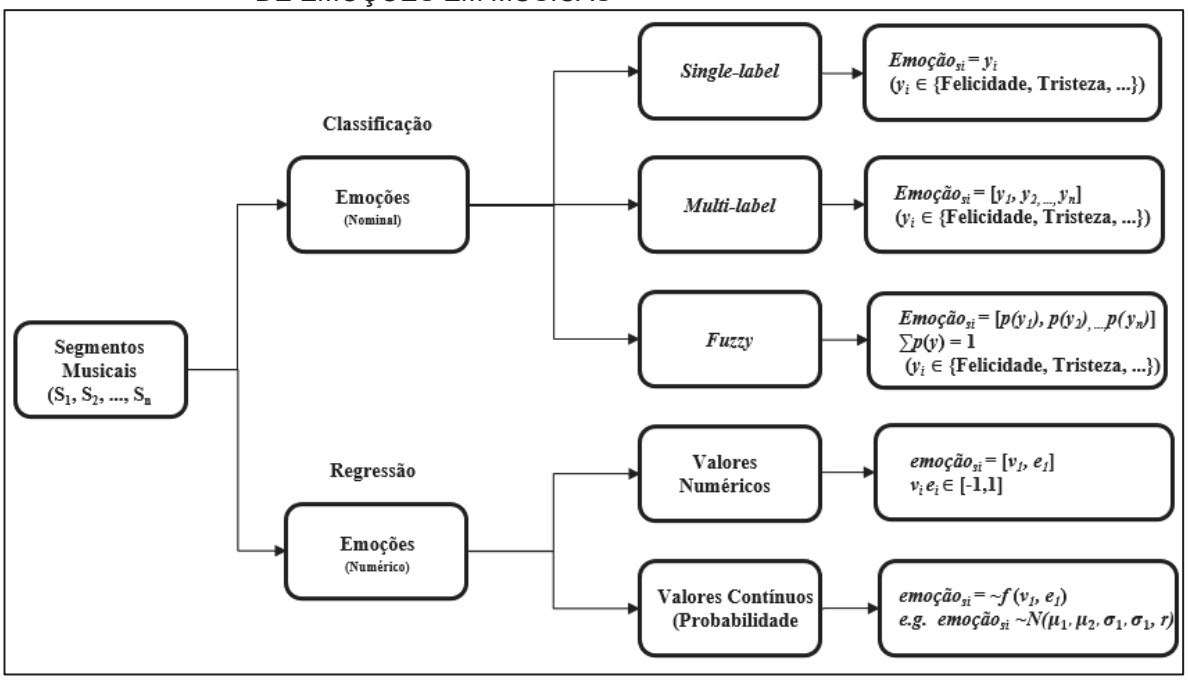
FONTE: Adaptado de Hu e Downie (2007, p.2).

Esta abordagem customizada é composta por 29 rótulos emocionais organizados em cinco *clusters* com termos semanticamente semelhantes (Hu; Downie, 2007; Dufour; Tzanetakis, 2018). A definição desses rótulos é resultado de uma análise estatística das anotações dos metadados de humor presentes no *site AllMusic* (Hu; Downie, 2007; Dufour; Tzanetakis, 2018).

2.4.4 Algoritmos empregados para o reconhecimento de emoções em músicas

Normalmente, as abordagens relacionadas ao REM apresentam-se como problemas de classificação ou regressão (Kim *et al*., 2010; Wu *et al*., 2014; Yang; Dong; LI, 2017), conforme observa-se na FIGURA 13.

FIGURA 13 - POSSIBILIDADES DE ABORDAGENS PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS



FONTE: Adaptado de Yang, Dong e Li (2017, p. 375).

Em relação às abordagens de REM cujo objetivo manifesta-se em problemas de classificação, observa-se que estas podem ser divididas em três grupos: a) *single-label*; b) *multi-label*; e c) *fuzzy* (Yang; Dong; LI, 2017). Um problema classificado como *single-label* consiste em um experimento no qual um segmento musical é associado a um único rótulo emocional, correspondente a um adjetivo (felicidade, tristeza etc.) (Trohidis *et al*., 2008; Yang; Dong; Li, 2017). Nesta categoria, por exemplo, os algoritmos empregados representam métodos tradicionais de aprendizado de máquina, além de arquiteturas de *deep learning* (QUADRO 19).

QUADRO 19 - EXEMPLOS DE ALGORITMOS EMPREGADOS (CLASSIFICAÇÃO *SINGLE-LABEL*)

Algoritmos	**Empregado em**
CNN	Chaudhary; Singh e Singh (2020)
Convolutional Long Short-Term Memory Deep Neural Network (CLDNN)	Hizlisoy, Yildirim e Tufekci (2021)
RF	Wang e Sun (2019); Hizlisoy, Yildirim e Tufekci (2021)
k-NN	Bai *et al*. (2017); Wang e Sun (2019); Hizlisoy, Yildirim e Tufekci (2021)
Logistic Regression	Wang e Sun (2019)
Long Short-Term Memory (LSTM)	Hizlisoy, Yildirim e Tufekci (2021)
SVM	Bai *et al*. (2017); Panda, Malheiro e Paiva (2020a); Wang e Sun (2019); Chaudhary; Singh e Singh (2020); Hizlisoy, Yildirim e Tufekci (2021)

FONTE: O autor (2021).

Concernente à classificação *multi-label*, este tipo de abordagem possibilita a classificação de um segmento musical em uma série de categorias de emoção (Trohidis *et al*., 2008; Yang; Dong; Li, 2017), por meio de um conjunto de algoritmos específicos como os listados no QUADRO 20.

QUADRO 20 - EXEMPLOS DE ALGORITMOS EMPREGADOS (CLASSIFICAÇÃO *MULTI-LABEL*)

Algoritmos	**Empregado em**
2BR	Trohidis *et al*. (2011)
Binary Relevance (BR)	Trohidis *et al*. (2008, 2011); Wu *et al*. (2014)
Calibrated Label Ranking (CLR)	Trohidis *et al*. (2011)
HMER	Wu *et al*. (2014)
Label Powerset (LP)	Trohidis *et al*. (2008, 2011); Wu *et al*. (2014)
M3LDA	Wu et al. (2014)
Multi-label back-propagation (BP-MLL)	Trohidis et al. (2011)
Multi-label k-nearest neighbor (ML-k-NN)	Trohidis et al. (2008, 2011)
Random K-Labelsets (RAKEL)	Trohidis et al. (2008, 2011); Wu et al. (2014)
Random	Wu et al. (2014)

FONTE: O autor (2021).

O uso da lógica *fuzzy* em trabalhos de REM apresenta-se como um tipo específico de classificação *multi-label*, com a característica de representar a pertinência de um segmento musical a uma série de categorias de emoção (Yang; Liu; Chen, 2006; Yang; Dong; Li, 2017). Relacionado aos algoritmos *fuzzy* já empregados em estudos anteriores, uma breve relação é apresentada no QUADRO 21.

QUADRO 21 - EXEMPLOS DE ALGORITMOS EMPREGADOS (CLASSIFICAÇÃO *FUZZY*)

Algoritmos	**Empregado em**
Fuzzy k-NN (FK-NN)	Yang, Liu e Chen (2006); Bai *et al*. (2017)
Fuzzy Nearest-Mean (FNM)	Yang, Liu e Chen (2006)
Neuro-Fuzzy Networks Classification (NFNC)	Bai *et al*. (2017)

FONTE: O autor (2021).

Para problemas de REM baseados na tarefa de regressão, duas possibilidades são visualizadas: a) regressão destinada a valores numéricos; e b) regressão de distribuição de probabilidade contínua (Yang; Dong; Li, 2017). No tocante aos algoritmos utilizados nessas abordagens, alguns exemplos são apresentados no QUADRO 22.

QUADRO 22 - EXEMPLOS DE ALGORITMOS EMPREGADOS PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS (REGRESSÃO)

Tipo	**Algoritmos**	**Trabalhos**
Numérica	*AdaBoost.RT* (BoostR)	Yang *et al*. (2008)
	Bayesian Ridge	Xu *et al*. (2020)
	k-NN	Xu *et al*. (2020)
	Multiple Linear Regression (MLR)	Yang *et al*. (2008); Wang *et al*. (2012); Xu *et al*. (2020)
	Random Forest	Xu *et al*. (2020)
	SVR	Yang *et al*. (2008); Wang *et al*. (2012, 2015); Zhang, K. *et al.* (2018); Xu *et al*. (2020)
Probabilidade	*Acoustic Emotion Gaussians* (AEG)	Wang *et al*. (2012, 2015)

FONTE: O autor (2021).

Como diferenças entre as duas abordagens, Yang, Dong e Li (2017) afirmam que: a) a busca por valores numéricos como resultado evita a ambiguidade dos modelos categóricos de emoções, contudo, os resultados podem ser afetados pela subjetividade existente no processo de anotação; e b) trabalhar com valores de probabilidade tende a capturar as diferentes experiências relacionadas à emoção que

as pessoas possuem ao escutar a mesma música, além de reduzir os aspectos subjetivos e auxiliar na construção de sistemas com maior grau de personalização.

Finalizado o tema reconhecimento de emoções em músicas, a próxima temática analisada consiste nas diferentes arquiteturas de *deep learning*.

2.5 *DEEP LEARNING*

Deep learning (em português, aprendizado profundo) é um subcampo específico do aprendizado de máquina (Chollet, 2018; Hosseini *et al*., 2020) baseado em redes neurais artificiais (Aldweesh; Derhab; Emam, 2020; Khan *et al*., 2021). As aplicações dos métodos oriundos desse subcampo podem ser visualizadas em diferentes áreas de pesquisa, como reconhecimento de fala e imagem, processamento de linguagem natural, filtragem de redes sociais, sistemas de recomendação, visão computacional (Aldweesh; Derhab; Emam, 2020; Rezaeianjouybari; Shang, 2020; Zhou; Liu; Gravelle, 2020), entre outras aplicações (Aldweesh; Derhab; Emam, 2020).

O princípio básico desse subcampo consiste no uso de arquiteturas dotadas de camadas sucessivas com capacidade de aprendizagem automática (Lecun; Bengio; Hinton, 2015; Wan *et al*., 2019), destinadas ao reconhecimento de vários níveis de representações de dados (Chollet, 2018; Pouyanfar *et al*., 2018; Young *et al*., 2018). Nesse aspecto, essas arquiteturas podem apresentar dezenas ou centenas de camadas (Chollet, 2018), que aumentam o potencial de aprendizagem, melhorando o desempenho quanto à precisão em problemas de classificação e à redução do erro em problemas de regressão (Kamilaris; Prenafeta-Boldú, 2018).

Em termos de características, é possível descrever as arquiteturas de *deep learning* a partir das seguintes particularidades: a) a necessidade de um grande volume de dados para o treinamento; b) a dependência de boa estrutura de *hardware*, como o uso de *Graphics Processing Unit* (GPU), uma vez que essas arquiteturas realizam muitas operações baseadas em matrizes; c) a redução do esforço para extração de recursos para a criação de modelos; d) a resolução de problemas de ponta a ponta (*end-to-end*); e) tempo elevado para o treinamento dos modelos; e f) o baixo grau de interpretabilidade dos resultados (Xi *et al*., 2018).

Em comparação com as abordagens de aprendizado de máquina, que apresentam capacidade limitada de extração e processamento de características para

representar dados (Lecun; Bengio; Hinton, 2015), as arquiteturas de *deep learning* podem se aproveitar de uma grande quantidade de exemplos para extrair recursos de alto nível (Zhang, Q. *et al*., 2018), permitindo que determinadas arquiteturas superem o que existe de mais moderno em tarefas de reconhecimento de padrões (Zhou; Liu; Gravelle, 2020).

Em relação à taxonomia de arquiteturas de *deep learning* existentes, visualiza-se várias abordagens desenvolvidas para atender a diferentes propósitos (Akkus *et al*., 2017). Todavia, é possível agrupar as diferentes arquiteturas em três grandes classes: a) arquiteturas discriminativas; b) arquiteturas generativas; e c) arquiteturas híbridas (Fadlullah *et al*., 2017; Aldweesh; Derhab; Emam, 2020; Rezaeianjouybari; Shang, 2020).

Arquiteturas discriminativas são aquelas que modelam a probabilidade condicional de um resultado (*output*), baseando-se nos dados observados (*inputs*) (Rezaeianjouybari; Shang, 2020). Essas arquiteturas apresentam-se como adequadas para problemas de classificação ou reconhecimento de padrões (Fadlullah *et al*., 2017). Por sua vez, as arquiteturas generativas verificam as propriedades de correlação de alta ordem dos dados de entrada para fins de síntese (Fadlullah *et al*., 2017), por exemplo, além de possibilitarem a geração de instâncias aleatórias (Rezaeianjouybari; Shang, 2020). Por fim, as arquiteturas híbridas combinam as particularidades das arquiteturas discriminativas e generativas (Fadlullah *et al*., 2017; Rezaeianjouybari; Shang, 2020).

Tendo em vista a necessidade de compreender as diferentes arquiteturas de *deep learning*, na sequência um conjunto de arquiteturas empregadas em abordagens baseadas em *deep learning* é apresentado (Aldweesh; Derhab; Emam, 2020; Fadlullah *et al*., 2017; Pouyanfar *et al*., 2018; Zhang, Q. *et al*., 2018; Grigorescu *et al*. 2020; Gao, J. *et al*., 2020; Wan *et al*., 2020; Zhou; Liu; Gravelle, 2020; Rezaeianjouybari; Shang, 2020).

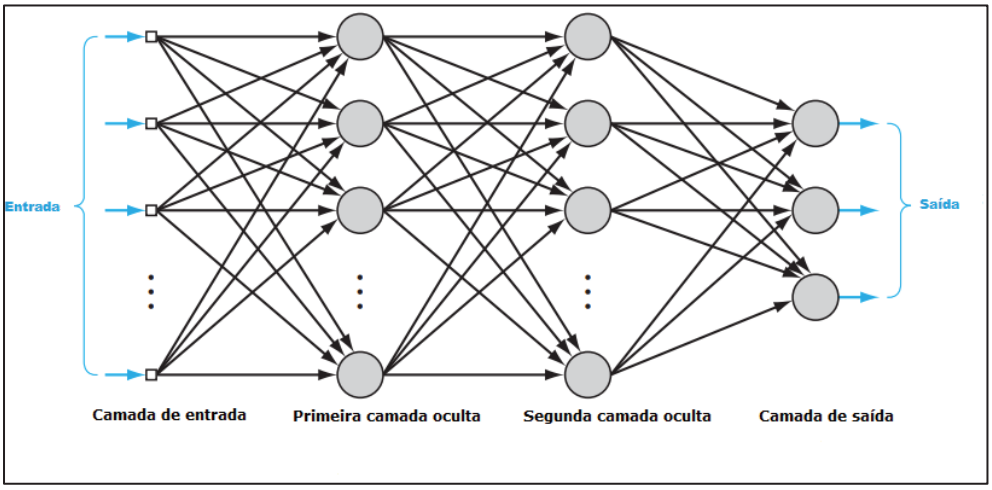
2.5.1 *Multilayer Perceptron*

A arquitetura *perceptron* multicamadas (*multilayer perceptron* (MLP)) é considerada a arquitetura de *deep learning* mais simples existente (Rezaeianjouybari; Shang, 2020), cujas características básicas residem em

> a) cada neurônio na rede inclui uma função de ativação não linear que é diferenciável;
> b) a rede contém uma ou mais camadas ocultas dos nós de entrada e saída;
> c) a rede exibe um alto grau de conectividade, cuja extensão é determinada pelos pesos sinápticos da rede (Haykin, 2009, p. 123, tradução nossa).

Quanto à sua estrutura (FIGURA 14), uma arquitetura MLP apresenta uma camada de entrada com neurônios que atuam como receptores; uma ou mais camadas ocultas de neurônios que computam os dados e passam por iterações, e, por fim, a última camada representa a saída que prevê o resultado (Desai; Shah, 2021).

FIGURA 14 – ARQUITETURA MLP COM DUAS CAMADAS OCULTAS



FONTE: Adaptado de Haykin (2009, p. 124).

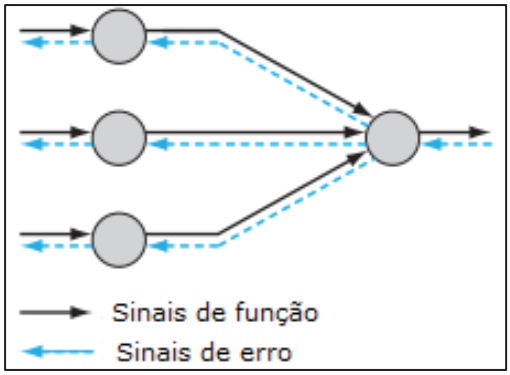
Com base na FIGURA 14, observa-se uma arquitetura totalmente conectada, com cada neurônio de uma camada conectado a todos os neurônios da camada anterior, e os sinais da rede progridem em uma direção direta, da esquerda para a direita e camada por camada (Haykin, 2009).

Outrossim, uma MLP consiste em uma rede neural do tipo *feedforward*, e que emprega a retropropagação (*backpropagation*) no seu aprendizado (Desai; Shah, 2021), que ocorre de maneira superviosionada (Turkoglu; Kaya, 2020). Além disso, o processo é iterativo, visando a minimização do erro e ajustes de pesos (Kalteh; Berndtsson, 2007; Khan *et al*., 2021).

De maneira detalhada, o treinamento de uma MLP ocorre em duas fases: na primeira, denominada fase direta (*forward phase*), os pesos sinápticos da rede são fixos e o sinal de entrada é propagado pela rede, camada por camada, até atingir a saída; na segunda, a fase de retorno (*backward phase*), um sinal de erro é produzido comparando a saída da rede com uma resposta desejada. O sinal de erro resultante é propagado através da rede, novamente camada por camada, mas, desta vez, a propagação é realizada no sentido inverso. Nesta segunda fase, são feitos sucessivos ajustes nos pesos sinápticos da rede (Haykin, 2009).

Com relação aos sinais produzidos na MLP, visualizam-se dois tipos: a) sinal de função (*function signals*), que se propaga para frente, neurônio por neurônio, através da rede; e b) sinal de erro (*error signals*), que se origina em um neurônio de saída da rede e se propaga para trás (camada por camada) através da rede (Haykin, 2009). A representação gráfica desses sinais é apresentada na FIGURA 15.

FIGURA 15 - SINAIS (MLP)



FONTE: Adaptado de Haykin (2009, p. 125).

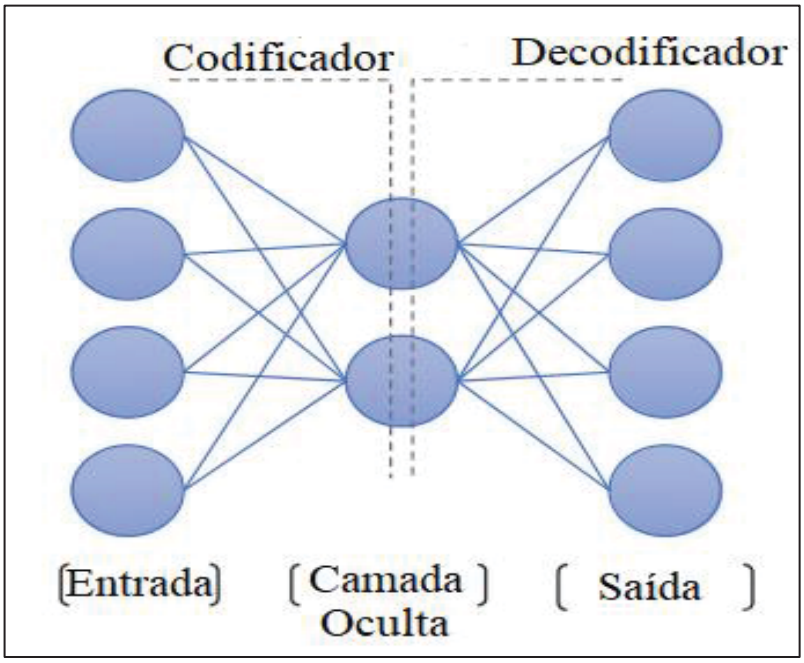
Finalmente, quanto às funções dos neurônios ocultos em uma MLP, esses são responsáveis pelo: a) cálculo do sinal de função que aparece na saída de cada neurônio, que é expresso como uma função não linear contínua do sinal de entrada e dos pesos sinápticos associados a esse neurônio; e b) cálculo de uma estimativa do vetor gradiente (ou seja, os gradientes da superfície de erro em relação aos pesos

conectados às entradas de um neurônio), necessária para a passagem para trás pela rede (Haykin, 2009).

2.5.2 *Autoencoders*

Autoencoders (AEs) são arquiteturas compostas por uma camada de entrada denominada codificador (*encoder*), responsável pela conversão dos dados de entrada em uma representação diferente, e uma camada de saída nomeada como decodificador (*decoder*) (FIGURA 16), cuja função é converter a representação gerada pelo *encoder* para o formato original (Goodfellow; Bengio; Courville, 2016; Gao, J. *et al*., 2020). Além disso, entre essas camadas há uma camada oculta destinada à representação de recursos de baixa dimensão (Aldweesh; Derhab; Emam, 2020).

FIGURA 16 - REPRESENTAÇÃO BÁSICA DE UM *AUTOENCODER*



FONTE: Adaptado de Aldweesh, Derhab e Emam (2020, p. 4).

O processo de codificação realizado por um AE consiste na conversão de um vetor de entrada x pela camada oculta h por meio de uma matriz de ponderação ω; na sequência, ocorre o processo de decodificação, no qual há a conversão da camada oculta h para o formato original, obtendo, dessa maneira, $\tilde{x}$ com uma nova matriz de ponderação ω', teoricamente transposta de ω (Liu *et al*., 2017).

Uma das principais vantagens de um AE consiste na capacidade de extrair recursos úteis de forma contínua durante a propagação e filtrar as informações inúteis (Liu *et al*., 2017). Desse modo, normalmente, um AE é empregado para a redução de dimensionalidade (Aldweesh; Derhab; Emam, 2020; Khan *et al*., 2021), possibilitando a preservação do máximo de informação possível ao longo do processo (Goodfellow; Bengio; Courville, 2016).

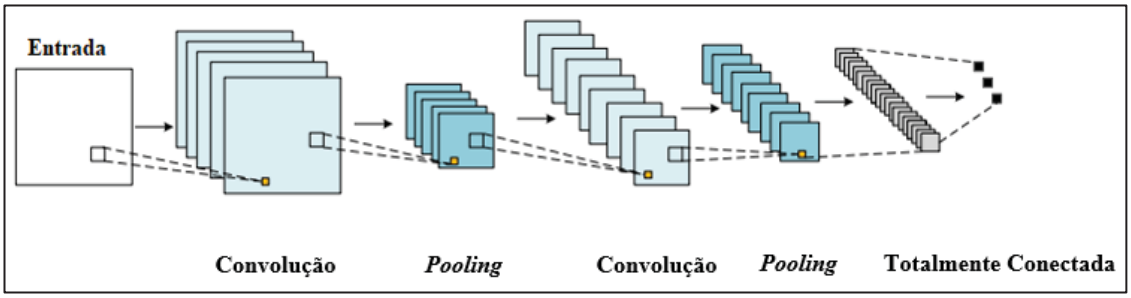
Por fim, existem diferentes variações de AEs, como o *Stacked Autoencoder* (SAE), uma variação que corresponde ao empilhamento de vários *autoencoders* (Aldweesh; Derhab; Emam, 2020; Gao, J. *et al*., 2020), e o *Denoising Autoencoder* (DAE), que adiciona intencionalmente ruído aos dados de treinamento e, posteriormente, realiza o treinamento com esses dados corrompidos, tendo como o objetivo a recuperação de uma versão sem ruído dos dados, como forma de obter uma representação final refinada (Liu *et al*., 2017; Gao, J. *et al*., 2020).

2.5.3 *Convolutional Neural Network*

Uma Rede Neural Convolucional (*Convolutional Neural Network* (CNN)) é uma arquitetura de *deep learning* com grande aplicação no estudo de imagens (Aldweesh; Derhab; Emam, 2020; Hosseini *et al*., 2020), como segmentação e classificação (Akkus *et al.*, 2017), além de abordagens de visão computacional (Dhillon; Verma, 2020). Outrossim, verifica-se o uso dessa arquitetura em áreas como reconhecimento de caligrafia, detecção de rostos, reconhecimento de fala, processamento de linguagem natural e sistemas de recomendação (Liu *et al*., 2017; Hosseini *et al*., 2020).

Trata-se de uma arquitetura do tipo *feedforward* (Wan *et al*., 2020) - a informação é direcionada para a próxima camada -, composta, basicamente, por três tipos de camadas: a) convolucional; b) *pooling*; e c) totalmente conectada (WAN *et al*., 2017; Pouyanfar *et al*., 2018; Zhang, Q. *et al*., 2018; Rezaeianjouybari; Shang, 2020). Uma representação básica de uma arquitetura CNN pode ser visualizada na FIGURA 17.

FIGURA 17 - ESTRUTURA BÁSICA DE UMA CNN



FONTE: Adaptado de Khan *et al*. (2021, p. 103).

A arquitetura CNN recebe este nome por conta da camada convolucional realizar uma operação matemática denominada convolução[6] (Goodfellow; Bengio; Courville, 2016; Zhang, Q. *et al*., 2018; Dhillon; Verma, 2020) em vez de uma multiplicação geral de matrizes (Goodfellow; Bengio; Courville, 2016; Dhillon; Verma, 2020; Hosseini *et al*., 2020). Para mais, o uso da convolução potencializa recursos que tendem a melhorar o desempenho de um modelo: a) interações esparsas; e b) compartilhamento de parâmetros (Goodfellow; Bengio; Courville, 2016; Hosseini *et al*., 2020).

As interações esparsas são responsáveis por transformar o *kernel* do modelo em um elemento menor do que a entrada, como, por exemplo, em modelos elaborados para classificação de imagens em alta resolução, que podem apresentar milhões de *pixels* (Goodfellow; Bengio; Courville, 2016; Hosseini *et al*., 2020). Neste caso, o *kernel* será ajustado para trabalhar com características mais representativas, como contraste e bordas (Goodfellow; Bengio; Courville, 2016; Hosseini *et al*., 2020). Deste modo, haverá redução dos requisitos de memória do modelo e incremento da sua eficiência estatística (Goodfellow; Bengio; Courville, 2016).

O compartilhamento de parâmetros corresponde ao uso do mesmo parâmetro para mais de uma função em um modelo (Goodfellow; Bengio; Courville, 2016). No caso de uma CNN, é possível dizer que os pesos são “amarrados” (*tied weights*), ou seja, o valor do peso de uma entrada encontra-se atrelado (“amarrado”) ao peso estabelecido em outra entrada (Goodfellow; Bengio; Courville, 2016; Hosseini *et al*., 2020).

[6] De maneira concisa, a convolução refere-se a uma operação matemática em duas funções que produz uma terceira função que expressa como a forma de uma é modificada pela outra. Este conceito é amplamente utilizado em redes neurais convolucionais para análise de imagens (Jiménez-Luna; Grisoni; Schneider, 2020).

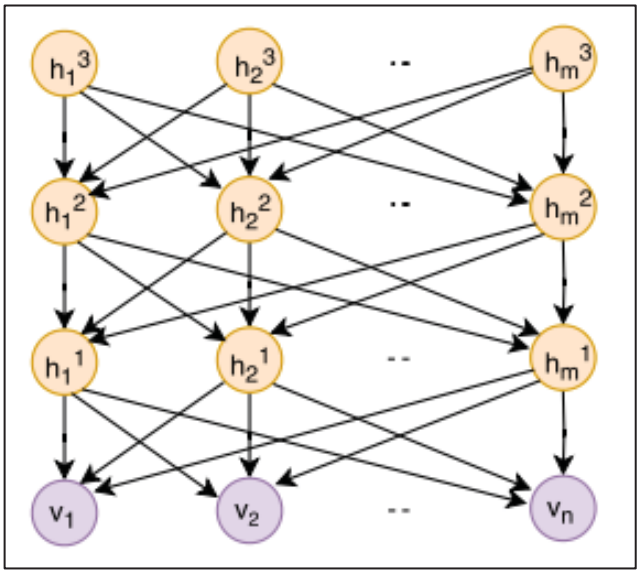
A camada de *pooling* é responsável por realizar uma subamostragem, cujo principal objetivo é reduzir a dimensão do mapa de características, visando, dessa forma, diminuir o número de parâmetros e cálculos efetuados, minimizando o risco de ocorrência de sobreajuste (*overfitting*) (Pouyanfar *et al*., 2018; Dhillon; Verma, 2020). Em termos de operações realizadas nessa camada, é possível selecionar o valor médio (*average pooling*) ou o máximo (*max pooling*) da região analisada (Zhang, Q. *et al*., 2018), ou optar por uma média ponderada (*weighted average pooling*) (Goodfellow; Bengio; Courville, 2016), por exemplo.

Por fim, em uma CNN, as camadas totalmente conectadas são acrescentadas próximas à saída do modelo (Kamilaris; Prenafeta-Boldú, 2018), comportando-se como classificadores (Kamilaris; Prenafeta-Boldú, 2018; Zhang, Q. *et al*., 2018). Nessas camadas, pode-se recorrer a uma função - *softmax*, por exemplo - ou a um algoritmo de aprendizado de máquina, como o SVM, para gerar as pontuações de classificação, de tal modo que cada pontuação corresponda à probabilidade de uma determinada classe para uma determinada instância (Pouyanfar *et al*., 2018).

2.5.4 *Deep Belief Network*

A arquitetura conhecida como *Deep Belief Network* (DBN) (FIGURA 18) caracteriza-se por ser uma arquitetura generativa composta por várias camadas de variáveis latentes (Goodfellow; Bengio; Courville, 2016; Liu *et al*., 2017; Ozbayoglu; Gudelek; Sezer, 2020), sem conexão intra-camadas (Goodfellow; Bengio; Courville, 2016).

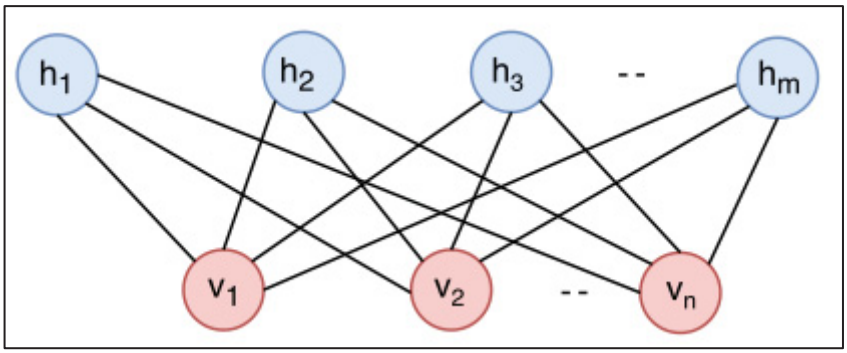
FIGURA 18 - ESTRUTURA BÁSICA DE UMA DBN



FONTE: Ozbayoglu, Gudelek e Sezer (2020, p. 5).

DBNs são arquiteturas compostas por uma pilha de camadas correspondentes a uma arquitetura de rede neural artificial conhecida como *Restricted Boltzmann Machine* (RBM) (Zhang, Q. *et al*., 2018; Ozbayoglu; Gudelek; Sezer, 2020), que consiste em um modelo bipartido não direcionado com duas camadas, sendo uma visível (v) e outra oculta (h) (Zhang, Q. *et al*., 2018; Ozbayoglu; Gudelek; Sezer, 2020), também sem conexão intra-camadas (Ozbayoglu; Gudelek; Sezer, 2020), conforme visualiza-se na FIGURA 19.

FIGURA 19 - ESTRUTURA BÁSICA DE UMA RBM



FONTE: Ozbayoglu, Gudelek e Sezer (2020, p. 5).

Em uma DBN, apesar de não existirem conexões intra-camadas, há conexões entre as unidades ocultas situadas em camadas separadas (Goodfellow; Bengio; Courville, 2016). Neste sentido, a cada duas camadas adjacentes verifica-se a

formação de uma RBM, de tal forma que, a camada visível de cada RBM, tem-se uma conexão com a camada oculta da RBM anterior (Liu *et al*., 2017).

Em termos de direcionamento, as conexões presentes entre as duas camadas superiores não são direcionadas; por sua vez, as demais conexões entre todas as outras camadas são direcionadas, sendo representadas por setas apontadas para a camada mais próxima dos dados (Goodfellow; Bengio; Courville, 2016).

Referente ao treinamento de uma arquitetura DNB, este é conduzido de forma sequencial: primeiro, são treinadas as camadas RBMs inferiores; posteriormente, há o treinamento para as camadas RBMs superiores (Liu *et al*., 2017). Este processo é dividido em duas fases: a) a fase de pré-treinamento; e b) a fase de ajuste fino (Liu *et al*., 2017; Zhang, Q. *et al*., 2018).

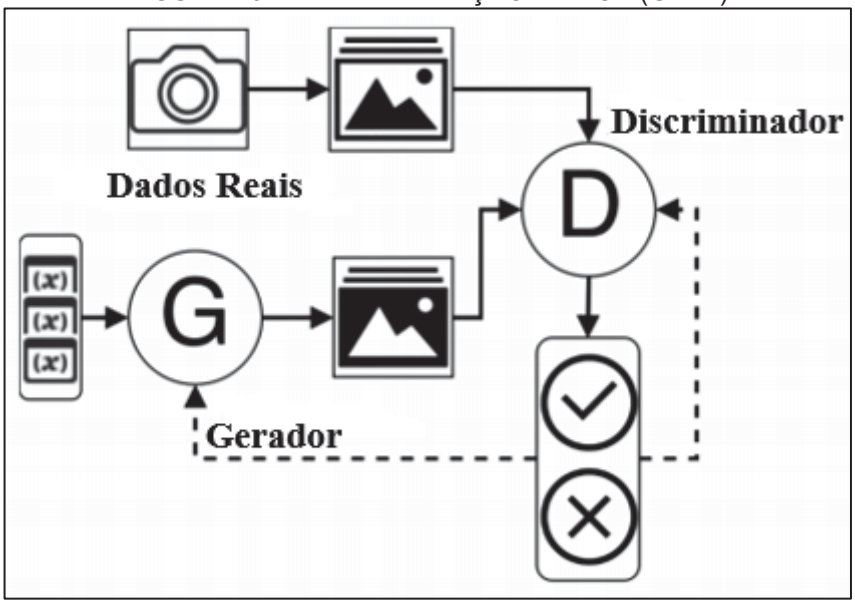
A fase de pré-treinamento é responsável pela definição dos parâmetros iniciais, com abordagem não supervisionada (Liu *et al*., 2017; Zhang, Q. *et al*., 2018). O treinamento é realizado de maneira *down-up*, e objetiva extrair recursos a partir da estrutura dos dados de entrada (Liu *et al*., 2017). Deste modo, a primeira camada é modelada como uma RBM, recebendo os dados de treinamento como entrada; na sequência, a próxima camada oculta - também modelada como uma RBM - é treinada a partir da saída resultante da primeira camada (Gao, J. *et al*., 2020)

Este processo repete-se até que a última camada oculta seja pré-treinada (Gao, J. *et al*., 2020). Posteriormente, é realizado o ajuste fino (*fine-tuning*) por meio de aprendizado supervisionado e estratégia *up-down*, com o objetivo de ajustar os parâmetros em relação às amostras rotuladas (Liu *et al*., 2017; Zhang, Q. *et al*., 2018).

2.5.5 *Generative Adversarial Networks*

Redes Adversárias Generativas (*Generative Adversarial Networks* (GANs)) (FIGURA 20) consistem em arquiteturas que têm como principal característica a presença de dois elementos em sua estrutura, denominados gerador (G) e discriminador (D) (Aldweesh; Derhab; Emam, 2020; Rezaeianjouybari; Shang, 2020), treinados de forma não supervisionada e competitiva (Habimana *et al*., 2020).

FIGURA 20 - REPRESENTAÇÃO BÁSICA (GANs)



FONTE: Adaptado de Pouyanfar *et al*. (2018, p. 92:7).

Para compreender as arquiteturas GANs, pode-se atribuir a uma das redes o papel de um falsificador, enquanto a outra assume o papel de um especialista. No caso de problemas relacionados à imagem, por exemplo, o falsificador (G) irá criar imagens - falsificações - como se fossem imagens verdadeiras; por sua vez, o especialista (D) terá ao seu alcance as imagens geradas por G, bem como imagens autênticas, e a sua função será diferenciá-las (Creswell, *et al*., 2018).

De forma matemática, o treinamento das arquiteturas GANs assemelha-se a um jogo MiniMax, no qual ocorre a competição entre as redes geradora e discriminadora (Pouyanfar *et al*., 2018; Aldweesh; Derhab; Emam, 2020; Rezaeianjouybari; Shang, 2020). Sendo assim, enquanto G captura a distribuição p_g sobre os dados reais t localmente; D tenta diferenciar uma amostra oriunda dos dados de modelagem m em vez de p_g, representada por p_m (Pouyanfar *et al*., 2018).

Neste caso, a competição entre essas redes estabelece uma função de valor $V(G, D)$, representada pela EQUAÇÃO 1 (Pouyanfar *et al*., 2018).

$$min_G max_D V(G,D) = E_{t\sim pdados}[logD(t)] + E_{m\sim pm}(m)[logD(1 - D(G(m)))] \quad (1)$$

Onde $D(t)$ representa a probabilidade de que t é originária dos dados em vez de p_g, e p_{dados} é a distribuição dos dados do mundo real (Pouyanfar *et al*., 2018). Neste processo de treinamento, a cada rodada, G produz amostras aleatórias de

ruído, enquanto D recebe as amostras geradas por G junto com as amostras reais para tentar distingui-las (Aldweesh; Derhab; Emam, 2020).

Finalmente, o modelo será considerado estável quando G e D atingirem o ponto no qual nenhum deles possa ser melhorado, como de p_g = p_{dados}. Ou seja, D não tem mais condições de diferenciar as duas distribuições (Pouyanfar *et al*., 2018).

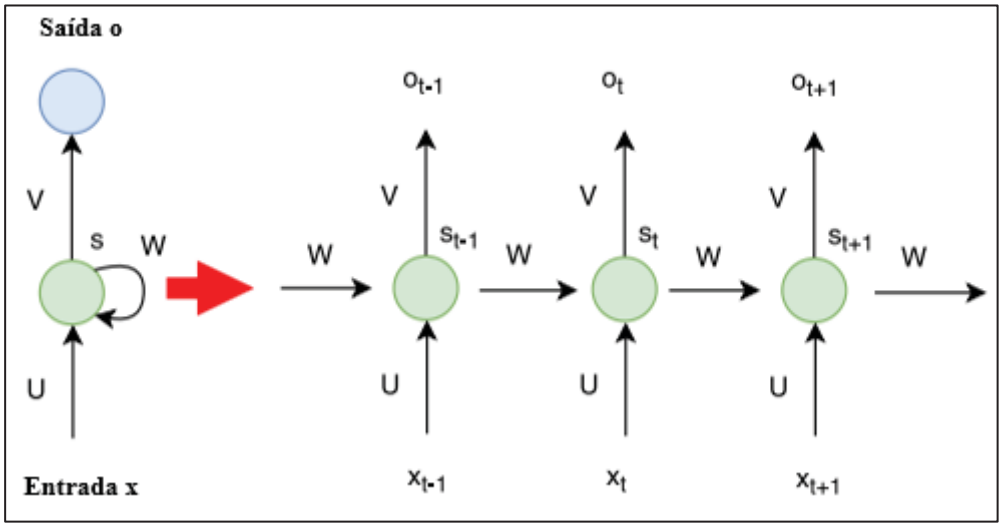
2.5.6 *Recurrent Neural Network*s

Redes Neurais Recorrentes (*Recurrent Neural Networks* (RNNs)) consistem em arquiteturas destinadas, principalmente, para abordagens que envolvem dados sequenciais, como, por exemplo, dados de séries temporais, áudio, fala e idioma. (Ozbayoglu; Gudelek; Sezer, 2020). Nessas arquiteturas, o termo “recorrente” corresponde ao seu comportamento, que se manifesta na execução do mesmo cálculo em cada *token* da sequência de dados, com cada etapa dependendo dos cálculos e resultados anteriores (Young *et al*., 2018).

Neste sentido, normalmente, RNNs processam sequencialmente a série de entrada uma a uma, durante sua operação (Ozbayoglu; Gudelek; Sezer, 2020). Ademais, trata-se de arquiteturas compostas por uma “memória” de entradas anteriores armazenadas no seu estado interno, responsável pela aprendizagem de recursos (Zhang, Q. *et al*., 2018).

Para efeitos de compreensão, uma representação básica de uma arquitetura RNN é apresentada na FIGURA 21.

FIGURA 21 - ESTRUTURA BÁSICA DE UMA RNN



FONTE: Adaptado de Ozbayoglu, Gudelek; Sezer (2020, p. 4).

Com base na FIGURA 21, verifica-se que o fluxo de informação na camada oculta da RNN é dividido em tempos discretos, com o estado do nó S em tempos diferentes de t representado como S_t. Por seu turno, o valor de entrada x em tempos diferentes é x_t, enquanto o valor de saída o em tempos diferentes é representado como o_t (Ozbayoglu; Gudelek; Sezer, 2020). No final, em relação aos parâmetros U, W e V, estes são sempre utilizados no mesmo passo (Ozbayoglu; Gudelek; Sezer, 2020) e representam os parâmetros de peso (Gao, J. *et al*., 2020).

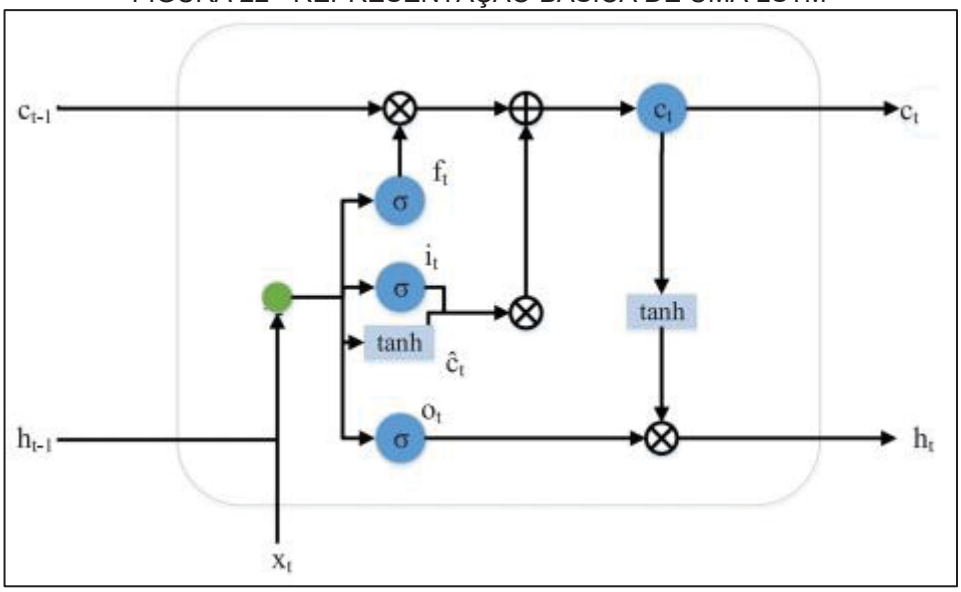
Outrossim, uma RNN contém *loops* de *feedback* para lembrar as informações das unidades anteriores (Rezaeianjouybari; Shang, 2020). Desse modo, essa arquitetura captura a dependência entre a amostra atual x_t e a anterior x_{t-1}, e integra a representação oculta anterior s_{t-1} na passagem para frente (Zhang, Q. *et al*., 2018).

A respeito das variações de RNNs, duas arquiteturas denominadas como *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU) se destacam (Aldweesh; Derhab; Emam, 2020; Rezaeianjouybari; Shang, 2020). A arquitetura LSTM é capaz de adequar-se para aprender dependências de longo prazo (Yamak; Yujian; Gadosey; 2019; Gao, S. *et al*., 2020) e tem como característica a presença de duas variáveis de estados: a primeira, conhecida como estado oculto (h); a segunda, conhecida como estado da célula (c) (Ayzel; Heistermann, 2021).

Para mais, em sua estrutura, uma LSTM contempla três “portões” nomeados como entrada (*input gate*), esquecimento (*forget gate*) e saída (*output gate*) (FIGURA 22) (Young *et al*., 2018; Yamak; Yujian; Gadosey; 2019; Ayzel; Heistermann, 2021),

destinados a avaliar quais informações podem ou não chegar ao estado da célula (Yamak; Yujian; Gadosey; 2019; Wan *et al*., 2020).

FIGURA 22 - REPRESENTAÇÃO BÁSICA DE UMA LSTM



FONTE: Gao *et al*. (2020b, p. 5).

Quanto ao funcionamento de cada portão, o de esquecimento (f_t) é responsável por determinar quais informações serão destinadas para longe do estado da célula (c_t); o de entrada (i_t) tem como papel principal definir quais novas informações serão armazenadas no estado da célula (c_t); por fim, o de saída (o_t) especifica quais informações do estado da célula são usadas como saída (o_t) (Gao, S. *et al*., 2020). Do ponto de vista matemático, as operações conduzidas dentro de uma LSTM são apresentadas pelo conjunto de equações a seguir:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2)$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (3)$$

$$\hat{c}_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \hat{c}_t) \quad (5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (6)$$

$$h_t = o_t tanh \otimes c_t \quad (7)$$

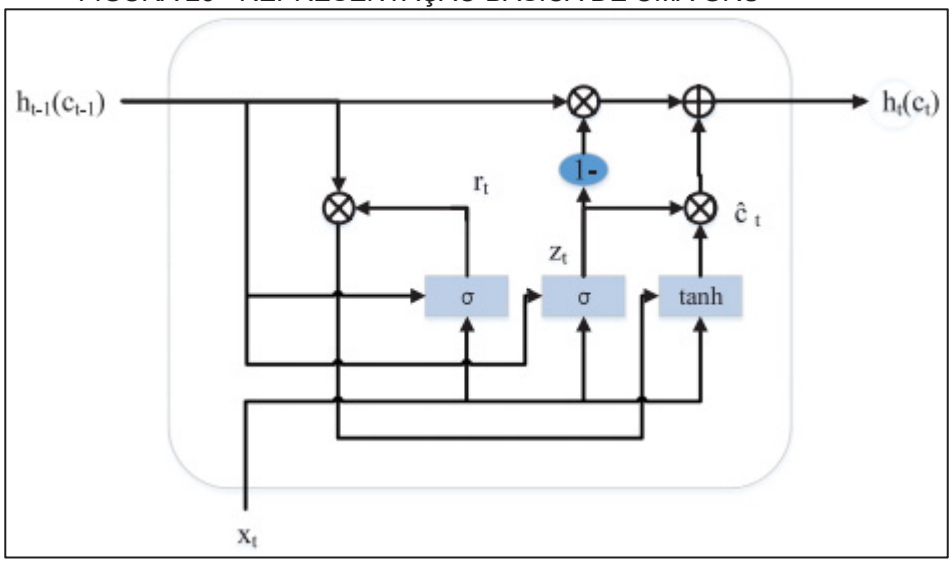
Onde σ corresponde a uma função sigmoide e $tanh$ representa uma função tangente hiperbólica. $\otimes$ indica a multiplicação elemento a elemento de dois vetores,

e W_{xi}, W_{hi}, W_{xf}, W_{hf}, W_{xo}, W_{ho}, W_{xc} e W_{hc} são as matrizes de pesos da rede. Enquanto b_i, b_f, b_o e b_c representam os vieses. Por fim, f_t, i_t e o_t são, respectivamente, os portões de esquecimento, entrada e saída (Gao, S. *et al*., 2020).

Em termos de aplicação, verifica-se a construção de sistemas baseados em LSTM em tarefas como: a) tradução de idiomas; b) robótica; c) análise de imagens; d) sumarização de documentos; e) reconhecimento de voz; f) reconhecimento de imagem; e g) reconhecimento de caligrafia (Wan *et al*., 2019).

Em relação à arquitetura GRU (FIGURA 23), trata-se de uma tentativa de simplificar a LSTM, unindo as duas variáveis de estado (oculto e célula) em apenas uma variável oculta (h) (Ayzel; Heistermann, 2021), além de uma estrutura com dois portões, definidos como redefinir (r) (*reset gate*) e atualizar (z) (*update gate*) (Young *et al*., 2018; Ayzel; Heistermann, 2021).

FIGURA 23 - REPRESENTAÇÃO BÁSICA DE UMA GRU



FONTE: Gao, S. *et al*. (2020, p. 5).

Em uma GRU, o portão de atualização (z) é projetado para controlar o grau em que a informação de estado $h_{t-1}(c_{t-1})$ na etapa de tempo anterior $t-1$ será destinada para a etapa de tempo atual t. Desse modo, quanto maior o valor para o portão de atualização, mais as informações de estado da etapa de tempo anterior são obtidas. Quanto ao portão de redefinição (r), esse controla quanta informação do estado anterior é gravada no conjunto atual. Dessa forma, quanto menor for o valor para esse portão, menos informações do estado anterior serão gravadas (Gao, S. *et al*., 2020).

Matematicamente, as operações conduzidas dentro de uma GRU são apresentadas pelo conjunto de equações a seguir:

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \quad (8)$$

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \quad (9)$$

$$\hat{c}_t = tanh(W_{xc}x_t + W_{hc}(r_t \otimes h_{t-1}) + b_c) \quad (10)$$

$$c_t = (1 - z_t) \otimes c_{t-1} + z_t \otimes \hat{c}_t \quad (11)$$

$$h_t = c_t \quad (12)$$

Onde W_{xr}, W_{hr}, W_{xz}, e W_{hz} são as matrizes de pesos da rede. Enquanto b_r e b_z representam os vieses. Por fim, r_t e z_t são, respectivamente, os portões de atualização e redefinir (Gao, S. *et al*., 2020).

2.6 *DEEP LEARNING* PARA DADOS TABULARES

Embora destaquem-se em abordagens envolvendo visão computacional, classificação de imagens, séries temporais e dados sequenciais, arquiteturas de *deep learning* não são as principais técnicas para dados tabulares - dados estruturados em linhas e colunas -, para os quais, algoritmos tradicionais de aprendizado de máquina, como Regressão Logística, k-NN, SVM e *Random Forest* (Alvi *et al*., 2021), além de modelos mais robustos, baseados em técnicas como o *Gradient Boosting* (Friedman, 2001), tais como o *CatBoost* (Prokhorenkova *et al*., 2018), o *LightGBM* (Ke *et al*., 2017), e o *XGBoost* (Chen; Guestrin, 2016) são os protagonistas (Gorishniy *et al*., 2021; Shwartz-Ziv; Armon, 2022).

Arquiteturas de *deep learning*, quando aplicadas em cenários com dados tabulares, enfrentam desafios como a esparsidade de dados (valores ausentes), variáveis heterogêneas (numéricas, ordinais e categóricas) e a falta de conhecimento prévio sobre a estrutura do conjunto de dados (Shwartz-Ziv; Armon, 2022). Contudo, tendo em vista que podem oferecer um desempenho potencialmente maior, o uso de arquiteturas de *deep learning* em dados tabulares torna-se atraente, uma vez que permite o trabalho em cenários “mistos”, nos quais pode-se agregar dados tabulares e outras estruturas, como imagens e áudio (Gorishniy *et al*., 2021).

Nesse contexto, nos últimos anos, arquiteturas de *deep learning* para dados tabulares foram desenvolvidas (Popov; Morozov; Babenko, 2020; Kadra *et al*., 2021),

dentre as quais, por exemplo, elencam-se: a) *FT-Transformer* (Gorishniy *et al*., 2021); b) *Neural Oblivious Decision Ensembles* (NODE) (Popov; Morozov; Babenko, 2020); c) *TabNet* (Arik; Pfister, 2021); e d) *TabTransformer* (Huang *et al*., 2020).

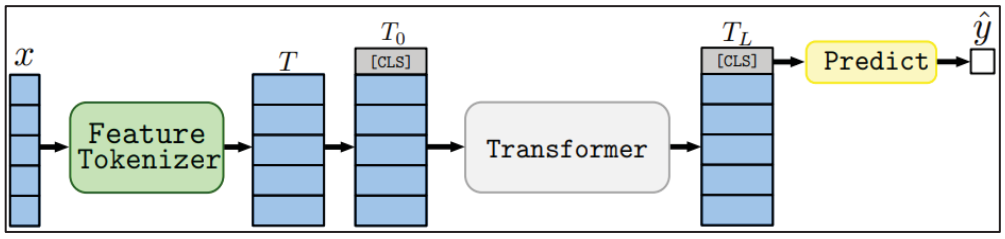
Para mais, além de arquiteturas específicas, problemas com dados tabulares permitem a adaptação de arquiteturas de *deep learning* tradicionais, como uma CNN (1D-CNN) (Shwartz-Ziv; Armon, 2022), ou, ainda, o emprego de métodos mais "simples", como o caso do MLP (Gorishniy *et al*., 2021).

Na sequência, para melhor compreensão das arquiteturas para dados tabulares acima mencionadas, os principais fundamentos dessas são exemplificados.

2.6.1 *FT-Transformer*

A arquitetura denominada *FT-Transformer* (*Feature Tokenizer* + *Transformer*) (FIGURA 24) foi desenvolvida por Gorishniy *et al*. (2021) e consiste em uma adaptação para dados tabulares da arquitetura *Transformer* (Vaswani et al., 2017). De maneira concisa, a *FT-Transformer* transforma os atributos de entrada, categóricos e numéricos, em *embeddings*[7] para, na sequência, aplicar uma pilha de camadas *Transformer*.

FIGURA 24 - AQUITETURA *FT-TRANSFORMER*

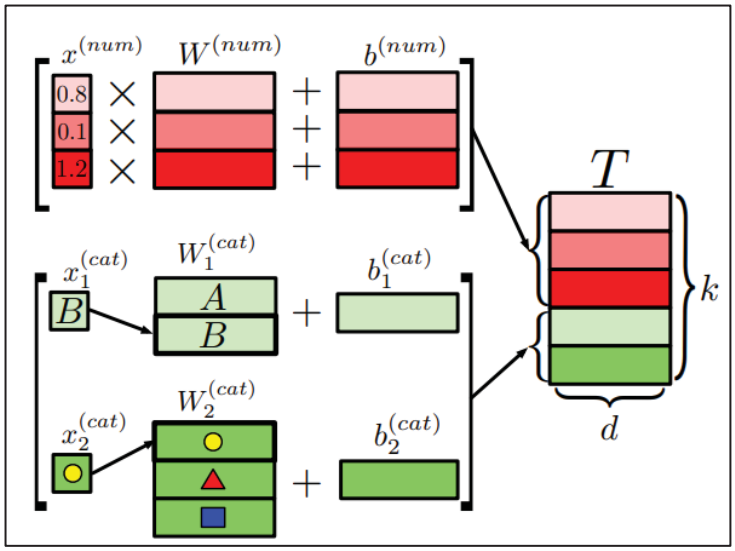


FONTE: Gorishniy *et al*. (2021, p. 4).

O primeiro módulo da arquitetura *FT-Transformer* é denominado *Feature Tokenizer* (FIGURA 25), e tem como função transformar as entradas em *embeddings* (Gorishniy *et al*., 2021).

[7] Representação vetorial contínua de atributos categóricos. De forma específica, é uma maneira de representar cada atributo como uma matriz de *n* elementos preenchida com números flutuantes diferentes de zero (Hagiwara, 2021).

FIGURA 25 - *FT-TRANSFORMER (FEATURE TOKENIZER)*



FONTE: Gorishniy *et al*. (2021, p. 4)

O módulo *Feature Tokenizer* transforma as entradas x em *embeddings* $T \in \mathbb{R}^{kxd}$. Já a incorporação de um recurso x_j é calculada pela seguinte equação (Gorishniy *et al*., 2021):

$$T_j = b_j + f_j(x_j) \in \mathbb{R}^d \quad f_j : \mathbb{x}_j \rightarrow \mathbb{R}^d \tag{13}$$

Onde: b_j representa o viés $j - \text{th}$; $f_j^{(num)}$ é implementado como a multiplicação elemento a elemento com o vetor $W_j^{(num)} \in \mathbb{R}^d$; e $f_j^{(cat)}$ é implementando como a tabela de busca $W_j^{(cat)} \in \mathbb{R}^{S_j \, x \, d}$ para as variáveis categóricas (Gorishniy *et al*., 2021). Ademais:

$$T_j^{(num)} = b_j^{(num)} + x_j^{(num)} . W_j^{(num)} \quad \in \mathbb{R}^d, \tag{14}$$

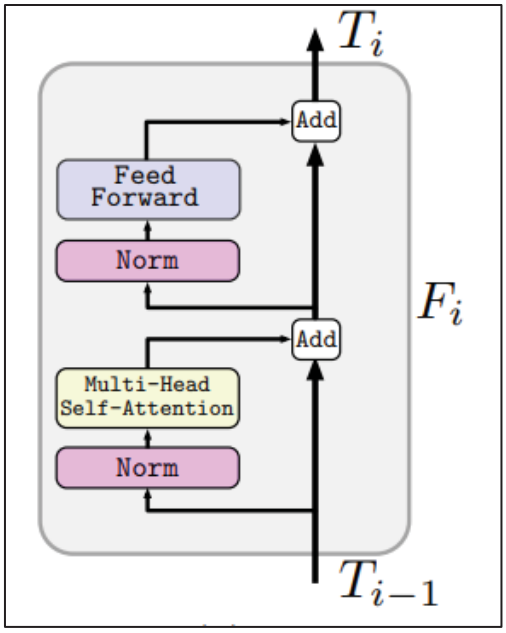
$$T_j^{(cat)} = b_j^{(cat)} + e_j^T W_j^{(cat)} \quad \in \mathbb{R}^d, \tag{15}$$

$$T = pilha \left[T_1^{(num)}, \ldots, T_{k^{(num)}}^{(num)}, T_1^{(cat)}, \ldots T_{k^{(cat)}}^{(cat)} \right] \quad \in \mathbb{R}^{kxd}. \tag{16}$$

Onde: e_j^T é um vetor *one-hot* para a variável categórica correspondente (Gorishniy *et al*., 2021).

Após o módulo *Feature Tokenizer*, o próximo elemento da *FT-Transformer* é o módulo *Transformer* (FIGURA 26), cuja função é processar os *embeddings* oriundos do módulo anterior (Gorishniy *et al*., 2021).

FIGURA 26 - FT-TRANSFORMER (TRANSFORMER)



FONTE: Gorishniy *et al*. (2021, p. 4)

No módulo *Transformer*, a incorporação do *token* de classificação [CLS] é anexado a T, e L camadas de transformadores F_1, ... F_L são aplicadas (Gorishniy *et al*., 2021):

$$T_0 = pilha\left[[\mathrm{CLS}], T\right] \; T_i = F_i(T_{i-1}). \tag{17}$$

Por fim, para a predição, a versão final do *token* de classificação [CLS] é dada pela EQUAÇÃO 18 (Gorishniy *et al*., 2021):

$$\hat{y} = Linear\left(ReLU\left(LayerNorm\left(T_L^{[\mathrm{CLS}]}\right)\right)\right) \tag{18}$$

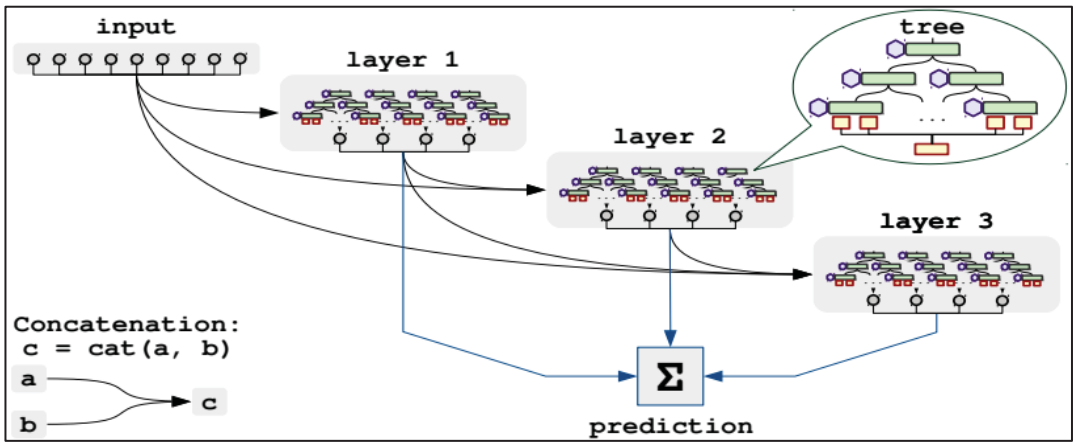
Onde: $Linear$ representa uma transformação linear ($y = x\,A^t + b$); $ReLU$ indica a função de ativação *Rectified Linear Unit* ($f(x) = x^+ = \max(0, x)$); e $LayerNorm$

representa a normalização de camadas em cada um dos minilotes (*mini-batch*) de entrada.

2.6.2 *Neural Oblivious Decision Ensembles*

A arquitetura denominada *Neural Oblivious Decision Ensembles* (NODE) (FIGURA 27) foi desenvolvida por Popov, Morozov e Babenko (2020), e tem como característica principal basear-se em uma classe específica de árvore de decisão conhecida como *Oblivious Decision Trees* (ODT) (Popov; Morozov; Babenko, 2020; Shwartz-Ziv; Armon, 2022).

FIGURA 27 - ARQUITETURA *NODE*

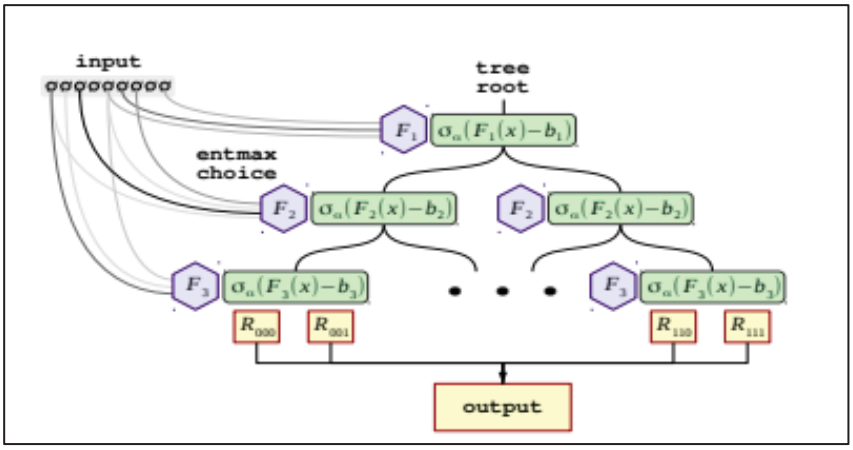


FONTE: Popov, Morozov e Babenko (2020, p. 5).

Resumidamente, uma ODT é uma tabela de decisão que divide os dados ao longo de d recursos de divisão e compara cada recurso a um limite aprendido (Popov; Morozov; Babenko, 2020). Posteriormente, a árvore retorna uma das 2^d respostas possíveis, correspondente ao resultado das comparações. Neste contexto, cada ODT é completamente determinada por suas características de divisão $f \in \mathbb{R}^d$, limiares de divisão $b \in \mathbb{R}^d$ e um tensor d- dimensional de respostas $R \in \mathbb{R}^{d\overbrace{2x2x2}}$ (Popov; Morozov; Babenko, 2020).

A partir do comportamento de uma ODT, tem-se uma camada denominada *NODE* (FIGURA 28), composta por m árvores ODT de igual profundidade d, cuja na entrada, todas as m árvores recebem um vetor comum $x \in \mathbb{R}^n$, contendo n variáveis numéricas (Popov; Morozov; Babenko, 2020).

FIGURA 28 – ARQUITETURA *NODE* (CAMADA *NODE* – EXEMPLO)



FONTE: Popov, Morozov e Babenko (2020, p. 5).

Dentro da camada *NODE*, os recursos de divisão e os limites de divisão são compartilhados em todos os nós internos da mesma profundidade, e a saída é a soma das respostas das folhas dimensionadas pelos pesos escolhidos (Popov; Morozov; Babenko, 2020).

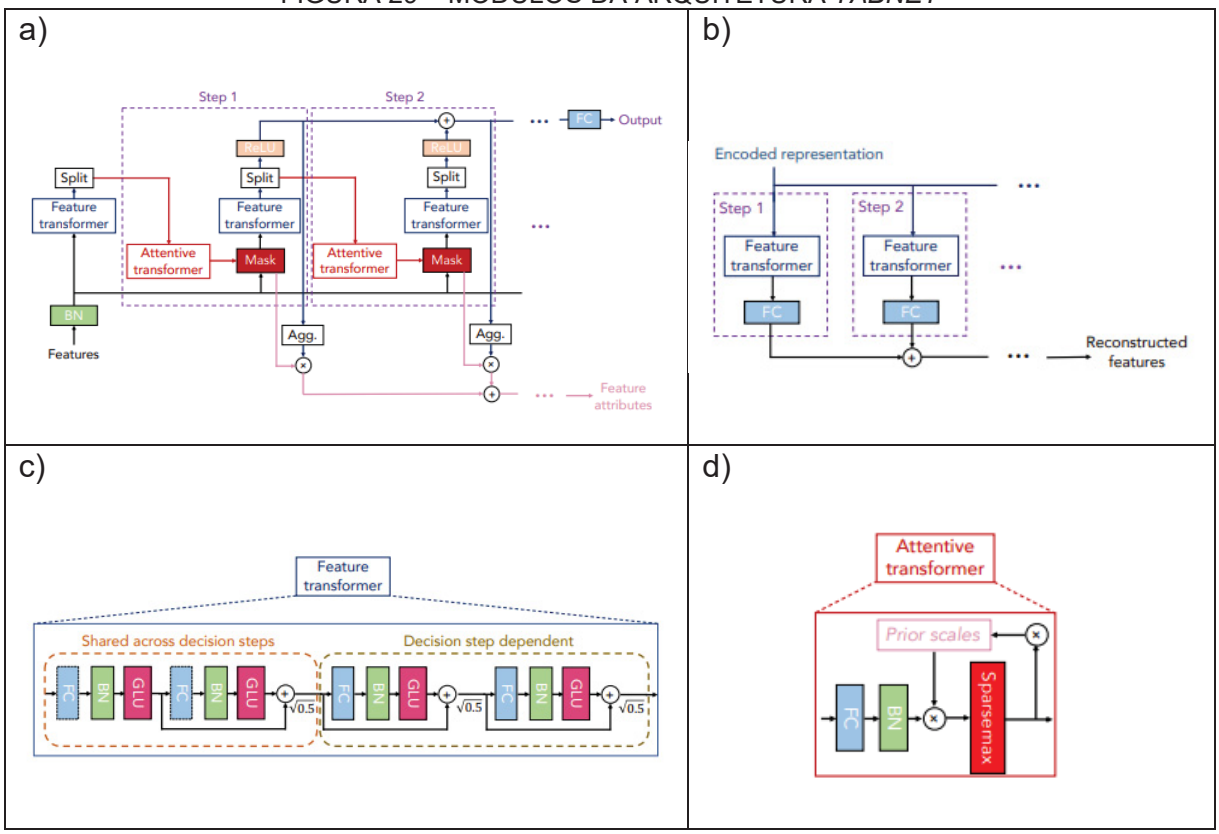
Por fim, a arquitetura *NODE* é composta por camadas densamente conectadas, em que cada camada contém várias árvores com saídas concatenadas e que servem como entrada para a camada subsequente. A previsão final é obtida pela média das saídas de todas as árvores de todas as camadas (Popov; Morozov; Babenko, 2020).

2.6.3 *TabNet*

A arquitetura *TabNet* foi desenvolvida por Arik e Pfister (2021) e caracteriza-se por utilizar atenção sequencial para escolher quais recursos priorizar em cada etapa de decisão, o que permite maior interpretabilidade e um aprendizado mais eficiente, tendo em vista que a capacidade de aprendizado é usada para os recursos mais importantes (Arik; Pfister; 2021).

Em relação aos componentes da arquitetura *TabNet*, esta apresenta dois módulos principais, denominados codificador (*encoder*) (FIGURA 29a) e decodificador (FIGURA 29b), respectivamente. Ademais, a arquitetura *TabNet* apresenta dois blocos secundários, intitulados *feature transformer* (FIGURA 29c) e *attentive transformer* (FIGURA 29d) (Arik; Pfister, 2021).

FIGURA 29 – MÓDULOS DA ARQUITETURA *TABNET*



FONTE: Adaptado de Arik e Pfister (2021, p. 6.682).

O módulo *encoder* é composto por três blocos: a) *feature transformer*; b) *attentive transformer*; e c) *feature masking*. Nesse módulo, um bloco divisor divide a representação processada para ser usada pelo *attentive transformer* da etapa subsequente, bem como para a saída geral. Para cada etapa, a máscara de seleção de recursos (*feature selection mask*) fornece informações interpretáveis sobre a funcionalidade do modelo, e as máscaras podem ser agregadas para obter atribuição importante de recursos globais (Arik; Pfister, 2021).

Quanto ao módulo *decoder*, este é composto por um bloco com *feature transformer* em cada etapa, seguido por camadas totalmente conectadas (FC) em cada etapa de decisão. Finalmente, as saídas são somadas para obter as características reconstruídas (Arik; Pfister, 2021).

Com relação ao bloco *feature transformer*, este inclui camadas que são compartilhadas em todas as etapas de decisão, uma vez que os mesmos recursos são inseridos em diferentes etapas de decisão, além de camadas dependentes da etapa de decisão (Arik; Pfister, 2021). Tomando como exemplo a FIGURA 29c, observa-se uma rede com quatro camadas, sendo duas compartilhadas em todas as etapas de decisão, e duas camadas dependentes da etapa de decisão. Nesta

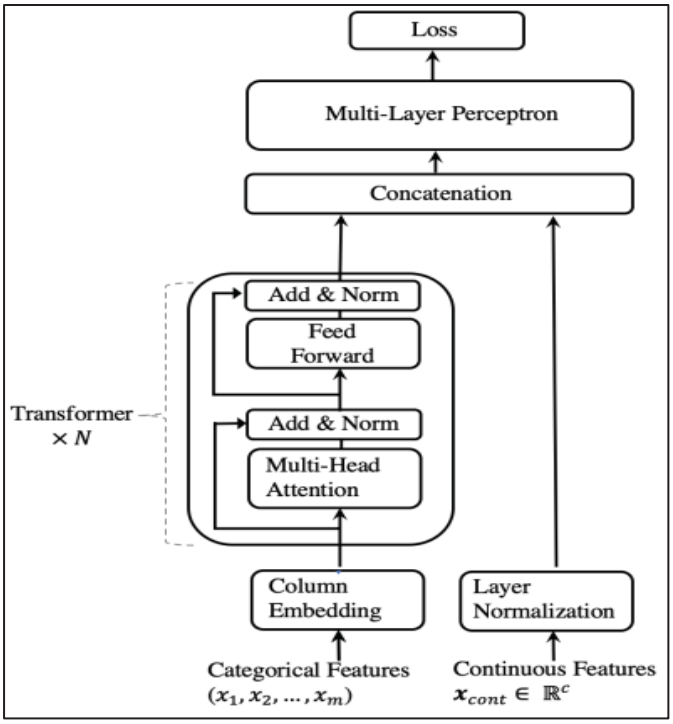
representação, cada camada é composta por uma camada totalmente conectada (FC), normalização de lotes (*batch normalization*) e uma operação denominada *gated linear unit* (GLU) (Arik; Pfister, 2021).

Por fim, na FIGURA 29d, consta uma exemplificação de um bloco *attentive transformer*, onde um mapeamento de camada única é modulado com informação de escala anterior que agrega o quanto cada recurso foi usado antes da etapa de decisão atual. Para a normalização dos coeficientes, uma função *sparsemax* é aplicada (Arik; Pfister, 2021).

2.6.4 *TabTransformer*

A arquitetura *Tab Transformer* (FIGURA 30) foi desenvolvida por Huang *et al*. (2020) e compreende, em sua estrutura, uma camada de incorporação de colunas (*embedding*), uma pilha de N camadas *Transformer* (Vaswani *et al*., 2017) e um *perceptron* multicamadas (Huang *et al*., 2020). Nesta estrutura, cada camada *Transformer* consiste em uma camada de atenção com várias cabeças (*multi-head self-attention*) seguida por uma camada *feedforward* de posição.

FIGURA 30 - ARQUITETURA *TAB TRANSFORMER*



FONTE: Huang *et al*. (2020, p. 2).

Com base na FIGURA 30, seja (x, y) um par de variáveis-alvo, onde $x \equiv \{x_{cat}, x_{cont}\}$. x_{cat} representa as variáveis categóricas e $x_{cont} \in \mathbb{R}^c$ indica todas as c variáveis contínuas. Outrossim, seja $x_{cat} \equiv \{x_1, x_2, \ldots, x_m\}$ com cada x_i representando cada variável categórica, para $i \in \{1, \ldots, m\}$ (Huang *et al*., 2020).

Na sequência, há a incorporação paramétrica de dimensão d para cada uma das x_i variáveis categóricas, considerando uma *column embedding*. Nesta etapa, entende-se que $\{e_{\emptyset_1}(x_1), \ldots, e_{\emptyset_m}(x_m)\}$ seja o conjunto de *embeddings* para todas as variáveis categóricas (Huang *et al*., 2020).

Posteriormente, os *embeddings* paramétricos $E_{\emptyset}(x_{cat})$ são inseridos na primeira camada *Transformer*. A saída da primeira camada *Transformer* é utilizada como entrada para a segunda camada *Transformer* e assim sucessivamente. Cada *embedding* paramétrico é transformando em um *embeddings* contextual quando enviado para a camada *Transformer* superior, por meio de agregação sucessiva de contexto oriunda de outros *embeddings* (Huang *et al*., 2020).

Nesse cenário, a sequência de camadas *Transformer* é representada como uma função f_θ. A função f_θ opera nos *embeddings* paramétricos $\{e_{\emptyset_1}(x_1), \ldots, e_{\emptyset_m}(x_m)\}$ e retorna os *embeddings* contextuais correspondentes $\{h_1, \ldots, h_m\}$, onde $h_1 \in \mathbb{R}^d$ para cada $i \in \{1, \ldots, m\}$ (Huang *et al*., 2020).

Adiante, os *embeddings* contextuais $\{h_1, \ldots, h_m\}$ são concatenados às variáveis contínuas x_{cont} para formarem um vetor de dimensão $(d \times m + c)$. Este vetor é inserido em um *Multilayer Perceptron* (MLP), representado por g_ψ para a predição do alvo y. Para tanto, seja H a entropia cruzada para as tarefas de classificação e o erro quadrático médio para as tarefas de regressão, tem-se a minimização da seguinte função de perda $\mathcal{L}(x, y)$ para aprender todos os parâmetros da arquitetura *TabTransformer* em um aprendizado ponta a ponta por métodos de gradiente de primeira ordem. Os parâmetros da *TabTransfomer* incluem $\emptyset$ para a *embedding column*, θ para as camadas *Transformer*, e ψ para a camada superior MLP (EQUAÇÃO 19) (Huang *et al*., 2020).

$$\mathcal{L}(x, y) \equiv H\left(g_\psi\left(f_\theta\left(E_\emptyset(x_{cat})\right),\ x_{cont}\right), y\right). \tag{19}$$

Com relação ao detalhamento das camadas *Transformer* e para *embedding column*, recomenda-se o estudo na totalidade da pesquisa de Huang *et al*. (2020).

Com a exposição das diferentes arquiteturas de *deep learning* para dados tabulares, julga-se que os conceitos necessários para a compreensão desta pesquisa foram apresentados ao longo do referencial teórico, desta forma, encerra-se esta seção, e na sequência apresentam-se os encaminhamentos metodológicos realizados nesta tese.

3 ENCAMINHAMENTOS METODOLÓGICOS

Esta seção apresenta a caracterização da proposta quanto à sua natureza, ao seu nível, ao seu delineamento e à sua finalidade. Além disso, apresenta os procedimentos de análise e coleta de dados determinados. Para mais, demonstra-se as estratégias adotadas para o planejamento e a construção da base de dados.

3.1 CARACTERIZAÇÃO DA PESQUISA

Esta tese consiste em uma pesquisa de natureza quantitativa (Perdigão, 2012; SampierI; Collado; Lucio, 2013), com sua caracterização estabelecida de acordo com os critérios básicos apresentados por Vergara (2004), que consistem em classificar uma pesquisa quanto aos fins e aos métodos. Neste sentido, quanto aos fins, esta pesquisa é exploratória (Martins Júnior, 2012; Sampieri; Collado; Lucio, 2013), pois objetiva “aumentar a familiaridade com o problema de pesquisa” (Martins Júnior, 2012, p. 28) e os seus desdobramentos podem “identificar conceitos ou variáveis promissoras, estabelecer prioridades para pesquisas futuras ou sugerir afirmações e postulados” (Sampieri; Collado; Lucio, 2013, p. 107).

Ainda em relação aos fins, trata-se de uma pesquisa aplicada (Perdigão, 2012), tendo em vista o “propósito prático e específico para gerar conhecimento ou avanço do conhecimento” (Perdigão, 2012, p. 110). Com relação ao segundo critério estabelecido por Vergara (2004), este estudo apresenta-se como uma pesquisa bibliográfica (Lakatos; Marconi, 2017), pois utiliza-se para a sua sistematização o conhecimento presente em artigos científicos, anais de eventos, livros, dissertações e teses que abordam os conteúdos necessários ao estudo.

Ademais, o uso da pesquisa bibliográfica fornece as condições necessárias para atingir os objetivos específicos propostos, uma vez que, para cada objetivo, analisam-se *surveys*, artigos de revisão e revisões sistemáticas condizentes com os procedimentos propostos em cada um dos objetivos definidos. Para mais, recorre-se à revisão integrativa da literatura, na qual se adota a recomendação PRISMA (Moher *et al*., 2009) para identificar, nas pesquisas multimodais sobre reconhecimento de emoções em músicas, as modalidades necessárias para compor a base de dados desta pesquisa.

Ademais, este trabalho é uma pesquisa documental (Lakatos; Marconi, 2017), caracterizada pela manipulação de arquivos de músicas (.mp3, .wav, MIDI etc.) diretamente de plataformas apropriadas. Por fim, utiliza-se de questionário (Lakatos; Marconi, 2017) para a condução da revisão das músicas selecionadas que compõem o modelo discreto identificado.

3.2 COLETA E ANÁLISE DE DADOS

Os procedimentos de coleta e análise de dados desta pesquisa, estes são apresentados no QUADRO 23.

QUADRO 23 - PROCEDIMENTOS DE COLETA E ANÁLISE DOS DADOS

(continua)

Objetivo Específico	**Procedimentos**	**Como**	**Principais Autores**
Identificar o modelo de emoções a ser utilizado	Analisar a literatura sobre emoções e músicas	Pesquisa bibliográfica	Juslin e Laukka (2003, 2004); Bigand *et* al. (2005); Eerola e Vuoskoski (2011); Nunes-Silva *et* al. (2016); Juslin (2013b); Warrenburg, (2020)
Identificar quais modalidades empregar para o reconhecimento de emoções em músicas	Analisar *surveys*, artigos de revisão e revisões sistemáticas sobre reconhecimento de emoções em músicas	Pesquisa bibliográfica	Kim *et al*. (2010); Yang e Chen (2012); Panda, Malheiro e Paiva (2020b)
	Analisar pesquisas multimodais sobre reconhecimento de emoções em músicas	Revisão integrativa de literatura	Lu *et al*. (2010); Schuller, Dorfner e Rigoll (2010); Schuller *et al*. (2010); Zhao, Yang e Chen (2010); Schuller, Weninger e Dorfner (2011); Panda *et al*. (2013); Imbrasaite (2015); Przybysz (2016)

FONTE: O autor (2021).

QUADRO 23 - PROCEDIMENTOS DE COLETA E ANÁLISE DOS DADOS

(continua)

Objetivo Específico	Procedimentos	Como	Principais Autores
Construir uma base multimodal composta por características relacionadas às modalidades identificadas	Identificar as músicas relacionadas ao modelo de emoções considerado	Pesquisa documental Busca nas plataformas *AllMusic* e *Spotify*	Juslin e Laukka (2003); Yang e Lee (2009); Panda *et al*. (2013); Chen *et al*. (2017)
	Coletar as músicas identificadas	Pesquisa documental Coleta dos áudios diretamente do *Youtube*	Moreira (2019); Santana *et al*. (2020)
	Padronizar a duração das músicas	Padronizar a duração das músicas em 30 segundos	Yang e Chen (2011); Chen *et al*. (2015); Russo *et al*. (2020)
	Submeter as músicas identificadas para avaliação/revisão	Teste de escolha forçada	Ramos (2008)
		Kappa de Cohen Kappa de Fleiss	Cohen (1960); Fleiss (1971); Landis e Koch (1977); Hallgreen (2012)
	Extrair as características identificadas	Características relacionadas aos acordes	Mauch (2010); Przybysz (2016)
		Características relacionadas ao áudio	McKay (2010); Bogdanov *et al*. (2013); McFee *et al*. (2015); Panda, Malheiro e Paiva (2020a, 2020b)
		Características relacionadas às letras	Malheiro *et al*. (2018); Hasan e Matin (2021); Özçift *et al*. (2021); Rustan *et al*. (2021); Ahanin e Ismail (2022)
		Características simbólicas	Cuthbert e Ariza (2010); Mckay (2010); Wu, Chen e Su (2019; 2020); Wu *et al*. (2021)
		Características relacionadas à voz	El Ayadi, Kamel e Karray (2011); Fernandes *et al*. (2017) Akçay e Oğuz (2020); Hennequin *et al*. (2020)

FONTE: O autor (2021).

QUADRO 23 - PROCEDIMENTOS DE COLETA E ANÁLISE DOS DADOS

(conclusão)

Objetivo Específico	Procedimentos	Como	Principais Autores
Comparar diferentes arquiteturas de *deep learning* para o reconhecimento de emoções na base elaborada	Identificar e aplicar arquiteturas de *deep learning* para o problema proposto	Pesquisa Bibliográfica Realização de cenários de classificação com os métodos selecionados	Schmidt e Kim (2011), Li, J. *et al*. (2015); Thammasan, Fukui e Numao (2016); Gao, J. *et al*. (2020) Hizlisoy, Yildirim e Tufekci (2021)
Comparar o desempenho de classificação obtido com as abordagens multimodais desenvolvidas com o desempenho proporcionado por abordagens unimodais, construídas a partir de algoritmos de aprendizado de máquina	Avaliar os resultados obtidos	Métricas de avaliação empregadas em problemas de classificação	Kamilaris e Prenafeta-Boldú (2018); Chicco e Jurman (2020); Kok, Azween e Jhanjhi (2020); Hizlisoy, Yildirim e Tufekci (2021)

FONTE: O autor (2021).

A seguir, apresenta-se uma breve explicação acerca dos experimentos piloto e ajustado conduzidos para a construção da base de dados deste estudo.

3.3 CONSTRUÇÃO DOS EXPERIMENTOS

Antes de detalhar os passos relacionados à construção da base de dados desta tese, julga-se necessário esclarecer a lógica empregada para o desenvolvimento dos dois experimentos conduzidos, aqui denominados como experimento piloto e experimento ajustado, para a avaliação/revisão das músicas selecionadas.

Em primeiro lugar, desenvolveu-se um estudo considerando um modelo discreto de emoções composto por seis emoções, sendo essas: a) amor/ternura; b) calma; c) felicidade; d) medo; e) raiva; e f) tristeza. Este conjunto foi definido a partir da análise de estudos que investigaram as emoções mais “comuns” associadas à música. Para mais, os trabalhos conduzidos por Juslin e Laukka (2003, 2004) serviram como referencial para a escolha dessas emoções.

A partir da identificação do modelo discreto, a plataforma *AllMusic* foi selecionada como fonte primária para a coleta das músicas representativas das emoções-alvo, tendo em vista a sua aplicação em estudos correlatos, bem como por

apresentar um processo de rotulação controlado. Dessa forma, para o experimento piloto, foi definido como objetivo revisar as músicas identificadas na plataforma, com o intuito de garantir que a percepção emocional dos avaliadores selecionados correspondesse às emoções oriundas da plataforma *AllMusic*.

Com base nos resultados obtidos nesta etapa[8], visualizou-se baixa concordância entre os voluntários e a plataforma (14,09%). De forma semelhante, ao se verificar a concordância entre os voluntários, desconsiderando as emoções observadas na plataforma *AllMusic*, o resultado geral não apresentou resultados robustos de concordância (35,07%), mesmo que, para essa concordância, os resultados obtidos tenham sido estatisticamente significativos ao se considerar as avaliações/revisões por duplas (p-valor < 0,0001; IC 95%).

Como consequência, a quantidade de registros na base de dados (919) ficou aquém do esperado para um problema com arquiteturas de *deep learning*. Outrossim, a relação de músicas por emoções ficou desbalanceada, apresentando indícios de possíveis dificuldades no que concerne à distinção das emoções por parte dos voluntários, quer seja pelo fato de a base não ter sido revisada em sua totalidade por todos os participantes, ou por não existir definições claras sobre os significados de cada uma das emoções consideradas.

Sendo assim, optou-se por ajustar determinados passos empregados no experimento piloto e conduzir um novo processo de avaliação/revisão (experimento ajustado), considerando: a) um novo modelo discreto de emoções, com um menor número de emoções; b) número maior de músicas para revisão; c) revisão da base conduzida por todos os participantes; e d) um processo de pós-avaliação junto aos voluntários, visando identificar possíveis dificuldades associadas à percepção emocional dos voluntários.

A comparação entre os experimentos piloto e ajustado é apresentada no QUADRO 24.

[8] Os resultados obtidos para o processo de revisão conduzido no experimento piloto são elucidados na seção 4.1.

QUADRO 24 - PROCEDIMENTOS DOS EXPERIMENTOS PILOTO E AJUSTADO

Características	**Experimento piloto**	**Experimento ajustado**
Emoções	Amor/ternura; calma; felicidade; medo; raiva e tristeza	Alegria; raiva/medo; serenidade e tristeza
Músicas	2.620	4.797
Fonte primária	*AllMusic*	*AllMusic* e *Spotify*
Número de avaliadores	Oito	Cinco
Especialistas e não especialistas	Sim	Sim
Esquema de avaliação	Duplas avaliando apenas um conjunto de músicas	Dois voluntários avaliando o conjunto inteiro de músicas; três voluntários avaliando apenas conjuntos de músicas
Tipo de escolha	Escolha forçada	Escolha forçada
Pós-avaliação	Não	Sim

FONTE: O autor (2022).

Na sequência, os procedimentos metodológicos empregados, bem como as justificativas para cada experimento são detalhados.

3.4 DEFINIÇÃO DO MODELO DE EMOÇÕES

Antes de apresentar os modelos de emoções selecionados para esta pesquisa, verificou-se que não há consenso acerca de qual taxonomia ou número preciso de rótulos emocionais podem ser experimentados por um usuário (Russo *et al.*, 2020; Warrenburg, 2020). Porém, em virtude do presente estudo ser um problema de classificação do tipo *single-label* (Yang; Dong; Li, 2017), visualizou-se a necessidade em se adotar um conjunto relacionado à abordagem categórica/discreta das emoções.

Além do tipo de problema associado à pesquisa, a opção por uma abordagem categórica/discreta encontra-se alinhada aos estudos sobre cognição musical, nos quais essa abordagem é dominante (Warrenburg, 2020). Similarmente, em comparação com os modelos dimensionais, os modelos discretos produzem desempenho de classificação semelhante ao se considerar a percepção emocional em músicas (Eerola; Vuoskoski, 2011; Warrenburg, 2020).

Todavia, no âmbito musical, é preciso ressaltar que os pesquisadores tendem a não apresentarem os seus modelos discretos como representações diretas das emoções básicas (Juslin, 2013b; Warrenburg, 2020), evitando, por exemplo, argumentações sobre determinadas características relacionadas às emoções básicas, como a existência de um passado evolutivo e o uso de substratos neurais comuns (Warrenburg, 2020).

Dessa forma, nesta pesquisa, adota-se a expressão “modelo discreto” para representar as emoções consideradas, mesmo que, porventura, haja relação direta do conjunto selecionado com taxonomias relacionadas às emoções básicas. Para mais, adverte-se que, mesmo que as emoções básicas possam ser “privilegiadas” quanto à expressão musical, isso não significa que outras emoções também não possam ser transmitidas na música (Juslin, 2013b).

Diante do contexto apresentando, como ponto de partida, a primeira opção para o modelo emocional deste estudo consistiu no conjunto composto pelas emoções felicidade, medo, nojo, raiva, surpresa e tristeza, uma vez que essas são conhecidas na literatura como “emoções universais” (Saarimäki *et al*., 2016; An *et al*., 2017; Wang *et al*., 2019; Wingenbach *et al*., 2020).

Além do aspecto “universal”, recorreu-se ao estudo elaborado por Ekman (2016) que obteve, junto a uma amostra de pesquisadores que estudam emoções, valores relevantes em termos de aceitação para os adjetivos do conjunto das “emoções universais”, sendo os respectivos valores correspondentes a: a) 91,00% (raiva); b) 90,00% (medo); c) 86,00% (nojo); d) 80,00% (tristeza); e e) 76,00% (felicidade). Por fim, para a emoção surpresa, a taxa de aceitação situou-se entre 40,00% a 50,00%.

No que consiste à aceitação para as emoções listadas por Ekman (2016), é necessário ressaltar que esta representa a escolha dos voluntários da pesquisa ao serem questionados, a partir de uma lista com 18 adjetivos, quais eram aqueles que deveriam ser considerados estabelecidos empiricamente (Ekman, 2016).

Ao se transpor o conjunto das “emoções universais” para o campo musical, contudo, percebeu-se que há emoções que, de certa forma, são inadequadas ou difíceis de serem representadas ou percebidas, dentre as quais é possível citar o nojo e a surpresa (Kallinen, 2005; Eerola; Vuoskoski, 2013; Argstatter, 2015; Siedlecka; Denson, 2018).

Com relação ao nojo, Kallinen (2005) aponta que essa emoção raramente é desenvolvida na música ocidental, tendo em vista que a tradição de composição se destina a situações de diversão, prazer estético e fins práticos, como missas, funerais e festivais. De modo semelhante, Argstatter (2016) descreve que o nojo é um instinto moldado no curso da socialização cultural, sendo, portanto, extremamente difícil encontrar um “código musical” internacional para representá-lo.

Siedlecka e Denson (2018), ao revisarem a literatura em busca de evidências sobre a indução das “emoções universais" em cinco técnicas experimentais comuns de indução de emoções - estímulos visuais, música, recordação autobiográfica, procedimentos situacionais e imagens -, não conseguiram identificar estudos nos quais a música tenha sido empregada para a indução de nojo.

Da mesma forma, para surpresa, os autores não encontraram evidências suficientes de que a música é capaz de induzir essa emoção (Siedlecka; Denson, 2018). Por sua vez, Kallinen (2005) afirma que a surpresa, na própria concepção de emoção, é de difícil avaliação, por conta da sua natureza, que pode ser tanto negativa, quanto positiva.

Sendo assim, em virtude das emoções nojo e surpresa não apresentarem indícios assegurando as suas relações com a música, optou-se por desconsiderá-las e dar sequência na definição do modelo discreto do estudo considerando as emoções restantes (felicidade, medo, raiva e tristeza). Outrossim, como forma de complementar esse conjunto, recorreu-se a estudos que tiveram como objetivo elencar possíveis emoções com maior “facilidade” de serem percebidas por ouvintes ou comunicadas por intérpretes (QUADRO 25).

QUADRO 25 - RELAÇÃO DOS ESTUDOS CONSULTADOS SOBRE EMOÇÕES PERCEBIDAS OU COMUNICADAS NA MÚSICA

Estudo	**Emoções consideradas/ Adjetivos considerados**	**Principais emoções/adjetivos**
Wells e Hakanen (1991)	Amor, confiança, esperança, excitação, felicidade, medo, orgulho, paixão, pesar, raiva e tristeza	Excitação, felicidade e amor
Gabrielsson e Juslin (1996)	Afetuoso, amedrontado, feliz, raivoso, sem expressão, solene e triste	Feliz, raivoso e triste
Balkwill e Thompson (1999)	Alegria, paz, raiva e tristeza	Alegria, raiva e tristeza
Laukka e Gabrielsson (2000)	Afetuoso, amedrontado, feliz, raivoso, sem expressão, solene e triste	Amedrontado, feliz, raivoso e triste
Bigand *et al*. (2005)	Felicidade, medo/raiva, serenidade e tristeza	Felicidade, medo/raiva, serenidade e tristeza
Kreutz *et al*. (2008)	Felicidade, medo, paz, raiva e tristeza	Felicidade e paz
Vieillard *et al*. (2008)	Ameaça, felicidade, tranquilidade e tristeza	Ameaça, felicidade e tristeza
Fritz *et al*. (2009)	Assustado/temeroso, feliz e triste	Assustado/temeroso, feliz e triste
Bresin e Friberg (2011)	Calma, felicidade medo e tristeza	Calma, felicidade medo e tristeza
Mohn, Argstatter e Wilker (2011)	Felicidade, medo, nojo, raiva, surpresa e tristeza	Felicidade e tristeza
Quinto, Thompson e Taylor (2013)	Felicidade, medo, neutra, raiva, ternura e tristeza	Felicidade, medo, raiva e tristeza
Saarikallio, Vuoskoski e Luck (2014)	Alegria/felicidade, desejo, esperança, medo, poder, raiva, ternura e tristeza	Alegria/felicidade, medo, raiva, ternura e tristeza

FONTE: O autor (2021).

Com base no QUADRO 25, verifica-se que felicidade (alegria), medo (ameaça, amedrontado, assustadora, temeroso), tristeza (triste) e raiva (raivoso) foram as emoções mais frequentemente consideradas e com maior número de indícios positivos de relacionarem-se com a música, no que diz respeito à percepção dos ouvintes ou à intenção do intérprete. Além disso, mesmo que não tão frequentes, emoções como calma (pacífica, paz, serenidade, tranquilidade) e ternura (afetuoso) também se apresentaram associadas à música.

3.4.1 Modelo de emoções considerado para o experimento piloto

Com base nos resultados presentes no QUADRO 26, observa-se que as emoções calma, felicidade, medo, raiva, ternura e tristeza já foram destacadas em outros estudos, como o conduzido por Juslin e Laukka (2004), no qual os autores consultaram 141 ouvintes e, ao questioná-los acerca de quais emoções a música poderia expressar, obtiveram como principais respostas: a) alegria (99,00%); b) tristeza (91,00%); c) amor (90,00%); d) calma (87,00%); e) raiva (82,00%); e f) ternura (82,00%) (Juslin; Laukka, 2004).

Em estudo anterior, os mesmos autores, ao analisarem 104 estudos de expressão vocal e 41 estudos sobre *performance* musical, elencaram cinco categorias de emoções com evidências de serem comumente comunicadas e percebidas em termos de expressão vocal e *performance* musical, sendo essas: a) raiva; b) medo; c) felicidade; d) tristeza; e e) amor/ternura (Juslin; Laukka, 2003).

Para formularem esse conjunto, Juslin e Laukka (2003), com base na dificuldade existente para definir precisamente os termos empregados para representar as emoções, agruparam diferentes adjetivos, de diferentes idiomas, com base na “proximidade” dos conceitos representados (ex: tristeza (depressão, pesar etc.); felicidade (alegria, euforia etc.)) (Juslin; Laukka, 2003).

Comparando os dois estudos (Juslin; Laukka, 2003, 2004), percebe-se que as cinco emoções listadas anteriormente estão presentes, porém, há presença de uma sexta emoção, calma, cujo percentual obtido é maior do que os das emoções raiva e ternura. Outrossim, das quatro “emoções universais” (felicidade, medo, raiva e tristeza), consideradas para compor o modelo discreto desta tese, a única abaixo do primeiro quartil foi o medo (63,00%) (Juslin; Laukka, 2004). Contudo, ressalta-se que a emoção ansiedade (*anxiety*), tratada por Juslin e Laukka (2003) como parte da “família emocional” do medo, teve taxa de concordância de 75,00%.

Finalmente, as emoções identificadas por Juslin e Laukka (2003, 2004) assemelham-se a outras propostas de emoções relacionadas à música, como, por exemplo, a escala GEMS (Zentner; Grandjean; Scherer, 2008), na qual visualiza-se termos como alegria, calma, ternura e tristeza. Desse modo, a partir do exposto nesta subseção, definiu-se, para compor o modelo discreto para o experimento piloto, as seguintes emoções: a) amor/ternura; b) calma; c) felicidade; d) medo; e) raiva; e f) tristeza.

3.4.2 Modelo de emoções considerado para o experimento ajustado

Conforme relatado no experimento piloto, seis emoções foram definidas e consideradas para a revisão das músicas, de acordo com um levantamento de estudos sobre emoções “mais facilmente” comunicadas ou percebidas na música. Todavia, com base nos resultados obtidos, levantou-se a possibilidade de que as emoções identificadas não estavam bem delimitadas, o que pôde ter dificultado a tarefa de avaliação por parte dos voluntários, resultando em baixos valores de concordância e, consequentemente, na redução do conjunto de dados (músicas).

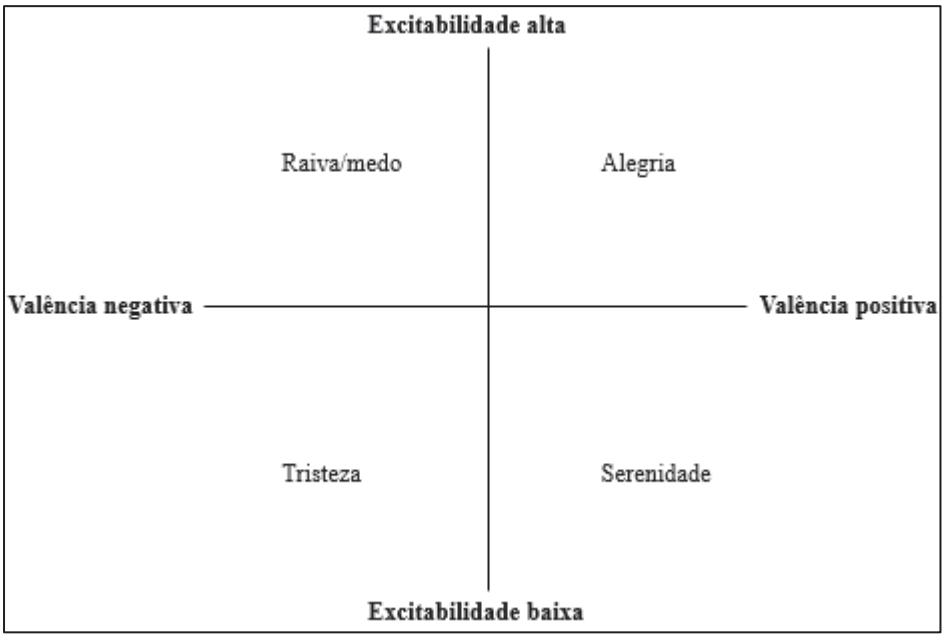
Ressalta-se, conforme descrito anteriormente, que as músicas da plataforma *AllMusic* têm a possibilidade de apresentarem mais de uma emoção. Além disso, como forma de auxiliar os voluntários, apenas os termos de busca considerados como “sinônimos” das seis emoções-alvo definidas foram disponibilizados. Sendo assim, não houve indicativos do que cada emoção poderia representar ao estar inserida no contexto musical.

Buscando contornar essas limitações, julgou-se prudente optar por um modelo no qual fosse possível, por exemplo, obter um conjunto de emoções distintas uma das outras, o que consiste em um requisito chave em abordagens construídas a partir de modelos discretos (Russo *et al*., 2020). Neste contexto, como modelo para o experimento ajustado, empregou-se o modelo proposto por Bigand *et al*. (2005), no qual os autores estabeleceram as emoções alegria, raiva/medo, serenidade e tristeza.

O modelo proposto por Bigand *et al*. (2005) pode ser interpretado, de certa forma, como uma abordagem emocional “mista”, tendo em vista que apresenta as emoções de forma discreta e, ao mesmo tempo, as dispõem em quadrantes, considerando as dimensões excitabilidade e valência, adequando-se, por exemplo, ao modelo circumplexo proposto por Russell (1980).

A representação do modelo proposto por Bigand *et al*. (2005) é apresentada na FIGURA 31.

FIGURA 31 - REPRESENTAÇÃO DO MODELO DE EMOÇÕES PROPOSTO POR BIGAND *et al*. (2005)



FONTE: Adaptado de Nunes-Silva *et al*. (2016, p. 716).

Com base na FIGURA 31, observa-se uma relação de antagonismo entre as emoções: a) alegria (valência positiva/excitabilidade alta) x tristeza (valência negativa/excitabilidade baixa); e b) raiva/medo (valência negativa/excitabilidade alta) x serenidade (valência positiva/excitabilidade baixa).

Visualiza-se, além, que as emoções “raiva” e “medo” são tratadas como uma única categoria (raiva/medo), tendo em vista, por exemplo, que ocupam a mesma dimensão em relação à excitabilidade e à valência (Russell, 1980; Juslin; Laukka, 2004), bem como resultados em outros estudos apontarem que essas emoções costumam ser confundidas entre si (Robazza; Macaluso; D'urso, 1994; Mohn; Argstatter; Wilker, 2011).

Com relação ao significado considerado para as quatro emoções do modelo proposto, principalmente no que tange à relação com a música, adotou-se como referência o conjunto de sensações e sentimentos desenvolvido por Nunes-Silva *et al*. (2016), nomeado como Protocolo de Avaliação de Percepções Subjetivas (PAPS) (QUADRO 26).

QUADRO 26 - PROTOCOLO DE AVALIAÇÃO DE PERCEPÇÕES SUBJETIVAS (PAPS)

Alegria	**Raiva/Medo**	**Serenidade**	**Tristeza**
Disposto	Agressividade	Bem-estar	Apatia
Alegria	Medo	Relaxado	Tédio
Empolgado	Irritação	Aliviado	Tristeza
Euforia	Impaciente	Corpo leve	Angústia
Descontração	Raiva	Serenidade	Abafamento
Vivacidade	Estressado	Calma	Pessimista
Otimista	Assustado	Tranquilidade	Memórias negativas
Imagens positivas	Nervoso	Plenitude	Vazio existencial
Memórias positivas	Corpo pesado	Paz	Solidão
Criatividade	Aflição	Sentimento de religiosidade	Imagens negativas
Ânimo	Ansiedade	Sentimento de elevação	Abatimento
Bem-humorado	Apreensivo	Confiança	Amargura
Satisfeito	Preocupado	Seguro	Sem energia
Êxtase	Tenso	Confortável	Desamparado
Contentamento	Desconfortável	Sossegado	Desanimado

FONTE: Adaptado de Nunes-Silva *et al*. (2016, p. 717).

De acordo com as autoras, o PAPS é uma abordagem composta por 60 sentimentos ou sensações, sendo 15 para cada uma das emoções do modelo proposto por Bigand *et al*. (2005), e “foi desenvolvido com intuito de apreender a percepção subjetiva dos sentimentos e sensações dos participantes em relação às músicas escutadas” (Nunes-Silva *et al*., 2016, p. 715).

3.5 ESTRUTURA E CONSTRUÇÃO DAS BASES DE DADOS

Embora existam conjuntos de dados musicais devidamente rotulados quanto ao aspecto emocional (Zhang *et al*., 2018), a reutilização destes em novos estudos nem sempre é viável, devido, por exemplo, às questões de direitos autorais das canções (Aljanaki; Yang; Soleymani, 2017; Panda, 2019; Chaturvedi *et al*., 2021). Deste modo, graças ao aspecto privado de alguns conjuntos de dados (Panda, 2019), muitas vezes não é possível acessar e manipular diretamente os arquivos musicais

utilizados pelos pesquisadores para replicar os procedimentos de extração de características.

Sendo assim, para o presente estudo decidiu-se por criar uma base de dados própria, em razão das seguintes necessidades: a) utilização de músicas relacionadas às emoções do modelo discreto definido (piloto e ajustado); b) presença de letras, em inglês, para todos os registros; c) adoção de intervalo de duração padronizado para segmentação das músicas; d) manipulação dos arquivos musicais para extrair as características; e e) adoção de um processo de revisão conduzido por terceiros.

Além dessas necessidades, analisando pesquisas que compilaram bases de dados famosas no campo de REM, como as elaboradas por Panda (2019) e Gómez-Cañón *et al*. (2021), é possível visualizar elementos que dificultam o reaproveitamento dessas bases nesta pesquisa. Em sua tese, por exemplo, Panda (2019, p. 121-125) analisou 17 bases de dados tradicionalmente consideradas em trabalhos de REM, das quais 11 apresentam menos de 1.000 registros e nove conjuntos utilizam modelos dimensionais (Panda, 2019, p. 121-125).

Das bases presentes no trabalho de Panda (2019), apenas duas contam com mais de duas modalidades: a) *Greek Music Dataset* (GMD), composta por áudio, letras e MIDI (Makris; Karydis; Sioutas, 2015); e b) *Multi-modal MIREX-like*, também composta por áudio, letras e MIDI (Panda *et al*., 2013). Todavia, essas bases não apresentam o mesmo número de registros para todas as modalidades consideradas (Panda, 2019, p. 121-125).

De maneira semelhante ao levantamento elaborado por Panda (2019), Gómez-Cañón *et al*. (2021) compilaram um conjunto de bases de dados e disponibilizaram a relação em um repositório *on-line*[9]. Das 16 bases estudadas pelos autores, oito constam na pesquisa conduzida por Panda (2019), sendo sete com menos de 1.000 registros. Somadas a essas sete bases, Gómez-Cañón *et al*. (2021) apresentam mais cinco bases com menos de 1.000 registros. Por fim, nenhuma das bases compiladas pelos autores apresenta mais de uma modalidade.

A partir dos elementos apresentados acerca de alguns conjuntos de dados existentes para a tarefa de REM, nas próximas subseções são apresentados os passos adotados na construção das bases de dados multimodais desta pesquisa.

---

[9] Disponível em: https://github.com/juansgomez87/datasets_emotion#music-emotion-recognition-datasets. Acesso em: 05 nov. 2021.

3.5.1 Duração e segmentação das músicas

A estratégia de segmentação adotada neste estudo consistiu em selecionar, de maneira semiautomática, os 30 primeiros segundos com letras das canções. Entendeu-se que este era o mecanismo mais “simples” para garantir, minimamente, que todos os registros apresentassem conteúdo relacionados às letras e à voz, permitindo, assim, o caráter multimodal da proposta.

Como justificativa para a utilização de segmentos com 30 segundos, recorreu-se a estudos que discutiam sobre qual a “duração ideal” para o reconhecimento de emoções em músicas, seja no escopo desta tese - abordagem computacional - ou destinados a avaliar qual o tempo mínimo necessário para a produção de uma resposta emocional em voluntários. A síntese dos estudos consultados está listada no QUADRO 27.

QUADRO 27 - INDICATIVOS ACERCA DA DURAÇÃO DE SEGMENTOS PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS

(continua)

Estudo	**Área**	**Apontamentos**
Bigand, Filipic e Lalitte (2005)	Cognição	Resultados consistentes de que **menos de 1 s** de música é suficiente para incutir respostas emocionais elaboradas nos ouvintes.
Bachorik *et al*. (2009)	Cognição	O tempo médio para a formulação de uma resposta emocional variou de **8,31 a 11 segundos.**
Xiao *et al*. (2008)	MER	Melhores taxas de classificação são obtidas com os segmentos musicais de **8 segundos e 16 segundos.**
Hu *et al*. (2008)	MER	**30 segundos**: redução da carga dos avaliadores; redução no processamento de áudio; alivia as possíveis variações emocionais ao longo da canção.
MacDorman, Ough e Ho (2007)	MER	Classificações sobre **amostras** e **não** sobre a **música inteira;** Utilizar o segmento **mais curto possível.**

FONTE: O autor (2022).

QUADRO 27 - INDICATIVOS ACERCA DA DURAÇÃO DE SEGMENTOS PARA O RECONHECIMENTO DE EMOÇÕES EM MÚSICAS

(conclusão)

Estudo	Área	Apontamentos
Yang e Chen (2012)	MER	O levantamento da literatura revela que o uso de um segmento de **30 segundos parece ser comum**, talvez porque corresponde à duração típica de uma seção do refrão de música popular.
Aljanaki, Wiering e Veltkamp (2015)	MER	Ironicamente, as abordagens estáticas e dinâmicas adotam amostras com duração semelhantes (**15 segundos a 45 segundos**); O comprimento médio de um **segmento emocional estável** é, aproximadamente, de **20 segundos.**
Russo *et al*. (2020)	MER	Para diminuir a variação emocional de uma canção, autores costumam utilizar **20 a 30 segundos**, enquanto outros optam por extrair segmentos relacionados ao refrão ou optam por segmentos mais representativos da música manualmente.

FONTE: O autor (2022).

Conforme os estudos analisados, do ponto de vista cognitivo, os estudos apontam que uma resposta emocional pode ser formulada em questão de segundos ou até mesmo em durações menores (Bigand; Filipic; Latitte, 2005; Bachorik *et al*., 2009). Por seu turno, os estudos de cunho computacional direcionam para o uso de segmentos de curta duração, como 30 segundos, em vez da utilização de músicas com a duração total, em virtude de elementos como a redução da carga de trabalho dos voluntários, e a manipulação de um segmento emocional estável, por exemplo.

No que diz respeito à obtenção dos segmentos 30 segundos de cada música, o primeiro passo consistiu em identificar, manualmente, o início, em segundos, do segmento em que a letra da canção começava para cada música. Na sequência, desenvolveu-se um *script* com a biblioteca *pydub*[10] para acrescentar 30 segundos aos segmentos identificados, a partir do segundo inicial identificado para cada canção. Por fim, com a mesma biblioteca, realizou-se o “corte”, automaticamente, considerando os valores obtidos.

Com essa estratégia, garantiu-se a homogeneidade da duração das canções, uma vez que todos os registros contam com 30 segundos. Outro ponto a ser destacado é o fato de que não houve perda na quantidade de músicas, pois, por exemplo, caso a opção fosse pela extração dos segundos relacionados ao refrão, músicas sem este elemento seriam eliminadas.

[10] Disponível em: https://pytube.io/en/latest/. Acesso em: 09 nov. 2021.

Ademais, não seria possível garantir que a duração fosse homogênea, tendo em vista que cada canção pode apresentar duração específica para a parte do refrão, ou até mesmo mais de um refrão. De todo modo, é prudente reconhecer que a estratégia adotada apresenta limitações, como possíveis “quebras abruptas” no raciocínio do compositor e no sentido das letras, o que pode influenciar na percepção emocional das canções. Para mais, possíveis pistas emocionais presentes nas introduções das músicas foram ignoradas em prol da uniformidade das canções.

3.5.2 Identificação e coleta das músicas para o experimento piloto

Para identificar quais músicas empregar neste experimento, recorreu-se à plataforma *AllMusic*, em virtude da sua aplicação em estudos correlatos (Yang; Lee, 2009; Panda *et al*., 2013; Chen *et al*., 2015; Panda, 2019; Panda; Malheiro; Paiva, 2020a). Além do mais, trata-se de um ambiente com um conjunto controlado e pré-determinado de adjetivos emocionais, atribuídos às músicas por editores profissionais[11] (Hu; Downie, 2007) e especialistas (Panda, 2019), reduzindo possíveis ruídos como os relacionados, por exemplo, a ambientes como *Last.FM*, cuja rotulação é efetuada, livremente, pelos usuários da plataforma. Entretanto, ressalta-se que, como ponto negativo, o processo adotado na *AllMusic* não permite uma análise crítica das suas etapas (Panda *et al*., 2013; Panda, 2019).

Em relação aos termos empregados para identificar as músicas pertencentes às seis emoções do modelo discreto deste experimento, os adjetivos considerados são listados no QUADRO 28.

[11] Em solicitação direta ao fórum de esclarecimentos disponibilizado pela plataforma, por mais detalhes sobre o processo de rotular as músicas quanto aos adjetivos emocionais, obteve-se como resposta que todo o processo segue critérios estabelecidos por cada editor, conforme observa-se no ANEXO B.

QUADRO 28 - LISTA DE TERMOS EMPREGADOS NA PLATAFORMA *ALLMUSIC*

Emoção	Termos empregados	Referência
Amor/ternura	*Tender*	Juslin e Laukka (2003)
	Romantic, Precious, Reverent, Sentimental, Warm	Yang e Lee (2009)
Calma	*Laidback/Mellow, Gentle, Refined, Reserved, Restrained*	Yang e Lee (2009)
	Calm/Peaceful, *Relaxed*	*AllMusic*
Felicidade	*Happy*	Juslin e Laukka (2003)
	Cheerful	*AllMusic*
	Bright, Celebratory, Effervescent, Gleeful, Humorous, Irreverent, Rambunctious, Raucous, Rollicking, Silly, Sweet, Whimsical, Witty	Yang e Lee (2009)
	Joyous	*AllMusic*
Medo	*Manic, Paranoid, Spooky, Unsettling*	Yang e Lee (2009)
	Angst-Ridden, Scary, Tense/Anxious	*AllMusic*
Raiva	*Angry*	Juslin e Laukka (2003)
	Acerbic, Aggressive, Bitter, Fiery, Outraged, Rebellious, Snide	Yang e Lee (2009)
Tristeza	*Sad*	Juslin e Laukka (2003)
	Anguished/Distraught, Bleak, Somber, Poignant, Melancholy, Plaintive	Yang e Lee (2009)

FONTE: O autor (2021).

Para a formulação dos termos presentes no QUADRO 28, recorreu-se, além do agrupamento proposto por Juslin e Laukka (2003), à pesquisa conduzida por Yang e Lee (2009), que consistiu em refinar 168 adjetivos emocionais apresentados pela *AllMusic* em 23 adjetivos. Finalmente, como forma complementar, adotaram-se alguns dos sinônimos sugeridos pela própria *AllMusic*.

Após a definição dos termos de busca, desenvolveu-se um conjunto de *scripts* em *Python*, empregando as bibliotecas *BeautifulSoup*[12] e *Selenium*[13], para coletar os *links* das músicas diretamente das páginas da *AllMusic* de cada um dos termos listados no QUADRO 28. A partir dos *links*, os seguintes dados foram obtidos: a) título das canções; b) nome do artista; e c) todas as emoções associadas às músicas, uma vez que essas podem apresentar mais de uma emoção.

Como forma de aumentar o conjunto de músicas, tendo em vista que o retorno proporcionado pela primeira estratégia não foi satisfatório em termos de registros (1.270 músicas), elaborou-se uma segunda estratégia, considerando as músicas dos 500 melhores álbuns da história, elaborada pela revista *Rolling Stone*[14], edição 2020.

Atualizando os *scripts* da primeira estratégia, obteve-se os dados relacionados aos títulos das canções, aos nomes dos artistas e a todas as emoções associadas às canções, totalizando 4.487 registros. Cabe salientar que a escolha pela lista dos 500 melhores álbuns justifica-se por esta lista ser atual; ser composta quase exclusivamente por discos elaborados em inglês, e cobrir uma variedade de gêneros e artistas populares dos séculos XX e XXI.

Unindo os resultados das estratégias um e dois, o total de registros correspondeu a 5.757 músicas, que foram submetidas a três processos de triagem. No primeiro processo, adotaram-se os seguintes critérios de exclusão: a) músicas repetidas; b) músicas sem adjetivos emocionais; c) músicas com adjetivos emocionais que não constam no QUADRO 29. Como resultado, restaram 4.601 registros para a segunda fase de triagem.

Para o segundo processo de triagem, como cada música pode apresentar mais de uma emoção, os seguintes procedimentos foram realizados para selecionar a emoção “final” da música: a) conversão dos termos para a emoção principal (ex: *happy*, *joyous* → *happiness*, *happiness*); c) identificação da emoção “dominante” para representar a música (ex: música um: *happiness*, *happiness*, *sadness*, *fear*, *happiness* → ***happiness***).

A partir desses procedimentos, as músicas em que não se identificaram uma emoção “dominante” foram desconsideradas, reduzindo o total de registros para 4.183

[12] Disponível em: https://beautiful-soup-4.readthedocs.io/en/latest/. Acesso em: 08 nov. 2021.
[13] Disponível em: https://selenium-python.readthedocs.io. Acesso em: 08 nov. 2021.
[14] https://rollingstone.uol.com.br/noticia/os-500-melhores-discos-de-todos-os-tempos-da-rolling-stone/. Acesso em: 08 nov. 2021.

músicas. Por fim, no último processo de triagem, foram eliminadas: a) músicas ao vivo; b) músicas instrumentais; c) músicas cujas letras não foram gravadas em inglês; e d) músicas com menos de 45 segundos de duração.

Após os processos de triagem, restaram 2.620 músicas para compor a lista inicial das músicas deste estudo, distribuídas conforme apresenta-se na TABELA 2.

TABELA 2 - DISTRIBUIÇÃO DAS MÚSICAS POR EMOÇÃO

Emoção	Total de registros	Total de registros (%)
Amor/ternura	293	11,18%
Calma	347	13,24%
Felicidade	660	25,19%
Medo	210	8,02%
Raiva	450	17,18%
Tristeza	660	25,19%
Total	**2.620**	**100,00%**

FONTE: O autor (2021).

Neste processo, salienta-se que, para as músicas das emoções raiva (681), felicidade (1.756) e tristeza (906), devido à quantidade de registros em relação às demais emoções, realizou-se uma seleção aleatória dos registros para chegar aos totais apresentados na TABELA 2. Dessa forma, é possível que músicas descartadas possam atender aos critérios de inclusão definidos (músicas de estúdio, com letras em inglês e com duração superior a 45 segundos). Para as demais emoções, todos os registros foram conferidos.

A partir da seleção de quais músicas considerar para a construção da base de dados, a próxima etapa consistiu em criar *playlists* no *Youtube* para coletar os áudios dessas canções. Seis *playlists* foram elaboradas, considerando apenas "vídeos de áudio", priorizando o conteúdo disponibilizado pelos canais oficiais de cada artista. Após a criação das *playlists*, os áudios foram coletados por meio de *script* em *Python*, desenvolvido com a biblioteca *pytube*[15].

[15] Disponível em: https://pytube.io/en/latest/. Acesso em: 09 nov. 2021.

3.5.3 Revisão das músicas selecionadas para o experimento piloto

Por conta da rotulação conduzida pela plataforma *AllMusic* ser fechada e dependente, exclusivamente, de critérios estabelecidos por seus editores, entendeu-se que não era possível precisar sobre qual categoria de emoções - percebida, sentida ou transmitida - as músicas selecionadas correspondiam. Para mais, compreendeu-se que os procedimentos adotados pelo pesquisador para identificar a emoção “final” das músicas não estavam livres de vieses, bem como levantou-se a hipótese de que os segmentos extraídos - trinta segundos - poderiam não corresponder às emoções originais obtidas da *AllMusic*.

Dessa forma, esses segmentos foram submetidos a um processo de revisão, conduzido por voluntários, para fins de: a) minimizar vieses provocados ao longo do processo de associação das emoções às músicas; b) verificar se havia concordância entre as emoções oriundas da *AllMusic* e a percepção dos voluntários; c) avaliar se existia concordância entre as emoções percebidas pelos avaliadores; e d) garantir, minimamente, que os segmentos extraídos representassem a emoção percebida.

Ressalta-se que se adotou a emoção percebida em vez das emoções sentidas ou transmitidas, por conta de que o uso de modelos de inteligência artificial para a predição de emoções sentidas pode ser considerado raro (Xu *et al*., 2020); além do mais, a emoção percebida é, de certa forma, menos impactada por fatores situacionais, como condições de ambiente e aspectos relacionados ao humor (Yang; Chen, 2011).

Desse modo, oito voluntários foram recrutados para ouvirem os segmentos e apontarem a sua percepção emocional. Dos oito voluntários, foram selecionados quatro músicos/musicistas, com estudo formal em Música (graduação (bacharelado/licenciatura) ou conservatório), sendo três voluntárias e um voluntário, com idade média de 35,25 anos e desvio padrão igual a 8,42. Por sua vez, foram selecionados, para o grupo de participantes não músicos, três voluntárias e um voluntário. A média de idade para este grupo foi de 25,25 anos, com desvio padrão igual a 2,87.

Antes do acesso aos segmentos, todos os participantes receberam um Termo de Consentimento Livre e Esclarecido (TCLE) (APÊNDICE B) e instruções sobre a sua participação no estudo (APÊNDICE C). O processo de revisão ocorreu durante o período de 13/12/2021 a 06/02/2022. Durante este período, cada voluntário,

semanalmente, recebeu um conjunto de segmentos e uma planilha para indicar a emoção percebida para cada segmento.

Em relação à divisão dos voluntários e à condução do processo, as seguintes questões foram tomadas:

a) quatro duplas, compostas por um músico/musicista e por um voluntário não especialista;
b) em virtude dos segmentos das emoções raiva e medo, e amor/ternura e calma, serem menores do que os 660 registros para as emoções felicidade e tristeza, para balancear o número de segmentos avaliados por cada dupla, essas emoções foram agrupadas em dois novos grupos: 1) medo-raiva (660 segmentos); e 2) amor/ternura e calma (640 segmentos);
c) cada um dos três conjuntos com 660 segmentos (felicidade, medo-raiva e tristeza) foi revisado por uma dupla; e o conjunto com 640 registros (amor/ternura e calma) foi avaliado pela dupla restante;
d) um mesmo grupo de segmentos foi revisado por dois voluntários;
e) os voluntários não tiveram acesso às emoções originais para cada registro, nem dados como título da canção e o nome dos artistas; e
f) não houve interações entre os voluntários.

Com relação aos critérios para seleção dos segmentos, dois foram testados: a) igualdade entre a percepção emocional dos avaliadores e a emoção identificada na plataforma (emoção plataforma = emoção voluntário 1 = emoção voluntário 2); e b) igualdade na percepção emocional dos avaliadores (emoção voluntário 1 = emoção voluntário 2), independente dos rótulos obtidos junto à plataforma.

Os resultados desse processo são descritos na seção 4.1.

3.5.4 Identificação e coleta das músicas para o experimento ajustado

Com o intuito de aumentar o número de registros da base de dados, em virtude da redução oriunda do experimento piloto, recorreu-se à coleta dos títulos de músicas presentes em 29 *playlists* elaboradas pelo perfil oficial do *Spotify* (APÊNDICE D), e que se relacionavam com as emoções elencadas no modelo proposto por Bigand

et al. (2005). Nesta etapa, as *playlists* foram selecionadas a partir da busca pelos termos em inglês para cada uma das emoções, resultando em 2.966 músicas.

A partir da relação das músicas obtidas, os mesmos procedimentos de seleção aplicados na plataforma *AllMusic* (músicas em inglês; não instrumentais etc.) foram adotados, e 2.177 músicas restaram. Salienta-se, entretanto, que as músicas dessas *playlists*, diferente do que ocorre na plataforma *AllMusic*, não apresentam adjetivos emocionais associados, dessa forma, a associação inicial das músicas às emoções respeitou a emoção da *playlist* na qual a música foi recuperada.

Quanto à coleta das informações das canções, como títulos e artistas, essa deu-se por meio da elaboração de um *script* em *Python* considerando a biblioteca *Spotipy*[16]. Por fim, o áudio dessas canções foi coletado diretamente do *Youtube*, reaproveitando os *scripts* desenvolvidos para o experimento piloto.

3.5.5 Avaliação das músicas selecionadas para o experimento ajustado

Tendo em vista que o experimento piloto apresentou, em relação à concordância entre os voluntários, resultados no máximo razoáveis - embora estatisticamente significativos -, e o fato de a base completa não ter sido revisada por todos os avaliadores, optou-se por reconsiderar as 2.620 da plataforma *AllMusic* anteriormente analisadas para um novo processo avaliativo, somadas às 2.177 oriundas das *playlists* desenvolvidas pelo perfil oficial do *Spotify*. Dessa forma, o total de segmentos disponibilizados para a avaliação correspondeu a 4.797 registros.

No que tange às emoções associadas às canções recuperadas, tanto em relação a *AllMusic*, quanto para o *Spotify*, é prudente salientar que essas foram desconsideradas, e a “emoção final” para cada segmento consistiu nas anotações fornecidas pelos avaliadores. Outrossim, a avaliação foi conduzida sem que os participantes tivessem conhecimento dos títulos e dos nomes dos artistas. Finalmente, assim como no experimento piloto, não houve interações entre os participantes.

Para a seleção dos participantes, um *folder* de divulgação (APÊNDICE E), com informações básicas sobre a pesquisa, foi encaminhado para diversos cursos de graduação e pós-graduação brasileiros, sem restrições de área do conhecimento e formação (músico/musicista ou não, por exemplo).

---

[16] Disponível em: https://spotipy.readthedocs.io/en/2.19.0/. Acesso em: 06 jun. 2022.

É importante pontuar que a escolha por não delimitar os voluntários quanto às suas respectivas formações acadêmicas encontra-se alinhada a recomendações presentes na literatura correlata, na qual há trabalhos que incentivam o uso de participantes especialistas e não especialistas (Gómez-Cañón *et al*., 2021), bem como não haja limitações para o recrutamento de indivíduos, buscando o desenvolvimento de um sistema aplicável a todas as pessoas (Yang *et al*., 2008).

Apesar da estratégia de divulgação acima relatada, a adesão à pesquisa foi baixa, com apenas dois voluntários permanecendo como avaliadores até o final do processo, anotando, assim, todos os registros da base de dados (4.797). Por conta disso, mais três voluntários foram recrutados para complementar a avaliação, com cada novo participante responsável por rotular 1.599 registros. Sendo assim, ao final, cada música desta pesquisa foi anotada por três voluntários.

Do ponto de vista descritivo, a média de idade dos cinco voluntários foi igual 45,20 anos, com desvio padrão de 19,60 anos. Ademais, em relação ao gênero, dois voluntários identificaram-se como pertencentes ao gênero masculino e três voluntárias descreveram o seu gênero como feminino.

O processo de avaliação adotado para o experimento ajustado foi conduzido de maneira remota, com os arquivos para a audição disponibilizados aos avaliadores para que realizassem as anotações conforme as suas preferências de horário e local, durante o período de 04/08/2022 a 07/10/2022. Por fim, salienta-se que os voluntários que participaram do experimento piloto não foram considerados neste novo processo.

Quanto às etapas conduzidas neste novo processo de avaliação, compuseram este processo: a) apresentação do TCLE (APÊNDICE F); b) instruções acerca do experimento (APÊNDICE G); c) coleta de dados demográficos (gênero, idade, presença de problemas auditivos e formação musical); e d) suporte aos avaliadores durante o período de avaliação.

Como critérios de avaliação para definir quais músicas reter para a construção da base de dados, dois cenários foram adotados: a) concordância entre os avaliadores; e b) maioria observada. Para avaliar a confiabilidade da concordância entre os avaliadores do experimento ajustado, recorreu-se à estatística kappa de Fleiss (Fleiss, 1971; Hallgren, 2012), cuja aplicação destina-se a mensurar o nível de concordância observada em experimentos com mais de dois avaliadores e variáveis nominais (categóricas) (Hallgren, 2012).

Para o uso da maioria como critério de seleção, uma vez que a quantidade de avaliadores por músicas consistiu em um número ímpar, procedeu-se que, sempre que observada a concordância entre, no mínimo, dois dos três avaliadores, a música em questão foi considerada para compor a base de dados.

Dessa forma, a partir das métricas acima relatadas, optou-se pela construção de duas bases de dados, aqui denominadas como base um, construída a partir do consenso obtido entre os três avaliadores, e composta por 1.412 registros; por sua vez, a base obtida mediante a maioria observada, formada por 4.303 músicas, foi nomeada como base dois. Por fim, os resultados obtidos juntos aos voluntários, bem como os detalhes acerca das duas bases de dados, são descritos na seção 4.2.

3.5.6 Extração das características

As características empregadas neste estudo representam cinco modalidades: a) acordes; b) áudio; c) letras; d) simbólicas; e e) voz. A opção por essas modalidades deu-se, principalmente, pelos resultados identificados na revisão integrativa conduzida (QUADRO 1), nos quais essas modalidades foram empregadas, com destaque para áudio e letras. Para esta pesquisa, ressalta-se que apenas características relacionadas ao conteúdo musical foram consideradas e, por isso, recursos baseados em metadados, por exemplo, foram descartados.

É prudente salientar que, para obter os acordes e as características simbólicas, empregaram-se algoritmos e bibliotecas (Mauch, 2010; Mauch; Dixon, 2010; Wu *et al*., 2021) para estimar esses recursos. Dessa forma, evitaram-se consultas a fontes externas para a aquisição de dados, como cifras e arquivos MIDI, permitindo, assim, que não houvesse eliminação de registros em virtude da falta desses elementos.

Neste aspecto, entendeu-se, conforme apontaram Panda, Malheiro e Paiva (2020a), que a estimativa de determinadas características, como linhas melódicas predominantes, embora seja um processo que demande aperfeiçoamentos, ainda assim pode fornecer informações relevantes e que não são consideradas nos experimentos de REM (Panda; Malheiro; Paiva, 2020a).

Quanto às letras, estas foram coletadas, via *script* (*Python*), diretamente do *site Genius*, tendo em vista que as tentativas de transcrição automática ficaram aquém do esperado, apresentando inconsistências que inviabilizaram a automação deste

processo. Relativo à voz, a aquisição desta modalidade ocorreu por meio da biblioteca *spleeter* (Hennequin *et al*., 2020), que permitiu separar a parte instrumental da parte vocal das músicas.

Por fim, todos os procedimentos de extração consideraram apenas os 30 segundos extraídos e, com relação à padronização dos áudios, os seguintes parâmetros foram adotados: a) extensão (wav); b) taxa de amostragem (22.500 Hz); c) tamanho da amostra (16 *bits*); e d) taxa de *bits* (1411 *kilobits per second* (kbps)) (Yang *et al*., 2018; Griffiths *et al*., 2021).

Os detalhes acerca dos processos conduzidos para a extração das características para cada modalidade são apresentados nas próximas subseções.

3.5.6.1 Acordes

Para estimar os acordes, empregou-se a biblioteca *chord-extractor*[17], que consiste em uma implementação, em *Python*, do *plugin Chordino*[18], sendo este *plugin* frequentemente utilizado em pesquisas destinadas a estimar acordes (Aljanaki; Wiering; Veltkamp, 2014; Aljanaki, 2016; Deng; Kwok, 2018; Giesel; Vatolkin; Rudolph, 2020; Miller; Nicosia; Sandler, 2021).

As bases teóricas que sustentam o *Chordino* e, consequentemente, o processo para estimar os acordes conduzido neste estudo, derivam da tese elaborada por Mauch (2010), na qual o autor propôs uma estratégia baseada na aplicação do algoritmo *Non-Negative Least Squares* (NNLS) para melhorar o reconhecimento de acordes, a qual recebeu o nome de *NNLS Chroma* (Mauch, 2010; Mauch; Dixon, 2010).

Os acordes extraídos neste estudo foram obtidos considerando apenas a parte instrumental das músicas. Como parâmetros, adotaram-se os valores *default* da biblioteca utilizada, por serem recomendações presentes na própria documentação do *plugin Chordino*.

Quanto às características, inicialmente, duas técnicas de mineração de textos, denominadas *Bag-of-Words* (BOW) e *Term Frequency-Inverse Document Frequency* (TF-IDF) (Hasan; Matin, 2021; Rustan *et al*., 2021; Ahanin; Ismail, 2022) foram consideradas. Todavia, a partir da condução de experimentos de classificação,

[17] Disponível em: https://ohollo.github.io/chord-extractor/. Acesso em: 13 dez. 2021.
[18] Disponível em: http://www.isophonics.net/nnls-chroma. Acesso em: 13 dez. 2021.

constatou-se a necessidade de substituí-las por características “puramente numéricas”, tendo em vista que, devido à realização de procedimentos como balanceamento de classes e *data augmentation*, tendem a serem mais “simples” no que concerne à replicação de maneira sintética.

Dessa forma, como características para representar os acordes, recorreu-se à contagem e à proporção dos tipos de acordes, conforme apresenta-se no QUADRO 29.

QUADRO 29 - CARACTERÍSTICAS EXTRAÍDAS PARA OS ACORDES

Característica	Descrição
quantidadeAcordes	Quantidade de acordes
outros	Quantidade de acordes classificados como “outros”
setima	Quantidade de acordes de sétima
setimaMaior	Quantidade de acordes de sétima maior
setimaMenor	Quantidade de acordes de sétima menor
setimaMenorQuinta	Quantidade de acordes com sétima menor e quinta diminuta
sexta	Quantidade de acordes com sexta
sextaMenor	Quantidade de acordes com sexta menor
triadeAumentada	Quantidade de acordes com tríade aumentada
triadeDiminuta	Quantidade de acordes com tríade diminuta
triadeMaior	Quantidade de acordes com tríade maior
triadeMenor	Quantidade de acordes com tríade menor
acordesUnicos	Quantidade de acordes únicos
proporcaoUnicos	Proporção de acordes únicos
proporcaoTriadeMenor	Proporção de acordes com tríade menor
proporcaoTriadeMaior	Proporção de acordes com tríade maior
proporcaoTriadeDiminuta	Proporção de acordes com tríade diminuta
proporcaoTriadeAumentada	Proporção de acordes com tríade aumentada
proporcaoSextaMenor	Proporção de acordes com sexta menor
proporcaoSexta	Proporção de acordes com sexta
proporcaoSetima	Proporção de acordes de sétima
proporcaoSetimaMaior	Proporção de acordes de sétima maior
proporcaoSetimaMenor	Proporção de acordes de sétima menor
proporcaoSetimaMenorQuinta	Proporção de acordes de sétima menor e quinta diminuta
proporcaoOutros	Proporção de acordes classificados como “outros”

FONTE: O autor (2023).

Para a extração das características relatadas na QUADRO 29, cada acorde estimado foi substituído pela sua respectiva categoria, como, por exemplo: Fdim (tríade diminuta), C7 (sétima), C6 (sexta) etc.

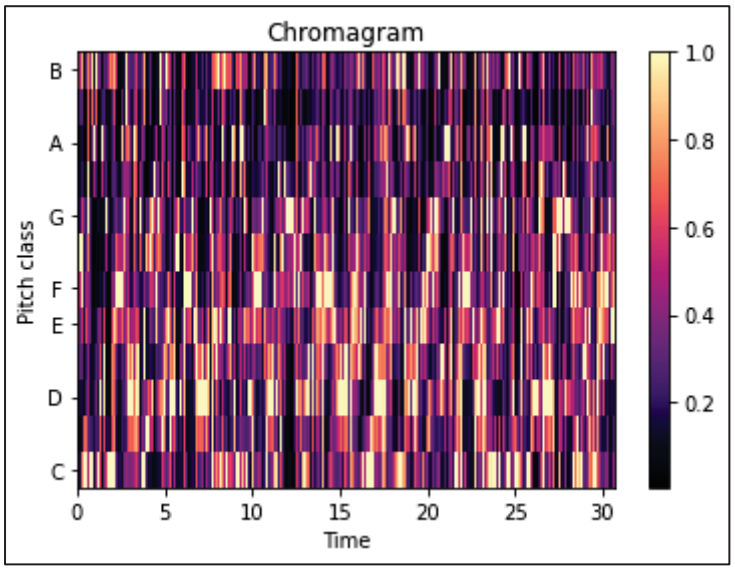
3.5.6.2 Áudio

As características extraídas para o áudio correspondem às propriedades básicas da música, sendo estas: a) dinâmica; b) harmonia; c) melodia; d) ritmo; e) timbre. Como base para definir quais características utilizar, os trabalhos de McKay (2010); Bogdanov *et al*. (2013); McFee *et al*. (2015); e Panda, Malheiro e Paiva (2020a, 2020b) foram consultados de maneira complementar aos estudos multimodais elencados no QUADRO 1.

A primeira propriedade abordada foi a dinâmica, para qual duas características foram extraídas: a) *Root Mean Square* (RMS): indica a energia média de um sinal (Mckay, 2010); e b) *Loudness*: fornece uma impressão subjetiva da intensidade de um determinado som (Panda; Malheiro; Paiva, 2020a).

Para representar a harmonia, a primeira característica consistiu na obtenção dos valores relacionados ao cromagrama (*chromagram* - em inglês) (FIGURA 32), que consiste em um vetor com 12 dimensões, em que cada dimensão representa a concentração de energia (intensidade) aproximada para 12 semitons (ex: C, C#, D, D#, E, F, F#, G, G#, A, A#, B) (Panda; Malheiro, Paiva, 2020b).

FIGURA 32 - CROMAGRAMA (EXEMPLO)



FONTE: O autor (2021).

Para a melodia, assim como para a dinâmica, duas características foram extraídas, sendo a primeira denominada *Pitch Salience*, cuja função recai em demonstrar o quão perceptível é a presença - ou sensação - do tom em um som. Neste

aspecto, tons puros apresentam valores médios próximos a 0, enquanto os sons compostos por vários harmônicos no espectro tendem a demonstrarem valores maiores (Panda; Malheiro; Paiva, 2020b).

Como segunda característica melódica, buscou-se extrair a frequência fundamental da melodia predominante das canções (*Predominant Pitch Melodia*), baseando-se em uma implementação do algoritmo *MELODIA*, do qual os conceitos remetem à pesquisa de Salamon e Gómez (2012). De forma resumida, a proposta de Salamon e Gómez (2012) faz uso da criação e da caracterização de contornos de tom, além de sequências contínuas de tempo considerando o agrupamento de tons por meio de pistas auditivas, estabelecendo, assim, a presença ou não de uma melodia predominante.

A quarta propriedade contemplada com características foi o ritmo, para o qual as características extraídas foram: a) *Beats per Minute* (BPM): representa o número de batidas por minuto de um sinal de áudio; b) *Danceability*: estima a "danceabilidade" de um determinado sinal de áudio, no qual, quanto maior o valor para esta característica, mais "dançante" a música é. Baseia-se no algoritmo *Detrended Fluctuation Analysis* (DFA) proposto por Streich e Herrera (2005); e c) *Predominant Local Pulse* (PLP): captura o andamento dominante, mesmo em músicas com elementos não percussivos (Panda; Malheiro; Paiva, 2020b). Deriva da pesquisa desenvolvida por Grosche e Müller (2011).

Por último, para representar o timbre, as seguintes características foram extraídas:

a. ***Mel-Frequency Cepstral Coefficients* (MFCCs):** coeficientes baseados na escala Mel[19]. Consiste em uma série de bancos de filtros triangulares de baixa a alta frequência para filtrar o sinal de ruído de entrada (Jin *et al*., 2021).

b. ***Spectral Centroid*:** indica, de certa forma, o "brilho" do som (ex: "escuro" ou "brilhante") (Mckay, 2010);

c. ***Spectral Flux*:** reflete a quantidade de mudança espectral em um sinal quadro a quadro (Mckay, 2010);

d. ***Spectral Rolloff*:** frequentemente utilizado para indicar a assimetria das frequências presentes em uma janela (Panda; Malheiro; Paiva, 2020b);

[19] Descreve as características não lineares da percepção da frequência do ouvido humano (Jin *et al*., 2021).

e. ***Zero-crossing rate*** **(ZCR):** quantidade de vezes que um sinal alterna entre valores negativos e positivos. Fornece uma estimativa acerca do ruído do sinal (Mckay, 2010).

A síntese das características relacionadas ao áudio consideradas para esta tese, bem como a relação com as estatísticas consideradas para cada característica, além dos principais parâmetros e as ferramentas utilizadas são apresentadas no QUADRO 30.

QUADRO 30 - SÍNTESE DAS CARACTERÍSTICAS RELACIONADAS AO ÁUDIO CONSIDERADAS

Propriedade	**Característica**	**Atributos**	**Total de Valores**	**Ferramenta**
Dinâmica	RMS	Média e desvio padrão	2	*librosa*
	Loudness	Total	1	*Essentia*
Harmonia	*Chromagram*	Média e desvio padrão das 12 dimensões	24	*librosa*
Melodia	*Pitch Salience*	Total	1	*Essentia*
	Predominant Pitch Melodia	Média e desvio padrão	2	
Ritmo	*BPM*	Total	1	*Essentia*
	Danceability	Total	1	
	PLP	Média e desvio padrão	2	*librosa*
Timbre	MFCCs	Média e desvio padrão dos 13 primeiros coeficientes	26	*librosa*
	Spectral Centroid	Média e desvio padrão	2	
	Spectral Rolloff	Média e desvio padrão	2	
	ZCR	Média e desvio padrão	2	
	Spectral Flux	Total	1	*Essentia*

FONTE: O autor (2021).

Todo o processo de extração de características do áudio foi conduzido com as bibliotecas *librosa* (Mcfee *et al*., 2015) e *Essentia* (Bogdanov *et al*., 2013),

considerando as suas implementações em *Python*, com os todos os parâmetros *default,* excetuando-se o número de coeficientes para os MFCCs (13). Por fim, salienta-se que todo o processo de extração considerou apenas a parte instrumental das canções, tendo em vista que características específicas para a voz foram consideradas separadamente.

3.5.6.3 Letras

Ao considerar as letras como uma modalidade, o objetivo estabelecido era transcrever, automaticamente, o conteúdo a partir dos arquivos de áudio. Todavia, ao utilizar ferramentas específicas de transcrição, desenvolvidas, por exemplo, por empresas como *Microsoft* e *Google*, os resultados obtidos foram insatisfatórios (QUADRO 31), mesmo ao se tentar a transcrição utilizando apenas a voz.

QUADRO 31 - EXEMPLOS DE TRANSCRIÇÕES OBTIDAS AUTOMATICAMENTE

Original	***Speech to text* (Azure Microsoft)**	***Cloud Speech-to-Text* (Google)**
For fifty years they've been married *And they can't wait for their fifty-first to roll around* *Yeah, roll around* *Thirty years they've been married* *And now they're old and happy and they settle down* *Settle down, yeah* *Twenty years they've been*	*They've been very.* *Can't wait for the 1st.* *You roll around.* *Yeah. authorities say they've been married and now they're old and habits and they settle down.* *Settle down.* *20 years*	*Oh, please, please.* *They've been married thinking.* *Wait was in the first roll around. Yeah.* *I'm seriously pissed and now we're happy.* *What is it?*
I'm going out sleepwalking *Where mute memories start talking* *The boss that couldn't help but hurt you* *And the pretty thing you made desert you* *I'm going out like a baby* *A naive, unsatisfiable*	*Going out sleepwalking.* *Where Mutna Marie start talking.* *Was getting help with her cheese.* *And pretty thing he may desert you.* *I'm going out like a baby. And I even said*	*Going out. Sleepwalking* *Where music memory start talking* *back and helped her Chi.* *And pretty thing and made dessert, you.* *I'm going out like a baby.* *The naive onset of.*

FONTE: O autor (2021).

Para além das ferramentas listadas no QUADRO 31, recorreu-se à biblioteca *SpeechRecognition*[20] (*Python*), com a qual os resultados também não atingiram desempenho satisfatório. Outrossim, determinados módulos dessa biblioteca demandam a criação de credenciais e não são totalmente gratuitos.

Sendo assim, para não descartar as letras e reduzir o número de modalidades, a alternativa adotada consistiu no desenvolvimento de um *script* para coletar as letras diretamente do *site Genius*[21]. Por sua vez, o recorte para adequar o conteúdo das letras para os 30 segundos extraídos foi realizado manualmente.

Como características, as já citadas BOW e TF-IDF (Hasan; Matin, 2021; Rustan *et al*., 2021; Ahanin; Ismail, 2022) foram consideradas inicialmente para obter as frequências e os pesos das palavras em cada letra. Entretanto, da mesma maneira que ocorreu com os acordes, em virtude dos procedimentos para a geração de dados sintéticos, optou-se por duas categorias de características numéricas: a) frequências; e b) análise de sentimentos.

Para a primeira categoria, frequências simples foram extraídas, correspondendo ao número de palavras, caracteres e palavras únicas das letras, desconsiderando caracteres especiais (ex: @, #) e números. Ademais, a média e a proporção de palavras únicas (riqueza do vocabulário) foram adicionadas.

Com relação às métricas de análise de sentimentos, empregou-se o modelo léxico denominado *Valence Aware Dictionary for sEntiment Reasoning* (VADER) (Hutto; Gilbert, 2014), que proporcionou a extração de quatro métricas: a) positivo: o quão positivo o conteúdo é; b) negativo: o quão negativo o conteúdo é; c) neutro: o quão neutro o conteúdo é; e d) *compound*: métrica única que representa uma síntese dos valores negativos e positivos do conteúdo, onde valores próximos a -1 indicam conteúdo extremamente negativo, e valores próximos a 1, conteúdo extremamente positivo. Por fim, valores maiores que -0,05 e menores do que 0,05 são tratados como neutros.

A síntese das características associadas às letras é exemplificada no QUADRO 32.

[20] Disponível em: https://pypi.org/project/SpeechRecognition/. Acesso em 27 dez. 2021.
[21] Disponível em: https://genius.com. Acesso em: 28 dez. 2021.

QUADRO 32 - CARACTERÍSTICAS EXTRAÍDAS (LETRAS)

Característica	Definição
charCount	Quantidade de caracteres
wordCount	Quantidade de palavras
avgWord	Médias de palavras
uniqueWords	Quantidade de palavras únicas
uniqueWordsProp	Proporção de palavras únicas
neg	Valor negativo da letra
neu	Valor neutro da letra
pos	Valor positivo da letra
compound	Métrica síntese dos valores positivos, negativos e neutros

FONTE: O autor (2023).

Com relação ao pré-processamento das letras, o procedimento adotado consistiu em converter todas as palavras para caracteres minúsculos, tendo em vista a padronização do vocabulário. Desse modo, variações como *House*/*house*, e *Love*/*love*, por exemplo, foram analisadas somente como *house* e *love*, respectivamente. Somado a esse procedimento, caracteres especiais (ex: #, @, $, *), marcadores como *[chorus]* e *[bridge]*, e números foram descartados. Os procedimentos referentes às letras foram conduzidos com a biblioteca *Natural Language Toolkit*[22] (NLTK).

Nessa etapa, destaca-se que, por conta de os segmentos analisados possuírem poucas sentenças, optou-se por não remover as chamadas *stop words* - termos com alta frequência de ocorrência em um idioma (ex: *a, as, of, me, the, with*) e que pouco agregam em termos de conhecimento (Malheiro *et al*., 2018).

Finalizada a apresentação do processo de extração relacionado às letras, a próxima subseção contempla a extração das características simbólicas.

3.5.6.4 Características simbólicas

Antes de extrair as características simbólicas, foi necessário transcrever os arquivos de áudio para o formato MIDI. Todavia, salienta-se que a transcrição automática de músicas ainda se encontra em fase experimental (Benetos *et al*., 2019; Wu; Chen; Su, 2020), tendo em vista, por exemplo, a presença de fontes simultâneas no sinal (instrumentos e vocais), com diferentes alturas, volumes e timbres (Benetos *et al*., 2019; Wu *et al*., 2021).

Entretanto, apesar do aspecto experimental, decidiu-se pela transcrição do áudio para o formato MIDI de forma automática, em virtude da proposta desta tese em

[22] Disponível em: https://www.nltk.org. Acesso em: 30 abr. 2023.

minimizar consultas a fontes externas em busca de modalidades e, eventualmente, o descarte de registros por conta da falta de arquivos para todas as músicas consideradas na base de dados desenvolvida.

Dessa forma, recorreu-se à biblioteca *Omnizart*[23] (Wu *et al*., 2021) a qual permitiu a transcrição de todos os áudios da base de dados. A escolha por essa biblioteca deu-se por conta do seu processo ser construído a partir de módulos que utilizam *deep learning*, desenvolvidos para suportar abordagens com múltiplos instrumentos; para mais, trata-se de uma aplicação de fácil manipulação, tendo em vista que sua execução é simplificada por instruções via linha de comando (Wu *et al*., 2021).

Finalmente, o processo de transcrição adotado, com a biblioteca *Omnizart,* resultou em arquivos MIDI compostos por transcrições em piano das músicas, contemplando apenas a parte instrumental das canções. Os fundamentos deste processo derivam das abordagens desenvolvidas por Wu, Chen e Su (2019; 2020), nas quais os autores desenvolveram os seus modelos de transcrição baseados na arquitetura de *deep learning U-Net* (Ronneberger; Fischer; Brox, 2015).

Após o processo de transcrição, procedeu-se a extração das características simbólicas por meio da biblioteca *music21*[24] (Cuthbert; Ariza, 2010), considerando o módulo *music21.features.jSymbolic*, que implementa, em *Python*, parte das características disponíveis na ferramenta *jSymbolic*[25] (Mckay, 2010). A relação das características simbólicas utilizadas é listada no QUADRO 33.

[23] Disponível em: https://music-and-culture-technology-lab.github.io/omnizart-doc/index.html. Acesso em: 03 jan. 2022.
[24] Disponível em: https://web.mit.edu/music21/doc/index.html. Acesso em 03 jan. 2022.
[25] Disponível em: http://jmir.sourceforge.net/jSymbolic.html. Acesso em: 03 jan. 2022.

QUADRO 33 - CARACTERÍSTICAS SIMBÓLICAS CONSIDERADAS NESTA TESE

Característica	**Descrição**
Altura média (*Primary register*)	Altura média identificada para o arquivo MIDI
Média da duração das notas (*Average note duration*)	Média da duração das notas em segundos
Desvio padrão da duração das notas *(Variability of note duration*)	Desvio padrão da duração das notas em segundos
Duração máxima das notas (*Maximum note duration*)	Valor máximo da duração das notas em segundos
Duração mínima das notas (*Minimum note duration*)	Valor mínimo da duração das notas em segundos
Alcance *(Range*)	Diferença em semitons entre os tons mais altos e mais baixos
Densidade (*Note density*)	Número médio de notas por segundo
Tons baixos (*Importance of bass register*)	Fração de notas entre tons de MIDI 0 e 54
Tons médios (*Importance of middle register*)	Fração de notas entre os tons de MIDI 55 e 72
Tons altos (*Importance of high register*)	Fração de notas entre os tons de MIDI 73 e 127
Altura prevalente (*Most common pitch class prevalence*)	Fração de notas correspondentes à classe de afinação mais comum
Variedade de tons (*Pitch class variety*)	Número de tons usados pelo menos uma vez

FONTE: O autor (2022) com base em McKay (2010).

Diante do conteúdo apresentado no QUADRO 33, ressalta-se que cada característica é representada por um único valor, dessa forma, 12 características simbólicas foram consideradas nesta tese. Quanto à definição de quais características extrair, analisou-se a documentação da biblioteca *music21*, bem como a tese de Mckay (2010), em busca de características adequadas para a transcrição conduzida (piano), na qual, por exemplo, não permitiu o uso de características destinadas a outros instrumentos (ex: violão, violino, saxofone etc.).

3.5.6.5 Voz

A última modalidade considerada consistiu na voz, para a qual utilizou-se a biblioteca *spleeter*[26] (Hennequin *et al*., 2020) para separar a parte instrumental do conteúdo vocal das músicas. A biblioteca em questão é uma aplicação desenvolvida

[26] Disponível em: https://github.com/deezer/spleeter. Acesso em 08 jan. 2022.

por uma equipe do *streaming Deezer* e baseia-se no modelo proposto por Jansson *et al*. (2017), cuja principal característica é o uso da arquitetura de *deep learning U-net* (Ronneberger; Fischer; Brox, 2015) como mecanismo para a separação dos recursos.

Após obter as vozes para todas as músicas da base de dados, foi possível extrair três categorias de características concernentes à voz, sendo essas: a) características espectrais; b) características prosódicas - também nomeadas como contínuas; e c) características relacionadas à qualidade da voz (El Ayadi; Kamel; Karray, 2011; Akçay; Oğuz, 2020).

As características espectrais representam os recursos obtidos ao se transpor o sinal de áudio do domínio do tempo para o domínio da frequência, normalmente empregando a transformada de Fourier (Akçay; Oğuz, 2020). Para essa categoria, consideraram-se três características: a) formantes; b) *Gammatone Frequency Cepstral Coefficients* (GFCCs); e c) MFCCs.

Os formantes correspondem às várias regiões de frequência que apresentam intensidade relativamente grande em um espectro de som, e cuja causa deve-se à ressonância acústica do trato vocal humano (Liu *et al*., 2021). Além do mais, representam uma das formas para se caracterizar um fonema (Akçay; Oğuz, 2020).

Quanto aos GFCCs, estes coeficientes derivam da aplicação de bancos de filtros *Gammatone*, cuja principal função é "imitar" o comportamento da cóclea[27] (Patni *et al*., 2021). No que tange ao número de coeficientes, a extração de 12 coeficientes pode ser suficiente para descrever uma classe de emoção (Patni *et al*., 2021).

Em relação aos MFCCs, estes consistem nos mesmos coeficientes extraídos para representar o timbre nas características relacionadas ao áudio. Todavia, para a voz, optou-se por analisar os 16 primeiros coeficientes, tendo em vista que os seguintes são redundantes e diminuem a precisão de um sistema (Patni *et al*., 2021).

As características prosódicas são aquelas que podem ser percebidas pelo homem, como entonação e ritmo (Akçay; Oğuz, 2020). Para esta tese, as seguintes características prosódicas foram consideradas: a) duração; e b) frequência fundamental (F_0).

A duração refere-se à quantidade de tempo para construir vogais, palavras e construções semelhantes que estão presentes na fala, englobando, por exemplo, a taxa de duração das regiões com vozes ("vozeadas") e em silêncio ("não vocalizadas")

[27] "unidade auditiva onde as vias auditivas elétricas, que consistem em células ciliadas vibratórias internas e externas, se iniciam" (Tuhanioğlu *et al*., 2019, p. 283).

(Akçay; Oğuz, 2020). A respeito da frequência fundamental (F_0), essa é criada pelas vibrações nas cordas vocais, e produz características rítmicas e tonais da fala (Akçay; Oğuz, 2020).

Finalmente, as características relacionadas à qualidade da voz, nas quais a qualidade é determinada pelo resultado de propriedades físicas do trato vocal (Akçay; Oğuz, 2020). Nesta categoria, extraíram-se as seguintes características: a) *Harmonics to Noise Ratio* (HNR): mede a relação entre os componentes harmônicos e de ruído (Fernandes *et al*., 2017); b) *jitter*: medida da variação glótica entre os ciclos de vibração das cordas vocais (Fernandes *et al*., 2017), sendo um indicador da instabilidade da frequência fundamental (Akçay; Oğuz, 2020); e c) *shimmer*: relaciona-se com a variação de magnitude ao longo dos períodos glóticos (Fernandes *et al*., 2017), e indica a instabilidade da amplitude (Akçay; Oğuz, 2020).

A síntese das características vocais extraídas para esta tese é apresentada no QUADRO 34.

QUADRO 34 - CARACTERÍSTICAS RELACIONADAS À VOZ CONSIDERADAS NESTA TESE

Categoria	Características	Atributos	Total de Valores	Ferramenta
Espectrais	Formantes	Média e desvio padrão da frequência e da largura de banda (três primeiros formantes)	12	*openSMILE*[28]
	GFCC	Média e desvio padrão (12 coeficientes)	2	*Essentia*
	MFCC	Média e desvio padrão (16 primeiros coeficientes)	32	*librosa*
Prosódicas	Duração	Média e desvio padrão das regiões vozeadas e não vocalizadas	4	*openSMILE*
	F_0	Média e desvio padrão	2	
Qualidade	HNR	Média e desvio padrão	2	*openSMILE*
	Jitter	Média e desvio padrão	2	
	Shimmer	Média e desvio padrão	2	

FONTE: O autor (2022).

Quanto às bibliotecas, considerou-se as implementações em *Python*, com todos os parâmetros para a extração das características como *default*, exceto: a) número de MFCCs (16); e b) número de GFCCs (12).

Com a extração das características relacionadas à voz, o total de características para esta tese correspondeu a 169, conforme a distribuição relatada na TABELA 3.

TABELA 3 - TOTAL DE CARACTERÍSTICAS EXTRAÍDAS

Modalidade	Número de características
Acordes	25
Áudio	65
Letras	9
Simbólicas	12
Voz	58
Total	**169**

FONTE: O autor (2023).

Finalizada a etapa de extração de características, na sequência, as estratégias de classificação adotadas nesta pesquisa são apresentadas.

[28] Disponível em: https://audeering.github.io/opensmile-python/. Acesso em: 07 jan. 2022.

3.6 ESTRATÉGIAS DE CLASSIFICAÇÃO

Nesta subseção, detalhes acerca das arquiteturas de *deep learning* e dos algoritmos de aprendizado de máquina empregados, como os hiperparâmetros testados são fornecidos. Por fim, o protocolo de experimentos realizados é apresentado.

3.6.1 Arquiteturas de *deep learning* empregadas

Neste estudo, as arquiteturas de *deep learning* consistiram nas descritas na subseção 2.6, uma vez que essas foram projetadas para problemas compostos por dados tabulares. Para mais, com o intuito de enriquecer a proposta, duas arquiteturas "tradicionais" foram adaptadas e testadas, sendo essas: a) uma versão com três camadas da arquitetura MLP; e b) uma adaptação da arquitetura CNN, denominada como 1D-CNN, tendo como principal atributo uma camada totalmente conectada para criar um conjunto maior de atributos com características de localidade, seguida por várias camadas 1D-*Conv* (Shwartz-Ziv; Armon, 2022).

Quanto aos hiperparâmetros, esses foram definidos a partir de consultas à literatura e às documentações associadas às arquiteturas empregadas, e consistiram em elementos básicos conforme a relação abaixo:

a) ***batch size*:** número de exemplos de treinamento usados em uma única iteração do algoritmo de otimização durante o treinamento;
b) ***epochs*:** número de vezes que todo o conjunto de treinamento é apresentado ao modelo durante o treinamento;
c) ***optimizer*:** algoritmo de otimização usado para atualizar os pesos do modelo durante o treinamento;
d) ***dropout*:** técnica de regularização usada para evitar *overfitting* (sobreajuste);
e) ***num_layers***: número de camadas em uma rede neural;
f) ***num_trees***: número de árvores em um modelo;
g) ***max_epochs*:** número máximo de épocas permitido durante o treinamento de um modelo;

h) ***n steps*:** número de etapas usadas para atualizar os pesos do modelo durante o treinamento.

Finalmente, a síntese dos hiperparâmetros e dos respectivos valores e atributos testados em cada arquitetura de *deep learning* é apresentada no QUADRO 35. Outrossim, as estruturas projetadas para cada arquitetura, como número de camadas e funções de ativação, por exemplo, são listadas no APÊNDICE H.

QUADRO 35 - HIPERPARÂMETROS - ARQUITETURAS DE *DEEP LEARNING*

Arquitetura	Hiperparâmetros testados	Apoio teórico
1D-CNN	***batch size*:** [32, 64, 128] ***epochs*:** [50, 100, 150] ***optimizer*:** ['SGD', 'adam'] ***dropout*:** [0,2; 0,25; 0,3]	Shwartz-Ziv e Armon (2022)
FT-Transformer	***batch size*:** [32, 64, 128] ***epochs*:** [50, 100, 150]	Gorishniy *et al*. (2021)
MLP	***batch size*:** [32, 64, 128] ***epochs*:** [50, 100, 150] ***optimizer*:** ['SGD', 'adam'] ***dropout*:** [0,2; 0,25; 0,3]	Sahu *et al*. (2020)
NODE	***num_layers*:** [3, 5, 7, 9, 12, 15] ***num_trees*:** [64, 128, 256] ***batch size*:** [32, 64, 128] ***max_epochs*:** [50, 100, 150] ***optimizer*:** ['SGD', 'adam']	Shwartz-Ziv e Armon (2022)
TabNet	***n steps*:** [3, 5, 7, 9, 12, 15] ***batch size*:** [32, 64, 128] ***epochs*:** [50, 100, 150] ***optimizer*:** ['SGD', 'adam']	Shwartz-Ziv e Armon (2022)
TabTransform	***batch size*:** [32, 64, 128] ***epochs*:** [50, 100, 150] ***optimizer*:** ['SGD', 'adam']	Huang *et al.* (2020)

FONTE: O autor (2023).

Com relação aos algoritmos de aprendizado de máquina e os respectivos hiperparâmetros considerados, esses são detalhados na sequência.

3.6.2 Algoritmos de aprendizado de máquina empregados

Os algoritmos de aprendizado de máquina empregados nesta tese foram selecionados mediante o estudo de pesquisas correlatas, como as descritas no

QUADRO 1, da subseção 1.3 (justificativa), além de métodos considerados em trabalhos sobre arquiteturas de *deep learning* para dados tabulares, cujos objetivos consistiam em verificar o desempenho dessas arquiteturas diante de algoritmos "tradicionais".

Após esse estudo, seis algoritmos foram selecionados e testados ao longo da pesquisa e, para fins de compreensão, suscintas descrições dos algoritmos considerados nesta tese são fornecidas no QUADRO 36.

QUADRO 36 - ALGORITMOS DE APRENDIZADO DE MÁQUINA UTILIZADOS NA TESE

Algoritmo	**Descrição**	**Referência**
CatBoost	Algoritmo de *boosting* de gradiente que usa processamento paralelo para acelerar o treinamento e reduzir a necessidade de ajuste de hiperparâmetros.	Prokhorenkova *et al.* (2018)
k-Nearest Neighbors (k-NN)	Algoritmo de aprendizado supervisionado que classifica pontos de dados com base na classe da maioria dos seus vizinhos mais próximos em um espaço de recursos.	Cover e Hart (1967)
LightGBM	Algoritmo de *gradient boosting* que usa a técnica de exclusão de características baseada em histograma para acelerar o treinamento e reduzir o uso de memória.	Ke *et al*. (2017)
Random Forest	Algoritmo que constrói várias árvores de decisão independentes com diferentes amostras de dados e recursos e combina suas previsões para produzir uma saída final.	Breiman (2001)
Support Vector Machine (*SVM*)	Algoritmo de aprendizado supervisionado que encontra o hiperplano que melhor separa as classes em um espaço de recursos, maximizando a margem entre as classes.	Vapnik (1995)
XGBoost	Algoritmo de *boosting* de gradiente que usa a técnica de crescimento por árvore para construir árvores de decisão em cada iteração do treinamento.	Chen e Guestrin (2016)

FONTE: O autor (2023).

Quanto aos hiperparâmetros, esses foram definidos a partir de consultas à literatura e às documentações associadas às arquiteturas empregadas, e consistiram em elementos básicos conforme a relação abaixo:

- ***depth*:** número de camadas em uma árvore de decisão ou em um modelo de rede neural;
- ***learning rate*:** controla o tamanho dos passos durante o processo de ajuste dos pesos em um algoritmo de aprendizado de máquina;
- ***iterations*:** número de vezes que um algoritmo de aprendizado de máquina percorre todo o conjunto de dados durante o treinamento;
- ***n_neighbors*:** número de vizinhos mais próximos considerados em um algoritmo de classificação ou regressão baseado em vizinhos mais próximos (k-NN);
- ***weights*:** pesos atribuídos aos vizinhos mais próximos em um algoritmo k-NN.
- ***metric*:** métrica usada para medir a distância entre pontos em um algoritmo k-NN;
- ***num_leaves*:** número máximo de folhas em uma árvore de decisão ou em um modelo de floresta aleatória;
- ***min_data in leaf*:** número mínimo de exemplos de treinamento permitidos em uma folha em uma árvore de decisão ou em um modelo de floresta aleatória.
- ***max_depth*:** profundidade máxima permitida em uma árvore de decisão ou em um modelo de floresta aleatória;
- ***n_estimators*:** número de árvores em um modelo de floresta aleatória ou em um modelo de *boosting*;
- ***max_features*:** número máximo de recursos considerados em cada divisão em uma árvore de decisão ou em um modelo de floresta aleatória;
- ***min_samples_leaf*:** número mínimo de exemplos de treinamento permitidos em uma folha em uma árvore de decisão ou em um modelo de floresta aleatória;
- ***kernel*:** função usada em um algoritmo de SVM para transformar os dados de entrada em um espaço de maior dimensionalidade;
- ***C*:** hiperparâmetro em um algoritmo SVM que controla a penalidade por violação das margens de decisão;
- ***colsample bytree***: número de colunas (recursos) considerados em cada árvore em um modelo de *boosting*.

Finalmente, a síntese dos hiperparâmetros por algoritmos, e dos respectivos valores e atributos testados é apresentada no QUADRO 37.

QUADRO 37 - HIPERPARÂMETROS - ARQUITETURAS DE APRENDIZADO DE MÁQUINA

Modelo	Hiperparâmetros testados	Suporte teórico
CatBoost	***'depth'*:** [2, 3, 4, 5] ***'learning rate'*:** [0.0001, 0.001, 0.01, 0.1, 1.0] ***'iterations'*:** [10, 50, 100, 250, 500]	Hancock e Khoshgoftaar (2020)
K-NN	***'n_neighbors'*:** [5,7,9,11,13,15] ***'weights'*:** ['*uniform*','*distance*'] ***'metric'*:** ['*minkowski*','*euclidean*','*manhattan*']	Wazirali (2020)
LightGBM	***'num_leaves':*** [20, 100, 200, 500] ***'min_data in leaf':*** [10, 100, 200] ***'max_depth':*** [3, 6, 9, 12] ***learning_rate*:** [0.01, 0.1, 0.3]	Xiaolei, Mingxi e Zeqian (2020)
RF	***'n estimators'*:** [150, 200, 300, 400] ***'max features'*:** ['*auto*', '*sqrt*'] ***'min_samples_leaf'***: [2, 3, 4, 5], ***'max_depth'*:** [20, 30, 40, 50 *None*]	Joo *et al*. (2021)
SVM	***'kernel'*:** [*'linear', 'poly', 'rbf', 'sigmoid'*] ***'C'*:** [0.1, 1, 10, 100, 1000]	Mantovani *et al*. (2019)
XGBoost	***'max depth':*** [3, 6, 12] ***'learning rate':*** [0.1, 0.2, 0.3] ***'n estimators':*** [100, 200, 300] ***'colsample bytree':*** [0.4, 0.7, 1.0]	Pan *et al*. (2022)

FONTE: O autor (2023).

Expostos os hiperparâmetros empregados para cada algoritmo de aprendizado de máquina, a seguir, apresenta-se o protocolo com os experimentos conduzidos.

3.6.3 Protocolo de experimentos

Para a condução dos experimentos de classificação, um protocolo com as diferentes estratégias empregadas neste estudo foi desenvolvido. Neste protocolo, quatro cenários foram elaborados, sendo esses: a) classificação considerando todos os atributos/características e ajuste de hiperparâmetros (A); b) classificação mediante seleção de atributos/características e redução de dimensionalidade (B); c) classificação mediante balanceamento de classes (C); e d) classificação a partir de *data augmentation* (D).

No cenário número A, os experimentos de classificação foram realizados considerando todas as características extraídas para as bases de dados – um e dois – para a pesquisa. No caso das arquiteturas de *deep learning*, em virtude da proposta de multimodalidade, esse cenário ocorreu a partir da união das cinco modalidades – acordes, áudio, letras, simbólicas e voz – em um único arquivo; para os algoritmos de aprendizado de máquina, a classificação foi conduzida para cada modalidade.

Para mais, neste cenário, conduziram-se os ajustes dos hiperparâmetros relatados nas subseções 3.6.1 e 3.6.2, respectivamente, por meio de uma estratégia automática denominada *grid search*, responsável por conduzir uma busca exaustiva a partir dos parâmetros indicados para os métodos. Outrossim, nos cenários que envolviam métodos como o algoritmo de aprendizado de máquina SVM e arquiteturas de *deep learning*, devido ao tempo de processamento, o ajuste dos hiperparâmetros foi realizado de forma manual pelo pesquisador.

Para o segundo cenário, duas abordagens foram consideradas. Na primeira, que visou reduzir a dimensionalidade dos conjuntos de dados, aplicou-se a Análise de Componentes Principais (*Principal Component Analysis* (PCA)), determinado como percentual de variância 90,00% (Johnson; Smith; Adams, 1985; Alkhayrat; AljnidI; Aljoumaa, 2020).

Ao se trabalhar com PCA, tem-se que essa técnica é matematicamente definida como uma transformação linear ortogonal que transforma os dados em um novo sistema de coordenadas, de modo que a maior variação por alguma projeção dos dados venha a residir na primeira coordenada (primeiro componente principal), a segunda maior variação na segunda coordenada, e assim sucessivamente (Alkhayrat; Aljnidi; Aljoumaa, 2020).

Para a segunda abordagem, selecionou-se as melhores características para cada algoritmo e arquitetura, construída com o algoritmo boruta (Kursa; Rudnicki, 2010), cujo funcionamento consiste, basicamente, em remover iterativamente os recursos que são comprovados por um teste estatístico como menos relevantes do que versões aleatórias (Kursa; Rudnicki, 2010). Como algoritmo de classificação, acrescentou-se o *Random Forest* ao boruta.

No que tange ao terceiro cenário, os dados foram balanceados considerando duas estratégias: a) redução de exemplos de todas as classes, menos para os registros pertencentes à emoção “raiva/medo”, por meio da técnica *Tomek Links* (Tomek, 1976); e b) aumento dos registros da classe composta por dados da emoção

“raiva/medo”, aplicando o algoritmo *Synthetic Minority Oversampling Technique* (SMOTE) (Chawla *et al*., 2002).

Referente ao quarto e último cenário, este contemplou apenas a base um, em virtude dos resultados de classificação observados ao longo dos três cenários anteriores. Ademais, ressalta-se que o aumento dos registros, realizados de forma automática com a biblioteca *audiomentations*[29], foi conduzido apenas para a extração das características relacionadas às características pertencentes às modalidades áudio, simbólicas e voz.

Como justificativa para essa escolha, consideraram-se os resultados observados para essas modalidades durante os experimentos com os algoritmos de aprendizado de máquina, que indicaram melhor desempenho dessas sobre as outras duas características (acordes e letras). Por fim, como recursos para o aumento de dados, quatro foram testados: a) acréscimo de ruído gaussiano (*AddGaussianNoise*); b) alteração da velocidade (*TimeStretch*) ou a duração do sinal sem alterar o tom; c) mudança do tom (*PitchShift*), para cima e para baixo, sem comprometer o ritmo; e d) alternar a ordem do sinal (*Shift*).

A síntese dos cenários acima descritos é apresentada no protocolo de experimentos, listado no QUADRO 38.

[29] Disponível em: https://github.com/iver56/audiomentations. Acesso em: 21 jan. 2023.

QUADRO 38 - PROTOCOLO DE EXPERIMENTOS

#	Cenário	Descrição
A	Considerando todos os atributos extraídos e ajuste hiperparâmetros	***Deep learning*:** união dos atributos de todas as modalidades; ***Machine learning***: atributos de cada modalidade (ex: áudio; letras etc.).
		Automática: hiperparâmetros ajustados por meio da estratégia *grid search*; **Manual:** quando necessário, devido à complexidade computacional da abordagem automática em alguns métodos (ex: SVM; arquiteturas de *deep learning*).
B	Empregando seleção de atributos e redução de dimensionalidade	**Seleção de atributos:** seleção automática por meio do algoritmo boruta (Kursa; Rudnicki, 2010).
		Redução de dimensionalidade: PCA, considerando 90,00% de variância (Johnson; Smith; Adams, 1985; Alkhayrat; Aljnidi; Aljoumaa, 2020).
C	Aplicando balanceamento de classes	***Undersampling:*** remoção de exemplos de todas as classes, exceto a classe minoritária, empregando a abordagem *Tomek Links* (Tomek, 1976). Parâmetros *default*, com exceção à estratégia de amostragem (*sampling_strategy*), empregada como '*auto*' ***Oversampling*:** acréscimo de exemplos para a classe minoritária, empregando o algoritmo SMOTE (Chawla *et al.* 2002). Parâmetros *default*, com exceção à estratégia de amostragem (*sampling_strategy*), empregada como '*minority*'
D	Recorrendo a técnicas de *data augmentation*	Aumento de registros para a base um, utilizando a biblioteca *audiomentations*.

FONTE: O autor (2023).

Concernente ao pré-processamento, para todos os cenários realizou-se a padronização dos dados de treinamento, considerando $z = (x - u)$ / s, onde x representa uma amostra; u indica a média das amostras de treinamento; e s significa o desvio padrão das amostras de treinamento.

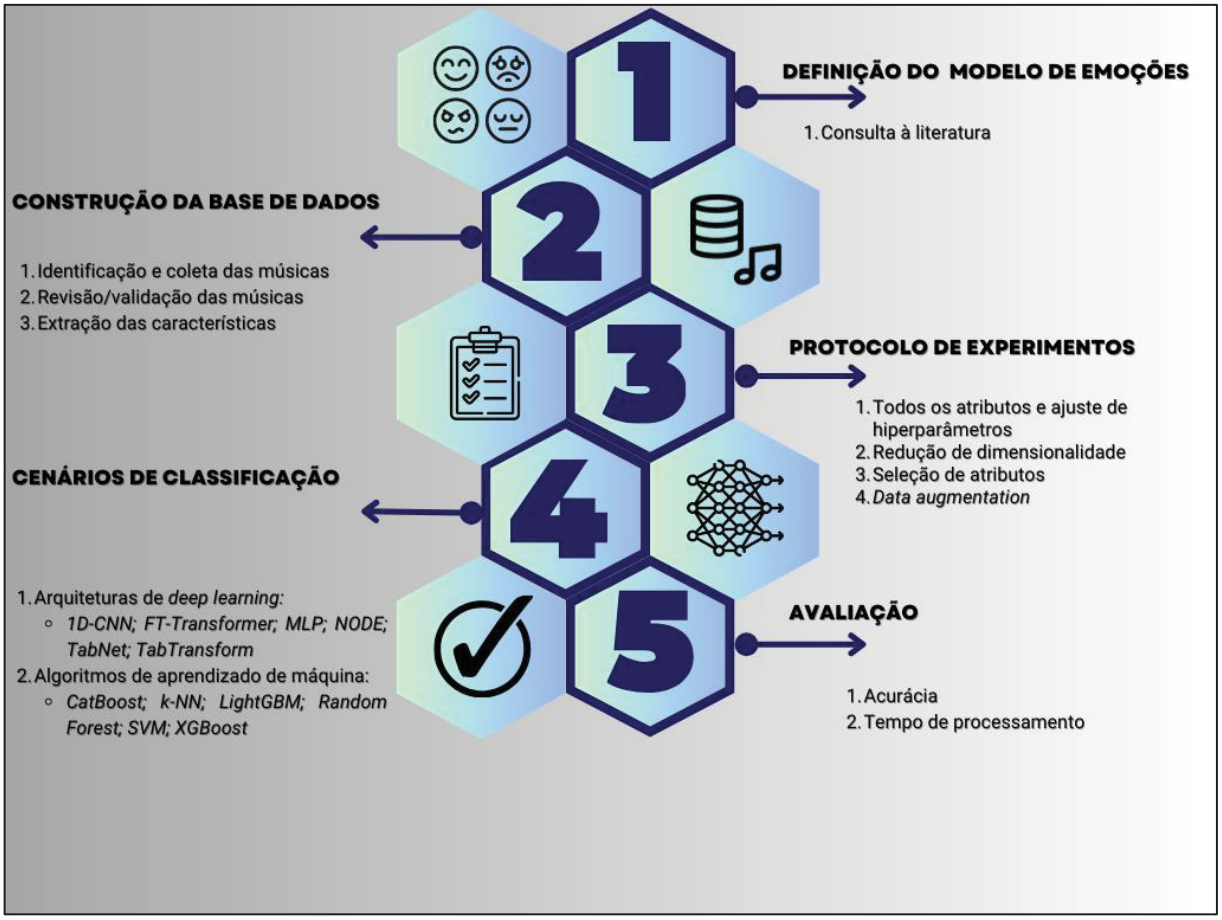
Quanto à separação dos conjuntos de treino e teste, respeitou-se a proporção 75,00% dos registros para treinamento e os 25,00% restantes para o conjunto de teste, com a separação de forma estratificada, considerando a função *train_test_split* da biblioteca *Scikit-learn* (Pedregosa *et al*., 2011). No tocante à avaliação dos

experimentos, a métrica selecionada correspondeu à acurácia obtida nos conjuntos de testes, considerando a seguinte equação: $\frac{total\ de\ acertos}{total\ de\ registros} * 100$.

3.7 SÍNTESE DA METODOLOGIA DESENVOLVIDA

Esta subseção fornece elementos de síntese acerca dos encaminhamentos metodológicos adotados nesta tese, visando facilitar a compreensão desses. Neste sentido, um resumo visual (FIGURA 33) e um quadro abarcando todas as ferramentas empregadas neste estudo foram elaborados.

FIGURA 33 - RESUMO VISUAL DOS ENCAMINHAMENTOS METODOLÓGICOS



FONTE: O autor (2023).

Com relação à lista de ferramentas utilizadas ao longo desta pesquisa, é prudente salientar que os objetivos associados à identificação do modelo de emoções a ser utilizado e a quais modalidades empregar para o reconhecimento de emoções em músicas não necessitaram de ferramentas específicas, tendo em vista que foram

conduzidos por meio de pesquisa bibliográfica e revisões de literatura. Para os demais, a relação com as ferramentas é apresentada no QUADRO 39.

QUADRO 39 - SÍNTESE DAS PRINCIPAIS FERRAMENTAS EMPREGADAS NA TESE

Objetivo	**Procedimento**	**Ferramentas**
Construir uma base multimodal composta por características relacionadas às modalidades identificadas	Identificar as músicas relacionadas ao modelo de emoções considerado	*BeautifulSoup* *Selenium* *Spotipy*
	Coletar as músicas identificadas	*pytube*
	Padronizar a duração das músicas	*pydub*
	Submeter as músicas identificadas para a revisão	-
	Extrair as características identificadas	*chord-extractor* *Essentia* *librosa* *music21* *NLTK* *Omnizart* *openSMILE* *spleeter*
Comparar diferentes arquiteturas de *deep learning* para o reconhecimento de emoções na base elaborada	Realização de cenários de classificação com os métodos selecionados	*Tensorflow* *PyTorch Tabular* *Keras* *pytorch-tabnet*
Comparar o desempenho de classificação obtido com a abordagem desenvolvida, com os desempenhos proporcionados por abordagens com uma modalidade, construídas a partir de algoritmos de aprendizado de máquina	Avaliar os resultados obtidos	*CatBoost* *LightGBM* *scikit-learn* *xgboost*

Fonte: elaborado pelo autor (2023).

Finalizada a seção acerca dos encaminhamentos metodológicos, a próxima seção contempla os resultados deste estudo.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Esta seção contempla os resultados para os processos de revisão e avaliação dos segmentos conduzidos junto aos voluntários, bem como discutem-se os resultados obtidos para as classificações consideradas as arquiteturas de *deep learning* e os algoritmos de aprendizado de máquina, além da comparação entre os resultados dessas duas abordagens. Por fim, as publicações obtidas até o momento são listadas.

4.1 RESULTADOS DA REVISÃO DAS MÚSICAS (EXPERIMENTO PILOTO)

Conforme relatado na subseção 3.4.3, o processo de revisão dos segmentos foi conduzido com o auxílio de oito voluntários, no qual, em um primeiro momento, almejou-se avaliar se as emoções apontadas pelos voluntários correspondiam aos termos retirados da plataforma *AllMusic*. Este cenário é apresentado na TABELA 4.

TABELA 4 - CONCORDÂNCIA ENTRE VOLUNTÁRIOS E PLATAFORMA

Emoção	Registros	Concordância	Concordância (%)
Amor/ternura	293	118	40,27%
Calma	347	61	17,58%
Felicidade	660	49	7,42%
Medo	210	10	4,76%
Raiva	450	83	18,44%
Tristeza	660	48	7,27%
Total	**2.620**	**369**	**14,09%**

FONTE: O autor (2022).

Ao observar-se a TABELA 4, percebe-se que o percentual geral de concordância entre os avaliadores e a plataforma foi de apenas 14,09%, com o maior percentual de concordância para os segmentos da emoção amor/ternura (40,27%). Com base nesses resultados, entendeu-se que não havia condições de sustentar que os segmentos classificados pelos voluntários correspondiam às emoções originais oriundas da *AllMusic*.

Neste sentido, relembra-se que o processo de rotulação adotado pela plataforma é fechado e não informa quais critérios os editores empregam para a

classificação das canções. Além do mais, não é possível saber por quantos editores cada canção é avaliada, tampouco não se pode precisar sobre qual duração e segmento foram atribuídos os rótulos emocionais.

Assim sendo, os rótulos originais, obtidos da plataforma *AllMusic* e tratados pelo pesquisador, foram desconsiderados. A partir desse descarte e, tendo em vista que o processo de revisão por parte dos avaliadores foi conduzido de forma cega, ou seja, sem estes terem conhecimento a respeito de qual emoção estava atrelada aos segmentos analisados, decidiu-se por considerar para a base de dados somente aqueles segmentos em que houve igualdade nas respostas dos avaliadores (emoção voluntário 1 = emoção voluntário 2).

Como exemplificação, um segmento foi rotulado com a emoção calma, caso os dois avaliadores que o escutaram (ex: A7 e A2) elegeram essa emoção; de maneira semelhante, caso os avaliadores 3 e 4 (A3 e A4) identificassem um segmento como relacionado à tristeza, esta foi adotada como a sua emoção. Nesta abordagem, a base de dados é a soma de todas as concordâncias identificadas nos conjuntos avaliados pelas duplas.

A redistribuição dos segmentos conforme o processo adotado é apresentada na TABELA 5.

TABELA 5 - REDISTRIBUIÇÃO DOS SEGMENTOS A PARTIR DA CONCORDÂNCIA DOS VOLUNTÁRIOS

Avaliadores	**Amor/ternura**	**Calma**	**Felicidade**	**Medo**	**Raiva**	**Tristeza**	**Total**
A1 e A2	4	7	38	29	96	65	239
A3 e A4	89	13	49	14	30	45	240
A5 e A6	35	16	8	23	20	48	150
A7 e A8	118	61	29	12	20	50	290
Total	**246**	**97**	**124**	**78**	**166**	**208**	**919**

FONTE: O autor (2022).

Conforme observa-se na TABELA 5, o total de registros para a base de dados foi de 919 registros. Para avaliar se as concordâncias obtidas poderiam ser sustentadas estatisticamente, empregou-se o coeficiente Kappa de Cohen (κ) (Cohen; 1960; Landis; Koch, 1977), conforme apresentado na TABELA 6.

TABELA 6 - AVALIAÇÃO DA CONCORDÂNCIA ENTRE OS AVALIADORES DE ACORDO COM O COEFICIENTE KAPPA DE COHEN (κ)

Avaliadores	***κ***	**Concordância (%)**	**p-valor (<)**	**Resultado**
A1 e A2	0,2127 [IC 95%: 0,17 - 0,25]	36,21%	0,0001	Razoável concordância estatisticamente significativa
A3 e A4	0,2224 [IC 95%: 0,17 - 0,26]	36,36%	0,0001	Razoável concordância estatisticamente significativa
A5 e A6	0,086 [IC 95%: 0,05 - 0,12]	22,73%	0,0001	Leve concordância estatisticamente significativa
A7 e A8	0,2898 [IC 95%: 0,24 - 0,34]	45,31%	0,0001	Razoável concordância estatisticamente significativa

FONTE: O autor (2022).

A definição da força da concordância foi realizada com base nos valores de *κ* (TABELA 6), e utilizou-se, como referência, o trabalho de Landis e Koch (1977). De acordo com esse estudo, as seguintes classificações podem ser efetuadas com base no valor de *κ*: a) < 0,00 (ruim); b) 0,00-0,20 (leve); c) 0,21-0,40 (razoável); d) 0,41-0,60 (moderado); e) 0,61-0,80 (substancial); e f) 0,81-1,00 (quase perfeito).

Por fim, ressalta-se, mesmo que, no máximo razoáveis, as concordâncias entre os avaliadores apresentaram-se como estatisticamente significativas (p-valor < 0,0001), considerando um intervalo de confiança de 95,00%.

4.2 RESULTADOS DA AVALIAÇÃO DAS MÚSICAS (EXPERIMENTO AJUSTADO)

Para o experimento ajustado, conforme exemplificado na seção 3.5.5, duas abordagens de validação foram adotadas: a primeira, considerando a concordância entre os voluntários; a segunda, concentrando-se na emoção mais indicada para cada música. Para a primeira abordagem, o total de músicas com concordância observada foi de 1.412, e o percentual de emoções por emoções é demonstrado na TABELA 7.

TABELA 7 - PERCENTUAL DE REGISTROS POR EMOÇÕES (BASE UM)

Emoção	Frequência	Frequência (%)
Alegria	442	31,30%
Raiva/Medo	236	16,71%
Serenidade	305	21,60%
Tristeza	429	30,38%
Total	**1412**	**100,00%**

FONTE: O autor (2023).

Como observa-se na TABELA 7, as emoções alegria e tristeza destacaram-se durante o processo avaliativo, concentrando pouco mais de 60,00% dos registros para a base um.

Concernente ao nível de concordância observada entre os avaliadores, recorreu-se ao Kappa de Fleiss (Fleiss, 1971; Hallgren, 2012) e os resultados obtidos são listados na TABELA 8.

TABELA 8 – RESULTADOS PARA A CONCORDÂNCIA ENTRE OS AVALIADORES

Indicador	Avaliador (1,2,3)	Avaliador (1,2,4)	Avaliador (1,2,5)
Percentual	21,80%	26,70%	39,90%
Kappa	0,2301	0,2343	0,4226
Concordância	Razoável	Razoável	Moderada
p-valor	0,0001	0,0001	0,0001
IC (95%)	(0,2136 - 0,2467)	(0,2159 - 0,2527)	(0,4057 - 0,4395)
Músicas	349	426	637

FONTE: elaborado pelo autor (2023).

Com base na análise da TABELA 8, observa-se que as concordâncias foram, minimamente, razoáveis, com exceção ao resultado alcançado para os registros avaliados pelos voluntários um, dois e cinco, cujo valor do Kappa de Fleiss indica uma concordância moderada. Por fim, ressalta-se que todos os valores observados são estatisticamente significativos ao se considerar um intervalo de confiança de 95,00%.

Quanto à base composta pelos registros para os quais maioria foi estabelecida como critério, 4.303 músicas foram selecionadas, conforme elenca-se na TABELA 9.

TABELA 9 - PERCENTUAL DE REGISTROS POR EMOÇÕES (BASE DOIS)

Emoção	Frequência	Frequência (%)
Alegria	1.237	28,75%
Raiva/Medo	602	13,99%
Serenidade	1.210	28,12%
Tristeza	1.254	29,14%
Total	**4.303**	**100,00%**

FONTE: O autor (2023).

A partir da TABELA 9, visualiza-se que a base composta pela maioria observada apresenta maior equilíbrio entre as emoções do que a base um, todavia, mesmo neste cenário, a emoção “raiva/medo”, assim como na base anterior, apresentou-se como a classe com o menor número de registros (602). Finalmente, quanto aos percentuais de músicas retidas em cada conjunto avaliado, esses são listados na TABELA 10.

TABELA 10 - PERCENTUAIS DE MÚSICAS RETIDAS (MAIORIA OBSERVADA)

Avaliadores	Percentual	Músicas
1,2,3	85,30%	1.364
1,2,4	90,30%	1.443
1,2,5	93,62%	1.496

Fonte: elaborado pelo autor (2023).

Para a abordagem considerando a maioria observada entre os avaliadores, nota-se alta taxa de retenção de segmentos, com o percentual médio igual a 89,75%. Para mais, no que tange às estatísticas descritivas das bases de dados desenvolvidas, estas estão organizadas no APÊNDICE I, para a base um, e no APÊNDICE J, para a base dois.

4.3 RESULTADOS PARA O CENÁRIO DE CLASSIFICAÇÃO A

Nesta subseção, detalham-se os resultados obtidos para o cenário de classificação A composto por classificações considerando todas as características extraídas e pelo ajuste dos hiperparâmetros descritos na subseção 3 .6 (estratégias de classificação). Neste aspecto, a partir do ajuste de hiperparâmetros, os melhores valores foram aplicados para o desenvolvimento dos experimentos nos seguintes cenários (B, C e D).

4.3.1 Resultados alcançados com as abordagens multimodais

Para as abordagens multimodais, considerando as arquiteturas de *deep learning*, os resultados obtidos considerando todos os atributos e o ajuste de hiperparâmetros são sintetizados na TABELA 11.

TABELA 11 - RESULTADOS PARA TODOS OS ATRIBUTOS E AJUSTE DE HIPERPARÂMETROS

Base dois			**Base um**		
Arquitetura	**Acurácia**	**Tempo (s)**	**Arquitetura**	**Acurácia**	**Tempo (s)**
TabNet	65,89%	91,07	*FT-Transformer*	81,02%	28,78
FT-Transformer	65,70%	50,75	*TabNet*	81,01%	44,94
NODE	63,47%	84,37	*NODE*	79,89%	151,52
TabTransform	63,38%	4,50	*MLP*	79,32%	21,14
MLP	63,10%	10,96	*TabTransform*	77,05%	3,39
1D-CNN	59,11%	124,23	*1D-CNN*	67,99%	82,04

FONTE: O autor (2023).

A partir da análise da TABELA 11, nota-se que as acurácias das arquiteturas melhoraram consideravelmente na base um em comparação com a base dois, com a diferença, em média, de 15,58%. Tal comportamento indica que a base um tende a representar melhor o comportamento dos dados em relação à base composta pela maioria identificada.

Quanto ao desempenho das arquiteturas, para a base dois, a arquitetura *TabNet* obteve a maior acurácia, com 65,89%, enquanto a arquitetura *1D-CNN* apresentou a menor acurácia, com 59,11% de acerto. Quanto à base um, as arquiteturas *FT-Transformer* e *TabNet* apresentaram as maiores acurácias, com 81,02% e 81,01%, respectivamente, enquanto a *1D-CNN* apresentou a menor acurácia, com 67,99%.

Em relação ao tempo de execução, assim como observado quanto à acurácia, de maneira geral, os tempos de execução para a base um foram inferiores aos da base dois. Neste sentido, percebe-se que a arquitetura NODE apresentou o maior tempo de execução na base um, com 151,52 segundos, enquanto na base dois, o maior tempo de execução (124,23 segundos) foi registrado para a arquitetura *1D-CNN*.

Por fim, os respectivos hiperparâmetros selecionados para a obtenção dos resultados aqui relatados são apresentados no APÊNDICE K.

4.3.2 Resultados obtidos com os algoritmos de aprendizado de máquina

Considerando a modalidade áudio, na base dois, a melhor acurácia foi obtida com o algoritmo *CatBoost* (todos os atributos), que apresentou 62,91% de acerto. Na base um, o algoritmo *CatBoost* (hiperparâmetros) obteve a maior acurácia com 77,33%. Em termos de tempo de execução, o algoritmo K-NN consistiu no mais rápido para ambas as bases de dados, com tempos de execução próximos a zero segundos.

Quanto aos experimentos envolvendo os acordes, o algoritmo com melhor desempenho na base dois foi o *SVM* (hiperparâmetros), com 42,10% de acurácia; em relação à base um, o *SVM* (hiperparâmetros), com 50,42% de acurácia, sobressaiu-se. Ademais, assim como para o áudio, os menores tempo de execução foram registrados para o algoritmo K-NN nas duas bases de dados.

Para as letras, na base dois, o modelo com melhor acurácia foi o XGBoost (hiperparâmetros) com 45,72%, e na base um, o SVM (hiperparâmetros) obteve o melhor desempenho com 54,67%. Novamente, quanto ao tempo de execução, o K-NN apresentou-se como o algoritmo mais rápido para ambas as bases de dados.

Concernente às características simbólicas, na base dois, o algoritmo *CatBoost* (hiperparâmetros), com 56,13% de acerto, apresentou-se como o melhor método. Na base um, o *SVM* (hiperparâmetros) obteve o melhor desempenho, alcançando 68,27% de acurácia. Em relação ao tempo de processamento, assim como observado até o momento, o K-NN teve o tempo de execução mais rápido para ambas as bases de dados.

Finalmente, em relação à voz, o modelo com melhor acurácia na base dois foi o *CatBoost* (todos os atributos) com 60,40%, e na base um, o *CatBoost* (hiperparâmetros) obteve a maior acurácia com 72,23%. O modelo com tempo de execução mais rápido – assim como para as demais modalidades – foi o K-NN, com o melhor desempenho observado nas bases deste estudo.

A comparação detalhada da acurácia e do tempo de processamento para cada modalidade e algoritmo é apresentada, respectivamente, no APÊNDICE L (todos os atributos) e no APÊNDICE M (hiperparâmetros), enquanto os hiperparâmetros selecionados para os algoritmos de aprendizado de máquina são listados no APÊNDICE N.

4.4 RESULTADOS PARA O CENÁRIO DE CLASSIFICAÇÃO B

Nesta subseção, detalham-se os resultados obtidos para o cenário de classificação B, composto por classificações empregando redução de dimensionalidade e seleção de atributos.

4.4.1 Resultados alcançados com as abordagens multimodais

Para as abordagens multimodais, elaboradas a partir da aplicação das arquiteturas de *deep learning*, os hiperparâmetros que apresentaram os melhores resultados para o cenário um foram replicados (APÊNDICE K). Como primeira estratégia metodológica neste cenário, recorreu-se ao uso da técnica PCA, com 90,00% de variância, que resultou em 72 componentes para a base dois, e 67 componentes para a base um. Os resultados dessa estratégia são listados na TABELA 12.

TABELA 12 - RESULTADOS OBTIDOS COM O USO DE PCA, COM 90,00% DE VARIÂNCIA

Base dois			**Base um**		
Arquitetura	**Acurácia**	**Tempo (s)**	**Arquitetura**	**Acurácia**	**Tempo (s)**
TabNet	63,47%	62,56	*MLP*	77,90%	4,35
MLP	61,15%	17,36	*TabNet*	75,92%	27,00
1D-CNN	60,13%	142,01	*NODE*	75,35%	147,25
NODE	53,15%	96,96	*FT-Transformer*	73,75%	28,11
FT-Transformer	28,71%	26,12	*1D-CNN*	73,65%	11,00
TabTransform	21,56%	94,62	*TabTransform*	73,37%	4,64

FONTE: O autor (2023).

Com base na TABELA 12, no que diz respeito ao desempenho das arquiteturas de *deep learning*, verifica-se variação de acordo com a base de dados utilizada. No caso da base dois, observa-se que a arquitetura *TabNet* apresentou a maior acurácia (63,47%), seguida pelas arquiteturas MLP e 1D-CNN, com acurácias de 61,15% e 60,13%, respectivamente. Ainda quanto à base dois, é importante destacar os resultados obtidos com as arquiteturas *FT-Transformer* e *TabTransform*, cujas taxas de acerto ficaram abaixo dos 30,00%.

Para a base um, a arquitetura MLP apresentou aumento consistente na acurácia, apresentando o melhor desempenho para essa base, com 77,90% de acerto, seguida pelas arquiteturas *TabNet* (75,92%) e NODE (75,35%), cujos

desempenhos também se apresentaram como consistentes ao serem comparados com os alcançados na base dois.

Além das acurácias, é importante ressaltar que o tempo de execução das arquiteturas também variou consideravelmente, com destaque para a arquitetura MLP que apresentou tempos de execução inferiores às demais arquiteturas em ambas as bases, destacando-se os 4,35 segundos registrados na base um. Por outro lado, a arquitetura NODE registrou tempos elevados, em ambas as bases, como os 147,25 segundos apresentados na base um.

Nos resultados alcançados por meio da seleção de atributos, salienta-se que esses foram alcançados com o número de características igual a 118, para a base dois, e 123, para a base um, além da aplicação padronização nos dados. As taxas de acerto obtidas são representadas na TABELA 13.

TABELA 13 - RESULTADOS OBTIDOS COM O USO DE SELEÇÃO DE ATRIBUTOS

Base dois			**Base um**		
Arquitetura	**Acurácia**	**Tempo (s)**	**Arquitetura**	**Acurácia**	**Tempo (s)**
TabNet	66,45%	87,07	*TabNet*	79,04%	38,62
TabTransform	65,89%	4,36	*NODE*	77,90%	153,48
NODE	65,42%	91,07	*TabTransform*	76,77%	3,00
FT-Transformer	64,21%	30,02	*MLP*	67,42%	6,97
MLP	60,50%	18,76	*1D-CNN*	66,85%	41,08
1D-CNN	59,57%	52,88	*FT-Transformer*	58,64%	10,84

FONTE: O autor (2023).

Observa-se na TABELA 13 que a arquitetura *TabNet* obteve o melhor desempenho em ambas as bases de dados, com acurácia de 66,45% na base dois e 79,04% na base um. Além disso, essa arquitetura apresentou tempos de processamento razoáveis nas duas bases (87,07 segundos na base dois e 38,62 segundos na base um).

Os resultados acima representados (TABELA 13) demonstram que, na base dois, além da arquitetura *TabNet,* as arquiteturas *TabTransform* e *NODE*, apresentaram acurácias levemente superiores aos 65,00% de acerto. De maneira semelhante, condizente à base um, essas arquiteturas apresentaram desempenho semelhante à arquitetura *TabNet*, com valores iguais 77,90% e 76,77%de acurácia, respectivamente.

Além da acurácia apresentada pelas arquiteturas de *deep learning*, observa-se que, em relação à base dois, a arquitetura *TabTransform* atingiu o menor tempo de

processamento, finalizando o seu treinamento em 4,36 segundos. Do mesmo modo, quanto à base um, essa mesma arquitetura também apresentou o menor tempo de processamento, com 3 segundos.

Indo além, com base nos resultados obtidos, pode-se concluir que a abordagem de seleção de atributos apresentou melhores resultados para a base dois, com arquiteturas mais rápidas e com acurácias levemente superiores em comparação à abordagem PCA. Nesta comparação, observa-se que a acurácia das arquiteturas variou entre 21,56% (*TabTransform*) a 63,47% (*TabNet*), sendo que a maioria das arquiteturas apresenta uma acurácia inferior a 61,00%. Concernente ao tempo de processamento, este oscilou de 17,36 segundos (*MLP*) a 142,01 segundos (*1D-CNN*).

Por outro lado, ao avaliar a tabela dos resultados com seleção de atributos, observa-se que a acurácia das arquiteturas é sutilmente maior, variando entre 59,57% (*1D-CNN*) e 66,45% (*TabNet*). Neste aspecto, cabe destacar que a arquitetura *TabNet* apresentou a maior acurácia tanto na abordagem PCA quanto na seleção de atributos. Finalmente, ao ser comparado, o tempo de processamento com seleção de atributos é menor se comparado ao obtido junto ao PCA, variando entre 4,36 segundos (*TabTransform*) a 91,07 segundos (*NODE*).

Adotando a mesma análise para a base um, visualize-se que os resultados apresentados nesta base indicam que as arquiteturas obtiveram um desempenho melhor do que na base dois em termos de acurácia, com as arquiteturas MLP (77,90%) e *TabNet* (79,04%) apresentando-se como as arquiteturas mais assertivas em suas bases: base dois, para a primeira, e base um, para a segunda.

Em relação ao tempo de processamento, nota-se que a arquitetura *NODE* apresentou o maior tempo de processamento em ambas as bases e abordagens - 147,25 (base dois) e 153,48 (base um), enquanto a *TabTransform* apresentou os menores tempos para as duas bases: 3,00 segundos para a base um e 4,64 segundos para a base dois.

Ao comparar as duas abordagens – PCA e seleção de atributos -, percebe-se que ambas produzem resultados até certo ponto próximos em relação à acurácia, com leve vantagem para o uso de PCA, cuja média de acerto girou em torno de 74,99% contra 71,10% para a seleção de atributos, considerando a base um, cujas taxas de acerto foram superiores às oriundas da classificação desempenhada na base dois.

4.4.2 Resultados obtidos com os algoritmos de aprendizado de máquina

Analisando a base dois classificada a partir da aplicação de PCA - cuja relação do número de componentes para cada base e modalidade é listada na TABELA 14 - tem-se que, para o áudio, a melhor acurácia nesta modalidade foi obtida com o algoritmo *LightGBM,* cujo valor atingiu 62,45% de acerto. Para os acordes, apenas o *XGBoost* obteve acurácia superior a 40,00% (40,14%). Considerando as letras como modalidade principal, o *LightGBM* obteve a melhor acurácia, com 43,77%. Concernente às características simbólicas, para esta modalidade, dois algoritmos apresentaram 51,57% de acerto, sendo esses o *Random Forest* e o *XGBoost*. Finalmente, considerando o cenário de classificação composto pelas características associadas à voz, o algoritmo SVM obteve o melhor desempenho, alcançando 59,10% de acurácia.

TABELA 14 - COMPONENTES PRINCIPAIS (BASES MAIORIA E CONCORDÂNCIA)

Modalidade	**Base dois**	**Base um**
Áudio	29	28
Acordes	12	11
Letras	4	4
Simbólicas	7	7
Voz	27	26

FONTE: O autor (2023).

Acerca do tempo de execução, o algoritmo mais rápido para todas as características foi o *K-NN*, com tempos de processamento próximos a 0,00 segundos; por sua vez, o algoritmo que apresentou o maior tempo de processamento foi o SVM, que levou 84,76 segundos para treinar os dados relacionados às características que representavam a modalidade voz.

Ao considerar os resultados para a base um empregando redução de dimensionalidade, ao utilizar o áudio como modalidade, o algoritmo SVM obteve a melhor acurácia, com 76,48% de acerto. No que diz respeito ao uso dos acordes, o algoritmo *CatBoost* teve melhor desempenho, com 47,59%. Para o cenário considerando as letras como recurso principal, o algoritmo *K-NN* obteve a melhor acurácia, com 53,54%.

Nas características simbólicas, o melhor algoritmo foi o *LightGBM*, cuja acurácia foi de 60,91%. Em conclusão, para as características representativas da voz, o algoritmo SVM obteve a melhor acurácia, com 72,52%. Quanto ao tempo de

execução, novamente o algoritmo K-NN, com processamentos próximos a 0,00 segundos, apresentou-se como o algoritmo mais rápido em todas as características, na base um.

Adentrando-se nos resultados da seleção de atributos, cujo número de características selecionadas para cada modalidade é elencado na TABELA 15, na base dois, os modelos com melhor desempenho em relação às características de áudio foram o *CatBoost* e o *SVM*, com acurácias de 62,45% e 60,68%, respectivamente. Em relação aos acordes, o *LightGBM* foi o modelo com melhor desempenho, atingindo uma acurácia de 33,36%. Já em relação às letras, o *XGBoost* teve o melhor desempenho, com acurácia de 50,28%. Nas características simbólicas, o *CatBoost* obteve a melhor acurácia, com 56,04%. Em relação à voz, o algoritmo *SVM* teve um desempenho superior, com uma acurácia de 59,85%.

TABELA 15 - NÚMERO DE CARACTERÍSTICAS SELECIONADAS

Modalidade	**Base dois**	**Base um**
Áudio	54	64
Acordes	1	5
Letras	5	7
Simbólicas	8	9
Voz	51	51

FONTE: O autor (2023).

Na base um, o modelo *CatBoost* obteve a melhor acurácia em relação à característica de áudio, atingindo 77,90%. Em relação aos acordes, o modelo *XGBoost* obteve uma acurácia de 47,03%. Para as letras, o algoritmo *SVM* obteve o melhor desempenho, com uma acurácia de 54,67%. Em relação às características simbólicas, o modelo *SVM* obteve o melhor desempenho, com uma acurácia de 67,42%. Por fim, na característica pertencentes à voz, o modelo SVM obteve novamente a melhor acurácia, com uma acurácia de 72,80%.

Em relação ao tempo de execução dos modelos, percebe-se variação entre os diferentes modelos e características. Na base dois, o modelo SVM apresentou o tempo de execução mais alto para a base, com 23,36 segundos de processamento nas características simbólicas, enquanto que na base um, o modelo *CatBoost* apresentou o tempo de execução mais alto para a característica de áudio, com 17,02 segundos. No geral, os modelos que tiveram um tempo de execução mais rápido foram o *K-NN* e o *XGBoost*, com tempos abaixo de 5 segundos em ambas as bases.

A comparação detalhada da acurácia e do tempo de processamento para cada modalidade e algoritmo, considerando o uso de PCA, é apresentada no APÊNDICE O, e, para a seleção de atributos, os indicadores constam no APÊNDICE P.

4.5 RESULTADOS PARA O CENÁRIO DE CLASSIFICAÇÃO C

Nesta subseção, detalham-se os resultados obtidos para o cenário de classificação C, composto por classificações aplicando balanceamento de classes.

4.5.1 Resultados alcançados com as abordagens multimodais

Os resultados alcançados para a base dois (TABELA 16), considerando a estratégia de remoção de registros das classes, demonstram que a arquitetura *FT-Transformer* obteve a maior acurácia, atingindo 65,70%. Além dessa arquitetura, ressaltam-se os desempenhos das arquiteturas *TabNet*, com acurácia de 65,25% e *TabTransform*, cuja acurácia foi de 62,54%, obtida em apenas 9,21 segundos de treinamento. Como arquitetura mais lenta em relação ao tempo para processamento, o posto pertenceu a arquitetura 1D-CCN, com 142,11 segundos.

TABELA 16 - RESULTADOS PARA A ESTRATÉGIA *UNDERSAMPLING*

Base dois			**Base um**		
Arquitetura	**Acurácia**	**Tempo (s)**	**Arquitetura**	**Acurácia**	**Tempo (s)**
FT-Transformer	65,70%	38,38	*TabTransform*	79,60%	2,61
TabNet	65,25%	91,64	*NODE*	78,47%	178,37
TabTransform	62,54%	9,21	*TabNet*	78,18%	45,38
NODE	62,36%	79,57	*FT-Transformer*	76,48%	21,43
MLP	58,27%	21,11	*MLP*	70,82%	6,01
1D-CNN	57,43%	142,41	*1D-CNN*	56,09%	12,69

FONTE: O autor (2023).

Na base um, as arquiteturas apresentaram resultados mais satisfatórios do que na base dois, sugerindo que a concordância entre os exemplos influenciou positivamente a capacidade de classificação dos modelos. Para esta base, a arquitetura *TabTransform* obteve a maior acurácia na categoria, com 79,60% de acerto. Além do mais, apenas a arquitetura *1D-CNN* não ultrapassou os 70,00% de acerto. Em relação ao tempo de execução, assim como na base dois, a *TabTransform*

foi a arquitetura mais rápida (2,61 segundos) e a arquitetura *NODE* a mais lenta, com 178,37 segundos.

Para a estratégia de acréscimo de registros (TABELA 17) – *oversampling* – para a classe minoritária – raiva/medo -, na base dois, a arquitetura *TabNet* novamente obteve a maior acurácia, alcançando 66,72%. Todavia, essa arquitetura apresentou o maior tempo de treinamento, com 117,98 segundos. Nesta base, como resultado negativo, tem-se o desempenho da arquitetura *TabTransform*, cuja acurácia foi de apenas 35,03%. Ademais, a arquitetura com melhor desempenho considerando o tempo de processamento consistiu na MLP, com 21,39 segundos.

TABELA 17 - RESULTADOS PARA A ESTRATÉGIA *OVERSAMPLING*

Base dois			**Base um**		
Arquitetura	**Acurácia**	**Tempo (s)**	**Arquitetura**	**Acurácia**	**Tempo (s)**
TabNet	66,72%	117,98	*FT-Transformer*	79,89%	43,69
NODE	65,15%	102,01	*NODE*	79,04%	249,79
FT-Transformer	63,75%	84,90	*MLP*	79,04%	5,75
MLP	63,38%	21,39	*TabNet*	77,05%	28,82
1D-CNN	57,80%	103,34	*TabTransform*	76,77%	2,67
TabTransform	35,03%	104,95	*1D-CNN*	70,25%	21,07

FONTE: O autor (2023).

A estratégia *oversampling* na base um resultou na arquitetura *FT-Transformer* como aquela com a maior acurácia, atingindo 79,89% de acerto. Em segundo lugar, visualiza-se empate entre as arquiteturas *NODE* e *MLP*, ambas com uma acurácia de 79,04%. Entretanto, é prudente destacar que a arquitetura *NODE* é aquele com o tempo mais longo de processamento, com 249,79 segundos. Por sua vez, a *TabTransform* é a arquitetura mais rápida, levando apenas 2,67 segundos para concluir o seu treinamento.

4.5.2 Resultados obtidos com os algoritmos de aprendizado de máquina

Na base dois, os resultados para a estratégia *undersampling* apresentaram acurácias variando entre 35,78% e 60,96%, com a maioria dos modelos apresentando acurácia abaixo de 60,00%. Em relação ao tempo de execução dos modelos, é possível notar que os modelos baseados em árvores (*XGBoost*, *CatBoost*, *LightGBM* e RF) apresentaram tempos de execução maiores que os outros algoritmos.

A maior acurácia na base dois foi obtida pelo algoritmo *XGBoost* ao ser aplicado para classificação considerando as características relacionados ao áudio. Neste cenário, esse algoritmo atingiu 60,96% de acerto.

Nota-se baixo desempenho em todos os modelos empregados para classificar acordes e letras, cujas acurácias estiveram abaixo de 50,00%, enquanto as características simbólicas e as referentes à voz apresentaram acurácia acima de 50,00% para todos os modelos, com destaque para o *XGBoost* (55,20% - simbólicas) e *SVM* (56,97% - voz). Além disso, é interessante notar que o *K-NN* apresentou tempos de execução menores ao longo dos experimentos se comparado aos demais algoritmos.

No tocante à base um, as acurácias variam de 43,90% a 76,77%, enquanto os tempos de execução variam de 0,03 segundos a 16,70 segundos. Na classificação da modalidade áudio, o modelo *CatBoost* obteve a maior acurácia com 76,77%. Para a classificação de acordes, o modelo *XGBoost*, com 52,12% de acurácia destacou-se, enquanto o algoritmo *CatBoost*, para as letras, alcançou 53,82% de acurácia.

Para as modalidades restantes, o algoritmo *Random Forest* atingiu 66,28% de acerto para a classificação baseada nas características simbólicas, enquanto o modelo *XGBoost* teve a maior acurácia para as características extraídas da voz, com 72,23%. Finalizando, em relação ao tempo de execução, o algoritmo com melhor desempenho foi o K-NN, com tempos abaixo dos dez segundos.

Considerando a estratégia de *oversampling*, os algoritmos de aprendizado de máquina na base dois, no que diz respeito à classificação da modalidade áudio, o algoritmo que apresentou melhor desempenho foi o *CatBoost*, com acurácia de 60,96%. Já para a classificação de acordes e letras, todos os modelos apresentaram desempenho ruim, com acurácia abaixo de 45,00%.

Para a classificação das características simbólicas, dois modelos obtiveram os melhores desempenhos, sendo esses o *CatBoost* e o *XGBoost*, ambos com acurácia de 52,69%. Por fim, para a classificação da modalidade voz, todos os modelos apresentaram desempenho relativamente semelhante, com destaque para o *CatBoost* e o *XGBoost*, empatados como os melhores resultados, alcançando acurácia de 55,20%. No que se refere ao tempo de execução, é possível observar que os modelos *LightGBM* e *K-NN*, em geral, apresentaram tempos menores de execução, enquanto os modelos SVM e *CatBoost* apresentaram tempos de execução maiores.

Na classificação com *oversampling* na base um, a modalidade áudio teve o maior resultado, proporcionado pelo algoritmo *CatBoost* cuja acurácia atingiu 80,17%. Entretanto, esse modelo foi o responsável pelo maior tempo de processamento (18,95 segundos). Na classificação por meio dos acordes, nenhum dos modelos obteve uma acurácia muito alta, sendo o modelo *Random Forest* o melhor com uma taxa de acerto de 52,97%.

Para a classificação das características baseadas em letras, novamente o modelo *XGBoost* obteve a maior acurácia (52,12%). Para a tarefa de classificação das características simbólicas, os algoritmos SVM e *LightGBM* apresentaram a maior acurácia (65,72%). Finalmente, ao se considerar a voz, houve um empate triplo entre os algoritmos *XGBoost*, *Random Forest* e *LightGBM*, todos com 70,53% de acerto para essa modalidade.

A comparação detalhada da acurácia e do tempo de processamento para cada modalidade e algoritmo, considerando a estratégia *undersampling*, é apresentada no APÊNDICE Q, e, para a estratégia *oversampling*, os indicadores constam no APÊNDICE R.

4.6 RESULTADOS PARA O CENÁRIO DE CLASSIFICAÇÃO D

Referente ao quarto e último cenário, este destinou-se apenas à base um, em virtude dos resultados de classificação observados ao longo dos três cenários anteriores, e foi conduzido apenas considerando a extração das características relacionadas às modalidades áudio, simbólicas e voz, tendo em vista que essas se sobressaíram durante os experimentos com os algoritmos de aprendizado de máquina.

Neste cenário, os 1.412 registros da base um foram enriquecidos com os registros aumentados que obtiveram os melhores resultados com os algoritmos de aprendizado de máquina, posto que, com esses algoritmos, cada modalidade foi testada de forma isolada. Para tanto, os melhores cenários foram obtidos, respectivamente considerando o uso de: a) *shift* para a modalidade áudio; b) *shift* para as características simbólicas; e c) *pitch* para as características relacionadas à voz.

Desse modo, a base um, em sua versão “aumentada”, contou com mais 1.412 registros, totalizando 2.824 músicas. Todavia, é necessário informar que os dados “aumentados” só foram inseridos para os conjuntos de treinamento, selecionado

esses registros a partir dos ids, a fim de evitar o vazamento de dados, ou seja, que os dados originais, armazenados no conjunto de treinamento pudessem “reconhecer” as suas cópias, inflacionando a acurácia dos experimentos.

Para as abordagens multimodais desenhadas a partir do uso das arquiteturas de *deep learning*, as modalidades letras e acordes foram desconsideradas, e o caracter multimodal foi sustentado pelas modalidades áudio, características simbólicas e voz. A síntese dos resultados para esse cenário é apresentada na TABELA 18.

TABELA 18 – RESULTADOS: ARQUITETURAS DE *DEEP LEARNING* (*DATA AUGMENTATION*)

Arquitetura	**Acurácia**	**Tempo (s)**
TabNet	80,73%	55,40
NODE	78,18%	382,96
FT-Transformer	77,90%	80,39
TabTransform	75,92%	19,24
MLP	75,35%	21,18
1D-CNN	70,82%	30,66

FONTE: elaborado pelo autor (2023).

Neste cenário, o melhor desempenho foi obtido considerando a aplicação da arquitetura *TabNet*, com acurácia igual a 80,73%, e um tempo de treinamento moderado de 55,40 segundos. Outro resultado a ser destacado consiste no obtido com a arquitetura *NODE*, que apresentou a segunda melhor acurácia, 78,18%, mas com um tempo de treinamento elevado - 382,96 segundos - ao ser comparado aos obtidos pelas outras arquiteturas.

Em relação aos resultados alcançados com algoritmos de aprendizado de máquina (TABELA 19), pode-se observar que a acurácia média desses algoritmos variou entre 60,00% e 78,00%, mais ou menos.

TABELA 19 – RESULTADOS: ALGORITMOS DE APRENDIZADO DE MÁQUINA (*DATA AUGMENTATION)*

Áudio			**Simbólicas**			**Voz**		
Modelo	**Acurácia**	**Tempo (s)**	**Modelo**	**Acurácia**	**Tempo (s)**	**Modelo**	**Acurácia**	**Tempo (s)**
CatBoost	78,75%	16,24	*SVM*	69,68%	4,34	*LightGBM*	75,64%	1,03
SVM	76,48%	0,25	*CatBoost*	64,87%	7,98	*SVM*	73,08%	0,30
LightGBM	73,93%	14,78	*XGBoost*	64,30%	0,64	*CatBoost*	71,11%	1,64
XGBoost	73,65%	1,37	*LightGBM*	64,02%	0,79	*XGBoost*	70,82%	1,96
RF	72,80%	5,95	*RF*	62,60%	2,19	*RF*	69,68%	3,55
K-NN	72,22%	0,09	*K-NN*	60,90%	0,09	*K-NN*	68,55%	0,07

FONTE: O autor (2023).

Considerando o áudio, o algoritmo *CatBoost* apresentou a melhor acurácia com 78,75%. Em relação ao tempo de treinamento, o algoritmo *K-NN* foi o mais rápido, levando apenas 0,09 segundos para finalizar o seu treinamento. Para a classificação empregando as características simbólicas, a melhor acurácia foi obtida com o algoritmo SVM, com valor igual a 69,68%. A respeito do tempo de treinamento, o algoritmo K-NN foi novamente o mais rápido, com os mesmos 0,09 segundos de execução. Finalmente, para a voz, o *LightGBM* apresentou a melhor acurácia com 75,64% de acerto. Quanto ao tempo de treinamento, assim como observado nas outras modalidades, o K-NN apresentou-se como o método mais rápido, treinando a sua base em um tempo de 0,06 segundos.

Finalizada a seção dos resultados obtidos para os cenários elaborados, a seguir é estabelecida a comparação entre esses, a fim de sintetizar os principais pontos observados nesta pesquisa.

4.7 COMPARAÇÃO E SÍNTESE DOS RESULTADOS

Com base nos resultados relatados nas seções anteriores, tem-se que, para a base dois, a arquitetura *TabNet* obteve a maior acurácia em quatro das cinco abordagens desenhadas para esta base (todos os atributos e hiperparâmetros (TA/HI), seleção de atributos (SA), PCA, redução de registros (*UNDER*) e aumento de registros (*OVER*)). Ademais, essa arquitetura apresentou o maior valor de acerto, 66,72%, na abordagem de aumento dos registros das classes (*OVER*). Para esta acurácia, a classificação para cada emoção é apresentada na TABELA 20.

TABELA 20 - MELHOR CENÁRIO DE *DEEP LEARNING* (BASE DOIS)

Emoção	Alegria	Raiva/medo	Serenidade	Tristeza
Alegria	**67,64%**	13,27%	17,15%	1,94%
Raiva/medo	26,00%	**66,00%**	2,67%	7,33%
Serenidade	11,22%	1,65%	**59,41%**	27,72%
Tristeza	3,82%	2,55%	20,38%	**73,25%**

FONTE: O autor (2023).

Analisando a TABELA 20, nota-se que a emoção tristeza apresenta a maior taxa de acerto, com 73,25% das amostras corretamente classificadas, enquanto serenidade possui uma taxa de acerto de 59,41%. A emoção alegria possui a segunda melhor taxa de acerto, com 67,64%. No entanto, é importante ressaltar que há uma porcentagem considerável de amostras dessa classe sendo erroneamente classificadas como raiva/medo (13,27%) e serenidade (17,15%).

A classe raiva/medo apresenta 66,00% das amostras corretamente classificadas. Todavia, também apresenta uma proporção não negligenciável de amostras classificadas incorretamente como alegria (26,00%) e tristeza (7,33%). Tais pontos indicam que a arquitetura *TabNet* pode ter dificuldades em distinguir entre as emoções alegria e raiva/medo, assim como entre alegria e serenidade.

Quanto às demais arquiteturas, nota-se estabilidade quanto à acurácia, com os valores dessa métrica girando em torno dos 60,00% de acerto. Todavia, sinaliza-se o desempenho das arquiteturas *FT-Transformer* e *TabTransform* com a aplicação de PCA, tendo em vista os baixos valores apresentados, sendo esses iguais a 28,71% e 21,56%, respectivamente. Outrossim, a arquitetura *TabTransform* também apresentou desempenho abaixo na abordagem com acréscimo de registros para a classe minoritária (*OVER*), com apenas 35,03% de acerto.

A comparação gráfica entre as arquiteturas de *deep learning* considerando as abordagens desenhadas para a base dois é apresentada no APÊNDICE S.

Ao transpor a análise para os resultados alcançados para base um, visualize-se os dois melhores resultados desta tese, obtidos com as arquiteturas *FT-Transformer* (81,02%) e *TabNet* (81,01%), oriundos das classificações realizadas considerando a abordagem TA/HI. Além do mais, para essa base, os resultados fornecidos pelas outras arquiteturas foram, em grande parte, superiores aos obtidos por essas na base dois.

Para essa base, a classificação para cada emoção, considerando o resultado obtido com a arquitetura *FT-Transformer* é apresentada na TABELA 21.

TABELA 21 - MELHOR CENÁRIO DE *DEEP LEARNING* (BASE UM)

Emoção	Alegria	Raiva/medo	Serenidade	Tristeza
Alegria	**89,20%**	4,50%	6,30%	0,00%
Raiva/medo	15,26%	**84,74%**	0,00%	0,00%
Serenidade	9,21%	0,00%	**73,69%**	17,10%
Tristeza	3,74%	0,93%	19,63%	**75,70%**

FONTE: O autor (2023).

A partir da TABELA 21, verifica-se que, assim como na base dois, a emoção tristeza apresenta a maior taxa de acerto, com 75,70% das amostras corretamente classificadas. A emoção raiva/medo também possui uma taxa de acerto considerável, com 84,74% das amostras corretamente classificadas. A emoção alegria teve uma taxa 89,20% das amostras corretamente classificadas, com o restante das suas instâncias erroneamente classificadas como raiva/medo (4,50%) e serenidade (6,30%).

Por fim, serenidade apresentou 73,69% das amostras corretamente classificadas, entretanto, há uma proporção de amostras classificadas incorretamente como tristeza (17,10%). Desse modo, com base nesses resultados, assim como para a base dois, há indicativos de dificuldades em distinguir as emoções alegria e raiva/medo, assim como as emoções serenidade e tristeza. A comparação gráfica entre as arquiteturas de *deep learning* considerando as abordagens desenhadas para a base um é apresentada no APÊNDICE T.

Passando para os algoritmos de aprendizado de máquina, na base dois, o melhor resultado foi obtido para a modalidade áudio, cujo algoritmo *CatBoost* atingiu 62,91% de acerto ao se considerar todos os atributos extraídos para essa modalidade. Além desse cenário, a modalidade áudio revelou-se aquela com as melhores acurácias, com os algoritmos apresentando taxas de acerto acima de 50,00%.

Em contrapartida, as modalidades acordes e letras apresentaram todos os cenários com acurácias abaixo dos 50,00%, com exceção para a classificação das letras considerando a seleção de atributos (50,28%). Para as modalidades restantes – simbólicas e voz – a acurácia residiu em cenários próximos a 50,00% de acerto. Por

fim, a síntese da classificação das emoções para os melhores algoritmos em cada modalidade, considerando a base dois, é apresentada na TABELA 22.

TABELA 22 - DESEMPENHO POR EMOÇÃO - ALGORITMOS DE APRENDIZADO DE MÁQUINA (BASE DOIS)

Modelo	Acurácia	Estratégia	Alegria	Raiva/medo	Serenidade	Tristeza
SVM (Acordes)	42,10%	HI	45,31%	18,00%	18,81%	72,93%
CatBoost (Áudio)	62,91%	TA	76,70%	44,67%	54,79%	65,92%
XGBoost (Letras)	50,28%	SA	64,08%	34,67%	39,93%	54,14%
CatBoost (Simbólicas)	56,13%	HI	76,70%	44,67%	54,79%	65,92%
CatBoost (Voz)	60,40%	TA	74,43%	55,33%	47,85%	61,15%

FONTE: O autor (2023).

Com base nos resultados apresentados na TABELA 22, observa-se que, para os acordes, a emoção tristeza apresentou a maior taxa de acerto (72,93%), enquanto as demais emoções ficaram abaixo dos 50,00%. Ao analisar o desempenho considerando o áudio, alegria, com 76,70% de registros corretamente classificados, seguida pela emoção tristeza (65,92%). Por seu turno, raiva/medo e serenidade apresentaram taxas de acerto intermediárias, com 44,67% e 54,79%, respectivamente.

De maneira semelhante, ao empregar as letras como modalidade, alegria e tristeza se sobressaíram, com taxas de acerto de 64,08% e 54,14%, respectivamente. Para as características simbólicas, cuja acurácia alcançou 56,13%, alegria teve a maior taxa de acerto (76,70%), seguida pela classe Tristeza (65,92%). Finalmente, para a voz, novamente alegria (74,43%) e tristeza (61,15%) apresentaram as maiores taxas de acerto. As matrizes de confusão geradas para cada um desses cenários estão elencadas no APÊNDICE U.

Adentrando nos resultados alcançados pelos algoritmos de aprendizado de máquina, na base um, o melhor resultado novamente consistiu na modalidade áudio, com algoritmo *CatBoost* apresentando 79,32% de acerto ao se considerar o acréscimo de registros (*OVER*). Este valor, além do melhor resultado para os algoritmos de aprendizado de máquina, apresenta-se como o resultado mais próximo dos obtidos com as arquiteturas de *deep learning*, colocando-se como um dos melhores deste estudo.

As modalidades acordes e letras, que na base dois apresentaram acurácia abaixo dos 50,00% apresentaram leve melhora, contudo, ainda distante das demais modalidades, principalmente do áudio. Outro ponto a se destacar consiste que, para as modalidades voz e simbólicas, o uso de *data augmentation* resultou nos dois

melhores resultados para essas modalidades ao se considerar os algoritmos de aprendizado de máquina.

Finalmente, a síntese da classificação das emoções para os melhores algoritmos em cada modalidade, considerando a base um, é apresentada na TABELA 23.

TABELA 23 - DESEMPENHO POR EMOÇÃO - ALGORITMOS DE APRENDIZADO DE MÁQUINA (BASE UM)

Modelo	Acurácia	Estratégia	Alegria	Raiva/medo	Serenidade	Tristeza
Random Forest (Acordes)	52,97%	*OVER*	56,76%	55,74%	12,99%	76,92%
CatBoost (Áudio)	80,17%	*OVER*	87,39%	80,33%	64,94%	83,65%
SVM (Letras)	54,67%	HI	72,07%	42,37%	1,32%	81,31%
SVM (Simbólicas)	69,68%	Aumentado	80,18%	49,15%	50,00%	84,11%
LigthGBM (Voz)	75,64%	Aumentado	87,39%	79,66%	48,69%	80,37%

FONTE: O autor (2023).

A partir da TABELA 23, visualiza-se, para os acordes, a correta classificação de 76,92% dos registros pertencentes à tristeza; ao mesmo tempo, é perceptível o baixo acerto das instâncias de serenidade (12,99%). Ao analisar os resultados para o áudio, nota-se que todas as emoções tiveram altas taxas de acerto, com destaque para alegria (87,39%) e tristeza (83,65%). De modo semelhante ao observado para os acordes, ao ter como modalidade as letras das canções, a emoção serenidade apresentou baixo percentual de acerto, com apenas 1,32%. Tal desempenho está muito aquém do obtido para as emoções tristeza (81,31%) e alegria (72,07%).

Para as características simbólicas, as emoções tristeza e alegria apresentaram desempenho semelhante, com 84,11% e 80,18% de acerto, respectivamente. Por seu turno, para a modalidade voz, apenas a emoção serenidade ficou abaixo dos 75,00% de acerto, apresentando apenas 48,69% de registros corretamente classificados. As matrizes de confusão geradas para cada um desses cenários estão elencadas no APÊNDICE V.

Finalmente, comparando os resultados obtidos com as arquiteturas de *deep learning* e com os algoritmos de aprendizado de máquina, tem-se que, dos seis melhores cenários (TABELA 24), todos foram obtidos diretamente usando a base um e, excetuando-se o algoritmo *CatBoost* (80,17%), os maiores valores de acurácia pertencem às arquiteturas de *deep learning*.

Embora dominantes, as arquiteturas de *deep learning* não se distanciaram tanto desse algoritmo, por exemplo, com a diferença residindo em cerca de um ponto

percentual, considerando o melhor resultado de ambas as abordagens, sendo 81,02% para a arquitetura *FT-Transformer* e 80,17% para o algoritmo *CatBoost*.

TABELA 24 - MELHORES RESULTADOS OBTIDOS NESTA TESE

Arquitetura/Modelo	Abordagem/Modalidade	Acurácia
FT-Transformer	TA/HI	81,02%
TabNet	TA/HI	81,01%
TabNet	Aumentado	80,73%
CatBoost (Áudio)	*OVER*	80,17%
FT-Transformer - NODE	OVER - TA/HI	79,89%

FONTE: O autor (2023).

Para complementar a síntese realizada ao longo desta seção, optou-se por comparar esta tese frente a outras pesquisas multimodais que recorreram à classificação como tarefa principal. O comparativo é apresentado no QUADRO 40.

QUADRO 40 - COMPARAÇÃO ENTRE PESQUISAS MULTIMODAIS

Autores	Modalidades	Emoções	Quantidade de Músicas	Principal Método	Principal Resultado
Esta tese (MOREIRA, 2023)	Acordes, áudio, letras, simbólicas e voz	Alegria, raiva/medo, serenidade e tristeza	1.412	*FT-Transformer*	81,02% (acurácia)
Lu *et al*. (2010)	Áudio, letras e MIDI.	Feliz, triste, raiva e relaxante	500	AdaBoost	72,40% (acurácia)
Zhao, Yang e Chen (2010)	Áudio, letras e MIDI.	Alegria, contentamento, raiva e depressão	981	SVM	61,60% (acurácia)
Panda *et al.* (2013)	Áudio, letras e MIDI.	Cinco *clusters* (MIREX)	193	SVM	61,20% (F1-*score*)
Przybysz (2016)	Áudio, letras e cifras.	Positiva, neutra e negativa.	784	SVM	60,60% (F1-*score*)

FONTE: O autor (2023).

Como observa-se no QUADRO 40, mesmo que com objetos diferentes (métricas, músicas, estratégias), o melhor resultado desta tese apresenta-se como o maior resultado ao ser comparado com pesquisas correlatas, apresentando, aproximadamente, nove pontos percentuais a mais do que a melhor taxa de acurácia observada. Para mais, o presente estudo contempla mais músicas (registros) e mais modalidades empregadas em conjunto, além de ser o único com a utilização de arquiteturas de *deep learning* como método de classificação.

Ademais, concentrando-se nos resultados obtidos neste estudo, ao se comparar as duas bases desenhadas e analisadas, as arquiteturas de *deep learning,* ao serem aplicadas à base de dados de concordância obtiveram resultados superiores em praticamente todos os cenários elaborados, indicando que as arquiteturas se beneficiarem da consistência entre as anotações dos juízes/avaliadores. De modo semelhante, os algoritmos de aprendizado de máquina também se beneficiaram dessa consistência.

É imprescindível destacar que, como um desdobramento da avaliação conduzida no experimento ajustado, uma etapa subsequente de pós-avaliação foi empreendida, facultando aos cinco voluntários a expressão de suas percepções acerca do experimento. Durante esse processo, os avaliadores foram indagados sobre quais elementos musicais específicos exerceram influência sobre suas respostas. Em relação a esse questionamento, o ritmo emergiu como o elemento preponderante, sendo mencionado de forma unânime por todos os participantes (5/5), seguido pela melodia (4/5). A harmonia e a voz também se sobressaíram, com três dos cinco participantes enfatizando sua relevância. Em contrapartida, letras e instrumentos receberam menor destaque, obtendo duas menções cada, conforme relatado pelos próprios voluntários.

Além dos elementos previamente citados, os avaliadores proporcionaram uma descrição pormenorizada das combinações emocionais (por exemplo: raiva/medo x serenidade; alegria x serenidade, etc.) que consideraram mais desafiadoras de discernir. Para esse questionamento, a combinação raiva/medo x tristeza e a combinação serenidade x tristeza foram identificadas como as mais complexas, recebendo três seleções cada. Notavelmente, a combinação tristeza x alegria não foi apontada como problemática.

Ainda que fundamentados em uma amostra limitada, os resultados oriundos do processo de pós-avaliação ecoam paralelos com estudos anteriores publicados. Eles corroboram a dificuldade em diferenciar entre composições musicais serenas e tristes (Hong; Chau; Horner, 2017; Taruffi *et al*., 2017), fornecendo indicações para a compreensão dos resultados de classificação relatados neste trabalho. Ademais, oferecem direções potenciais para futuras explorações a fim de evitar confusões entre determinadas emoções e também orientam a atenção dos pesquisadores para certas características e modalidades específicas. A literatura destaca, por exemplo, o ritmo (Schellenberg; Krysciak; Campbell, 2000; Trost; Labbé; Grandjean, 2017) e o

andamento (Fernández-Sotos; Fernández-Caballero; Latorre, 2016) como elementos de relevância.

Finalizado o relato dos resultados obtidos, tanto para as arquiteturas de *deep learning*, quanto para os algoritmos de aprendizado de máquina, a seguir são apresentadas as publicações produzidas no decorrer deste estudo.

4.8 PUBLICAÇÕES

No decorrer da pesquisa, até o presente momento, quatro estudos foram publicados, nos quais é possível observar elementos que dialogam com o tema desta tese. Os estudos em questão consistem em: a) artigo em periódico internacional, com classificação A1 no quadriênio 2017-2020 (1); artigo em periódico nacional, avaliado como B2 no quadriênio 2017-2020 (2); c) um capítulo de livro (3); e d) artigo em periódico nacional, avaliado como A2 no quadriênio 2017-2020 (4).

A relação dos estudos é apresentada abaixo:

1. MOREIRA, P. S. C.; TSUNODA, D. F. Recognition of emotions in music through the Adaptive-Network-Based Fuzzy (ANFIS). **Journal of New Music Research**, v. 50, n. 4, p. 342-354, 2021. Disponível em: https://doi.org/10.1080/09298215.2021.1977339. Acesso em: 01 fev. 2022.

2. MOREIRA, P. S. C.; TSUNODA, D. F. Proposta de base de dados para classificação automática de gêneros musicais brasileiros. **Revista Brasileira de Educação em Ciência da Informação**, v. 8, p. 1-13, 2021. Disponível em: https://portal.abecin.org.br/rebecin/article/view/234/212. Acesso em: 01 fev. 2022.

3. MOREIRA, P. S. C.; TSUNODA, D. F. Recuperação de informação musical: panorama das pesquisas brasileiras sobre o tema. *In*: ZALAFON, Z. R. *et al*. (Org.). **Percursos de pesquisa em Ciência da Informação: produção científica e discussões teórico-conceituais**. São Paulo: Abecin Editora, 2021, Cap. 13, p. 195-219.

4. MOREIRA, P. S. C.; TSUNODA, D. F. Música e emoções: um estudo altmétrico da produção científica de 1970 a 2019. **Em Questão**, v. 28, n. 1, p. 209–233, jan/mar. 2022. Disponível em: https://www.seer.ufrgs.br/EmQuestao/article/view/111905. Acesso em: 18 mar. 2023.

Além das pesquisas já publicadas, um artigo (5) encontra-se aceito para publicação em periódico nacional, avaliado como A3 no quadriênio 2017-2020, com previsão de publicação posterior à defesa desta tese.

5. NUNES-SILVA, M.; MOREIRA, P. S. C.; EYER, P. Modelos cognitivos de emoções musicais. **PERCEPTA - REVISTA DE COGNIÇÃO MUSICAL**, 2023.

Com a apresentação das publicações associadas à tese, encerram-se a apresentação e a discussão dos resultados e, na sequência, as considerações finais desta pesquisa são relatadas.

5 CONSIDERAÇÕES FINAIS

Esta pesquisa foi conduzida com o entendimento de que a relação entre música e emoção desempenha um papel importante na vida humana, destacando-se em atividades que vão desde o lazer até estudos científicos. Além disso, verificou-se que o impacto emocional das músicas é relevante para a organização do conteúdo musical em serviços de *streaming*, rádios digitais e serviços de recomendação, por ser a música um elemento onipresente na sociedade.

Embora a organização de músicas do ponto de vista emocional seja natural para os seres humanos, o reconhecimento automático de emoções em músicas ainda enfrenta desafios significativos, tornando-se um tema complexo na área de recuperação de informação musical.

Diante deste contexto, o objetivo geral desta tese foi investigar se a adoção de abordagens multimodais, utilizando informações de diversas fontes e arquiteturas de *deep learning*, é capaz de superar o desempenho de classificação obtido por abordagens unimodais, construídas com o emprego de algoritmos de aprendizado de máquina. Esse objetivo foi motivado pela carência de estratégias multimodais na área e pelas evidências de melhoria dos resultados de classificação relatados em pesquisas correlatas.

Para alcançar o objetivo proposto e responder à questão norteadora da pesquisa - para o reconhecimento de emoções em músicas, abordagens multimodais, somadas ao uso de arquiteturas de *deep learning*, são capazes de superar o desempenho de classificação de abordagens unimodais e construídas a partir de algoritmos de aprendizado de máquina? - foram estabelecidos cinco objetivos específicos.

O primeiro objetivo específico buscou identificar um modelo cognitivo de emoções a ser utilizado, sendo selecionado o modelo proposto por Bigand *et al*. (2005), que combina abordagens emocionais discretas e quadrantes com as dimensões excitabilidade e valência. Desse modo, o modelo adotado adequa-se, por exemplo, ao modelo circumplexo proposto por Russell (1980), além de apresentar uma relação de antagonismo entre as emoções: a) alegria (valência positiva/excitabilidade alta) x tristeza (valência negativa/excitabilidade baixa); e b) raiva/medo (valência negativa/excitabilidade alta) x serenidade (valência positiva/excitabilidade baixa). Por fim, a partir do trabalho de Nunes-Silva *et al*. (2016),

foi possível adotar significados para as quatro emoções do modelo proposto, auxiliando o processo de rotulação conduzido por voluntários.

O segundo objetivo específico da pesquisa concentrou-se na identificação das modalidades a serem usadas no reconhecimento de emoções em músicas, incluindo a extração de características de cinco modalidades diferentes: acordes, áudio, letras, simbólicas e voz. A combinação dessas modalidades, que não havia sido explorada em estudos anteriores, nos permitiu obter resultados abrangentes, o que ajuda a avançar o conhecimento na área. Esses resultados forneceram uma visão mais completa do processo de reconhecimento de emoções em músicas, o que pode ser útil para futuras pesquisas e desenvolvimentos nesta área.

O terceiro objetivo específico envolveu a construção de duas bases de dados multimodais, contendo segmentos musicais com duração de 30 segundos e letras em inglês correspondentes a essa duração, abrangendo diferentes gêneros musicais e artistas dos séculos XX e XXI. A construção dessas bases foi conduzida com o auxílio de voluntários em experimentos piloto e ajustado, garantindo a significância estatística dos resultados de avaliação/revisão. Ao final, as bases desenvolvidas apresentam 1.412 e 4.303 registros, respectivamente.

O quarto objetivo específico consistiu na comparação de diferentes arquiteturas de *deep learning* aplicadas às bases de dados elaboradas, considerando diversos cenários de classificação. Seis arquiteturas projetadas para problemas tabulares foram empregadas, possibilitando uma análise abrangente do desempenho de cada uma delas. Durante os experimentos, as arquiteturas *TabNet* e *FT-Transformer* destacaram-se, uma vez que, a primeira arquitetura apresentou-se como a arquitetura que obteve os melhores resultados em quatro dos cinco cenários de classificação envolvendo a base dois, enquanto a segunda, ao ser aplicada à base um, alcançou 81,02% de acerto, fornecendo o melhor resultado para este estudo.

Por fim, o quinto objetivo específico visou comparar o desempenho das abordagens multimodais desenvolvidas com o desempenho das abordagens unimodais construídas com algoritmos de aprendizado de máquina. Foram utilizados seis algoritmos de aprendizado de máquina, e os resultados obtidos com as abordagens multimodais mostraram-se superiores em praticamente todos os cenários de classificação elaborados em comparação com as estratégias unimodais.

Como resumo dos resultados, as melhores taxas de acerto obtidas para esta pesquisa correspondem aos seguintes valores: 1) *FT-Transformer* (TA/HI) - 81,02%;

2) *TabNet* (TA/HI) - 81,01%; 3) *TabNet* (Aumentado) - 80,73%; 4) *CatBoost* (Áudio - *OVER*) - 80,17%; 5) *FT-Transformer* (*OVER*) - 79,89%; e 6) *NODE* (TA/HI) - 79,89%. Destaca-se que todos os melhores resultados foram alcançados ao se aplicar os métodos à base um, composta pelo consenso obtido junto aos avaliadores.

Desta forma, a pesquisa respondeu positivamente à questão norteadora, constatando que, considerando as bases, características e algoritmos e arquiteturas empregados, a combinação de múltiplas fontes de informação (multimodalidade) e arquiteturas de *deep learning* foi capaz de proporcionar resultados superiores em reconhecimento de emoções em músicas, quando comparada às estratégias unimodais.

Além disso, ao contrapor os resultados obtidos com trabalhos multimodais já publicados, verificou-se que a pesquisa alcançou resultados notáveis, apresentando-se como uma das primeiras a realizar uma comparação entre arquiteturas de *deep learning* e algoritmos de aprendizado de máquina para o reconhecimento de emoções em músicas.

No entanto, a pesquisa também identificou pontos que requerem atenção dos pesquisadores da área. Destacou-se a importância de direcionar esforços para a definição de modelos emocionais que capturem as sutilezas das emoções associadas à música, bem como a cautela ao utilizar metadados e informações de plataformas *online* como fonte de dados rotulados, a fim de evitar vieses e ruídos indesejados.

5.1 CONTRIBUIÇÕES

Esta tese apresenta importantes contribuições para a área de reconhecimento de emoções em músicas, destacando-se nos seguintes tópicos:

a) **bases de dados multimodais:** a construção de duas bases de dados multimodais representando segmentos musicais com duração de 30 segundos, contendo informações de diversas fontes, foi uma das principais contribuições desta pesquisa. As bases denominadas como base um e base dois abrangem diferentes gêneros musicais e artistas, proporcionando um conjunto de dados abrangente e representativo para o estudo do reconhecimento de emoções em músicas;

b) **arquiteturas de *deep learning* para dados tabulares:** outra contribuição desta tese é a avaliação de seis arquiteturas de *deep learning* projetadas para problemas com dados tabulares. A aplicação dessas arquiteturas às bases de dados multimodais possibilitou a análise comparativa de seu desempenho no reconhecimento de emoções em músicas, proporcionando indícios sobre a eficácia dessas abordagens em cenários musicais específicos. Além do mais, essas arquiteturas conseguiram desempenho superiores a tradicionais algoritmos de aprendizado de máquina, possibilitando que novas pesquisas possam explorar ainda mais o potencial dessas arquiteturas em outros campos de pesquisa;

c) **protocolo de experimentos:** a elaboração de um protocolo de experimentos abrangente permitiu uma investigação sistemática das abordagens multimodais e unimodais no reconhecimento de emoções em músicas. Com esse protocolo, foi possível comparar os resultados obtidos em diferentes cenários de classificação, considerando, por exemplo, técnicas de seleção de atributos, além de *data augmentation* específico para áudio;

d) **contribuição para a área de cognição musical:** a pesquisa contribui para reforçar a necessidade de se compreender a relação entre música e emoção antes da aplicação de recursos computacionais. Tópicos como definição de modelos “adequados”, escolha de músicas, determinação da duração e de qual parte da música selecionar para analisar, por exemplo, abordados nesta tese, podem ser estudados com maiores detalhes em áreas cuja a relação entre música e emoção seja o objeto principal de pesquisa. Com os procedimentos metodológicos desenvolvidos ao longo da pesquisa, pesquisadores da área de cognição musical podem, por exemplo, adaptar as características extraídas das modalidades para direcionarem as suas pesquisas, envolvendo experimentos em laboratórios com a participação de voluntários, coleta de indicadores relacionados à frequência cardíaca, à pressão arterial, à frequência respiratória e às expressões faciais, que não foram contemplados neste estudo;

e) **publicações e divulgação científica:** elementos associados à tese já foram publicados em artigos científicos e capítulo de livro, além do mais, como próximos passos, novas publicações, contemplando os resultados e as bases de dados, estão previstas, visando contribuir para o avanço do conhecimento

no campo do reconhecimento de emoções em músicas e incentivar a continuidade de pesquisas nessa temática.

Com relação aos trabalhos para a continuação da temática abordada nesta tese, a próxima seção contempla os trabalhos futuros.

5.2 TRABALHOS FUTUROS

Com base nas descobertas e contribuições apresentadas nesta tese, surgem diversas possibilidades para trabalhos futuros que podem aprofundar e expandir o conhecimento no campo do reconhecimento de emoções em músicas. Algumas sugestões de temas promissores para futuras pesquisas incluem:

a) **uso de *deep learning* para a extração de características:** uma abordagem interessante para futuros estudos seria explorar o uso de técnicas avançadas de *deep learning* para a extração de características das modalidades presentes nas bases de dados multimodais. O desenvolvimento de modelos mais complexos e específicos para cada modalidade pode proporcionar representações mais ricas e detalhadas das informações contidas na música, levando a melhorias significativas no reconhecimento de emoções;
b) **trabalho com um modelo dinâmico de emoções:** considerando a natureza dinâmica e fluida das emoções em músicas, trabalhos futuros podem investigar a implementação de modelos que levem em conta a variação temporal das emoções ao longo da música. Modelos dinâmicos podem capturar nuances e mudanças nas emoções de forma mais precisa, resultando em classificações mais acuradas e realistas;
c) **testar diferentes durações e partes da música:** outro aspecto a ser explorado é a análise das variações no reconhecimento de emoções ao utilizar diferentes durações e partes da música como entrada para os modelos de classificação. Investigar se diferentes trechos ou segmentos musicais influenciam o desempenho do reconhecimento de

emoções pode fornecer *insights* valiosos sobre a importância de determinadas partes da música na evocação emocional;

d) **abordagem baseada em música brasileira e letras em Português:** expandir a pesquisa para incluir uma abordagem específica com músicas brasileiras e letras em português pode proporcionar uma análise mais aprofundada e contextualizada para o público brasileiro. Considerar aspectos culturais e características únicas da música brasileira pode melhorar a precisão do reconhecimento de emoções para esse contexto específico;
e) **implementação da abordagem em ambientes reais:** finalmente, trabalhos futuros podem se concentrar na implementação prática das abordagens propostas em ambientes reais, como serviços de *streaming* de música ou aplicações de recomendação. Avaliar a eficácia e aplicabilidade da abordagem em cenários do mundo real pode fornecer uma visão mais abrangente sobre o impacto dessa tecnologia em serviços musicais e a experiência do usuário.

O campo do reconhecimento de emoções em músicas ainda apresenta diversas oportunidades para pesquisas e inovações. Com a adoção contínua de abordagens multimodais com *deep learning* e a exploração de novas técnicas e cenários, espera-se que a interação entre música e emoção seja cada vez mais aprimorada, proporcionando experiências musicais mais personalizadas e significativas para os ouvintes.

REFERÊNCIAS

ACHEAMPONG, F. A.; WENYU, C.; NUNOO-MENSAH, H. Text-based emotion detection: advances, challenges, and opportunities. **Engineering Reports**, v. 2, n. 7, p. 1-24, maio 2020. Disponível em: https://doi.org/10.1002/eng2.12189. Acesso em: 17 abr. 2021.

ADOLPHS, R.; ANDERSON, D. J. **The neuroscience of emotion:** a new synthesis, Princeton: Princeton University Press, 2018.

AHANIN, Z.; ISMAIL, M. A. A multi-label emoji classification method using balanced pointwise mutual information-based feature selection. **Computer Speech & Language**, v. 73, p. 1-20, 2022. Disponível em: https://doi.org/10.1016/j.csl.2021.101330. Acesso em: 17 dez. 2021.

AKÇAY, M. B.; OĞUZ, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. **Speech Communication**, v. 116, p. 56-76, 2020.

AKKUS, Z. *et al*. Deep learning for brain MRI segmentation: state of the art and future directions. **Journal of Digital Imaging**, v. 30, n. 4, p. 449-459, jun. 2017. Disponível em: https://doi.org/10.1007/s10278-017-9983-4. Acesso em: 04 jun. 2021.

ALDRIGHI, J. D. *et al*. The experiences of pregnant women at an advanced maternal age: an integrative review. **Revista da Escola de Enfermagem da USP**, v. 50, p. 0512-0521, maio-jun. 2016. Disponível em: https://doi.org/10.1590/S0080-623420160000400019. Acesso em: 27 jul. 2021.

ALDWEESH, A.; DERHAB, A.; EMAM, A. Z. Deep learning approaches for anomaly-based intrusion detection systems: a survey, taxonomy, and open issues. **Knowledge-Based Systems**, v. 189, p. 1-19, fev. 2020. Disponível em: https://doi.org/10.1016/j.knosys.2019.105124. Acesso em: 26 maio 2021.

ALJANAKI, A. **Emotion in music: representation and computational modeling**. 2016. 149 f. Tese (Doutorado). Utrecht University, Utreque, 2016. Disponível em: https://dspace.library.uu.nl/handle/1874/339995. Acesso em: 15 dez. 2021.

ALJANAKI, A.; WIERING, F.; VELTKAMP, R. C. Computational modeling of induced emotion using GEMS. *In*: CONFERENCE OF THE INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL (ISMIR 2014), 15., 2014, Taipei. **Proceedings [...] .** Taipei: ISMIR, 2014. p. 373-378. Disponível em: https://zenodo.org/record/1415192#.Ybn5K-rMK00. Acesso em: 15 dez. 2021.

ALJANAKI, A.; WIERING, F.; VELTKAMP, R. C. Emotion based segmentation of musical audio. In: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONFERENCE (ISMIR 2015), 16., Málaga. **Proceedings [...] .** Málaga: ISMIR, 2015. p. 770-776. Disponível em: https://archives.ismir.net/ismir2015/paper/000265.pdf. Acesso em: 26 maio 2022.

ALJANAKI, A.; WIERING, F.; VELTKAMP, R. C. Studying emotion induced by music through a crowdsourcing game. **Information Processing & Management**, v. 52, n. 1, p. 115-128, jan. 2016. Disponível em: https://doi.org/10.1016/j.ipm.2015.03.004. Acesso em: 05 maio 2021.

ALJANAKI, A.; YANG, Y.; SOLEYMANI, M. Developing a benchmark for emotional analysis of music. **Plos One**, v. 12, n. 3, p.1-22, mar. 2017. Disponível em: https://doi.org/10.1371/journal.pone.0173392. Acesso em: 24 jul. 2020.

ALKHAYRAT, M.; ALJNIDI, M.; ALJOUMAA, K. A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. **Journal of Big Data**, v. 7, p. 1-23, fev. 2020. Disponível em: https://doi.org/10.1186/s40537-020-0286-0. Acesso em: 26 mar. 2023.

ALVI, R. H. *et al*. Deep learning approach on tabular data to predict early-onset neonatal sepsis. **Journal of Information and Telecommunication**, v. 5, n. 2, p. 226-246, 2021.Disponível em: https://doi.org/10.1080/24751839.2020.1843121. Acesso em: 11 maio 2022.

AN, S. *et al*. Two sides of emotion: exploring positivity and negativity in six basic emotions across cultures. **Frontiers in Psychology**, v. 8, p. 1-14, abr. 2017. Disponível em: https://doi.org/10.3389/fpsyg.2017.00610. Acesso em: 06 abr. 2021.

ANDERSSON, P. K. *et al*. Let the music play or not: the influence of background music on consumer behavior. **Journal of Retailing and Consumer Services**, v. 19, n. 6, p. 553-560, nov. 2012. Disponível em: https://doi.org/10.1016/j.jretconser.2012.06.010.

ANDRADE, N. C. **Nas partituras das emoções**: processamento de estímulos afetivos musicais e visuais em crianças e adolescentes com síndrome de Williams. 2017. 174 f. Tese (Doutorado) - Curso de Psicologia, Universidade de São Paulo, São Paulo, 2017. Disponível em: https://teses.usp.br/teses/disponiveis/47/47132/tde-18042018-160210/pt-br.php. Acesso em: 23 jul. 2020.

ARGSTATTER, H. Perception of basic emotions in music: Culture-specific or multicultural?. **Psychology of Music**, v. 44, n. 4, p. 674-690, jul. 2016. Disponível em: https://doi.org/10.1177%2F0305735615589214. Acesso em: 09 ago. 2021.

ARIK, S. O.; PFISTER, T. Tabnet: Attentive interpretable tabular learning. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 35, n.8, maio 2021. p. 6679-6687. Disponível em: https://ojs.aaai.org/index.php/AAAI/article/view/16826. Acesso em: 17 fev. 2022.

ARNOLD, M. B. **Emotion and personality**, Nova Iorque: Columbia University Press, 1960.

AYZEL, G.; HEISTERMANN, M. The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: a case study for six basins from the CAMELS dataset. **Computers & Geosciences**, v. 149,

p. 1-12, abr. 2021. Disponível em: https://doi.org/10.1016/j.cageo.2021.104708. Acesso em: 16 jun. 2021.

BACHORIK, J. P. *et al*. Emotion in motion: investigating the time-course of emotional judgments of musical stimuli. **Music Perception**, v. 26, n. 4, p. 355-364, 2009. Disponível em: https://doi.org/10.1525/mp.2009.26.4.355. Acesso em: 26 maio 2022.

BAI, J. *et al*. Music emotions recognition by cognitive classification methodologies. *In*: INTERNATIONAL CONFERENCE ON COGNITIVE INFORMATICS & COGNITIVE COMPUTING (ICCI*CC), 16., 2017, Oxford. **Proceedings[...] .** Oxford: IEEE, 2017. p. 121-129. Disponível em: https://doi.org/10.1109/ICCI-CC.2017.8109740. Acesso em: 23 maio 2021.

BAIRD, A.; SAMSON, S. Music and dementia. In: ALTENMÜLLER, E.; FINGER, S.; BOLLER, F. (Ed.). **Progress in Brain Research**. Amsterdã: Elsevier, 2015. Cap. 11. p. 207-235.

BALASUBRAMANIAN, G. *et al*. Music induced emotion using wavelet packet decomposition—An EEG study. **Biomedical Signal Processing and Control**, v. 42, n. 1, p. 115-128, abr. 2018. Disponível em: https://doi.org/10.1016/j.bspc.2018.01.015. Acesso em: 24 jul. 2020.

BALKWILL, L.L.; THOMPSON, W. F. A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. **Music Perception**, v. 17, n. 1, p. 43-64, out. 1999. Disponível em: https://doi.org/10.2307/40285811. Acesso em: 25 ago. 2021.

BALTRUŠAITIS, T.; AHUJA, C.; MORENCY, L. Multimodal machine learning: a survey and taxonomy. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 41, n. 2, p. 423-443, jan. 2018. Disponível em: https://ieeexplore.ieee.org/document/8269806. Acesso em: 17 dez. 2020.

BARRETT, L. F. Are emotions natural kinds? **Perspectives on Psychological Science**, v. 1, n. 1, p. 28-58, mar. 2006. Disponível em: https://doi.org/10.1111%2Fj.1745-6916.2006.00003.x. Acesso em: 06 abr. 2021.

BARRETT, L. F. Was Darwin wrong about emotional expressions? **Current Directions in Psychological Science**, v. 20, n. 6, p. 400-406, dez. 2011. Disponível em: https://doi.org/10.1177%2F0963721411429125. Acesso em: 15 fev. 2021.

BARRETT, L. F. **How emotions are made:** the secret life of the brain, Nova Iorque: Houghton Mifflin Harcourt, 2017.

BAYLE, Y.; ROBINE, M.; HANNA, P. SATIN: A persistent musical database for music information retrieval and a supporting deep learning experiment on song instrumental classification. **Multimedia Tools and Applications**, v. 78, n. 3, p. 2703-2718, fev. 2018. Disponível em: https://doi.org/10.1007/s11042-018-5797-8. Acesso em: 24 set. 2020.

BHASKAR, J.; SRUTHI, K.; NEDUNGADI, P. Hybrid approach for emotion classification of audio conversation based on text and speech mining. **Procedia Computer Science**, v. 46, p. 635-643, 2015. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050915001763. Acesso em: 23 jul. 2020.

BHATTI, A. M. *et al*. Human emotion recognition and analysis in response to audio music using brain signals. **Computers in Human Behavior**, v. 65, p. 267-275, dez. 2016. Disponível em: https://doi.org/10.1016/j.chb.2016.08.029. Acesso em: 04 set. 2020.

BENETOS, E. *et al*. Automatic music transcription: An overview. **IEEE Signal Processing Magazine**, v. 36, n. 1, p. 20-30, jan. 2019. Disponível em: https://doi.org/10.1109/MSP.2018.2869928. Acesso em: 03 jan. 2022.

BIGAND, E. *et al*. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. **Cognition & Emotion**, v. 19, n. 8, p. 1113-1139, 2005. Disponível em: https://doi.org/10.1080/02699930500204250. Acesso em: 25 ago. 2021.

BIGAND, E.; FILIPIC, S.; LALITTE, P. The time course of emotional responses to music. **Annals of the New York Academy of Sciences**, v. 1060, n. 1, p. 429-437, 2005. Disponível em: https://doi.org/10.1196/annals.1360.036. Acesso em: 26 maio 2022.

BO, H. *et al*. Music-evoked emotion recognition based on cognitive principles inspired EEG temporal and spectral features. **International Journal of Machine Learning and Cybernetics**, v. 10, n. 9, p. 2439-2448, out. 2018. Disponível em: https://link.springer.com/article/10.1007/s13042-018-0880-z. Acesso em: 30 jul. 2020.

BOGDANOV, D. *et al*. ESSENTIA: an open-source library for sound and music analysis. *In*: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 21., 2013, Barcelona. **Proceedings[...] .** Barcelona: ACM, 2013. p. 855-858. Disponível em: https://doi.org/10.1145/2502081.2502229. Acesso em: 09 maio 2021.

BRADLEY, M. M. *et al*. Remembering pictures: pleasure and arousal in memory. **Journal of Experimental Psychology: Learning, Memory, and Cognition**, v. 18, n. 2, p. 379-390, mar. 1992. Disponível em: https://psycnet.apa.org/doi/10.1037/0278-7393.18.2.379. Acesso em: 12 abr. 2021.

BRADLEY, M. M. et al. Emotion and motivation I: defensive and appetitive reactions in picture processing. **Emotion**, v. 1, n. 3, p. 276-298, set. 2001. Disponível em: https://psycnet.apa.org/doi/10.1037/1528-3542.1.3.276. Acesso em: 15 abr. 2021.

BREIMAN, L. Random forests. **Machine learning**, v. 45, p. 5-32, out. 2001. Disponível em: https://doi.org/10.1023/A:1010933404324. Acesso em: 19 mar. 2022.

BRESIN, R.; FRIBERG, A. Emotion rendering in music: range and characteristic values of seven musical variables. **Cortex**, v. 47, n. 9, p. 1068-1081, out. 2011.

Disponível em: https://doi.org/10.1016/j.cortex.2011.05.009. Acesso em: 26 ago. 2021.

BRIOT, J.P.; PACHET, F. Deep learning for music generation: challenges and directions. **Neural Computing and Applications**, v. 32, n. 4, p. 981-993, out. 2018. Disponível em: https://doi.org/10.1007/s00521-018-3813-6. Acesso em: 28 set. 2020.

CABRERA, D.; FERGUSON, S.; SCHUBERT, E. 'PsySound3': software for acoustical and psychoacoustical analysis of sound recordings. *In*: INTERNATIONAL CONFERENCE ON AUDITORY DISPLAY, 13., 2007, Montreal. **Proceedings[...] .** Montreal: ICAD, 2007. p. 365-363. Disponível em: https://smartech.gatech.edu/handle/1853/49969. Acesso em: 11 maio 2021.

ÇANO, E.; MORISIO, M. Music mood dataset creation based on Last.fm tags. *In*: FOURTH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 4., 2017, Viena. **Proceedings[...] .** Viena: CSCP, 2017. p. 15-26. Disponível em: https://core.ac.uk/download/pdf/84251935.pdf. Acesso em: 06 maio 2021.

CATHARIN, L. G. *et al*. Multimodal classification of emotions in Latin Music. *In*: 2020 IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA (ISM), 22., 2020, Nápoles. **Proceedings[...] .** Nápoles: IEEE. p. 173-180. Disponível em: https://doi.org/10.1109/ISM.2020.00038. Acesso em: 12 maio 2021.

CELEGHIN, A. *et al*. Basic emotions in human neuroscience: neuroimaging and beyond. **Frontiers in Psychology**, v. 8, p. 1-13, ago. 2017. Disponível em: https://doi.org/10.3389/fpsyg.2017.01432. Acesso em: 27 jan. 2021.

CHATURVEDI, V. *et al*. Music mood and human emotion recognition based on physiological signals: a systematic review. **Multimedia Systems**, p. 1-24, abr. 2021. Disponível em: https://doi.org/10.1007/s00530-021-00786-6. Acesso em: 27 abr. 2021.

CHAUDHARY, D.; SINGH, N. P.; SINGH, S. Development of music emotion classification system using Convolution Neural Network. **International Journal of Speech Technology**, p. 1-10, 2020. Disponível em: https://doi.org/10.1007/s10772-020-09781-0. Acesso em: 23 maio 2021.

CHAWLA, N. V. *et al*. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 2002. Disponível em: https://www.jair.org/index.php/jair/article/view/10302. Acesso em: 26 mar. 2023.

CHEAH, K. H. *et al*. Convolutional neural networks for classification of music-listening EEG: comparing 1d convolutional kernels with 2d kernels and cerebral laterality of musical influence. **Neural Computing and Applications**, v. 32, n. 13, p. 8867-8891, jul. 2019. Disponível em: https://doi.org/10.1007/s00521-019-04166-0. Acesso em: 28 set. 2020.

CHEŁKOWSKA-ZACHAREWICZ, M.; JANOWSKI, M. Polish adaptation of the Geneva Emotional Music Scale: Factor structure and reliability. **Psychology of**

Music, p. 1-15, jul. 2020. Disponível em: https://doi.org/10.1177%2F0305735620927474. Acesso em: 18 maio 2021.

CHEN, S.H. *et al*. Music emotion recognition using deep Gaussian process. In: ASIA-PACIFIC SIGNAL AND INFORMATION PROCESSING ASSOCIATION ANNUAL SUMMIT AND CONFERENCE (APSIPA), 7., 2015, Hong Kong. **Proceedings[...] .** Hong Kong: IEEE, 2015a. p. 495-498. Disponível em: https://ieeexplore.ieee.org/document/7415321. Acesso em: 28 set. 2020.

CHEN, T.; GUESTRIN, C. XGBoost: a scalable tree boosting system. *In*: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, São Francisco. **Proceedings [...] .** São Francisco: ACM, 2016. p. 785-794. Disponível em: https://dl.acm.org/doi/abs/10.1145/2939672.2939785. Acesso em: 11 maio 2022.

CHEN, Y. A. *et al*. The AMG1608 dataset for music emotion recognition. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 40., 2015, South Brisbane. **Proceedings[...] .** South Brisbane: IEEE, 2015b. p. 63-68. Disponível em: https://doi.org/10.1109/ICASSP.2015.7178058. Acesso em: 07 maio 2021.

CHEN, Y. A. *et al*. Component tying for mixture model adaptation in personalization of music emotion recognition. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 25, n. 7, p. 1409-1420, abr. 2017. Disponível em: https://doi.org/10.1109/TASLP.2017.2693565. Acesso em: 31 ago. 2021.

CHENG, Z. *et al*. Exploring user-specific information in music retrieval. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 40., 2017, Tóquio. **Proceedings [...] .** Tóquio: ACM, 2017. p. 655-664. Disponível em: https://dl.acm.org/doi/10.1145/3077136.3080772. Acesso em: 10 ago. 2020.

CHICCO, D.; JURMAN, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. **BMC Genomics**, v. 21, n. 1, p. 1-13, jan. 2020. Disponível em:https://doi.org/10.1186/s12864-019-6413-7. Acesso em: 03 set. 2021.

CHO, H. K. The effects of music therapy-singing group on quality of life and affect of persons with dementia: a randomized controlled trial. **Frontiers in Medicine**, v. 5, p. 279, out. 2018. Disponível em: https://doi.org/10.3389/fmed.2018.00279. Acesso em: 26 jun. 2021.

CHO, Y. *et al*. Music emotion recognition using chord progressions. In: IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN, AND CYBERNETICS, 29., 2016, Budapeste. **Proceedings[...] .** Budapeste: IEEE, 2016. p. 002588 - 002593. Disponível em: https://ieeexplore.ieee.org/document/7844628. Acesso em: 24 jul. 2020.

CHOLLET, F. **Deep learning with Python**. Nova Iorque: Manning, 2018.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37-46, abr. 1960. Disponível em: https://doi.org/10.1177%2F001316446002000104. Acesso em: 26 jan. 2022.

COHEN-KALAF, M. *et al*. Movie emotion map: an interactive tool for exploring movies according to their emotional signature. **Multimedia Tools and Applications**, p. 1-22, abr. 2021. Disponível em: https://doi.org/10.1007/s11042-021-10803-5. Acesso em: 17 abr. 2021.

COHRDES, C.; GROLIG, L.; SCHROEDER, S. The development of music competencies in preschool children: effects of a training program and the role of environmental factors. **Psychology of Music**, v. 47, n. 3, p. 358-375, fev. 2018. Disponível em: https://doi.org/10.1177%2F0305735618756764. Acesso em: 10 ago. 2020.

COLOMBETTI, G. **The feeling body:** affective science meets the enactive mind. Cambridge: MIT Press, 2014.

CORNELIS, O. *et al*. Access to ethnic music: advances and perspectives in content-based music information retrieval. **Signal Processing**, v. 90, n. 4, p. 1008-1031, abr. 2010. Disponível em: https://doi.org/10.1016/j.sigpro.2009.06.020. Acesso em: 24 jul. 2020.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21-27, jan.1967. Disponível em: https://doi.org/10.1109/TIT.1967.1053964. Acesso em: 19 mar. 2023.

CRAYENCOUR, H. C.; CELLA, C. E. Learning, probability and logic: toward a unified approach for content-based music information retrieval. **Frontiers in Digital Humanities**, v. 6, p. 1-25, abr. 2019. Disponível em: https://doi.org/10.3389/fdigh.2019.00006. Acesso em: 10 ago. 2020.

CRESWELL, A. *et al*. Generative adversarial networks: an overview. **IEEE Signal Processing Magazine**, v. 35, n. 1, p. 53-65, jan. 2018. Disponível em: https://doi.org/10.1109/MSP.2017.2765202. Acesso em: 11 jun. 2021.

CRIVELLI, C.; FRIDLUND, A. J. Inside-out: from basic emotions theory to the behavioral ecology view. **Journal of Nonverbal Behavior**, v. 43, n. 2, p. 161-194, jan. 2019. Disponível em: https://doi.org/10.1007/s10919-019-00294-2. Acesso em: 19 fev. 2021.

CUTHBERT, M. S.; ARIZA, C. music21: a toolkit for computer-aided musicology and symbolic music data. *In*: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONFERENCE (ISMIR 2010), 11., 2010, Utrecht. **Proceedings [...] .** Utrecht: ISMIR, 2010. p. 637-642. Disponível em: https://dspace.mit.edu/handle/1721.1/84963. Acesso em: 04 jan. 2022.

DARWIN, C. R. **The expression of the emotions in man and animals**, Londres: John Murray, 1872.

DENG, J.; KWOK, Y. K. Large vocabulary automatic chord estimation using bidirectional long short-term memory recurrent neural network with even chance training. **Journal of New Music Research**, v. 47, n. 1, p. 53-67, 2018. Disponível em: https://doi.org/10.1080/09298215.2017.1367820. Acesso em: 14 dez. 2021.

DENG, J. J. *et al*. Emotional States Associated with Music. **ACM Transactions on Interactive Intelligent Systems**, v. 5, n. 1, p. 1-36, mar. 2015. Disponível em: https://dl.acm.org/doi/10.1145/2723575. Acesso em: 24 jul. 2020.

DENG, S. *et al*. Exploring user emotion in microblogs for music recommendation. **Expert Systems with Applications**, v. 42, n. 23, p. 9284-9293, dez. 2015. Disponível em: https://doi.org/10.1016/j.eswa.2015.08.029. Acesso em: 20 out. 2023.

DESAI, M.; SHAH, M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). **Clinical eHealth**, v. 4, p. 1-11, 2021. Disponível em: https://doi.org/10.1016/j.ceh.2020.11.002. Acesso em: 17 maio 2022.

DHILLON, A.; VERMA, G. K. Convolutional neural network: a review of models, methodologies and applications to object detection. **Progress in Artificial Intelligence**, v. 9, n. 2, p. 85-112, jun. 2020. Disponível em: https://doi.org/10.1007/s13748-019-00203-0. Acesso em: 08 jun. 2021.

DIAS, C.; CRUZ, J. F.; FONSECA, A. M. Emoções: passado, presente e futuro. **Psicologia**, v. 22, n. 2, p. 11-31, jul. 2008. Disponível em: http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S0874-20492008000200002&lng=pt&nrm=iso. Acesso em: 21 jan. 2021.

DIGIROLAMO, M. A.; RUSSELL, J. A. The emotion seen in a face can be a methodological artifact: the process of elimination hypothesis. **Emotion**, v. 17, n. 3, p. 538-546, 2017. Disponível em: https://psycnet.apa.org/doi/10.1037/emo0000247. Acesso em: 07 abr. 2021.

DIRIWÄCHTER, R. Remembering Wilhelm Wundt and the second Leipzig School of Psychology. **Human Arenas**, v. 4, n. 1, p. 5-19, jan. 2021. Disponível em: https://doi.org/10.1007/s42087-020-00158-y. Acesso em: 13 abr. 2021.

DONG, Y. *et al*. Bidirectional convolutional recurrent sparse network (BCRSN): an efficient model for music emotion recognition. **IEEE Transactions on Multimedia**, v. 21, n. 12, p. 3150-3163, maio 2019. Disponível em: https://doi.org/10.1109/TMM.2019.2918739. Acesso em: 24 set. 2020.

DUFOUR, I.; TZANETAKIS, G. Using circular models to improve music emotion recognition. **IEEE Transactions on Affective Computing**, dez. 2018. Disponível em: https://doi.org/10.1109/TAFFC.2018.2885744. Acesso em: 27 abr. 2021.

EEROLA, T.; VUOSKOSKI, J. K. A comparison of the discrete and dimensional models of emotion in music. **Psychology of Music**, v. 39, n. 1, p. 18-49, jan. 2011. Disponível em: https://doi.org/10.1177%2F0305735610362821. Acesso em: 27 ago. 2021.

EEROLA, T.; VUOSKOSKI, J. K. A review of music and emotion studies: approaches, emotion models, and stimuli. **Music Perception**: An Interdisciplinary **Journal**, v. 30, n. 3, p. 307-340, fev. 2013. Disponível em: https://doi.org/10.1525/mp.2012.30.3.307. Acesso em: 18 fev. 2021.

EKMAN, P. An argument for basic emotions. **Cognition and Emotion**, v. 6, n. 3-4, p. 169-200, 1992. Disponível em: https://doi.org/10.1080/02699939208411068. Acesso em: 06 abr. 2021.

EKMAN, P. What scientists who study emotion agree about. **Perspectives on Psychological Science**, v. 11, n. 1, p. 31-34, jan. 2016. Disponível em: https://doi.org/10.1177%2F1745691615596992. Acesso em: 06 ago. 2021.

EKMAN, P.; CORDARO, D. What is meant by calling emotions basic. **Emotion Review**, v. 3, n. 4, p. 364-370, set. 2011. Disponível em: https://doi.org/10.1177%2F1754073911410740. Acesso em: 06 abr. 2021.

EKMAN, P.; FRIESEN, W. V. Constants across cultures in the face and emotion. **Journal of Personality and Social Psychology**, v. 17, n. 2, p. 124-129, fev. 1971. Disponível em: https://psycnet.apa.org/doi/10.1037/h0030377 . Acesso em: 07 abr. 2021.

EKMAN, P., FRIESEN, W. V., ELLSWORTH, P. What emotion categories or dimensions can observers judge from facial behavior? *In*: EKMAN, P. **Emotion in the human face**. Nova Iorque: Cambridge University Press, 1982, p. 39-55.

ELBIR, A. *et al*. Short time Fourier transform based music genre classification. *In*: 2018 ELECTRIC ELECTRONICS, COMPUTER SCIENCE, BIOMEDICAL ENGINEERINGS' MEETING (EBBT), 4., 2018, Istambul. **Proceedings[...] .** Istambul: IEEE. p. 1-4. Disponível em: https://doi.org/10.1109/EBBT.2018.8391437. Acesso em: 14 maio 2021.

EL AYADI, M.; KAMEL, M. S.; KARRAY, F. Survey on speech emotion recognition: Features, classification schemes, and databases. **Pattern Recognition**, v. 44, n. 3, p. 572-587, mar. 2011. Disponível em: https://doi.org/10.1016/j.patcog.2010.09.020. Acesso em: 08 jan. 2022.

EYBEN, F.; WÖLLMER, M.; SCHULLER, B. Opensmile: the munich versatile and fast open-source audio feature extractor. *In*: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 18., 2010, Florença. **Proceedings[...] .** Florença: ACM, 2010. p. 1459-1462. Disponível em: https://doi.org/10.1145/1873951.1874246. Acesso em: 09 maio 2021.

FABIO, R. A. *et al*. Adult listening behaviour, music preferences and emotions in the mobile context. Does mobile context affect elicited emotions? **Cogent Engineering**, v. 6, n. 1, p. 1-2017, mar. 2019. Disponível em: https://doi.org/10.1080/23311916.2019.1597666. Acesso em: 09 ago. 2020.

FADLULLAH, Z. M. *et al*. State-of-the-art deep learning: evolving machine intelligence toward tomorrow’s intelligent network traffic control systems. **IEEE**

Communications Surveys & Tutorials, v. 19, n. 4, p. 2432-2455, maio 2017. Disponível em: https://doi.org/10.1109/COMST.2017.2707140. Acesso em: 04 jun. 2021.

FAN, J. *et al*. Ranking-Based emotion recognition for experimental music. In: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONFERENCE, 18., 2017, Suzhou. **Proceedings[...] .** Suzhou: ISMIR, 2017. p. 368 - 375. Disponível em: http://archives.ismir.net/ismir2017/paper/000217.pdf. Acesso em: 24 jul. 2020.

FANSELOW, M. S. Emotion, motivation and function. **Current Opinion in Behavioral Sciences**, v. 19, p. 105-109, fev. 2018. Disponível em: https://doi.org/10.1016/j.cobeha.2017.12.013. Acesso em: 29 jan. 2021.

FERREIRA, L. N.; WHITEHEAD, J. Learning to generate music with sentiment. *In*: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONFERENCE (ISMIR 2019), 20., 2019, Delft. **Proceedings [...] .** Delft: ISMIR, 2019, p. 384-390. Disponível em: https://archives.ismir.net/ismir2019/paper/000045.pdf. Acesso em: 20 out. 2023.

FERNANDES, J. *et al*. Parameters for vocal acoustic analysis-cured database. **Procedia Computer Science**, v. 164, p. 654-661, 2019. Disponível em: https://doi.org/10.1016/j.procs.2019.12.232. Acesso em: 12 jan. 2022.

FIELD, A.; MILES, J.; FIELD, Z. **Discovering statistics using R**. Londres: Sage, 2012.

FLEISS, J. L. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, v. 76, n. 5, p. 378, nov. 1971. Disponível em: https://psycnet.apa.org/doi/10.1037/h0031619. Acesso em: 24 out. 2022.

FONTAINE, J. R.J. *et al*. The world of emotions is not two-dimensional. **Psychological Science**, v. 18, n. 12, p. 1050-1057, dez. 2007. Disponível em: https://doi.org/10.1111%2Fj.1467-9280.2007.02024.x. Acesso em: 12 abr. 2021.

FRIEDMAN, J. H *et al*. Greedy function approximation: a gradient boosting machine. **Annals of Statistics**, v. 29, n. 5, p. 1189-1232, out. 2001. Disponível em: https://www.jstor.org/stable/2699986. Acesso em: 11 maio. 2022.

FRITZ, T. *et al*. Universal recognition of three basic emotions in music. **Current Biology**, v. 19, n. 7, p. 573-576, mar. 2009. Disponível em: https://doi.org/10.1016/j.cub.2009.02.058. Acesso em: 26 ago. 2021.

FUENTES, C.; HAGBERG, J.; KJELLBERG, H. Soundtracking: music listening practices in the digital age. **European Journal of Marketing**, v. 53, n. 3, p. 483-503, abr. 2019. Disponível em: http://dx.doi.org/10.1108/ejm-10-2017-0753. Acesso em: 09 ago. 2020.

GABRIELSSON, Alf; JUSLIN, Patrik N. Emotional expression in music performance: Between the performer's intention and the listener's experience. **Psychology of**

Music, v. 24, n. 1, p. 68-91, abr. 1996. Disponível em: https://doi.org/10.1177%2F0305735696241007. Acesso em: 25 ago. 2021.

GAO, J. *et al*. A survey on deep learning for multimodal data fusion. **Neural Computation**, v. 32, n. 5, p. 829-864, maio 2020. Disponível em: https://doi.org/10.1162/neco_a_01273. Acesso em: 04 jun. 2021.

GAO, S. *et al*. Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. **Journal of Hydrology**, v. 589, p. 1-12, out. 2020. Disponível em: https://doi.org/10.1016/j.jhydrol.2020.125188. Acesso em: 17 jun. 2021.

GARDNER, M. W.; DORLING, S. R. Artificial neural networks (the multilayer perceptron)— a review of applications in the atmospheric sciences. **Atmospheric Environment**, v. 32, n. 14-15, p. 2627-2636, ago. 1998. Disponível em: https://doi.org/10.1016/S1352-2310(97)00447-0. Acesso em: 17 maio 2022.

GARRIDO, S.; SCHUBERT, E. Music and people with tendencies to depression. **Music Perception: an interdisciplinary journal**, v. 32, n. 4, p. 313-321, abr. 2015. Disponível em: https://mp.ucpress.edu/content/32/4/313. Acesso em: 24 jul. 2020.

GINSEL, P.; VATOLKIN, I.; RUDOLPH, G. Analysis of structural complexity features for music genre recognition. *In*: 2020 IEEE CONGRESS ON EVOLUTIONARY COMPUTATION (CEC), 21., 2020, Glasgow. **Proceedings[...] .** Glasgow: IEEE, 2020. p. 1-8. Disponível em: https://doi.org/10.1109/CEC48606.2020.9185540. Acesso em: 15 dez. 2021.

GÓMEZ, E. Tonal description of polyphonic audio for music content processing. **INFORMS Journal on Computing**, v. 18, n. 3, p. 294-304, ago. 2006. Disponível em: https://doi.org/10.1287/ijoc.1040.0126. Acesso em: 21 dez. 2021.

GÓMEZ-CAÑÓN, J. S. *et al*. Music emotion recognition: toward new, robust standards in personalized and context-sensitive applications. **IEEE Signal Processing Magazine**, v. 38, n. 6, p. 106-114, out. 2021. Disponível em: https://doi.org/10.1109/MSP.2021.3106232. Acesso em: 07 nov. 2021.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016.

GORISHNIY, Y. *et al*. Revisiting deep learning models for tabular data. *In*: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS 2021), 35., 2021. **Proceedings [...] .** NeurIPS, 2021. p. 1-12. Disponível em: https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html. Acesso em: 11 maio 2022.

GRAY, J. A. **The neuropsychology of anxiety**. Oxford: Oxford University Press, 1982.

GREER, T. *et al*. Learning shared vector representations of lyrics and chords in

music. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 44., 2019, Brighton. **Proceedings[... .]** Brighton: IEEE, 2019. p. 3951 - 3955. Disponível em: https://ieeexplore.ieee.org/document/8683735. Acesso em: 24 jul. 2020.

GREKOW, J. Music emotion maps in the arousal-valence space. In: GREKOW, J. **From content-based music emotion recognition to emotion maps of musical pieces**. Cham: Springer, 2017. p. 95-106.

GREKOW, J. Audio features dedicated to the detection and tracking of arousal and valence in musical compositions. **Journal of Information and Telecommunication**, v. 2, n. 3, p. 322-333, abr. 2018. Disponível em: https://doi.org/10.1080/24751839.2018.1463749. Acesso em: 07 maio 2021.

GREKOW, J. Generating polyphonic symbolic emotional music in the style of Bach using convolutional conditional variational autoencoder. **IEEE Access**, ago. 2023. Disponível em: 10.1109/ACCESS.2023.3309639. Acesso em: 20 out. 2023.

GRIFFITHS, D. *et al*. A multi-genre model for music emotion recognition using linear regressors. **Journal of New Music Research**, v. 50, n. 4., p. 355-372, dez. 2021. Disponível em: https://doi.org/10.1080/09298215.2021.1977336. Acesso em: 13 dez. 2021.

GRIGORESCU, S. *et al*. A survey of deep learning techniques for autonomous driving. **Journal of Field Robotics**, v. 37, n. 3, p. 362-386, abr. 2020. Disponível em: https://doi.org/10.1002/rob.21918. Acesso em: 04 jun. 2021.

GROSCHE, P.; MULLER, M. Extracting predominant local pulse information from music recordings. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 19, n. 6, p. 1688-1701, ago. 2011. Disponível em: https://doi.org/10.1109/TASL.2010.2096216. Acesso em: 22 dez. 2021.

GROSS, D. M.; PRESTON, S. D. Darwin and the situation of emotion research. **Emotion Review**, v. 12, n. 3, p. 179-190, jun. 2020. Disponível em: https://doi.org/10.1177%2F1754073920930802. Acesso em: 15 fev. 2021.

GU, S. *et al*. An integrative way for studying neural basis of basic emotions with fMRI. **Frontiers in Neuroscience**, v. 13, p. 1-12, jun. 2019. Disponível em: https://doi.org/10.3389/fnins.2019.00628. Acesso em: 06 abr. 2021.

HABIMANA, O. *et al*. Sentiment analysis using deep learning approaches: an overview. **Science China Information Sciences**, v. 63, n. 1, p. 1-36, 2020. Disponível em: https://doi.org/10.1007/s11432-018-9941-6. Acesso em: 27 maio 2021.

HAGIWARA, M. **Real-world natural language processing**. Shelter Island: Manning Publications Co, 2021.

HALLGREN, K. A. Computing inter-rater reliability for observational data: an overview and tutorial. **Tutorials in Quantitative Methods for Psychology**, v. 8, n. 1, p. 23,

2012. Disponível em: https://pubmed.ncbi.nlm.nih.gov/22833776/. Acesso em: 22 set. 2022.

HAN, D. *et al*. A survey of music emotion recognition. **Frontiers of Computer Science**, v. 16, n. 6, p. 1-11, jan. 2022. Disponível em: https://doi.org/10.1007/s11704-021-0569-4.

HANCOCK, J.; KHOSHGOFTAAR, T. M. Performance of catboost and xgboost in medicare fraud detection. *In:* IEEE International Conference on Machine Learning and Applications (ICMLA), 19., Miami. **Proceedings [...] .** Miami: IEEE, 2020. p. 572-579. Disponível em: https://doi.org/10.1109/ICMLA51294.2020.00095. Acesso em: 25 mar. 2023.

HARNUM, J. **Basic music theory:** how to read, write, and understand written music. Sol Ut Press, 2013. 4 ed.

HASAN, T.; MATIN, A. Extract sentiment from customer reviews: a better approach of TF-IDF and BOW-based text classification using n-gram technique. *In*: INTERNATIONAL JOINT CONFERENCE ON ADVANCES IN COMPUTATIONAL INTELLIGENCE, 4., 2020. **Proceedings[...] .** Springer, 2021. p. 231–244. Disponível em: https://link.springer.com/chapter/10.1007/978-981-16-0586-4_19. Acesso em: 17 dez. 2021.

HAYKIN, S. **Neural networks and learning machines**, Nova Iorque: Pearson, 2009. 3. ed.

HENNEQUIN, R. *et al*. Spleeter: a fast and efficient music source separation tool with pre-trained models. **Journal of Open Source Software**, v. 5, n. 50, p. 1-4, jun. 2020. Disponível em: https://joss.theoj.org/papers/10.21105/joss.02154. Acesso em: 13 dez. 2021.

HEVNER, K. Experimental studies of the elements of expression in music. **The American Journal of Psychology**, v. 48, n. 2, p. 246-268, abr. 1936. Disponível em: https://www.jstor.org/stable/1415746. Acesso em: 08 set. 2020.

HINOJOSA, J. A. et al. Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions. **Behavior Research Methods**, v. 48, n. 1, p. 272-284, mar. 2016. Disponível em: https://doi.org/10.3758/s13428-015-0572-5. Acesso em: 15 abr. 2021.

HIZLISOY, S.; YILDIRIM, S.; TUFEKCI, Z. Music emotion recognition using convolutional long short term memory deep neural networks. **Engineering Science and Technology, an International Journal**, v. 24, n. 3, p. 760-767, 2021. Disponível em: https://doi.org/10.1016/j.jestch.2020.10.009. Acesso em: 27 abr. 2021.

HOSSAIN, M. S.; MUHAMMAD, G. Emotion recognition using deep learning approach from audio–visual emotional big data. **Information Fusion**, v. 49, p. 69-78, set. 2019. Disponível em: https://doi.org/10.1016/j.inffus.2018.09.008. Acesso em 24 set. 2020.

HOSSEINI, M. P. *et al*. Deep learning architectures. *In*: PEDRYCZ, W.; CHEN; S. M. (Ed.) **Deep learning: concepts and architectures**. Cham: Springer, 2020. p. 1-24.

HSU, J. L. *et al*. Affective content analysis of music emotion through EEG. **Multimedia Systems**, v. 24, n. 2, p. 195-210, mar. 2018. Disponível em: https://doi.org/10.1007/s00530-017-0542-0. Acesso em: 27 abr. 2021.

HU, X; DOWNIE, J. S. Exploring mood metadata: relationships with genre, artist and usage metadata. *In*: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL, 8., 2007, Viena. **Proceedings[...] .** Viena: ISMIR, 2007. p. 1-6. Disponível em: https://ismir2007.ismir.net/proceedings/ISMIR2007_p067_hu.pdf. Acesso em: 17 maio 2021.

HU, X. et al. The 2007 MIREX audio mood classification task: lessons learned. In: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONFERENCE (ISMIR 2008), 9., 2008, Philadelphia. **Proceedings [...] .** Philadelphia: ISMIR, 2008. p. 462-467. Disponível em: https://ismir2008.ismir.net/papers/ISMIR2008_263.pdf. Acesso em: 26 maio 2022.

HU, X.; YANG, Y. H. Cross-cultural mood regression for music digital libraries. In: IEEE/ACM JOINT CONFERENCE ON DIGITAL LIBRARIES, 14., 2014, Londres. **Proceedings[...] .** Londres: IEEE, 2014. p. 471-472. Disponível em: https://ieeexplore.ieee.org/document/6970230. Acesso em: 23 jul. 2020.

HUANG, J. *et al*. Integration of multimodal features for video scene classification based on HMM. In: IEEE THIRD WORKSHOP ON MULTIMEDIA SIGNAL PROCESSING, 3, 1999, Copenhage. **Proceedings[...] .** Copenhage: IEEE, 1999, p. 53 - 58. Disponível em: https://ieeexplore.ieee.org/document/793797. Acesso em: 17 dez. 2020.

HUANG, X. *et al*. **TabTransformer: tabular data modeling using contextual embeddings**. *Preprint*, submetido a 11 de dezembro de 2020. https://arxiv.org/abs/2012.06678. Acesso em: 12 maio 2022.

HULT, S. *et al*. Analysis of the effect of dataset construction methodology on transferability of music emotion recognition models. In: INTERNATIONAL CONFERENCE ON MULTIMEDIA RETRIEVAL, 10., 2020, Dublin. **Proceedings[...] .** Dublin: Acm, 2020. p. 316-320. Disponível em: https://dl.acm.org/doi/10.1145/3372278.3390733. Acesso em: 04 ago. 2020.

HUQ, A.; BELLO, J. P.; ROWE, R. Automated music emotion recognition: a systematic evaluation. **Journal of New Music Research**, v. 39, n. 3, p. 227-244, nov. 2010. Disponível em: https://doi.org/10.1080/09298215.2010.513733. Acesso em: 04 set. 2020.

HUTTO, C.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: **Proceedings of the international AAAI conference on web and social media**. maio 2014. p. 216-225. Disponível em: https://doi.org/10.1609/icwsm.v8i1.14550. Acesso em: 12 mar. 2023.

HUTTO, D. D.; ROBERTSON, I.; KIRCHHOFF, Michael D. A new, better BET: rescuing and revising basic emotion theory. **Frontiers in Psychology**, v. 9, p. 1-12, jul. 2018. Disponível em:https://doi.org/10.3389/fpsyg.2018.01217. Acesso em: 07 abr. 2021.

IMBRASAITE, V. **Continuous dimensional emotion tracking in music**. 2015. 158 f. Tese (Doutorado) - Curso de Ciência da Computação, Universidade de Cambridge, Cambridge, 2015. Disponível em: https://www.repository.cam.ac.uk/handle/1810/247917. Acesso em: 19 out. 2020.

IZARD, C.E. **The face of emotion**. Nova Iorque: Appleton-Century-Crofts, 1971.

JAKUBOWSKI, K.; GHOSH, A. Music-evoked autobiographical memories in everyday life. **Psychology of Music**, v. 0, n. 00, p. 1-19, dez. 2019. Disponível em: https://doi.org/10.1177%2F0305735619888803. Acesso em: 09 ago. 2020.

JAMES, W. What is an emotion? **Mind*,*** v. os-IX, n. 34, p. 188-205, abr. 1884. Disponível em: https://doi.org/10.1093/mind/os-IX.34.188. Acesso em: 05 fev. 2021.

JANSSON, A. *et al*. Singing voice separation with deep U-net convolutional networks. *In*: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONFERENCE, 18., 2017, Suzhou. **Proceedings[...] .** Suzhou: ISMIR, 2017. p. 323–332. Disponível em: https://ejhumphrey.com/assets/pdf/jansson2017singing.pdf. Acesso em: 08 jan. 2022.

JIMÉNEZ-LUNA, J.; GRISONI, F.; SCHNEIDER, G. Drug discovery with explainable artificial intelligence. **Nature Machine Intelligence**, v. 2, n. 10, p. 573-584, out. 2020. Disponível em: https://www.nature.com/articles/s42256-020-00236-4. Acesso em: 26 fev. 2022.

JIN, S. *et al*. Evaluation and modeling of automotive transmission whine noise quality based on MFCC and CNN. **Applied Acoustics**, v. 172, p. 1-11, jan. 2021. Disponível em: https://doi.org/10.1016/j.apacoust.2020.107562. Acesso em: 23 dez. 2021.

JOHNSON, P. E.; SMITH, M. O.; ADAMS, J. B. Quantitative analysis of planetary reflectance spectra with principal components analysis. **Journal of Geophysical Research: Solid Earth**, v. 90, n. S02, p. C805-C810, fev. 1985. Disponível em: https://doi.org/10.1029/JB090iS02p0C805. Acesso em: 26 mar. 2023.

JOO, C. *et al*. Development of physical property prediction models for polypropylene composites with optimizing random forest hyperparameters. **International Journal of Intelligent Systems**, v. 37, n. 6, p. 3625-3653, set. 2021. Disponível em: https://doi.org/10.1002/int.22700. Acesso em: 25 mar. 2023.

JOSEPH, C.; LEKAMGE, S. Machine learning approaches for emotion classification of music: A systematic literature review. In: INTERNATIONAL CONFERENCE ON ADVANCEMENTS IN COMPUTING, 2019, Malabe. **Proceedings[...] .** Malabe: IEEE, 2019. p. 334-339. Disponível em: https://ieeexplore.ieee.org/document/9103378. Acesso em: 31 jul. 2020.

JUSLIN, P. N. From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. **Physics of Life Reviews**, v. 10, n. 3, p. 235-266, set. 2013a. Disponível em: https://doi.org/10.1016/j.plrev.2013.05.008. Acesso em: 04 fev. 2021.

JUSLIN, P. N. What does music express? basic emotions and beyond. **Frontiers in Psychology**, v. 4, p. 1-14, set. 2013b. Disponível em: https://doi.org/10.3389/fpsyg.2013.00596. Acesso em: 09 ago. 2021.

JUSLIN, P. N. **Musical emotions explained**: unlocking the secrets of musical affect. Oxford: Oxford University Press, 2019.

JUSLIN, P. N.; LAUKKA, P. Communication of emotions in vocal expression and music performance: different channels, same code?. **Psychological Bulletin**, v. 129, n. 5, p. 770-814, 2003. Disponível em: https://psycnet.apa.org/doi/10.1037/0033-2909.129.5.770. Acesso em: 11 ago. 2021.

JUSLIN, P. N.; LAUKKA, P. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. **Journal of New Music Research**, v. 33, n. 3, p. 217-238, 2004. Disponível em: https://doi.org/10.1080/0929821042000317813. Acesso em: 30 jun. 2021.

KADRA, A. *et al*. Well-tuned simple nets excel on tabular datasets. Advances in Neural Information Processing Systems, *In*: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS 2021), 35., 2021. **Proceedings [...] .** NeurIPS, 2021. p. 1-114. Disponível em: https://proceedings.neurips.cc/paper/2021/hash/c902b497eb972281fb5b4e206db38ee6-Abstract.html. Acesso em: 12 maio 2022.

KALLINEN, K. Emotional ratings of music excerpts in the western art music repertoire and their self-organization in the Kohonen neural network. **Psychology of Music**, v. 33, n. 4, p. 373-393, out. 2005. Disponível em: https://doi.org/10.1177%2F0305735605056147. Acesso em: 09 ago. 2021.

KALTEH, A. M.; BERNDTSSON, R. Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). **Hydrological Sciences Journal**, v. 52, n. 2, p. 305-317, 2007. Disponível em: https://doi.org/10.1623/hysj.52.2.305. Acesso em: 17 maio 2022.

KAMILARIS, A.; PRENAFETA-BOLDÚ, F. X. Deep learning in agriculture: a survey. **Computers and Electronics in Agriculture**, v. 147, p. 70-90, abr. 2018. Disponível em: https://doi.org/10.1016/j.compag.2018.02.016. Acesso em: 27 maio 2021.

KAMINSKAS, M.; RICCI, F. Contextual music information retrieval and recommendation: state of the art and challenges. **Computer Science Review**, v. 6, n. 2-3, p. 89-119, maio 2012. Disponível em: https://doi.org/10.1016/j.cosrev.2012.04.002. Acesso em: 21 set. 2020.

KE, G. LightGBM: a highly efficient gradient boosting decision tree. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS

2017), 31., 2017, Long Beach. **Proceedings [...] .** Long Beach: NeurIPS, 2017. p. 3149-3157. Disponível em: https://dl.acm.org/doi/abs/10.5555/3294996.3295074. Acesso em: 11 maio 2022.

KELTNER, D. *et al*. Emotional expression: advances in basic emotion theory. **Journal of Nonverbal Behavior**, v. 43, p. 1-28, fev. 2019. Disponível em: https://doi.org/10.1007/s10919-019-00293-3. Acesso em: 19 fev. 2021.

KHAN, Z. Y. *et al*. Deep learning techniques for rating prediction: a survey of the state-of-the-art. **Artificial Intelligence Review**, v. 54, n. 1, p. 95-135, jan. 2021. Disponível em: https://doi.org/10.1007/s10462-020-09892-9. Acesso em: 26 maio 2021.

KIM, Y. E. *et al*. Music emotion recognition: a state of the art review. In: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONFERENCE, 11., 2010, Utrecht. **Proceedings[...] .** Utrecht: ISMIR, 2010. p. 255 - 266. Disponível em: http://ismir2010.ismir.net/proceedings/ismir2010-45.pdf. Acesso em: 24 jul. 2020.

KIM, J. *et al*. One deep music representation to rule them all? A comparative analysis of different representation learning strategies. **Neural Computing and Applications**, v. 32, n. 4, p. 1067-1093, mar. 2019. Disponível em: https://doi.org/10.1007/s00521-019-04076-1. Acesso em: 28 set. 2020.

KLEINGINNA, P. R.; KLEINGINNA, A. M. A categorized list of emotion definitions, with suggestions for a consensual definition. **Motivation and Emotion**, v. 5, n. 4, p. 345-379, dez. 1981. Disponível em: https://doi.org/10.1007/BF00992553. Acesso em: 27 jan. 2021.

KOELSCH, S. Brain correlates of music-evoked emotions. **Nature Reviews Neuroscience**, v. 15, n. 3, p. 170-180, fev. 2014. Disponível em: https://www.nature.com/articles/nrn3666. Acesso em: 30 jun. 2021.

KOK, S. H.; AZWEEN, A.; JHANJHI, N. Z. Evaluation metric for crypto-ransomware detection using machine learning. **Journal of Information Security and Applications**, v. 55, p. 1-9, dez. 2020. Disponível em: https://doi.org/10.1016/j.jisa.2020.102646. Acesso em: 03 set. 2021.

KREUTZ, G. *et al*. Using music to induce emotions: influences of musical preference and absorption. **Psychology of Music**, v. 36, n. 1, p. 101-126, jan. 2008. Disponível em: https://doi.org/10.1177%2F0305735607082623. Acesso em: 25 ago. 2021.

KRYKLYWY, J. H. *et al*. From architecture to evolution: multisensory evidence of decentralized emotion. **Trends in Cognitive Sciences**, v. 24, n. 11, p. 916-929, nov. 2020. Disponível em: https://doi.org/10.1016/j.tics.2020.08.002. Acesso em: 04 fev. 2021.

KURSA, M. B.; RUDNICKI, W. R. Feature selection with the Boruta package. **Journal of Statistical Software**, v. 36, p. 1-13, set. 2010. Disponível em: https://www.jstatsoft.org/article/view/v036i11. Acesso em: 26 mar. 2023.

LAKATOS, E. M.; MARCONI, M. A. **Fundamentos da metodologia científica**. 8. ed. São Paulo: Editora Atlas, 2017. 333 p.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159-174, mar. 1977. Disponível em: https://doi.org/10.2307/2529310. Acesso em: 25 jan. 2022.

LANGE, J. *et al*. Toward an integrative psychometric model of emotions. **Perspectives on Psychological Science**, v. 15, n. 2, p. 444-468, fev. 2020. Disponível em: https://doi.org/10.1177%2F1745691619895057. Acesso em: 03 fev. 2021.

LARTILLOT, O.; TOIVIAINEN, P.; EEROLA, T. A Matlab toolbox for music information retrieval. *In*: PREISACH, C. *et al*. (ed.). **Data Analysis, machine Learning and applications**. Berlim: Springer, 2008. p. 261-268.

LAUKKA, P.; GABRIELSSON, A. Emotional expression in drumming performance. **Psychology of Music**, v. 28, n. 2, p. 181-189, 2000.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436-444, maio 2015. Disponível em: https://doi.org/10.1038/nature14539. Acesso em: 02 jun. 2021.

LI, B. *et al*. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. **IEEE Transactions on Multimedia**, v. 21, n. 2, p. 522-535, fev. 2019. Disponível em: https://ieeexplore.ieee.org/document/8411155. Acesso em: 27 jul. 2020.

LI, C. *et al*. Multimodal deep support vector classification with homologous features and its application to gearbox fault diagnosis**. Neurocomputing**, v. 168, p. 119-127, nov. 2015. Disponível em: https://www.sciencedirect.com/science/article/pii/S0925231215008280. Acesso em: 17 dez. 2020.

LI, J. Music mood classification via deep belief network. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOP, 15., 2015, Atlantic City. **Proceedings[...] .** Atlantic City: IEEE, 2015. p. 1241-1245. Disponível em: https://ieeexplore.ieee.org/document/7395810. Acesso em: 28 set. 2020.

LI, M. *et al*. Emotion recognition from multichannel EEG signals using K-nearest neighbor classification. **Technology and Health Care**, v. 26, n. S1, p. 509-519, jul. 2018. Disponível em: https://dx.doi.org/10.3233%2FTHC-174836. Acesso em 24 jul. 2020.

LIN, Y. C.; YANG, Y. H.; CHEN, H. H. Exploiting online music tags for music emotion classification. **ACM Transactions on Multimedia Computing, Communications, and Applications**, v. 7, n. 1, p. 26:1-26:16, jan. 2011. Disponível em: https://doi.org/10.1145/2037676.2037683. Acesso em: 05 maio. 2021.

LIU, G.; THAN, Z. Research on multi-modal music emotion classification based on audio and lyirc. *In*: 2020 IEEE 4TH INFORMATION TECHNOLOGY, NETWORKING, ELECTRONIC AND AUTOMATION CONTROL CONFERENCE (ITNEC), 4., 2020, Xunquim. **Proceedings[...] .** Xunquim: IEEE, 2020. p. 2331-2335. Disponível em: https://ieeexplore.ieee.org/document/9084846. Acesso em: 04 ago. 2020.

LIU, W. *et al*. A survey of deep neural network architectures and their applications. **Neurocomputing**, v. 234, p. 11-26, abr. 2017. Disponível em: https://doi.org/10.1016/j.neucom.2016.12.038. Acesso em: 07 jun. 2021.

LIU, Z. T. *et al*. Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence. **Information Sciences**, v. 563, p. 309-325, jul. 2021. Disponível em: https://doi.org/10.1016/j.ins.2021.02.016. Acesso em: 10 jan. 2022.

LONSDALE, A. J. Emotional intelligence, alexithymia, stress, and people's reasons for listening to music. **Psychology of Music**, v. 47, n. 5, p. 680-693, jun. 2018. Disponível em: http://dx.doi.org/10.1177/0305735618778126. Acesso em: 09 ago. 2020.

LÖVHEIM, H. A new three-dimensional model for emotions and monoamine neurotransmitters. **Medical Hypotheses**, v. 78, n. 2, p. 341-348, fev. 2012. Disponível em: https://doi.org/10.1016/j.mehy.2011.11.016. Acesso em: 13 abr. 2021.

LU, Q. *et al*. Boosting for multi-modal music emotion. In: 11TH INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONFERENCE (ISMIR 2010), 11., 2010, Utreque. **Proceedings[...] .** Utreque: ISMIR, 2010. p. 105 - 110. Disponível em: http://ismir2010.ismir.net/proceedings/ismir2010-20.pdf. Acesso em: 07 set. 2020.

LYONS, W. The philosophy of cognition and emotion. In: DALGLEISH, T.; POWER, M. J. **Handbook of cognition and emotion**, Chichester: John Wiley & Sons, 1999. p. 21-44.

MACDORMAN, K. F.; OUGH, S.; HO, C. C. Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. **Journal of New Music Research**, v. 36, n. 4, p. 281-299, 2007. Disponível em: https://doi.org/10.1080/09298210801927846. Acesso em: 26 maio 2022.

MAKRIS, D.; KARYDIS, I.; SIOUTRAS, S. The greek music dataset. *In*: INTERNATIONAL CONFERENCE ON ENGINEERING APPLICATIONS OF NEURAL NETWORKS (INNS), 16., 2015, Rodes. **Proceedings[...] .** Rodes: ACM, 2015. p. 1-7. Disponível em: https://doi.org/10.1145/2797143.2797175. Acesso em: 07 nov. 2021.

MALIK, M. *et al*. Stacked convolutional and recurrent neural networks for music emotion recognition. In: SOUND AND MUSIC COMPUTING CONFERENCE, 14., 2017, Espoo. **Proceedings[...] .** Espoo: Aalto University, 2017. p. 208-213. Disponível em:

http://smc2017.aalto.fi/media/materials/proceedings/SMC17_p208.pdf. Acesso em: 28 set. 2020.

MALHEIRO, R. *et al*. Emotionally-relevant features for classification and regression of music lyrics. **IEEE Transactions on Affective Computing**, v. 9, n. 2, p. 240-254, abr.-jun. 2018. Disponível em: https://doi.org/10.1109/TAFFC.2016.2598569. Acesso em: 12 maio 2021.

MANSOORIZADEH, M.; CHARKARI, N. M. Multimodal information fusion application to human emotion recognition from face and speech. **Multimedia Tools and Applications**, v. 49, n. 2, p. 277-297, ago. 2010. Disponível em: https://link.springer.com/article/10.1007/s11042-009-0344-2. Acesso em: 17 dez. 2020.

MANTOVANI, R. G. *et al*. A meta-learning recommender system for hyperparameter tuning: Predicting when tuning improves SVM classifiers. **Information Sciences**, v. 501, p. 193-221, out. 2019. Disponível em: https://doi.org/10.1016/j.ins.2019.06.005. Acesso em: 25 mar. 2023.

MARTINS JÚNIOR, J. P. A pesquisa nas Ciências Humanas e Sociais. *In*: PERDIGÃO, D. M.; HERLINGER, M.; WHITE, O. M. (org.). **Teoria e prática da pesquisa aplicada**. Rio de Janeiro: Elsevier, 2012. Cap. 3. p. 23-31.

MAUCH, M. **Automatic chord transcription from audio using computational models of musical context**. 2010. 168 f. Tese (Doutorado) - Curso de Electronic Engineering and Computer Science, University of London, Londres, 2010. Disponível em: https://qmro.qmul.ac.uk/xmlui/handle/123456789/451. Acesso em: 14 dez. 2021

MAUCH, M.; DIXON, S. Approximate note transcription for the improved identification of difficult chords. *In*: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONFERENCE, 11., 2010, Utrecht. **Proceedings[...] .** Utrecht: ISMIR, 2010. p. 135-140. Disponível em: https://archives.ismir.net/ismir2010/paper/000025.pdf. Acesso em: 13 dez. 2021.

MCDOUGALL, W. **An introduction to social psychology**. Boston: Luce, 1926.

MCFEE, B. et al. librosa: audio and music signal analysis in Python. *In*: Python in science conference, 14., 2015, Austin. **Proceedings[...] .** Austin: SciPy, 2015. p. 18-25. Disponível em: https://conference.scipy.org/proceedings/scipy2015/pdfs/brian_mcfee.pdf. Acesso em: 09 maio 2021.

MCFERRAN, K. S. *et al*. Intentional music use to reduce psychological distress in adolescents accessing primary mental health care. **Clinical Child Psychology and Psychiatry**, v. 23, n. 4, p. 567-581, abr. 2018. Disponível em: https://doi.org/10.1177%2F1359104518767231. Acesso em: 26 jun. 2021.

MCKAY, C. **Automatic music classification with jMIR**. 2010. 600 f. Tese (Doutorado) - Curso de *Music Technology*, *Music*, McGill University, Montreal, 2010.

Disponível em: http://jmir.sourceforge.net/publications/PhD_Dissertation_2010.pdf. Acesso em: 11 maio 2021.

MED, B. **Teoria da música**. Brasília: Musimed, 2017. 5 ed.

MEDINA, Y. O.; BELTRÁN, J. R.; BALDASSARRI, S. Emotional classification of music using neural networks with the MediaEval dataset. **Personal and Ubiquitous Computing**, p. 1-13, abr. 2020. Disponível em: https://doi.org/10.1007/s00779-020-01393-4. Acesso em: 03 maio 2021.

MÉLO, D.; ANDRADE, N. Exploiting the importance of personalization when selecting music for relaxation. In: RO, Y. *et al* (ed.). **MultiMedia Modeling. MMM 2020. Lecture Notes in Computer Science**. Cham: Springer, 2019. p. 49-61. Disponível em: https://link.springer.com/chapter/10.1007%2F978-3-030-37731-1_5 . Acesso em: 09 ago. 2020.

MEYER, L. B. **Style and music:** theory, history, and ideology. Chicago: University of Chicago Press, 1996.

MILLER, M. **The complete idiot's guide to music theory**. Indianápolis: Penguin Random House, 2016. 3 ed.

MILLER, J.; NICOSIA, V.; SANDLER, M. Discovering common practice: using graph theory to compare harmonic sequences in musical audio collections. *In*: INTERNATIONAL CONFERENCE ON DIGITAL LIBRARIES FOR MUSICOLOGY, 8., 2021. **Proceedings[...] .** ACM, 2021. p. 93-97. Disponível em: https://doi.org/10.1145/3469013.3469025. Acesso em: 14 dez. 2021.

MOHER, D. *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. **PLoS med**, v. 6, n. 7, p. 1-6, jul. 2009. Disponível em: https://doi.org/10.1371/journal.pmed.1000097. Acesso em 07 set. 2020.

MOHN, C.; ARGSTATTER, H.; WILKER, F. W. Perception of six basic emotions in music. **Psychology of Music**, v. 39, n. 4, p. 503-517, out. 2011. Disponível em: https://doi.org/10.1177%2F0305735610378183. Acesso em: 26 out. 2021.

MOORS, A. Automatic constructive appraisal as a candidate cause of emotion. **Emotion Review**, v. 2, n. 2, p. 139-156, fev. 2010. Disponível em: https://doi.org/10.1177%2F1754073909351755. Acesso em: 17 fev. 2021.

MOORS, A. *et al*. Appraisal theories of emotion: state of the art and future development. **Emotion Review**, v. 5, n. 2, p. 119-124, mar. 2013. Disponível em: https://doi.org/10.1177%2F1754073912468165. Acesso em: 11 fev. 2021.

MOREIRA, P. S. C. **Reconhecimento de emoções em músicas por meio de sistemas neuro-*fuzzy***. 2019. 176 f. Dissertação (Mestrado) - Curso de Gestão da Informação, Universidade Federal do Paraná, Curitiba, 2019.

MORENCY, L-P.; MIHALCEA, R.; DOSHI, P. Towards multimodal sentiment

analysis: harvesting opinions from the web. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERFACES, 13, 2011, Alicante**. Proceedings[...] .** Alicante: ACM, 2011, p.169 - 176. Disponível em: https://dl.acm.org/doi/10.1145/2070481.2070509. Acesso em: 17 dez. 2020.

MOWRER, O. H. **Learning theory and behavior**. Nova Iorque: Wiley, 1960.

MÜELLER, M. **Fundamentals of music processing**: audio, analysis, algorithms, applications. Cham: Springer, 2015.

MÜELLER, M.; MCFEE, B.; KINNAIRD, K. Interactive learning of signal processing through music: making Fourier analysis concrete for students. **IEEE Signal Processing Magazine**, v. 38, n. 3, p. 73-84, maio 2021. Disponível em: https://doi.org/10.1109/MSP.2021.3052181. Acesso em: 14 maio 2021.

MULLIGAN, K.; SCHERER, K. R. Toward a working definition of emotion. **Emotion Review**, v. 4, n. 4, p. 345-357, out. 2012. Disponível em: https://doi.org/10.1177%2F1754073912445818. Acesso em 29 jan. 2021.

MURTHY, Y. V. S.; KOOLAGUDI, S. G. Content-based music information retrieval (CB-MIR) and its applications toward the music industry: a review. **ACM Computing Surveys**, v. 51, n. 3, p. 1-46, jul. 2018. Disponível em: https://dl.acm.org/doi/10.1145/3177849. Acesso em: 09 ago. 2020.

NGIAM, J. et al. Multimodal deep learning. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 28, 2011, Bellevue. **Proceedings[…] .** Bellevue: IMLS, 2011, p. 689-696. Disponível em: http://ai.stanford.edu/~ang/papers/icml11-MultimodalDeepLearning.pdf. Acesso em: 17 jul. 2020.

NUNES-SILVA, M. *et al*. Avaliação de músicas compostas para indução de relaxamento e de seus efeitos psicológicos. **Psicologia: Ciência e Profissão**, v. 36, p. 709-725, jul-set. 2016. Disponível em: https://doi.org/10.1590/1982-3703001672014. Acesso em: 23 maio 2022.

OATLEY, K.; JOHNSON-LAIRD, P. N. Towards a cognitive theory of emotions. **Cognition and Emotion**, v. 1, n. 1, p. 29-50, 1987. Disponível em: https://doi.org/10.1080/02699938708408362. Acesso em: 07 abr. 2021.

ORTONY, A.; TURNER, T. J. What's basic about basic emotions? **Psychological Review**, v. 97, n. 3, p. 315-331, 1990. Disponível em: https://doi.apa.org/doi/10.1037/0033-295X.97.3.315. Acesso em: 06 abr. 2021.

OZBAYOGLU, A. M.; GUDELEK, M. U.; SEZER, O. B. Deep learning for financial applications: a survey. **Applied Soft Computing**, p. 1-29, ago. 2020. Disponível em: https://doi.org/10.1016/j.asoc.2020.106384. Acesso em: 26 maio 2021.

ÖZÇIFT, Ak. *et al*. Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish. **Automatika**, v. 62, n. 2, p.

226-238, maio 2021. Disponível em: https://doi.org/10.1080/00051144.2021.1922150. Acesso em: 29 dez. 2021.

PAN, S. *et al*. An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. **Journal of Petroleum Science and Engineering**, v. 208, p. 109520, jan. 2022. Disponível em: https://doi.org/10.1016/j.petrol.2021.109520. Acesso em: 25 mar. 2023.

PANDA, R. E. S. **Emotion-based analysis and classification of audio music**. 2019. 373 f. Tese (Doutorado) - Curso de Engenharia Informática, Universidade de Coimbra, Coimbra, 2019. Disponível em: https://estudogeral.sib.uc.pt/handle/10316/87618. Acesso em: 07 set. 2020.

PANDA, R. *et al*. Multi-modal music emotion recognition: a new dataset, methodology and comparative analysis. In: INTERNATIONAL SYMPOSIUM ON COMPUTER MUSIC MULTIDISCIPLINARY RESEARCH, 10, 2013, Marseille. **Proceedings[…] .** Marseille: CNRS, 2013, p. 570-582. Disponível em: https://eden.dei.uc.pt/~ruipedro/publications/Conferences/CMMR2013_MultiModal.pdf. Acesso em: 07 set. 2020.

PANDA, R.; MALHEIRO, R.; PAIVA, R. P. Novel audio features for music emotion recognition. **IEEE Transactions on Affective Computing**, v. 11, n. 4, p. 614-626, mar. 2020a. Disponível em: https://doi.org/10.1109/TAFFC.2018.2820691. Acesso em: 27 abr. 2021.

PANDA, R.; MALHEIRO, R.; PAIVA, R. P. Audio features for music emotion recognition: a survey. **IEEE Transactions on Affective Computing**, 2020b. Disponível em: https://doi.org/10.1109/TAFFC.2020.3032373. Acesso em: 30 ago. 2021.

PANDA, R.; ROCHA, B; PAIVA, R. P. Music emotion recognition with standard and melodic audio features. **Applied Artificial Intelligence**, v. 29, n. 4, p. 313-334, abr. 2015. Disponível em: https://doi.org/10.1080/08839514.2015.1016389. Acesso em: 10 maio 2021.

PANDEYA, Y. R.; LEE, J. Deep learning-based late fusion of multimodal information for emotion classification of music video. **Multimedia Tools and Applications**, p. 1-19, set. 2020. Disponível em: https://doi.org/10.1007/s11042-020-08836-3. Acesso em: 28 set. 2020.

PANKSEPP, J. Toward a general psychobiological theory of emotions. **Behavioral and Brain Sciences**, v. 5, n. 3, p. 407-422, set. 1982. Disponível em: https://doi.org/10.1017/S0140525X00012759. Acesso em: 07 abr. 2021.

PANKSEPP, J. **Affective neuroscience:** the foundations of human and animal emotions. Nova Iorque: Oxford University Press, 1998.

PANWAR, S. *et al*. A deep learning approach for mapping music genres. In: SYSTEM OF SYSTEMS ENGINEERING CONFERENCE (SOSE), 12., 2017,

Waikoloa. **Proceedings[...] .** Waikoloa: IEEE, 2017. p. 1-5. Disponível em: https://doi.org/10.1109/SYSOSE.2017.7994970. Acesso em: 28 set. 2020.

PANWAR, S. *et al*. Are you emotional or depressed? Learning about your emotional state from your music using machine learning. **The Journal of Supercomputing**, v. 75, n. 6, p. 2986-3009, jul. 2018. Disponível em: https://link.springer.com/article/10.1007/s11227-018-2499-y. Acesso em: 24 jul. 2020.

PATNI, H. *et al*. Speech emotion recognition using MFCC, GFCC, Chromagram and RMSE features. *In*: INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING AND INTEGRATED NETWORKS (SPIN), 8., 2021, Noida. **Proceedings [...] .** Noida: IEEE, 2021. p. 892-897. Disponível em: https://doi.org/10.1109/SPIN52536.2021.9566046. Acesso em: 10 jan. 2022.

PATRA, B. G.; DAS, D.; BANDYOPADHYAY, S. Multimodal mood classification of Hindi and Western songs. **Journal of Intelligent Information Systems**, v. 51, n. 3, p. 579-596, fev. 2018. Disponível em: https://doi.org/10.1007/s10844-018-0497-4. Acesso em: 12 maio 2021.

PEDREGOSA, F. *et al*. Scikit-learn: machine learning in Python. **The Journal of Machine Learning research**, v. 12, p. 2825-2830, 2011.Disponível em: https://www.jmlr.org/papers/v12/pedregosa11a.html. Acesso em: 17 dez. 2021.

PERDIGÃO, D. M. Introdução à pesquisa quantitativa. *In*: PERDIGÃO, D. M.; HERLINGER, M.; WHITE, O. M. (org.). **Teoria e prática da pesquisa aplicada**. Rio de Janeiro: Elsevier, 2012. Cap. 11. p. 109-117.

PESEK, M. *et al*. The moodo dataset: integrating user context with emotional and color perception of music for affective music information retrieval. **Journal of New Music Research**, v. 46, n. 3, p. 246-260, jun. 2017. Disponível em: https://www.tandfonline.com/doi/full/10.1080/09298215.2017.1333518. Acesso em: 24 jul. 2020.

PILHOFER, M.; DAY, H. **Music theory for dummies**. Hoboken: John Wiley & Sons, 2019. 4 ed.

PLAMPER, J. **The history of emotions:** an introduction. Oxford: Oxford University Press, 2015.

PLUTCHIK, R. Emotions, evolution, and adaptive processes. *In:* ARNOLD, M. B. **Feelings and Emotions.** Nova Iorque: Academic Press, 1970. p. 3-24.

PLUTCHIK, R. A general psychoevolutionary theory of emotion. *In*: PLUTCHIK, R.; KELLERMAN, H. **Emotion:** Theory, research, and experience: Vol. 1. Theories of emotion. Nova Iorque: Academic Press, 1980, p. 3-31.

POPOV, S.; MOROZOV, S.; BABENKO, A. Neural oblivious decision ensembles for deep learning on tabular data. *In*: INTERNATIONAL CONFERENCE ON LEARNING

REPRESENTATIONS, 8., 2020. **Proceedings [...] .** ICLR, 2020. p. 1-12. Disponível em: https://iclr.cc/virtual_2020/poster_r1eiu2VtwH.html. Acesso em: 11 maio 2022.

PORIA, S. et al. Towards an intelligent framework for multimodal affective data analysis. **Neural Networks**, v. 63, p. 104-116, mar. 2015. Disponível em: https://www.sciencedirect.com/science/article/pii/S0893608014002342. Acesso em: 17 dez. 2020.

PORIA, S. et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. **Neurocomputing**, v. 174, p. 50-59, jan. 2016. Disponível em: https://www.sciencedirect.com/science/article/pii/S0925231215011297. Acesso em: 17 dez. 2020.

POUYANFAR, S. *et al*. A survey on deep learning: algorithms, techniques, and applications. **ACM Computing Surveys (CSUR)**, v. 51, n. 5, p. 1-36, set. 2018. Disponível em: https://doi.org/10.1145/3234150. Acesso em: 27 maio 2021.

PRIOLLI, M. L. M. **Princípios básicos da música para a juventude**. Rio de Janeiro: Casa Oliveira de Músicas, 2006. 48 ed.

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO DA INFORMAÇÃO (Curitiba). Regimento Interno – 2020 de Acordo com a Resolução CEPE – 32/2017. Dispõe sobre as normas internas do Programa de Pós-graduação em Gestão da Informação – PPGGI. **Regimento Interno**: Curitiba, p. 6, 17 ago. 2020. Disponível em: http://www.prppg.ufpr.br/site/ppggi/wp-content/uploads/sites/108/2020/09/2020-regimento-interno-ppggi-v3-correo-art--45.pdf. Acesso em: 24 jul. 2021.

PROKHORENKOVA, L. *et al*. CatBoost: unbiased boosting with categorical features. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2018), 32., 2018, Montreal. **Proceedings [...] .** Montreal: NeurIPS, 2018. p. 6639-6649. Disponível em: https://dl.acm.org/doi/abs/10.5555/3327757.3327770. Acesso em: 11 maio 2022.

PRZYBYSZ, A. L. **Classificação automática de emoções em músicas latinas utilizando diferentes fontes de informação**. 2016. 91 f. Dissertação (Mestrado) - Curso de Informática, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, 2016. Disponível em: http://repositorio.utfpr.edu.br/jspui/bitstream/1/2932/1/CP_PPGI_M_Przybysz%2c%20Andr%C3%A9%20Luiz_2016.pdf. Acesso em: 07 set. 2020.

PURWINS, H. *et al*. Deep learning for audio signal processing. **IEEE Journal of Selected Topics in Signal Processing**, v. 13, n. 2, p. 206-219, abr. 2019. Disponível em: https://ieeexplore.ieee.org/document/8678825. Acesso em: 29 set. 2020.

QUINTO, L.; THOMPSON, W. F.; TAYLOR, A. The contributions of compositional structure and performance expression to the communication of emotion in music. **Psychology of Music**, v. 42, n. 4, p. 503-524, jul. 2014. Disponível em: https://doi.org/10.1177%2F0305735613482023. Acesso em: 26 ago. 2021.

RAJESH, S.; NALINI, N. J. Musical instrument emotion recognition using deep recurrent neural network. **Procedia Computer Science**, v. 167, p. 16-25, 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050920306438?via%3Dihub. Acesso em: 23 jul. 2020.

RAMOS, D. **Fatores emocionais durante uma escuta musical afetam a percepção temporal de músicos e não-músicos?** 2008. 268 f. Tese (Doutorado) - Curso de Psicologia, Universidade de São Paulo, Ribeirão Preto, 2008. Disponível em: https://www.teses.usp.br/teses/disponiveis/59/59137/tde-08102008-013413/pt-br.php. Acesso em: 21 jan. 2022.

RATOVOHERY, S. *et al*. Music as a mnemonic strategy to mitigate verbal episodic memory in Alzheimer's disease: does musical valence matter? **Journal of Clinical and Experimental Neuropsychology**, v. 41, n. 10, p. 1060-1073, ago. 2019. Disponível em: https://www.tandfonline.com/doi/full/10.1080/13803395.2019.1650897. Acesso em: 24 jul. 2020.

REZAEIANJOUYBARI, B.; SHANG, Y. Deep learning for prognostics and health management: state of the art, challenges, and opportunities. **Measurement**, v. 163, p. 1-29, out. 2020. Disponível em: https://doi.org/10.1016/j.measurement.2020.107929. Acesso em: 04 jun. 2021.

ROBAZZA, C.; MACALUSO, C.; D'URSO, V. Emotional reactions to music by gender, age, and expertise. **Perceptual and Motor skills**, v. 79, n. 2, p. 939-944, out. 1994. Disponível em: https://doi.org/10.2466%2Fpms.1994.79.2.939. Acesso em: 25 maio 2022.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: convolutional networks for biomedical image segmentation. *In*: INTERNATIONAL CONFERENCE ON MEDICAL IMAGE COMPUTING AND COMPUTER-ASSISTED INTERVENTION, 18., 2015, Munique. **Proceedings[...] .** Munique: Springer, 2015. p. 234-241. Disponível em: https://doi.org/10.1007/978-3-319-24574-4_28. Acesso em: 04 jan. 2022.

ROSENWEIN, B. H.; CRISTIANI, R. **What is the history of emotions?** Cambridge: Polity Press, 2018.

RUBIN, D. C.; TALARICO, J. M. A comparison of dimensional models of emotion: evidence from emotions, prototypical events, autobiographical memories, and words. **Memory**, v. 17, n. 8, p. 802-808, 2009. Disponível em: https://dx.doi.org/10.1080%2F09658210903130764. Acesso em: 12 abr. 2021.

RUSSELL, J. A. A circumplex model of affect. **Journal of Personality and Social Psychology**, v. 39, n. 6, p. 1161-1178, dez. 1980. Disponível em: https://psycnet.apa.org/doi/10.1037/h0077714. Acesso em: 07 set. 2020.

RUSSELL, J. A.; MEHRABIAN, A. Evidence for a three-factor theory of emotions. **Journal of Research in Personality**, v. 11, n. 3, p. 273-294, set. 1977. Disponível em: https://doi.org/10.1016/0092-6566(77)90037-X. Acesso em: 13 abr. 2021.

RUSSO, M. *et al*. Cochleogram-based approach for detecting perceived emotions in music. **Information Processing & Management**, v. 57, n. 5, p. 1-17, set. 2020. Disponível em: http://dx.doi.org/10.1016/j.ipm.2020.102270. Acesso em: 02 ago. 2020.

RUSTAM, F. *et al*. A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. **Plos One**, v. 16, n. 2, p.1-23, fev. 2021. Disponível em: https://doi.org/10.1371/journal.pone.0245909. Acesso em: 17 dez. 2021.

SAARIMÄKI, H. *et al*. Discrete neural signatures of basic emotions. **Cerebral Cortex**, v. 26, n. 6, p. 2563-2573, jun. 2016. Disponível em: https://doi.org/10.1093/cercor/bhv086. Acesso em: 06 abr. 2021.

SAARIKALLIO, S.; VUOSKOSKI, J.; LUCK, G. Adolescents' expression and perception of emotion in music reflects their broader abilities of emotional communication. **Psychology of Well-Being**, v. 4, n. 1, p. 1-16, dez. 2014. Disponível em: https://doi.org/10.1186/s13612-014-0021-8. Acesso em: 26 ago. 2021.

SAHU, R. K. *et al*. Impact of input feature selection on groundwater level prediction from a multi-layer perceptron neural network. **Frontiers in Water**, v. 2, p. 1-15, nov. 2020. Disponível em: https://doi.org/10.3389/frwa.2020.573034. Acesso em: 18 mar. 2023.

SALAMON, J.; GÓMEZ, E. Melody extraction from polyphonic music signals using pitch contour characteristics. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 20, n. 6, p. 1759-1770, fev. 2012. Disponível em: https://doi.org/10.1109/TASL.2012.2188515. Acesso em: 22 dez. 2021.

SANA, S. K. et al. Facial emotion recognition based music system using convolutional neural networks**. Materials Today: Proceedings**, v. 62, p. 4699-4706, mar. 2022. Disponível em: https://doi.org/10.1016/j.matpr.2022.03.131. Acesso em: 20 out. 2023.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. P. B. **Metodologia de pesquisa**. 5. ed. Porto Alegre: Penso, 2013.

SANTANA, I. A. P. *et al*. Music4all: a new music database and its applications. In: INTERNATIONAL CONFERENCE ON SYSTEMS, SIGNALS AND IMAGE PROCESSING (IWSSIP), 27., 2020, Niterói. **Proceedings[...] .** Niterói: IEEE, 2020. p. 399-404. Disponível em: https://doi.org/10.1109/IWSSIP48289.2020.9145170. Acesso em: 31 ago. 2021.

SANTOS, C. L.; SILLA JR., C. N. The latin music mood database. **EURASIP Journal on Audio, Speech, and Music Processing**, v. 2015, n. 23, p. 1-11, ago. 2015. Disponível em: https://doi.org/10.1186/s13636-015-0065-6. Acesso em: 07 set. 2020.

SARKAR, R. *et al*. Recognition of emotion in music based on deep convolutional neural network. **Multimedia Tools and Applications**, v. 79, n. 1-2, p. 765-783, set.

2019. Disponível em: https://doi.org/10.1007/s11042-019-08192-x. Acesso em: 02 ago. 2020.

SCARANTINO, A. The philosophy of emotions and its impact on affective science. In: BARRETT, L. F.; LEWIS, M.; HAVILAND-JONES, J. M. **Handbook of emotions**, Nova Iorque: The Guilford Press, 2016. p. 3-47.

SCHEDL, M. *et al*. Current challenges and visions in music recommender systems research. **International Journal of Multimedia Information Retrieval**, v. 7, n. 2, p. 95-116, abr. 2018. Disponível em: https://link.springer.com/article/10.1007%2Fs13735-018-0154-2. Acesso em: 23 jul. 2020.

SCHEDL, M. Deep learning in music recommendation systems. **Frontiers in Applied Mathematics and Statistics**, v. 5, p. 44, ago. 2019. Disponível em: https://doi.org/10.3389/fams.2019.00044. Acesso em: 24 set. 2020.

SCHERER, K. R. Psychological models of emotion. *In:* BOROD, J. C. (ed.). **The neuropsychology of emotion.** Nova Iorque: Oxford University Press, 2000. p. 137-162.

SCHERER, K. R. What are emotions? And how can they be measured? **Social Science Information**, v. 44, n. 4, p. 695-729, dez. 2005. Disponível em: https://doi.org/10.1177%2F0539018405058216. Acesso em: 27 jan. 2021.

SCHLOSBERG, H. Three dimensions of emotion. **Psychological Review**, v. 61, n. 2, p. 81-88, 1954. Disponível em: https://psycnet.apa.org/doi/10.1037/h0054570. Acesso em: 12 abr. 2021.

SCHMEICHEL, B. J.; INZLICHT, M. Incidental and integral effects of emotions on self-control. In: ROBINSON, M. D; WATKINS, E. R.; HARMON-JONES, E. **Handbook of cognition and emotion**, Nova Iorque: The Guilford Press, 2013. p. 272-290.

SCHMIDT, E. M.; KIM, Y. E. Learning emotion-based acoustic features with deep belief networks. In: IEEE WORKSHOP ON APPLICATIONS OF SIGNAL PROCESSING TO AUDIO AND ACOUSTICS (WASPAA), 8., 2011, New Paltz. **Proceedings[...] .** New Paltz: IEEE, 2011. p. 65-68. Disponível em: https://ieeexplore.ieee.org/document/6082328. Acesso em: 28 set. 2020.

SCHUBERT, E. Update of the Hevner adjective checklist. **Perceptual and Motor Skills**, v. 96, p. 1117-1122, jun. 2003. Disponível em: https://doi.org/10.2466%2Fpms.2003.96.3c.1117. Acesso em: 17 maio 2021.

SCHULLER, B.; DORFNER, J.; RIGOLL, G. Determination of nonprototypical valence and arousal in popular music: features and performances. **EURASIP Journal on Audio, Speech, and Music Processing**, v. 2010, p. 1-19, abr. 2010. Disponível em: https://link.springer.com/article/10.1155/2010/735854. Acesso em: 07 set. 2020.

SCHULLER, B. *et al*. ‘Mister DJ, cheer me up!’: musical and textual features for automatic mood classification. **Journal of New Music Research**, v. 39, n. 1, p. 13-34, abr. 2010. Disponível em: https://doi.org/10.1080/09298210903430475. Acesso em: 07 set. 2020.

SCHULLER, B.; WENINGER, F.; DORFNER, J. Multi-modal non-prototypical music mood analysis in continuous space: reliability and performances. In: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL CONFERENCE, 12., 2011, Miami. **Proceedings[...] .** Miami: ISMIR, 2011. p. 1-6. Disponível em: https://mediatum.ub.tum.de/doc/1107308/file.pdf. Acesso em: 07 set. 2020.

SHACKMAN, A. J.; WAGER, T. D. The emotional brain: fundamental questions and strategies for future research. **Neuroscience Letters**, v.693, p. 68-74, fev. 2019. Disponível em: https://doi.org/10.1016/j.neulet.2018.10.012. Acesso em: 03 fev. 2021.

SHARIFF, A. F.; TRACY, J. L. What are emotion expressions for? **Current Directions in Psychological Science**, v. 20, n. 6, p. 395-399, dez. 2011. Disponível em: https://doi.org/10.1177%2F0963721411424739. Acesso em: 15 fev. 2021.

SHI, W.; FENG, S. Research on music emotion classification based on lyrics and audio. In: IEEE 3RD ADVANCED INFORMATION TECHNOLOGY, ELECTRONIC AND AUTOMATION CONTROL CONFERENCE, 3., 2018, Chongqing. **Proceedings[...] .** Chongqing: IEEE, 2018. p. 1154 - 1159. Disponível em: https://ieeexplore.ieee.org/document/8577944. Acesso em: 24 jul.2020.

SHISHEHGAR, S. *et al*. Health and socio-cultural experiences of refugee women: an integrative review. **Journal of Immigrant and Minority Health**, v. 19, n. 4, p. 959-973, ago. 2017. Disponível em: https://doi.org/10.1007/s10903-016-0379-1. Acesso em: 27 jul. 2021.

SHIRAISHI, N.; REILLY, J. Content analysis of the emotions affecting caregivers of relatives with schizophrenia. **Current Psychology**, p. 1-11, nov. 2020. Disponível em: https://doi.org/10.1007/s12144-020-01185-2. Acesso em: 17 abr. 2021.

SHWARTZ-ZIV, R.; ARMON, A. Tabular data: deep learning is not all you need. **Information Fusion**, v. 81, p. 84-90, maio 2022. Disponível em: https://doi.org/10.1016/j.inffus.2021.11.011. Acesso em: 17 fev 2022.

SIDOROV, G. *et al*. Syntactic n-grams as machine learning features for natural language processing. **Expert Systems with Applications**, v. 41, n. 3, p. 853-860, fev. 2014. Disponível em: https://doi.org/10.1016/j.eswa.2013.08.015. Acesso em: 29 dez. 2021.

SIEDLECKA, E.; DENSON, T. F. Experimental methods for inducing basic emotions: a qualitative review. **Emotion Review**, v. 11, n. 1, p. 87-97, mar 2018. Disponível em: https://doi.org/10.1177%2F1754073917749016. Acesso em: 09 ago. 2021.

SIMONETTA, F.; NTALAMPIRAS, S.; AVANZINI, F. Multimodal music information processing and retrieval: survey and future challenges. In: INTERNATIONAL WORKSHOP ON MULTILAYER MUSIC REPRESENTATION AND PROCESSING, 1., 2019, Milão. **Proceedings[...] .** Milão: IEEE, 2019. p. 10-18. Disponível em: https://ieeexplore.ieee.org/document/8665366. Acesso em: 23 jul. 2020.

SOLEYMANI, M. et al. A survey of multimodal sentiment analysis. **Image and Vision Computing**, v. 65, p. 3-14, set. 2017. Disponível em: https://www.sciencedirect.com/science/article/pii/S0262885617301191. Acesso em: 17 dez. 2020.

SOLOMON, J. W. **Music theory essentials**: a streamlined approach to fundamentals, tonal harmony, and post-tonal materials. Nova Iorque: Routledge, 2019.

SONG, Y.; DIXON, S.; PEARCE, M. Evaluation of musical features for emotion classification. *In*: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL, 13., 2012, Porto. **Proceedings[...] .** Porto: ISMIR, 2012. p. 523-528. Disponível em: https://ismir2012.ismir.net/event/papers/523_ISMIR_2012.pdf. Acesso em: 06 maio 2021.

STEWART, K.; KOH, H. E. Hooked on a feeling: the effect of music tempo on attitudes and the mediating role of consumers' affective responses. **Journal of Consumer Behaviour**, v. 16, n. 6, p. 550-564, jul. 2017. Disponível em: https://doi.org/10.1002/cb.1665. Acesso em: 26 jun. 2021.

STREICH, S.; HERRERA, P. Detrended Fluctuation Analysis of music signals: danceability estimation and further semantic characterization. *In*: AES 118TH CONVENTION, 118., 2005, Barcelona. **Proceedings[...] .** Barcelona: AES, 2005. p. 1-9. Disponível em: http://www.iua.upf.edu/mtg/publications/41c4a8-AES118-sstreich.pdf. Acesso em: 22 dez. 2021.

SUN, X.; LIU, M.; SIMA, Z. A novel cryptocurrency price trend forecasting model based on LightGBM. **Finance Research Letters**, v. 32, p. 101084, jan. 2020. Disponível em: https://doi.org/10.1016/j.frl.2018.12.032. Acesso em: 25 mar. 2023.

TALARICO, J. M.; LABAR, K. S.; RUBIN, D. C. Emotional intensity predicts autobiographical memory experience. **Memory & Cognition**, v. 32, n. 7, p. 1118-1132, out. 2004. Disponível em: https://doi.org/10.3758/BF03196886. Acesso em: 15 abr. 2021.

TERRY, P. C. *et al*. Effects of music in exercise and sport: a meta-analytic review. **Psychological Bulletin**, v. 146, n. 2, p. 91-117, fev. 2020. Disponível em: https://psycnet.apa.org/doi/10.1037/bul0000216. Acesso em: 26 jun. 2021.

THAMMASAN, N.; FUKUI, K. I.; NUMAO, M. Application of deep belief networks in EEG-based dynamic music-emotion recognition. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN), 26., 2016, Vancouver. **Proceedings[...] .** Vancouver: IEEE, 2016. p. 881-888. Disponível em: https://ieeexplore.ieee.org/document/7727292. Acesso em: 28 set. 2020.

THAMMASAN, N.; FUKUI, K.; NUMAO, M. Multimodal fusion of EEG and musical features in music-emotion recognition. In: AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, 31., 2017, São Francisco. **Proceedings[…] .** São Francisco: AAAI, 2017. p. 4991 - 4992. Disponível em: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14831/14237. Acesso em: 24 jul. 2020.

THAYER, R. E. **The biopsychology of mood and arousal**. Nova Iorque: Oxford University Press, 1989.

THAYER, R. E. Moods of energy and tension that. *In*: RYAN, R. M. (ed.). **The Oxford handbook of human motivation**. Nova Iorque: Oxford University Press, 2012. p. 408-419.

TOMEK, I. Two Modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics, v. SMC-6, n. 11, p. 769-772, nov. 1976. Disponível em: https://doi.org/10.1109/TSMC.1976.4309452. Acesso em: 26 mar. 2023.

TOMKINS, S. S. Affect theory. *In*: SCHERER, K. R.; EKMAN, P. **Approaches to emotion**. Hillsdale: Erlbaum, 1984, p. 163-195.
TROHIDIS, K. *et al*. Multi-label classification of music into emotions. *In*: INTERNATIONAL SOCIETY FOR MUSIC INFORMATION RETRIEVAL, 9., 2008, Filadélfia. **Proceedings[...] .** Filadélfia: ISMIR, 2008. p. 325-330. Disponível em: https://ismir2008.ismir.net/papers/ISMIR2008_275.pdf. Acesso em: 23 maio 2021.

TROHIDIS, K. *et al*. Multi-label classification of music by emotion. **EURASIP Journal on Audio, Speech, and Music Processing**, v. 2011, n. 1, p. 1-9, set. 2011. Disponível em: https://doi.org/10.1186/1687-4722-2011-426793. Acesso em: 23 maio 2021.

TUHANIOĞLU, B. *et al*. Efeito de ondas magnéticas pulsadas de dosagem extremamente baixa na cóclea. **Brazilian Journal of Otorhinolaryngology**, v. 85, n. 3, p. 282-289, maio-jun. 2019. Disponível em: https://www.scielo.br/j/bjorl/a/cF65RRB5FCcGyqvkDRnnVkc/?lang=pt. Acesso em: 10 jan. 2022.

TURKOGLU, B.; KAYA, E. Training multi-layer perceptron with artificial algae algorithm. **Engineering Science and Technology, an International Journal**, v. 23, n. 6, p. 1342-1350, dez. 2020. Disponível em: https://doi.org/10.1016/j.jestch.2020.07.001. Acesso em 18 maio 2022.

TYNG, C. M. *et al*. The influences of emotion on learning and memory. **Frontiers in psychology**, v. 8, p. 1-22, ago. 2017. Disponível em: https://doi.org/10.3389/fpsyg.2017.01454. Acesso em: 08 fev. 2021.

TZANETAKIS, G.; COOK, P. Marsyas: A framework for audio analysis. **Organised Sound**, v. 4, n. 3, p. 169-175, dez. 2000. Disponível em: https://doi.org/10.1017/S1355771800003071. Acesso em: 03 maio 2021.

VAPNIK, V. N. **The nature of statistical learning theory**. Nova Iorque: Springer-Verlag, 1995.

VASWANI, A. *et al*. Attention is all you need. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2017), 31., 2017, Long Beach. **Proceedings [...] .** Long Beach: NeurIPS, 2017. p. 1-11. Disponível em: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Acesso em: 13 maio 2022.

VERGARA, S. C. **Projetos e relatórios de pesquisa em Administração**. 5. ed. São Paulo: Atlas, 2004.

VIDAS, D.; NELSON, N. L.; DINGLE, G. A. Efficacy of the Tuned In music emotion regulation program in international university students. **Psychology & Health**, p. 1-21, mar. 2023. Disponível em: https://doi.org/10.1080/08870446.2023.2197007. Acesso em: 20 out. 2023.

VIEILLARD, S. *et al*. Happy, sad, scary and peaceful musical excerpts for research on emotions. **Cognition & Emotion**, v. 22, n. 4, p. 720-752, maio 2008. Disponível em: https://doi.org/10.1080/02699930701503567. Acesso em: 25 ago. 2021.

VILLAR, J. D. L.; LUJÁN, R. C. Reconocimiento de emociones musicales a través de datos y tecnologías digitales. **Comunicación y Hombre**, n. 16, p. 59-82, mar. 2020. Disponível em: https://doi.org/10.32466/eufv-cyh.2020.16.563.59-82. Acesso em: 10 ago. 2020.

VIPER, M.; THYRÉN, D.; HORWITZ, E.B. Music as consolation—the importance of music at farewells and mourning. **Omega - Journal of Death and Dying**, p. 1-23, jul. 2020. Disponível em: https://journals.sagepub.com/doi/10.1177/0030222820942391. Acesso em: 09 ago. 2020.

WAN, S. *et al*. Deep learning models for real-time human activity recognition with smartphones. **Mobile Networks and Applications**, v. 25, n. 2, p. 743-755, abr. 2020. Disponível em: https://doi.org/10.1007/s11036-019-01445-x. Acesso em: 04 jun. 2021.

WANG, J. *et al*. Enriching music mood annotation by semantic association reasoning. In: IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO, 11., 2010, Singapura. **Proceedings[...] .** Singapura: IEEE, 2010. p. 1445 - 1450. Disponível em: https://ieeexplore.ieee.org/document/5583243. Acesso em: 24 jul. 2020.

WANG, J. C. *et al*. Modeling the affective content of music with a Gaussian mixture model. **IEEE Transactions on Affective Computing**, v. 6, n. 1, p. 56-68, jan-mar. 2015. Disponível em: https://doi.org/10.1109/TAFFC.2015.2397457. Acesso em: 23 maio 2021.

WANG, J. C. *et al*. The acoustic emotion gaussians model for emotion-based music annotation and retrieval. *In*: ACM INTERNATIONAL CONFERENCE ON

MULTIMEDIA, 20., 2012, Nara. **Proceedings[...] .** Nara: ACM, 2012. p. 89-98. Disponível em: https://doi.org/10.1145/2393347.2393367. Acesso em: 23 maio 2021.

WANG, H.C.; SYU, S. W.; WONGCHAISUWAT, P. A method of music autotagging based on audio and lyrics. **Multimedia Tools and Applications**, p. 1-29, abr. 2021. Disponível em: https://doi.org/10.1007/s11042-020-10381-y. Acesso em: 12 maio 2021.

WANG, S. *et al*. Emo2Vec: learning emotional embeddings via multi-emotion category. **ACM Transactions on Internet Technology (TOIT)**, v. 20, n. 2, p. 1-17, mar. 2020. Disponível em: https://doi.org/10.1145/3372152. Acesso em: 20 abr. 2021.

WANG, Y. *et al*. Perceptual learning and recognition confusion reveal the underlying relationships among the six basic emotions. **Cognition and Emotion**, v. 33, n. 4, p. 754-767, 2019. Disponível em: https://doi.org/10.1080/02699931.2018.1491831. Acesso em: 07 abr. 2021.

WANG, Y.; SUN, S. Emotion recognition for internet music by multiple classifiers. *In*: INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION SCIENCE (ICIS), 18., 2019, Pequim. **Proceedings[...] .** Pequim: IEEE, 2019. p. 262-265. Disponível em: https://doi.org/10.1109/ICIS46139.2019.8940288. Acesso em: 23 maio 2021.

WATSON, D.; TELLEGEN, A. Toward a consensual structure of mood. **Psychological Bulletin**, v. 98, n. 2, p. 219-235, 1985. Disponível em: https://psycnet.apa.org/doi/10.1037/0033-2909.98.2.219. Acesso em: 12 abr. 2021.
WATSON, J. B. **Behaviorism**. Chicago: University of Chicago Press. 1930.

WARRENBURG, L. A. Comparing musical and psychological emotion theories. **Psychomusicology: Music, Mind, and Brain**, v. 30, n. 1, p. 1, 2020. Disponível em: https://doi.apa.org/doi/10.1037/pmu0000247. Acesso em: 03 ago. 2021.

WAZIRALI, R. An improved intrusion detection system based on K-NN hyperparameter tuning and cross-validation**. Arabian Journal for Science and Engineering**, v. 45, n. 12, p. 10859-10873, ago. 2020. Disponível em: https://doi.org/10.1007/s13369-020-04907-7. Acesso em: 25 mar. 2023.

WEINER, B.; GRAHAM, S. An attributional approach to emotional development. *In*: IZARD, C. E.; KAGAN, J.; ZAJONC, R. B. **Emotions, cognition, and behavior**. Nova Iorque: Cambridge University Press, 1984. p. 167-191.

WELLS, A.; HAKANEN, E. A. The emotional use of popular music by adolescents. **Journalism Quarterly**, v. 68, n. 3, p. 445-454, set. 1991. Disponível em: https://doi.org/10.1177%2F107769909106800315. Acesso em: 25 ago. 2021.

WINGENBACH, T. S.H. *et al*. Perception of discrete emotions in others: evidence for distinct facial mimicry patterns. **Scientific Reports**, v. 10, n. 1, p. 1-13, mar. 2020. Disponível em: https://doi.org/10.1038/s41598-020-61563-5. Acesso em: 06 abr. 2021.

WORSLEY, M.; BLIKSTEIN, P. Leveraging multimodal learning analytics to differentiate student learning strategies. In: INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS AND KNOWLEDGE, 5, 2015, Nova Iorque. **Proceedings[…] .** Nova Iorque: ACM, 2015, p. 360 - 367. Disponível em: https://dl.acm.org/doi/10.1145/2723576.2723624. Acesso em: 17 dez. 2020.

WU, B. *et al*. Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. *In*: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 22., 2014, Orlando. **Proceedings[...] .** Orlando: ACM, 2014. p. 117-126. Disponível em: https://doi.org/10.1145/2647868.2654904. Acesso em: 23 maio 2021.

WU, Y. T. *et al*. Omnizart: a general toolbox for automatic music transcription. **Journal of Open Source Software**, v. 6, n. 68, p. 1-5, dez. 2021. Disponível em: https://joss.theoj.org/papers/10.21105/joss.03391. Acesso em: 13 dez. 2021.

WU, Y. T.; CHEN, B.; SU, L. Polyphonic music transcription with semantic segmentation. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP 2019), 44., 2019, Londres. **Proceedings[...] .** Londres: IEEE, 2019. p 166-170. Disponível em: https://doi.org/10.1109/ICASSP.2019.8682605. Acesso em: 04 jan. 2022.

WU, Y. T.; CHEN, B.; SU, L. Multi-instrument automatic music transcription with self-attention-based instance segmentation. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 28, p. 2796-2809, out. 2020. Disponível em: https://doi.org/10.1109/TASLP.2020.3030482. Acesso em: 04 jan. 2022.

WUNDT, W. **Grundriss der Psychologie.** 15 ed. Leipzig: Alfred Kröner Verlag, 1922.

XIAO, Z. et al. What is the best segment duration for music mood analysis? In: INTERNATIONAL WORKSHOP ON CONTENT-BASED MULTIMEDIA INDEXING, 2., 2008, Londres. **Proceedings [...] .** Londres: IEEE, 2008. p. 17-24. Disponível em: https://doi.org/10.1109/CBMI.2008.4564922. Acesso em: 26 maio 2022.

XIE, B.; KIM, J. C.; PARK, C. H. Musical emotion recognition with spectral feature extraction based on a sinusoidal model with model-based and deep-learning approaches. **Applied Sciences**, v. 10, n. 3, p. 1-11, jan. 2020. Disponível em: https://doi.org/10.3390/app10030902. Acesso em: 24 set. 2020.

XIN, Y. *et al*. Machine learning and deep learning methods for cybersecurity. **IEEE Access**, v. 6, p. 35365-35381, maio 2018. Disponível em: https://doi.org/10.1109/ACCESS.2018.2836950. Acesso em: 02 jun. 2021.

XU, L. *et al*. Effects of individual factors on perceived emotion and felt emotion of music: based on machine learning methods. **Psychology of Music**, v. 0, n. 00, p. 1-19, jul. 2020. Disponível em: https://doi.org/10.1177%2F0305735620928422. Acesso em: 09 ago. 2020.

YADATI, K. *et al*. On the automatic identification of music for common activities. In: 2017 ACM ON INTERNATIONAL CONFERENCE ON MULTIMEDIA RETRIEVAL, 7.,

2017, Bucareste. **Proceedings [...] .** Bucareste: ACM, 2017. p. 192-200. Disponível em: https://dl.acm.org/doi/10.1145/3078971.3078997. Acesso em: 10 ago. 2020.

YAMAK, P. T.; YUJIAN, L.; GADOSEY, P. K. A comparison between ARIMA, LSTM, and GRU for time series forecasting. *In*: INTERNATIONAL CONFERENCE ON ALGORITHMS, COMPUTING AND ARTIFICIAL INTELLIGENCE, 2., 2019, Sanya. **Proceedings[...] .** Sanya: ACM, 2019. p. 49-55. Disponível em: https://doi.org/10.1145/3377713.3377722. Acesso em: 17 jun. 2021.

YANG, D.; LEE, W.S. Music emotion identification from lyrics. In: IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA, 11., 2009, San Diego. **Proceedings[...] .** San Diego: IEEE, 2009. p. 624-629. Disponível em: https://doi.org/10.1109/ISM.2009.123. Acesso em: 30 ago. 2021.

YANG, X.; DONG, Y.; LI, J. Review of data features-based music emotion recognition methods. **Multimedia Systems**, v. 24, n. 4, p. 365-389, ago. 2017. Disponível em: https://link.springer.com/article/10.1007/s00530-017-0559-4. Acesso em: 23 jul. 2020.

YANG, W. *et al*. Affective auditory stimulus database: an expanded version of the International Affective Digitized Sounds (IADS-E). **Behavior Research Methods**, v. 50, n. 4, p. 1415-1429, mar. 2018. Disponível em: https://doi.org/10.3758/s13428-018-1027-6. Acesso em: 13 dez. 2021.

YANG, Y. H. *et al*. A regression approach to music emotion recognition. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 16, n. 2, p. 448-457, fev. 2008. Disponível em: https://doi.org/10.1109/TASL.2007.911513. Acesso em: 23 maio 2021.

YANG, Y. H. *et al*. Music emotion recognition: the role of individuality. *In*: INTERNATIONAL WORKSHOP ON HUMAN-CENTERED MULTIMEDIA, 15., 2007, Augsburg. **Proceedings[...] .** Augsburg: ACM, 2007. p. 13-22. Disponível em: https://doi.org/10.1145/1290128.1290132. Acesso em: 11 maio 2021.

YANG, Y. H.; CHEN, H. H. **Music emotion recognition**. Boca Raton: CRC Press, 2011.

YANG, Y.; CHEN, H. H. Machine recognition of music emotion: a review. **ACM Transactions on Intelligent Systems and Technology**, v. 3, n. 3, p.1-30, maio 2012. Disponível em: https://doi.org/10.1145/2168752.2168754. Acesso em: 24 jul. 2020.

YANG, Y. H.; LIU, C. C; CHEN, H. H. Music emotion classification: a fuzzy approach. *In*: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 14., 2006, Santa Bárbara. **Proceedings[...] .** Santa Bárbara: ACM, 2006. p. 81-84. Disponível em: https://doi.org/10.1145/1180639.1180665. Acesso em: 23 maio 2021.

YOUNG, T. *et al*. Recent trends in deep learning based natural language processing. **IEEE Computational Intelligence Magazine**, v. 13, n. 3, p. 55-75, jul. 2018. Disponível em: https://doi.org/10.1109/MCI.2018.2840738. Acesso em: 20 jun. 2021.

YU, Y. *et al*. Deep attention based music genre classification. **Neurocomputing**, v. 372, p. 84-91, jan. 2020. Disponível em: https://doi.org/10.1016/j.neucom.2019.09.054. Acesso em: 04 ago. 2020.

ZAMACOIS, J. **Teoría de la música:** dividida en cursos. Madrid: Mundimúsica Ediciones, 2007.

ZENTNER, M.; GRANDJEAN, D.; SCHERER, Klaus R. Emotions evoked by the sound of music: characterization, classification, and measurement. **Emotion**, v. 8, n. 4, p. 494-521, 2008. Disponível em: https://psycnet.apa.org/doi/10.1037/1528-3542.8.4.494. Acesso em: 03 maio 2021.

ZHANG, J. L. *et al*. Feature selection and feature learning in arousal dimension of music emotion by using shrinkage methods. **Multimedia Systems**, v. 23, n. 2, p. 251-264, mar. 2017. Disponível em: https://doi.org/10.1007/s00530-015-0489-y. Acesso em: 03 maio.

ZHANG, K. *et al*. The PMEmo dataset for music emotion recognition. In: 2018 ACM ON INTERNATIONAL CONFERENCE ON MULTIMEDIA RETRIEVAL, 18., 2018, Yokohama. **Proceedings[...] .** Yokohama: ACM, 2018. p. 135-142. Disponível em: https://dl.acm.org/doi/10.1145/3206025.3206037. Acesso em: 30 jul. 2020.

ZHANG, L. *et al*. Using psychophysiological measures to recognize personal music emotional experience. **Frontiers of Information Technology & Electronic Engineering**, v. 20, n. 7, p. 964-974, jul. 2019. Disponível em: https://doi.org/10.1631/FITEE.1800101. Acesso em: 11 maio 2021.

ZHANG, Q. *et al*. A survey on deep learning for big data. **Information Fusion**, v. 42, p. 146-157, jul. 2018. Disponível em: https://doi.org/10.1016/j.inffus.2017.10.006. Acesso em: 31 maio 2021.

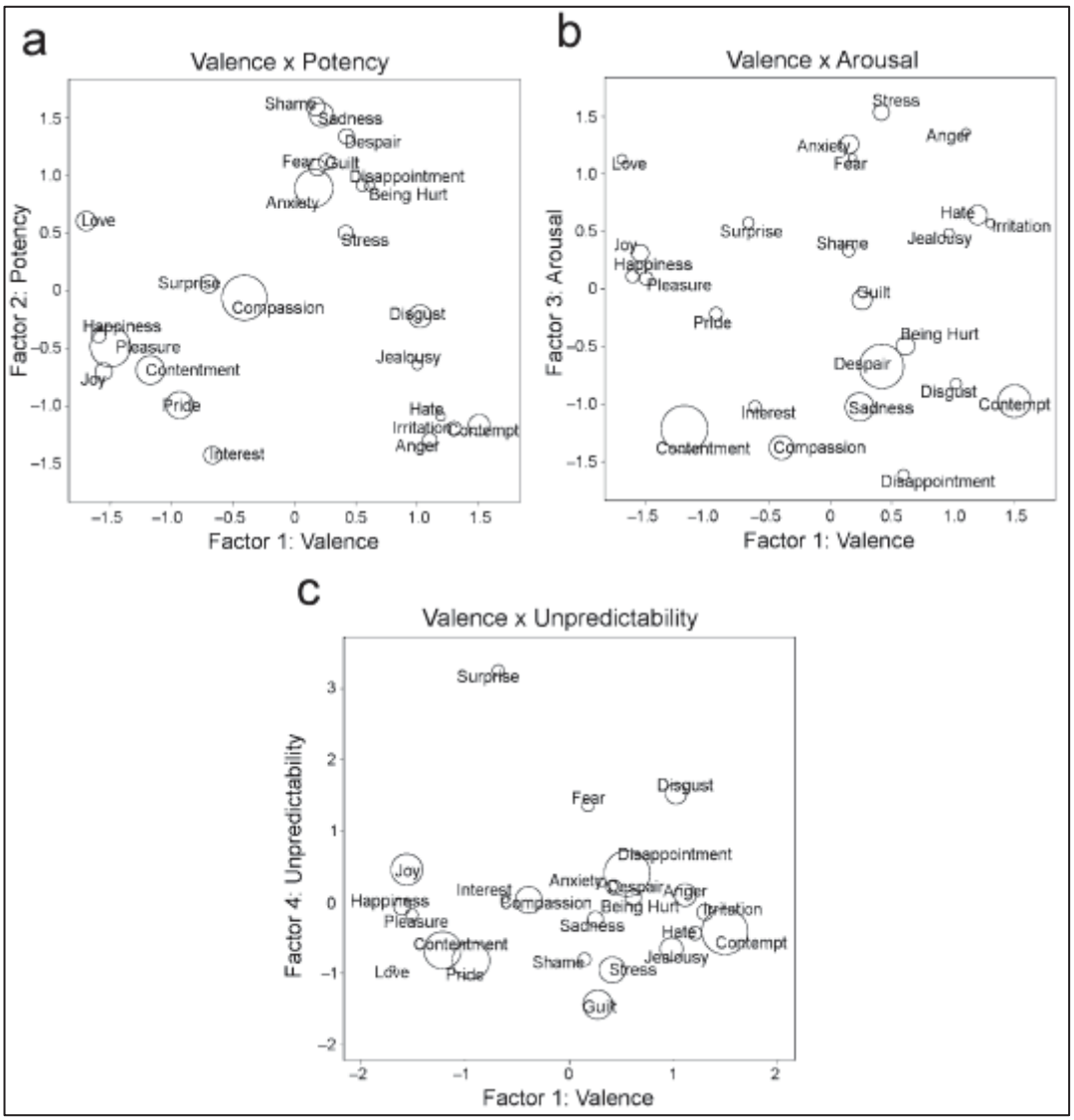
ZHAO, Y.; YANG, D.; CHEN, X. Multi-modal music mood classification using Co-Training. In: 2010 INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE AND SOFTWARE ENGINEERING, 3., 2010, Wuhan. **Proceedings [...] .** Wuhan: IEEE, 2010. p. 1-4. Disponível em: https://ieeexplore.ieee.org/document/5677056. Acesso em: 07 set. 2020.

ZHOU, J.; CHEN, X.; YANG, D. Multimodel music emotion recognition using unsupervised deep neural networks. In: CONFERENCE ON SOUND AND MUSIC TECHNOLOGY (CSMT), 6., 2019, Xiamen. **Proceedings[...] .** Xiamen: Springer, 2019. p. 27-39. Disponível em: https://doi.org/10.1007/978-981-13-8707-4_3. Acesso em: 28 set. 2020.

ZHOU, R.; LIU, F.; GRAVELLE, C. W. Deep learning for modulation recognition: a survey with a demonstration. **IEEE Access**, v. 8, p. 67366-67376, abr. 2020. Disponível em: https://doi.org/10.1109/ACCESS.2020.2986330. Acesso em: 29 maio 2021.

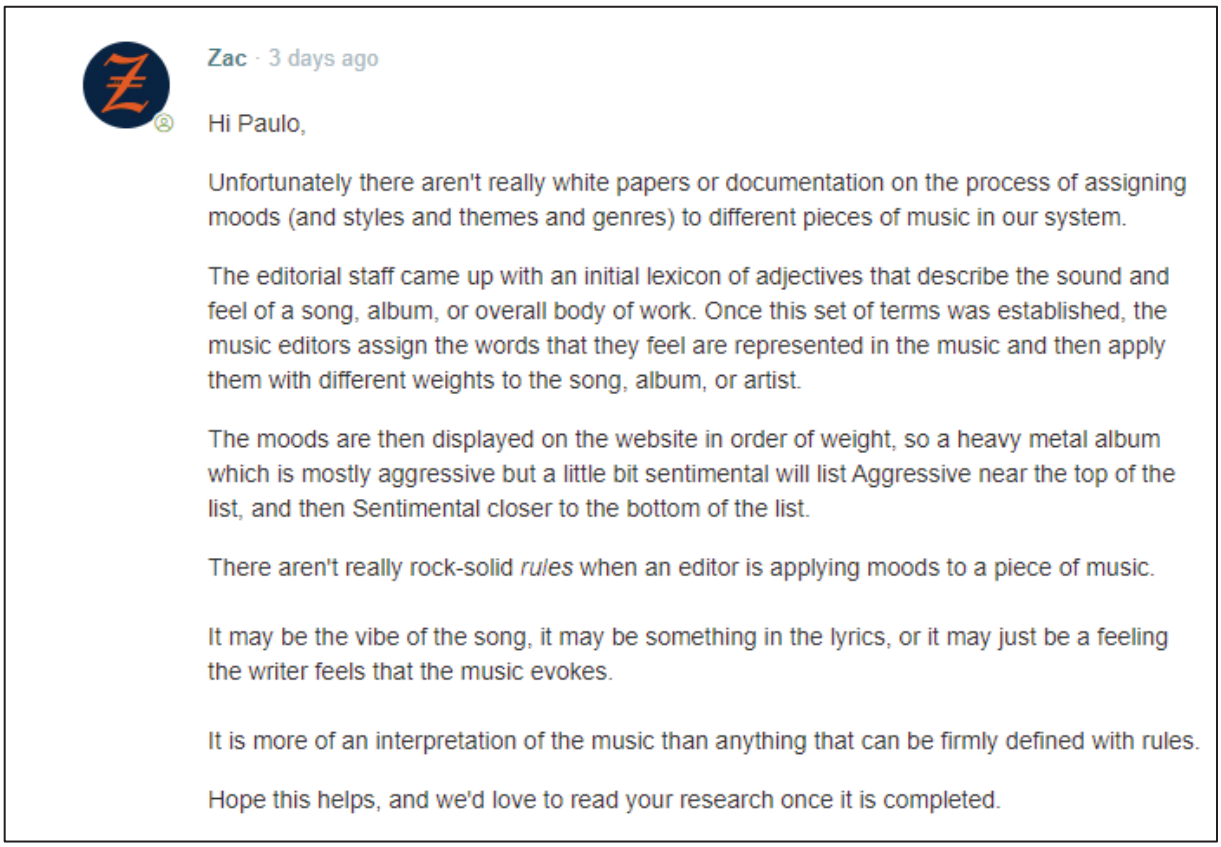
ZYCH, A. D.; GOGOLLA, N. Expressions of emotions across species. **Current Opinion in Neurobiology**, v. 68, p. 57-66, jun. 2021. Disponível em: https://doi.org/10.1016/j.conb.2021.01.003. Acesso em: 08 fev. 2021.

ANEXO A – MODELO QUADRIDIMENSIONAL (FONTAINE *et al*., 2007)



FONTE: Fontaine *et al*. (2007, p. 1055).

ANEXO B – CONTATO COM A PLATAFORMA *ALLMUSIC*



Zac · 3 days ago

Hi Paulo,

Unfortunately there aren't really white papers or documentation on the process of assigning moods (and styles and themes and genres) to different pieces of music in our system.

The editorial staff came up with an initial lexicon of adjectives that describe the sound and feel of a song, album, or overall body of work. Once this set of terms was established, the music editors assign the words that they feel are represented in the music and then apply them with different weights to the song, album, or artist.

The moods are then displayed on the website in order of weight, so a heavy metal album which is mostly aggressive but a little bit sentimental will list Aggressive near the top of the list, and then Sentimental closer to the bottom of the list.

There aren't really rock-solid *rules* when an editor is applying moods to a piece of music.

It may be the vibe of the song, it may be something in the lyrics, or it may just be a feeling the writer feels that the music evokes.

It is more of an interpretation of the music than anything that can be firmly defined with rules.

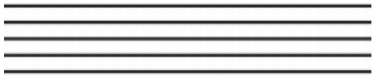
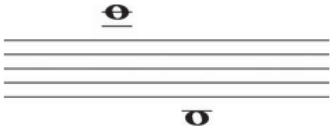
Hope this helps, and we'd love to read your research once it is completed.

FONTE: Fórum de esclarecimentos *AllMusic* (2021).

APÊNDICE A – COMPLEMENTO PARA OS FUNDAMENTOS DE TEORIA MUSICAL

A pauta musical representada possibilita a escrita de nove notas, sendo cinco nas linhas e quatro nos espaços (MILLER, 2016; SOLOMON, 2019), contudo, caso almeje-se representar notas mais agudas ou mais graves, faz-se necessário recorrer às linhas suplementares (MED, 2017), que podem ser acrescentadas acima ou abaixo da pauta (SOLOMON, 2019). Esses elementos são representados no QUADRO 41.

QUADRO 41 – REPRESENTAÇÃO DE PAUTA MUSICAL

Elemento	Representação	Fonte
Pauta musical		Harnum (2013, p.47).
Pauta musical com linhas suplementares		Solomon (2019, p.24).

FONTE: O autor (2021).

Entretanto, para a correta utilização da pauta musical, faz-se necessário acrescentar em seu início símbolos denominados claves (ZAMACOIS, 2007; MILLER, 2016), cuja função é estabelecer o nome e a altura das notas (PRIOLLI, 2006; MED, 2017). As claves empregadas são listadas no QUADRO 42.

QUADRO 42 - REPRESENTAÇÃO DAS CLAVES

Clave	Símbolo
Sol	
Fá	
Dó	

FONTE: Adaptado de Harnum (2013, p.191).

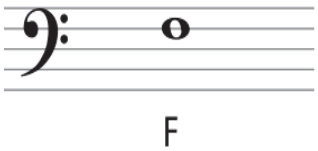
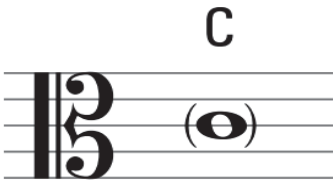
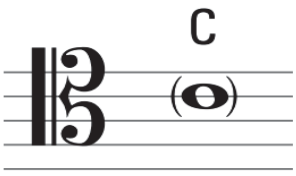
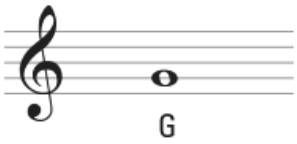
Em termos de posicionamento, a clave de Sol é escrita na segunda linha da pauta musical de baixo para cima (PRIOLLI, 2006; MED, 2017). Trata-se de uma clave destinada à representação das notas agudas (MILLER, 2016; SOLOMON, 2019), oriundas de instrumentos como, por exemplo, trompete, flauta e clarinete, além de partes relacionadas ao canto, como soprano e tenor (MILLER, 2016).

Quanto à clave de Fá, esta pode ser escrita na terceira (PRIOLLI, 2006; SOLOMON, 2019) ou na quarta linha (MILLER, 2016; MED, 2017) da pauta musical, de baixo para cima. O uso desta clave, normalmente, destina-se às notas graves (MED, 2017; PILHOFER; DAY, 2019), proporcionadas, por exemplo, por instrumentos como trombone, tuba e contrabaixo (MILLER, 2016; PILHOFER; DAY, 2019).

Por fim, em relação à clave de Dó, esta é um tipo especial de clave, cuja escrita pode ser realizada na primeira, segunda, terceira ou quarta linha da pauta musical (PRIOLLI, 2006). Todavia, as grafias mais comuns para esta clave consistem em representá-la na terceira linha e na quarta linha. Quando posicionada na terceira linha, a clave de Dó tem na viola o principal instrumento (PILHOFER; DAY, 2019; SOLOMON, 2019).

Ao ser grafada na quarta linha, a clave de Dó é geralmente empregada para instrumentos como o fagote e o trombone (PILHOFER; DAY, 2019; SOLOMON, 2019). A representação do posicionamento das claves é visualizada no QUADRO 43.

QUADRO 43 - EXEMPLO DE POSICIONAMENTO DAS CLAVES NA PAUTA MUSICAL

Elemento	Representação	Fonte
Clave de sol na pauta musical		Pilhofer e Day (2019, p.63).
Clave de fá na pauta musical		Pilhofer e Day (2019, p.64).
Clave de dó na terceira linha		Pilhofer e Day (2019, p.65)
Clave de dó na quarta linha		

FONTE: O autor (2021).

Posterior às claves, tem-se o conceito de valores - ou figuras – que são formas destinadas a representar as diferentes durações das notas musicais (PRIOLLI, 2006; PILHOFER; DAY, 2019). Nesse sentido, Med (2017) indica que o sistema musical atual apresenta sete figuras para representar os diferentes períodos de tempos de uma nota, sendo essas ilustradas na FIGURA 34.

FIGURA 34 - REPRESENTAÇÃO DE FIGURAS E PAUSAS

Nota	Semibreve	Mínima	Semínima	Colcheia	Semicolcheia	Fusa	Semifusa
Figura							
Pausa							

FONTE: Adaptado de Priolli (2006, p.13).

Cabe ressaltar que, além das figuras relacionadas à duração das notas, para cada uma dessas há um sinal denominado pausa, cujo objetivo é representar a duração dos silêncios (MILLER, 2016; SOLOMON, 2019). Em relação à duração proporcional das figuras, a semibreve, por apresentar atualmente a maior duração, é adotada como unidade de divisão (PRIOLLI, 2006; SOLOMON. 2019), conforme observa-se na FIGURA 35.

FIGURA 35 - DIVISÃO PROPORCIONAL DOS VALORES

Figura	1	2	4	8	16	32	64
Semibreve							
Mínima							
Semínima							
Colcheia							
Semicolcheia							
Fusa							
Semifusa							

FONTE: Adaptado de Med (2017, p.28).

Com base na FIGURA 35, é possível visualizar, por exemplo, que a semibreve engloba todas as demais figuras e, a partir desta, cada figura vale duas vezes a figura seguinte, com exceção à semifusa.

Somado aos valores, os acidentes - também denominados como alterações - são um conjunto de sinais grafados antes das notas (MED, 2017), cuja utilidade é aumentar ou diminuir a altura destas (PILHOFER; DAY, 2019). Os acidentes utilizados são apresentados no QUADRO 44.

QUADRO 44 - DESCRIÇÃO DOS ACIDENTES OU ALTERAÇÕES

Nome do acidente	Símbolo	Efeito
Sustenido	♯	Elevar a altura da nota em um semitom
Dobrado sustenido	𝄪	Elevar a altura da nota em dois semitons
Bequadro	♮	Anular os acidentes
Bemol	♭	Abaixar a altura da nota em um semitom
Dobrado bemol	𝄫	Abaixar a altura da nota em dois semitons

FONTE: Adaptado de Med (2017, p. 32-33).

Para a compreensão das alterações elencadas no QUADRO 44, é necessário definir os conceitos de tom e semitom. Neste sentido, semitom representa o “menor intervalo adotado entre duas notas na música ocidental (no sistema temperado)” (MED, 2017, p.30); por sua vez, tom refere-se ao “menor intervalo, entre dois sons, formado por dois semitons” (PRIOLLI, 2006, p.29).

Finalmente, o último conceito necessário para esta consiste na escala musical pode ser definida como uma ordem linear de notas (SOLOMON, 2019), ou, ainda, como “uma sucessão ascendente e descendente de notas diferentes consecutivas” (MED, 2017, p. 83). Em termos de classificação, as escalas musicais são divididas conforme apresenta-se no QUADRO 45.

QUADRO 45 - CLASSIFICAÇÃO DAS ESCALAS MUSICAIS

Quanto ao número de notas			
5 (pentatônica)	6 (hexacordal)	7 (heptatônica)	12 (artificial ou cromática)
Quanto à utilização			
Escalas naturais ou diatônicas	Escalas artificiais ou cromáticas	Escalas exóticas	Outras

FONTE: Adaptado de Med (2017, p.83).

Em relação às características, uma escala natural - ou diatônica - é uma escala com oito graus (QUADRO 46) – momento harmônico das notas que representam a escala -, sendo que o oitavo grau é a repetição do primeiro (PRIOLLI, 2006; MED, 2017).

QUADRO 46 – GRAUS DA ESCALA DIATÔNICA

Grau	Nome
I	Tônica
II	Supertônica
III	Mediante
IV	Subdominante
V	Dominante
VI	Superdominante
VII	Sensível/Subtônica
VIII	Tônica

FONTE: Adaptado de Med (2017, p. 85-86).

A respeito dos graus listados no QUADRO 46, é pertinente apresentar algumas particularidades: a) o primeiro grau (tônica) é o que dá nome à escala (PRIOLLI, 2006); b) os graus I, IV e V caracterizam o tom (MED, 2017); c) os graus III, VI e VII diferem escalas com a mesma tônica (MED, 2017); e d) o grau VII é denominado sensível quando encontra-se meio tom abaixo da tônica; quando está um tom abaixo da tônica, recebe o nome de subtônica (MED, 2017).

Além do mais, as escalas diatônicas são divididas em dois grupos: a) escalas antigas (modos); e b) escalas maiores e menores (MED, 2017). No que concerne à escala maior, esta caracteriza-se por possuir dois semitons (ST) e cinco tons (T) (PRIOLLI, 2006; SOLOMON, 2019).

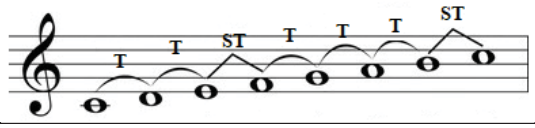
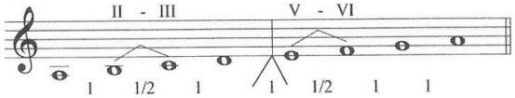
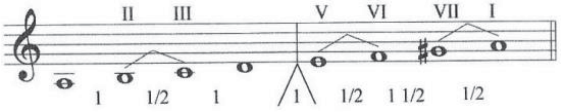
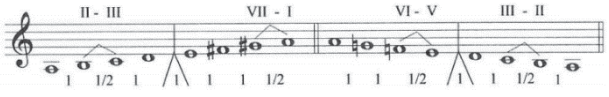
Em relação à escala menor, seu principal diferencial consiste em um intervalo de terça menor, localizado entre graus I e III (MED, 2017). Esta escala ainda se caracteriza por apresentar três formas: a) natural; b) harmônica; e c) melódica (MILLER, 2016; MED, 2017; PILHOFER; DAY, 2019). Uma escala menor natural é composta somente por notas naturais, e apresenta dois semitons, entre os graus: a) II e III; e b) V e VI. (MED, 2017).

A escala menor harmônica distingue-se da escala menor natural ao apresentar uma elevação de meio tom no grau VII (PRIOLLI, 2006; MILLER, 2016), ficando, desta maneira, com três semitons entre os graus: a) II e III; b) V e VI; e c) VII e VIII (I) (PRIOLLI, 2006; MED, 2017).

Por fim, quanto à escala menor melódica, esta escala apresenta diferenças com a escala menor natural nos graus VI e VII, que recebem uma elevação em meio tom na subida da escala (PRIOLLI, 2006; MED, 2017). Além desta elevação, na descida, esta escala apresenta a mesma forma da escala menor natural (PRIOLLI, 2006; MED, 2017).

Exemplos das escalas acima citadas são ilustrados no QUADRO 47.

QUADRO 47 - ESCALAS (MAIOR E MENOR (NATURAL, HARMÔNICA E MELÓDICA)

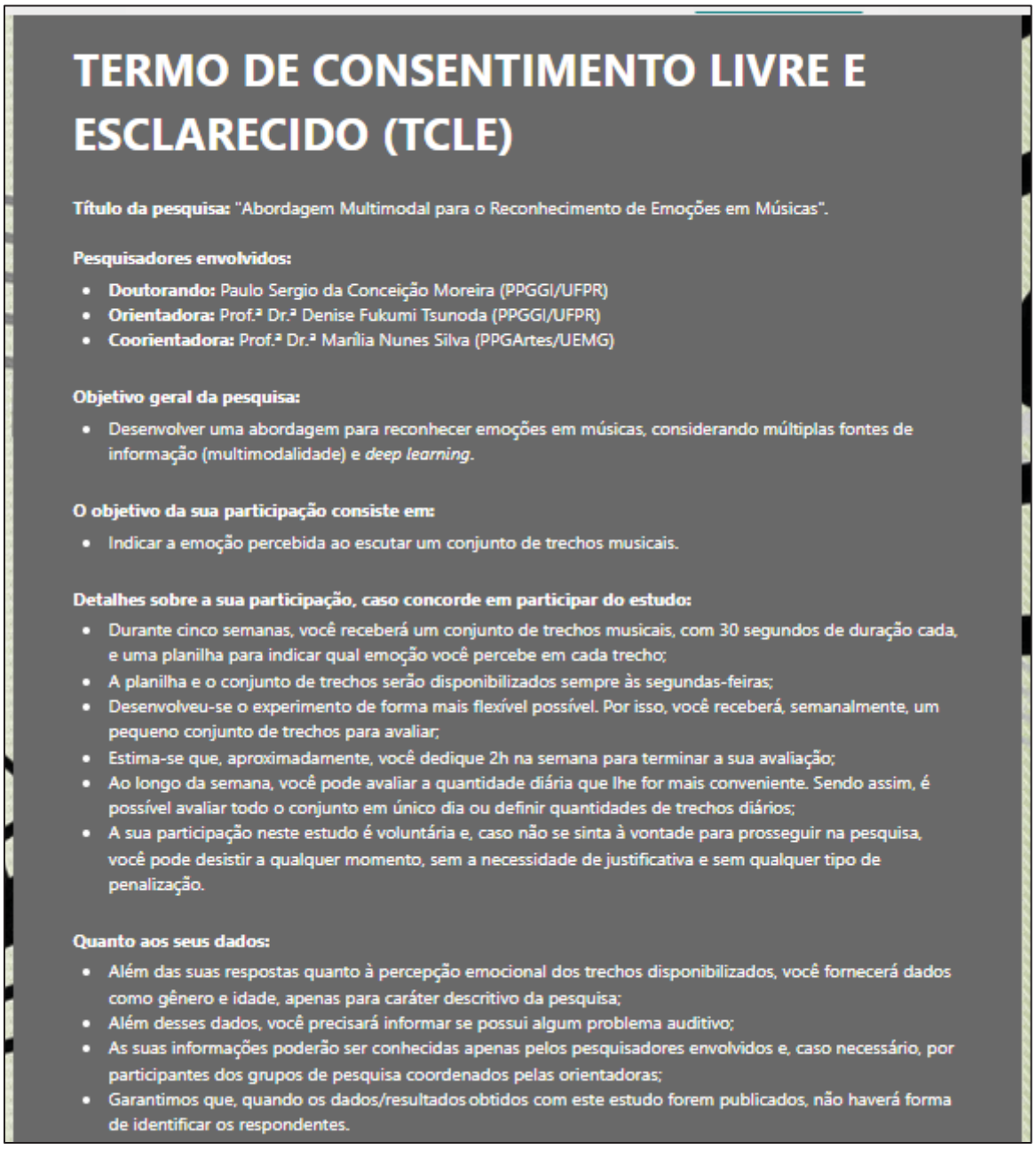
Escala	Representação	Fonte
Exemplo de escala maior		Adaptado de Solomon (2019, p.43)
Exemplo de escala menor natural		Med (2017, p.135)
Exemplo de escala menor harmônica		
Exemplo de escala menor melódica		Med (2017, p.137)

FONTE: O autor (2021).

Concernente aos outros tipos de escalas, as escalas artificiais - ou cromáticas - são aquelas que possuem 12 semitons em sequência (MED, 2017), e podem ser divididas em: a) escalas cromáticas; e b) escalas alteradas (MED, 2017). Dessas

escalas, uma escala cromática caracteriza-se por apresentar 12 divisões iguais de uma oitava (MILLER, 2016), que representam sete notas diatônicas e cinco notas cromáticas (MED, 2017). Por último, as escalas exóticas são formadas por estruturas singulares, dentre as quais citam-se: a) escalas ciganas; b) escalas pentatônicas ou chinesas; e c) escalas de tons inteiros (MED, 2017).

APÊNDICE B -TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (EXPERIMENTO PILOTO)



TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

Título da pesquisa: "Abordagem Multimodal para o Reconhecimento de Emoções em Músicas".

Pesquisadores envolvidos:

- **Doutorando:** Paulo Sergio da Conceição Moreira (PPGGI/UFPR)
- **Orientadora:** Prof.ª Dr.ª Denise Fukumi Tsunoda (PPGGI/UFPR)
- **Coorientadora:** Prof.ª Dr.ª Marília Nunes Silva (PPGArtes/UEMG)

Objetivo geral da pesquisa:

- Desenvolver uma abordagem para reconhecer emoções em músicas, considerando múltiplas fontes de informação (multimodalidade) e *deep learning*.

O objetivo da sua participação consiste em:

- Indicar a emoção percebida ao escutar um conjunto de trechos musicais.

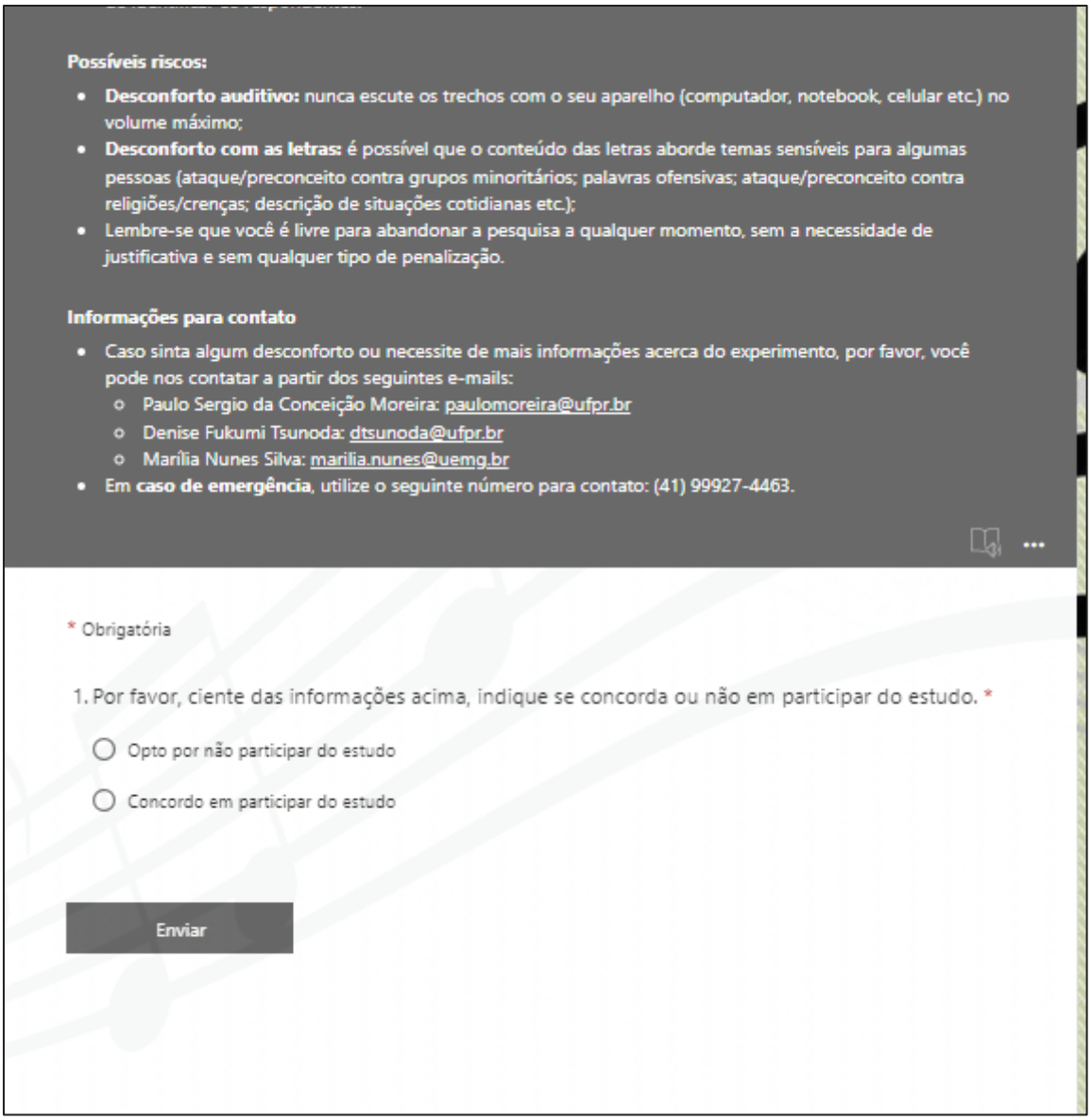
Detalhes sobre a sua participação, caso concorde em participar do estudo:

- Durante cinco semanas, você receberá um conjunto de trechos musicais, com 30 segundos de duração cada, e uma planilha para indicar qual emoção você percebe em cada trecho;
- A planilha e o conjunto de trechos serão disponibilizados sempre às segundas-feiras;
- Desenvolveu-se o experimento de forma mais flexível possível. Por isso, você receberá, semanalmente, um pequeno conjunto de trechos para avaliar;
- Estima-se que, aproximadamente, você dedique 2h na semana para terminar a sua avaliação;
- Ao longo da semana, você pode avaliar a quantidade diária que lhe for mais conveniente. Sendo assim, é possível avaliar todo o conjunto em único dia ou definir quantidades de trechos diários;
- A sua participação neste estudo é voluntária e, caso não se sinta à vontade para prosseguir na pesquisa, você pode desistir a qualquer momento, sem a necessidade de justificativa e sem qualquer tipo de penalização.

Quanto aos seus dados:

- Além das suas respostas quanto à percepção emocional dos trechos disponibilizados, você fornecerá dados como gênero e idade, apenas para caráter descritivo da pesquisa;
- Além desses dados, você precisará informar se possui algum problema auditivo;
- As suas informações poderão ser conhecidas apenas pelos pesquisadores envolvidos e, caso necessário, por participantes dos grupos de pesquisa coordenados pelas orientadoras;
- Garantimos que, quando os dados/resultados obtidos com este estudo forem publicados, não haverá forma de identificar os respondentes.

FONTE: O autor (2022).



FONTE: O autor (2022).

APÊNDICE C – INSTRUÇÕES E RECOMENDAÇÕES AOS VOLUNTÁRIOS (EXPERIMENTO PILOTO)

Instruções e recomendações

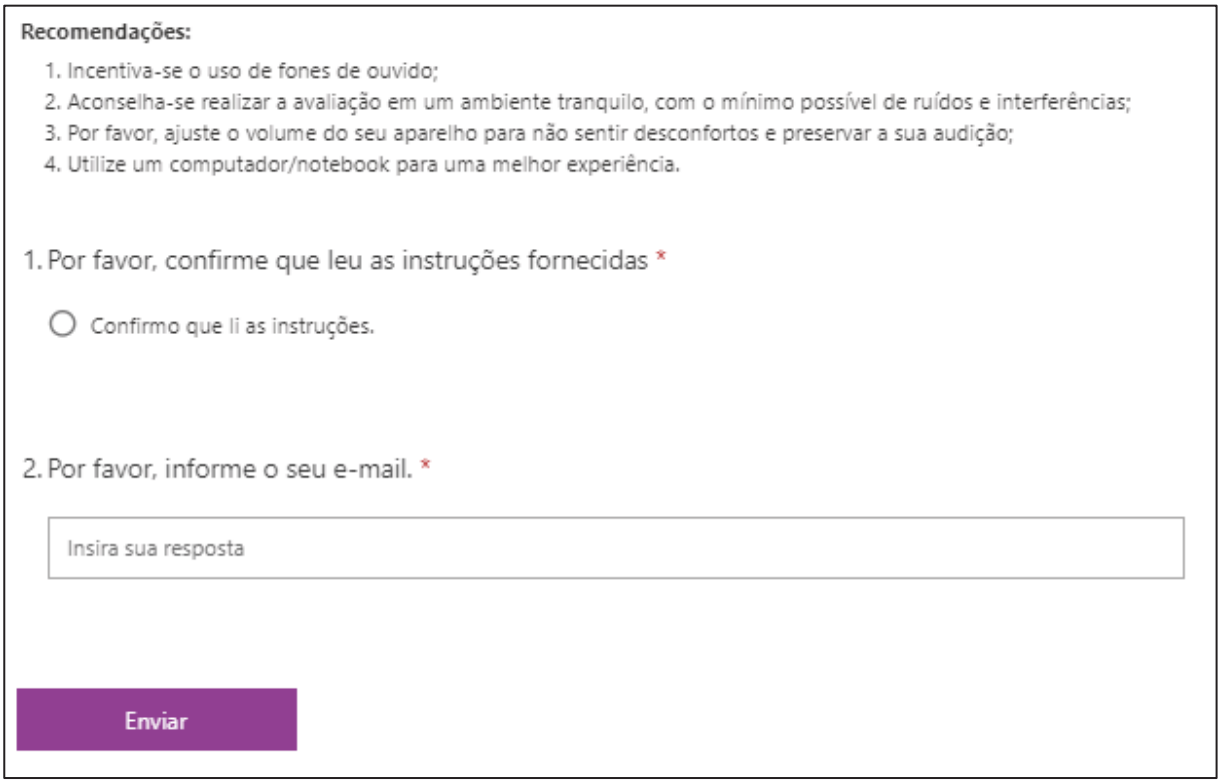
O objetivo deste experimento é a **emoção que você percebe** ao ouvir o segmento, o que, **não necessariamente**, representa o que **você sente** ao ouvir determinado segmento.

Ex: caso você **perceba que tristeza é a emoção mais pertinente**, por favor, **marque essa opção**, mesmo que **você não se sinta ou fique triste** ao ouvir o segmento.

Instruções:

1. Semanalmente, você receberá, no seu e-mail, um link para acessar e baixar os segmentos para avaliar, bem como a planilha para registrar a emoção que percebeu para cada trecho;
2. Cada trecho é composto por 30 segundos de duração e apresenta letras, em inglês;
3. O conjunto de segmentos corresponde a uma representação da música popular (Pop/Rock; R&B; Rap, Country etc.) produzida nos séculos XX e XXI;
4. Como a duração é de 30 segundos, é necessário boa concentração no momento da avaliação, pois há quebras repentinas entre cada segmento;
5. Na planilha de avaliação, há uma aba que apresenta os termos que representam as emoções consideradas. Consulte-os caso tenha dúvidas quanto ao significado de cada uma das emoções;
6. Cada trecho fornecido é numerado. Sendo assim, basta encontrar o ID correspondente na planilha fornecida e indicar qual emoção você percebeu;
7. Neste estudo, as seguintes emoções são consideradas: a) amor-ternura; b) calma; c) felicidade; d) medo; e) raiva; e f) tristeza;
8. Só é possível selecionar uma emoção para cada trecho;
9. Preste atenção em todos os elementos presentes nos segmentos (letras, harmonia, melodia, ritmo, instrumentos, voz etc.) para formular as suas respostas;
10. Você pode escutar cada trecho quantas vezes for necessário;
11. Por favor, revise suas respostas antes de devolver a planilha para o pesquisador (encaminhar para o e-mail: paulomoreira@ufpr.br);
12. Você deve excluir todos os trechos fornecidos após o fim da sua participação, em virtude dos direitos autorais relacionados às canções.

FONTE: O autor (2022).



FONTE: O autor (2022).

APÊNDICE D – ENDEREÇO DAS *PLAYLISTS* ANALISADAS

Alegria	
#	**Endereço (790 músicas)**
170	https://open.spotify.com/playlist/37i9dQZF1DX1WSnLRtI26o?si=9d8f086154f34b75
150	https://open.spotify.com/playlist/37i9dQZF1DWZKuerrwoAGz?si=f9ee946815ca449e
150	https://open.spotify.com/playlist/37i9dQZF1DX0UrRvztWcAU?si=ebe451194eaf460f
100	https://open.spotify.com/playlist/37i9dQZF1DXdPec7aLTmlC?si=f842a6eccc5f4333
70	https://open.spotify.com/playlist/37i9dQZF1DWW5Jdaphc81Z?si=5e3d43d71fb24ca2
50	https://open.spotify.com/playlist/37i9dQZF1DX84kJlLdo9vT?si=385dd1e50c274600
50	https://open.spotify.com/playlist/37i9dQZF1DXa9XlH8nDPns?si=207bebe2e7394e40
50	https://open.spotify.com/playlist/37i9dQZF1EIgWkLhbYXmZ0?si=883ab61a721e45d8

FONTE: O autor (2022).

Tristeza (757 músicas)	
#	**Endereço**
150	https://open.spotify.com/playlist/37i9dQZF1DWSqBruwoIXkA?si=c71c4a474923452e
150	https://open.spotify.com/playlist/37i9dQZF1DX8Vz2ROLXhTT?si=c83480b163e74f26
100	https://open.spotify.com/playlist/37i9dQZF1DWVV27DiNWxkR?si=ada1348bc93b40c1
73	https://open.spotify.com/playlist/37i9dQZF1DWZFicI79sEj9?si=9b77b0b60dbc4778
70	https://open.spotify.com/playlist/37i9dQZF1DXa39zZwdBPSN?si=ba06737ea22a4e66
68	https://open.spotify.com/playlist/37i9dQZF1DX7qK8ma5wgG1?si=de3ca6450ab84d21
54	https://open.spotify.com/playlist/37i9dQZF1DXaS0Bb5SeHIi?si=95e9a2dcb0e14237
52	https://open.spotify.com/playlist/37i9dQZF1DX6hO1vx5AApi?si=97e6f7ff9c05467a
40	https://open.spotify.com/playlist/37i9dQZF1DXchlyaSeZp0q?si=8704209bbaff44f4

FONTE: O autor (2022).

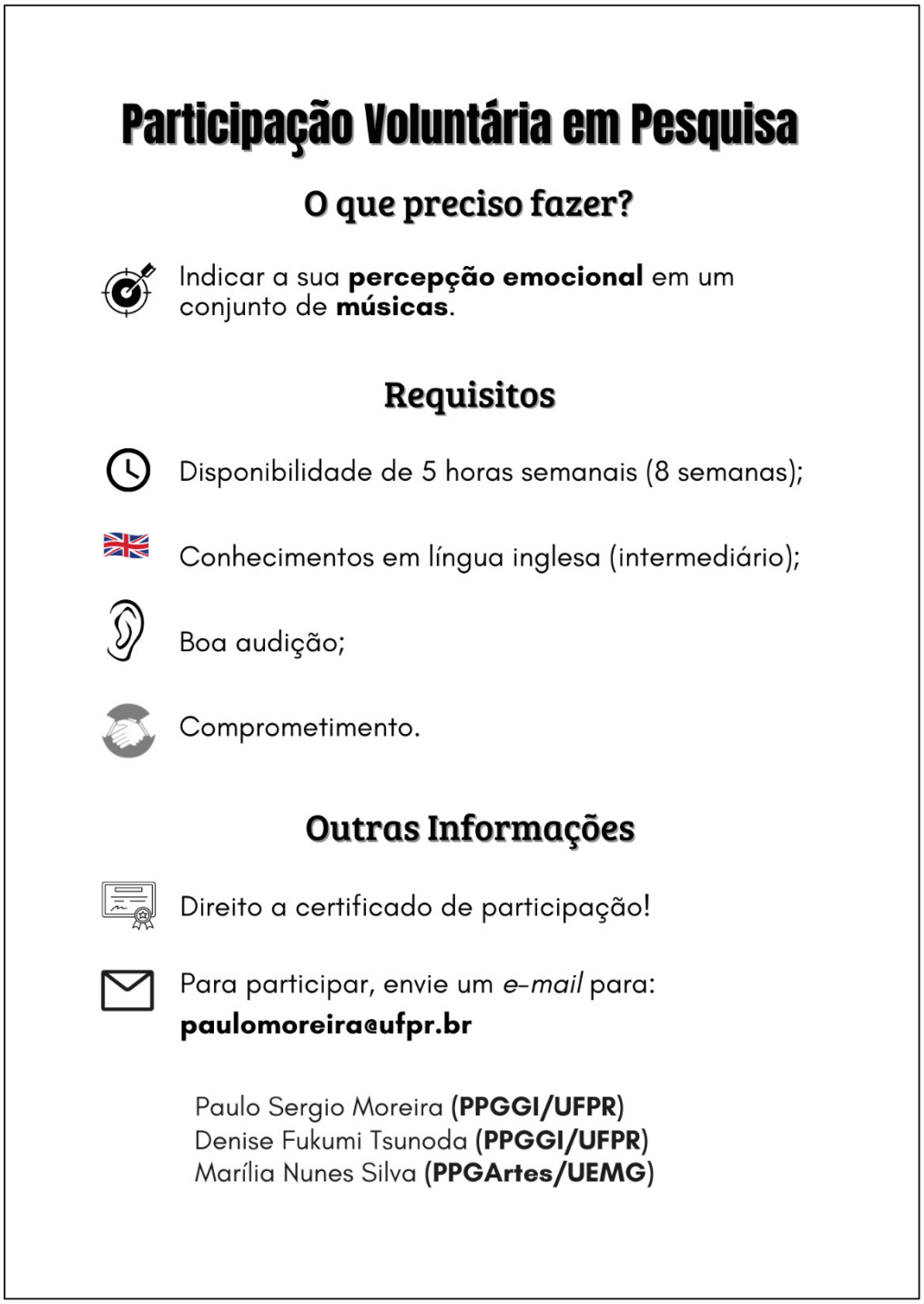
Raiva/Medo (719 músicas)	
#	**Endereço**
250	https://open.spotify.com/playlist/37i9dQZF1DWZVAVMhIe3pV?si=ffd0b91463ca4e3a
150	https://open.spotify.com/playlist/37i9dQZF1DWTXGqmP0bfT3?si=3703ff7be57545d0
110	https://open.spotify.com/playlist/37i9dQZF1DX06VUQ1GoKge?si=ec1e6f31d60a4f78
100	https://open.spotify.com/playlist/37i9dQZF1DWSVO64ma1pNY?si=bd9f6b1f154e468a
59	https://open.spotify.com/playlist/37i9dQZF1DX76t638V6CA8?si=2fbb2d511bb646c7
50	https://open.spotify.com/playlist/37i9dQZF1DWWPcvnOpPG3x?si=d56520f76b5440d9

FONTE: O autor (2022).

Serenidade (700 músicas)	
#	**Endereço**
250	https://open.spotify.com/playlist/37i9dQZF1DXcxacyAXkQDu?si=a4c1d0a938c04e1a
100	https://open.spotify.com/playlist/37i9dQZF1DXb5Mq0JeBbIw?si=72af76b662544d63
100	https://open.spotify.com/playlist/37i9dQZF1DX6e81LupkkgG?si=9b9c5064be3e46d4
100	https://open.spotify.com/playlist/37i9dQZF1DX5IDTimEWoTd?si=cfd009c6ac354cc6
100	https://open.spotify.com/playlist/37i9dQZF1DX50QitC6Oqtn?si=78b7db43554b43c6
50	https://open.spotify.com/playlist/37i9dQZF1DXcc6f6HRuPnq?si=d7d2fcd259d54783

FONTE: O autor (2022).

APÊNDICE E – *FOLDER* DE DIVUGALÇÃO DA PESQUISA



FONTE: O autor (2022).

APÊNDICE F – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (EXPERIMENTO AJUSTADO)

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

Título da pesquisa: "Reconhecimento de emoções em músicas considerando múltiplas fontes de informação e arquiteturas de *deep learning*".

Pesquisadores envolvidos:

- **Doutorando:** Paulo Sergio da Conceição Moreira (PPGGI/UFPR)
- **Orientadora:** Prof.ª Dr.ª Denise Fukumi Tsunoda (PPGGI/UFPR)
- **Coorientadora:** Prof.ª Dr.ª Marília Nunes Silva (PPGArtes/UEMG)

Objetivo geral da pesquisa:

- Desenvolver uma abordagem para reconhecer emoções em músicas, considerando múltiplas fontes de informação e arquiteturas de *deep learning*.

O objetivo da sua participação consiste em:

- Indicar a emoção percebida ao escutar um conjunto de trechos musicais.

Detalhes sobre a sua participação, caso concorde em participar do estudo:

- A sua **participação** neste estudo é **voluntária** e, caso não se sinta à vontade para prosseguir na pesquisa, **você pode desistir a qualquer momento**, sem a necessidade de justificativa e sem qualquer tipo de penalização.

Quanto aos seus dados:

- Além das suas respostas quanto à percepção emocional dos trechos disponibilizados, você fornecerá dados como gênero e idade, apenas para caráter descritivo da pesquisa;
- Além desses dados, você precisará informar se possui algum problema auditivo;
- As suas informações poderão ser conhecidas apenas pelos pesquisadores envolvidos e, caso necessário, por participantes dos grupos de pesquisa coordenados pelas orientadoras;
- Garantimos que, quando os dados/resultados obtidos com este estudo forem publicados, não haverá forma de identificar os respondentes.

Possíveis riscos:

- **Desconforto auditivo:** nunca escute os trechos com o seu aparelho (computador, notebook, celular etc.) no volume máximo;
- **Desconforto com as letras:** é possível que o conteúdo das letras aborde temas sensíveis para algumas pessoas (ataque/preconceito contra grupos minoritários; palavras ofensivas; ataque/preconceito contra religiões/crenças; descrição de situações cotidianas etc.);
- Lembre-se que **você é livre para abandonar a pesquisa a qualquer momento**, sem a necessidade de justificativa e sem qualquer tipo de penalização.

Informações para contato

- Caso sinta algum desconforto ou necessite de mais informações acerca do experimento, por favor, você pode nos contatar a partir dos seguintes *e-mails*:
  - Paulo Sergio da Conceição Moreira: paulomoreira@ufpr.br
  - Denise Fukumi Tsunoda: dtsunoda@ufpr.br
  - Marília Nunes Silva: marilia.nunes@uemg.br
- Em **caso de emergência**, utilize o seguinte número para contato: (41) 99927-4463.

FONTE: O autor (2022).

APÊNDICE G – INSTRUÇÕES E RECOMENDAÇÕES AOS VOLUNTÁRIOS (EXPERIMENTO AJUSTADO)

Instruções para o experimento

O objetivo deste experimento é a **emoção que você percebe** ao ouvir o segmento, o que, **não necessariamente**, representa o que **você sente** ao ouvir determinado segmento.

Ex: caso você **perceba que tristeza é a emoção mais pertinente**, por favor, **marque essa opção**, mesmo que **você não se sinta ou fique triste** ao ouvir o segmento.

Instruções:

1. No total, são **4.797 músicas**, cada qual com 30 segundos de duração e letras em inglês;
2. Neste estudo, as seguintes emoções são consideradas: **a) alegria; b) raiva/medo; c) serenidade; e d) tristeza**;
3. O conjunto de segmentos corresponde a uma representação da música popular (Pop/Rock; R&B; Rap, Country etc.) produzida nos séculos XX e XXI;
4. Só é possível selecionar uma emoção para cada trecho;
5. Você receberá, no seu e-mail, um link para acessar e baixar os segmentos para avaliar, bem como a planilha para registrar a emoção que percebeu para cada trecho;
6. O experimento será dividido em duas partes, em virtude da quantidade de registros. Sendo assim, ao receber o link, baixe os arquivos e faça as suas anotações de acordo com a sua disponibilidade, até a data determinada para a entrega;
7. Recomenda-se um total de 5 horas semanais para o experimento, contudo, vocês podem dedicar mais ou menos tempo, conforme os seus cronogramas. Todavia, é necessário devolver a planilha preenchida até a data estabelecida pelo pesquisador;
8. Como a duração é de 30 segundos, é necessário boa concentração no momento da avaliação, pois há quebras repentinas entre cada segmento;
9. Na planilha de avaliação, **há uma aba que apresenta os termos que representam as emoções consideradas.** Consulte-os para compreender o significado de cada uma das emoções;
10. Cada trecho fornecido é numerado. Sendo assim, basta encontrar o ID correspondente na planilha fornecida e indicar qual emoção você percebeu;
11. Preste atenção em todos os elementos presentes nos segmentos (letras, harmonia, melodia, ritmo, instrumentos, voz etc.) para formular as suas respostas;
12. Por favor, revise suas respostas antes de devolver a planilha para o pesquisador (encaminhar para o e-mail: paulomoreira@ufpr.br);
13. Você deve excluir todos os trechos fornecidos após o fim da sua participação, em virtude dos direitos autorais relacionados às canções.

1. Por favor, confirme que leu as instruções fornecidas *

◯ Confirmo que li as instruções.

2. Por favor, informe o seu e-mail. *

Insira sua resposta

FONTE: O autor (2022).

APÊNDICE H – ESTRUTURA DAS ARQUITETURAS DE *DEEP LEARNING*

Arquitetura	Estrutura das arquiteturas	Biblioteca principal
1D-CNN	**Camadas** Camadas: 3 (16, 8, 6) *kernel_size*: 5 *strides*: 1 *padding*: same Função de ativação: *relu* *Flatten*: 1 camada Função de ativação (última camada): *softmax* Função custo: *categorial_crossentropy*	*Tensorflow/keras*
	MaxPooling Camadas: 2 *pool_size*: 5 *strides*: 2 *padding: same* Função de ativação: *relu*	
FT-Transformer	*backbone: FTTransformerBackbone* *embeddings_layer: Embedding2dLayer* *head: LinearHead* *loss: CrossEntropyLoss*	*PyTorch Tabular*
MLP	Camadas: 3 (12, 8, 5) Função de ativação: *relu* Função de ativação da última camada: *softmax* Função custo: *categorial_crossentropy*	*Tensorflow/keras*
NODE	*backbone: NodeBackbone* *embeddings_layer: PreEncoded1dLayer* *head: Lambda* *loss: CrossEntropyLoss*	*PyTorch Tabular*
TabNet	*mask_type: sparsemax*	*PyTorch TabNet*
TabTransform	*backbone: TabTransformerBackbone* *dropout: Dropout* *output_layer: Linear* *loss: CrossEntropyLoss*	*PyTorch Tabular*

FONTE: O autor (2023).

APÊNDICE I - ESTATÍSTICAS DESCRITIVAS (BASE UM)

GÊNERO MUSICAL		
Gênero	**Total**	**Total (%)**
Rock	371	26,27%
Pop	337	23,87%
Heavy Metal	170	12,04%
R&B	159	11,26%
Folk	117	8,29%
Electronic	72	5,10%
Rap	61	4,32%
Country	48	3,40%
Punk	35	2,48%
Jazz	20	1,42%
Blues	10	0,71%
Reggae	4	0,28%
Avant-Garde	3	0,21%
Vocal	2	0,14%
Religious	2	0,14%
Easy Listening	1	0,07%
Total Geral	**1.412**	**100,00%**

FONTE: O autor (2023).

PRINCIPAIS ARTISTAS		
Artistas	**Total**	**Total (%)**
The Beatles	17	1,20%
Frank Sinatra	16	1,13%
Elvis Presley	15	1,06%
Merle Haggard	13	0,92%
Jerry Lee Lewis	13	0,92%
Ed Sheeran	11	0,78%
Bob Dylan	11	0,78%
Megadeth	9	0,64%
The Clash	8	0,57%
The Magnetic Fields	8	0,57%

FONTE: O autor (2023).

ARTISTAS E POR EMOÇÕES (SÍNTESE)		
Emoção	**Artistas (# registros)**	**Gêneros (# registros)**
Alegria	Elvis Presley (11); The Beatles (8); The Clash (7)	*Rock* (160); *Pop* (109); *R&B* (55)
Raiva/medo	Megadeth (9); Five Finger Death Punch (8); In Flames (6)	*Heavy Metal* (148); *Rap* (39); *Rock* (33)
Serenidade	Merle Haggard (10); Frank Sinatra (10); Bob Dylan (7)	*Roc*k (74); *R&B* (64); *Folk* (57)
Tristeza	Radiohead (6); Frank Sinatra (6); The Magnetic Fields (5)	*Pop* (174); *Rock* (104); *Folk* (55)

FONTE: O autor (2023).

ACORDES				
Característica	**Mínimo**	**Média**	**Desvio Padrão**	**Máximo**
quantidadeAcordes	1,0000	10,0935	4,9304	28,0000
outros	0,0000	0,6466	1,1030	8,0000
setima	0,0000	0,8102	1,2897	9,0000
setimaMaior	0,0000	0,8846	1,4107	17,0000
setimaMenor	0,0000	1,0227	1,6305	10,0000
setimaMenorQuinta	0,0000	0,0899	0,3909	5,0000
sexta	0,0000	0,6296	1,0586	9,0000
sextaMenor	0,0000	0,0786	0,3015	3,0000
triadeAumentada	0,0000	0,0999	0,4618	7,0000
triadeDiminuta	0,0000	0,0390	0,2394	3,0000
triadeMaior	0,0000	4,3895	3,6684	23,0000
triadeMenor	0,0000	1,4030	1,8764	14,0000
acordesUnicos	1,0000	6,3322	2,6393	19,0000
proporcaoUnicos	0,2000	0,7002	0,2087	1,0000
proporcaoTriadeMenor	0,0000	0,1284	0,1694	1,0000
proporcaoTriadeMaior	0,0000	0,4054	0,2792	1,0000
proporcaoTriadeDiminuta	0,0000	0,0036	0,0248	0,5000
proporcaoTriadeAumentada	0,0000	0,0090	0,0408	0,4286
proporcaoSextaMenor	0,0000	0,0120	0,0578	0,6667
proporcaoSexta	0,0000	0,0677	0,1227	1,0000
proporcaoSetima	0,0000	0,0958	0,1680	1,0000
proporcaoSetimaMaior	0,0000	0,0948	0,1540	1,0000
proporcaoSetimaMenor	0,0000	0,1012	0,1594	1,0000
proporcaoSetimaMenorQuinta	0,0000	0,0117	0,0635	1,0000
proporcaoOutros	0,0000	0,0702	0,1358	1,0000

FONTE: O autor (2023).

ÁUDIO				
Característica	**Mínimo**	**Média**	**Desvio Padrão**	**Máximo**
rmsMean	0,0044	0,0990	0,0626	0,3267
rmsSTD	0,0023	0,0418	0,0268	0,1670
chromaMean1	0,0087	0,3979	0,1493	0,9097
chromaSTD1	0,0345	0,2696	0,0641	0,4334
chromaMean2	0,0271	0,3814	0,1398	0,8748
chromaSTD2	0,0680	0,2500	0,0602	0,4348
chromaMean3	0,0162	0,4044	0,1515	0,8827
chromaSTD3	0,0496	0,2709	0,0644	0,4510
chromaMean4	0,0319	0,3848	0,1421	0,8822
chromaSTD4	0,0559	0,2516	0,0593	0,4119
chromaMean5	0,0318	0,4045	0,1507	0,8305
chromaSTD5	0,0503	0,2719	0,0633	0,4435
chromaMean6	0,0180	0,3863	0,1426	0,9500
chromaSTD6	0,0263	0,2634	0,0628	0,4633
chromaMean7	0,0451	0,3844	0,1412	0,8952
chromaSTD7	0,0478	0,2544	0,0601	0,4727
chromaMean8	0,0411	0,4131	0,1548	0,9001
chromaSTD8	0,0742	0,2748	0,0625	0,4569
chromaMean9	0,0376	0,4070	0,1477	0,8918
chromaSTD9	0,0571	0,2582	0,0610	0,4334
chromaMean10	0,0408	0,4248	0,1546	0,8569
chromaSTD10	0,0595	0,2792	0,0610	0,4620
chromaMean11	0,0246	0,3956	0,1464	0,9766
chromaSTD11	0,0576	0,2544	0,0617	0,4486
chromaMean12	0,0036	0,3994	0,1461	0,8881
chromaSTD12	0,0067	0,2627	0,0618	0,4453
pulseMean	0,0669	0,2177	0,0437	0,3017
pulseSTD	0,1207	0,2854	0,0415	0,3658
mfcc1Mean	-649,0789	-246,6576	136,7417	20,6493
mfcc1STD	9,7999	70,5059	29,8155	228,8510
mfcc2Mean	-35,9870	119,0124	40,3073	242,2672
mfcc2STD	6,9715	32,7477	10,9957	83,5963
mfcc3Mean	-94,1357	12,7615	30,7734	107,3422
mfcc3STD	6,7842	27,1680	8,5211	69,0640
mfcc4Mean	-78,8123	32,0119	19,2054	96,8024
mfcc4STD	5,3795	20,1589	5,4411	51,8807
mfcc5Mean	-41,0354	10,5772	13,9362	105,6812
mfcc5STD	5,7762	16,9875	4,5351	38,1140
mfcc6Mean	-41,2544	15,2418	12,0065	64,3013
mfcc6STD	5,8637	14,2490	3,7633	33,2084
mfcc7Mean	-31,3775	0,1712	8,4786	36,0982
mfcc7STD	5,3248	13,1838	3,0862	27,5982
mfcc8Mean	-34,1483	8,8613	9,5050	49,9314
mfcc8STD	5,2617	11,6246	2,6997	26,1492

mfcc9Mean	-32,4414	-2,5836	6,8373	27,0466
mfcc9STD	4,9951	11,0068	2,4146	23,3932
mfcc10Mean	-28,0078	4,2329	8,5002	30,1454
mfcc10STD	4,8367	10,4969	2,3412	19,3305
mfcc11Mean	-39,6084	-5,7822	6,0730	19,8067
mfcc11STD	5,2900	10,1326	2,3158	22,9680
mfcc12Mean	-23,5843	5,0460	7,3940	26,1753
mfcc12STD	4,4238	9,6745	2,1611	19,6235
mfcc13Mean	-38,4339	-5,8280	5,3240	15,1146
mfcc13STD	4,4416	9,3857	2,0827	18,7501
centroidMean	224,3094	1656,2467	802,6164	4703,4233
centroidSTD	54,2922	624,4764	334,7235	2300,4915
rollofMean	291,5393	3451,4901	1813,3506	7862,3420
rollofSTD	62,8573	1401,0601	690,5886	3445,8433
zcrMean	0,0093	0,0747	0,0410	0,3105
zcrSTD	0,0026	0,0438	0,0305	0,2216
loudness	3,1691	138,4820	106,3477	693,3931
salience	0,0114	0,2066	0,1461	0,7136
melodiaMean	3,7358	102,1139	61,9253	841,1295
melodiaSTD	19,9922	107,3818	73,8170	636,7784
bpm	60,0525	113,7829	25,6561	184,5703
danceability	0,7389	1,2917	0,3547	3,7589
flux	2,3650	37,5571	21,8750	131,8663

FONTE: O autor (2023).

LETRAS				
Característica	**Mínimo**	**Média**	**Desvio Padrão**	**Máximo**
charCount	25,0000	249,5035	107,8175	949,0000
wordCount	5,0000	54,3477	23,5238	209,0000
avgWord	2,4643	3,6360	0,4109	5,9773
uniqueWords	5,0000	37,3931	14,2213	100,0000
uniqueWordsProp	0,0602	0,7154	0,1310	1,0000
neg	0,0000	0,0941	0,0938	0,5700
neu	0,2640	0,7810	0,1190	1,0000
pos	0,0000	0,1249	0,1004	0,7360
compound	-0,9964	0,1332	0,6382	0,9975

FONTE: O autor (2023).

SIMBÓLICAS				
Característica	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**
averageNoteDuration	0,4610	0,1898	0,1688	1,5093
maximumNoteDuration	2,3467	1,1724	0,5000	14,6250
minimumNoteDuration	0,0806	0,0600	0,0000	0,1667
variabilityOfNoteDuration	0,3705	0,1839	0,0849	1,7119
noteDensity	5,7922	1,5752	1,5556	11,1455
importanceOfBassRegister	0,6152	0,2483	0,0000	1,0000
importanceOfMiddleRegister	0,2674	0,2096	0,0000	0,9524
importanceOfHighRegister	0,1174	0,1284	0,0000	0,8727
mostCommonPitchClassPrevalence	0,2900	0,0882	0,1339	0,7547
primaryRegister	51,3059	8,8584	31,2405	85,4455
pitchClassVariety	10,5085	1,4422	6,0000	12,0000
range	62,5446	13,3025	17,0000	85,0000

FONTE: O autor (2023).

VOZ				
Característica	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**
f1Mean	752,8774	153,4999	442,5658	1295,5975
f1STD	0,4194	0,0844	0,1693	0,7377
f1BandMean	1052,7984	143,4170	548,0128	1393,9761
f1BandSTD	0,3119	0,0767	0,1536	0,6924
f2Mean	1791,3116	153,0973	1433,9149	2272,8340
f2STD	0,1926	0,0303	0,1005	0,2906
f2BandMean	842,3283	99,8270	514,8638	1139,0696
f2BandSTD	0,4442	0,0799	0,2232	0,7214
f3Mean	2839,1388	135,6148	2440,4143	3250,3291
f3STD	0,1186	0,0168	0,0724	0,1800
f3BandMean	759,2945	85,9638	462,1772	1037,0382
f3BandSTD	0,5259	0,1120	0,2726	0,9619
voicedMean	0,8378	0,6050	0,1196	9,7533
voiceSTD	0,9476	0,6189	0,1031	9,4006
unvoicedMean	0,2743	0,2452	0,0353	2,1929
unvoicedSTD	0,4283	0,3968	0,0172	2,8710
f0Mean	40,4620	4,7232	26,2099	55,8514
f0STD	0,1521	0,0425	0,0516	0,3446
hnrMean	6,4837	3,2340	-3,8280	14,5467
hnrSTD	1,0394	9,6107	-195,1137	264,1762
jitterMean	0,0310	0,0161	0,0069	0,1008
jitterSTD	1,8671	0,4642	0,8416	3,7210
shitterMean	1,0800	0,2895	0,4547	2,0258
shitterSTD	0,6743	0,1786	0,3333	1,2456
mfcc1MeanVoice	-336,7978	79,4043	-610,8442	-98,3101
mfcc1STDVoice	122,2070	29,6769	52,0559	196,6398
mfcc2MeanVoice	33,3741	28,1598	-75,9802	127,2500

mfcc2STDVoice	50,6038	9,0867	27,2650	97,0370
mfcc3MeanVoice	-30,6804	26,9105	-142,5249	48,9045
mfcc3STDVoice	40,7499	9,9646	17,4497	78,2539
mfcc4MeanVoice	-1,2901	20,0537	-94,6529	65,9728
mfcc4STDVoice	33,4899	7,5977	15,3823	65,7400
mfcc5MeanVoice	-17,6797	13,7293	-64,1682	39,8565
mfcc5STDVoice	28,4200	6,0261	12,5986	53,0148
mfcc6MeanVoice	-14,3072	12,3033	-59,9610	31,4351
mfcc6STDVoice	22,3372	4,9005	9,9089	37,5017
mfcc7MeanVoice	-19,3381	10,2458	-59,3508	13,8447
mfcc7STDVoice	21,9938	4,5027	9,5845	38,3824
mfcc8MeanVoice	-10,0000	8,7687	-48,1863	29,1585
mfcc8STDVoice	17,6231	3,6955	7,4368	31,1243
mfcc9MeanVoice	-20,6206	9,5602	-54,4491	14,3570
mfcc9STDVoice	18,9113	3,3073	9,9331	31,3700
mfcc10MeanVoice	-3,5904	7,2819	-41,6568	22,6523
mfcc10STDVoice	16,1731	3,0378	8,7420	26,7509
mfcc11MeanVoice	-15,3929	8,1642	-46,4560	34,4364
mfcc11STDVoice	16,2500	2,8174	8,1710	30,5143
mfcc12MeanVoice	-6,5639	5,9702	-25,3922	12,1241
mfcc12STDVoice	13,6100	2,7914	6,1023	27,5247
mfcc13MeanVoice	-14,0327	6,6654	-40,6988	6,0279
mfcc13STDVoice	14,3680	2,5197	7,7298	26,4854
mfcc14MeanVoice	-3,5164	6,0362	-27,2170	29,1720
mfcc14STDVoice	13,3085	3,0781	6,4735	32,1520
mfcc15MeanVoice	-9,6886	6,2238	-31,7973	14,5058
mfcc15STDVoice	13,5594	3,3025	6,4147	29,9652
mfcc16MeanVoice	-2,9771	6,5783	-22,6448	40,8556
mfcc16STDVoice	13,3343	4,0512	6,1277	31,9655
gfccsMean	17,8298	16,6833	-25,5376	56,2674
gfccsSTD	210,2212	17,6435	136,9457	251,0072

FONTE: O autor (2023).

APÊNDICE J - ESTATÍSTICAS DESCRITIVAS (BASE DOIS)

GÊNERO MUSICAL		
Gênero	**Total**	**Total (%)**
Rock	1.353	31,44%
Pop	848	19,71%
R&B	558	12,97%
Heavy Metal	344	7,99%
Folk	323	7,51%
Electronic	232	5,39%
Rap	211	4,90%
Country	159	3,70%
Punk	112	2,60%
Jazz	79	1,84%
Blues	37	0,86%
Reggae	18	0,42%
Vocal	13	0,30%
Avant-Garde	10	0,23%
Religious	3	0,07%
Easy Listening	1	0,02%
Classical	1	0,02%
Latin	1	0,02%
Total Geral	**4.303**	**100,00%**

FONTE: O autor (2023).

PRINCIPAIS ARTISTAS		
Artistas	**Total**	**Total (%)**
The Beatles	64	1,49%
Frank Sinatra	62	1,44%
Bob Dylan	41	0,95%
Merle Haggard	39	0,91%
Elvis Presley	38	0,88%
Stevie Wonder	28	0,65%
Radiohead	28	0,65%
Jerry Lee Lewis	28	0,65%
Bruce Springsteen	27	0,63%
The Magnetic Fields	26	0,60%

FONTE: O autor (2023).

ARTISTAS E POR EMOÇÕES (SÍNTESE)		
Emoção	**Artistas (# registros)**	**Gêneros (# registros)**
Alegria	The Beatles (28); The Clash (21); Elvis Presley (17)	*Rock* (453); *Pop* (255); *R&B* (166)
Raiva/medo	Megadeth (15); Nine Inch Nails (14); Rage Against the Machine (11)	*Heavy Metal* (261); *Rock* (147); *Rap* (124)
Serenidade	Frank Sinatra (33); Merle Haggard (28); Bob Dylan (22)	*Rock* (337); *R&B* (247); *Pop* (222)
Tristeza	Frank Sinatra (29); Radiohead (17); Bruce Springsteen (14)	*Rock* (416); *Pop* (364); *Folk* (146)

FONTE: O autor (2023).

ACORDES				
Característica	**Mínimo**	**Média**	**Desvio Padrão**	**Máximo**
quantidadeAcordes	1,0000	10,1861	4,8771	29,0000
outros	0,0000	0,6577	1,0773	9,0000
setima	0,0000	0,7992	1,2934	13,0000
setimaMaior	0,0000	0,8817	1,3761	17,0000
setimaMenor	0,0000	1,0286	1,6136	15,0000
setimaMenorQuinta	0,0000	0,0881	0,3826	5,0000
sexta	0,0000	0,6507	1,1001	9,0000
sextaMenor	0,0000	0,0899	0,3550	4,0000
triadeAumentada	0,0000	0,0920	0,3963	7,0000
triadeDiminuta	0,0000	0,0476	0,2646	3,0000
triadeMaior	0,0000	4,4590	3,7052	23,0000
triadeMenor	0,0000	1,3916	1,8179	14,0000
acordesUnicos	1,0000	6,4139	2,7059	20,0000
proporcaoUnicos	0,1667	0,6988	0,2068	1,0000
proporcaoTriadeMenor	0,0000	0,1286	0,1666	1,0000
proporcaoTriadeMaior	0,0000	0,4121	0,2821	1,0000
proporcaoTriadeDiminuta	0,0000	0,0045	0,0300	1,0000
proporcaoTriadeAumentada	0,0000	0,0083	0,0368	0,5000
proporcaoSextaMenor	0,0000	0,0113	0,0525	0,6667
proporcaoSexta	0,0000	0,0676	0,1248	1,0000
proporcaoSetima	0,0000	0,0900	0,1592	1,0000
proporcaoSetimaMaior	0,0000	0,0911	0,1489	1,0000
proporcaoSetimaMenor	0,0000	0,1030	0,1618	1,0000
proporcaoSetimaMenorQuinta	0,0000	0,0102	0,0535	1,0000
proporcaoOutros	0,0000	0,0733	0,1390	1,0000

FONTE: O autor (2023).

ÁUDIO				
Característica	**Mínimo**	**Média**	**Desvio Padrão**	**Máximo**
rmsMean	0,0963	0,0578	0,0001	0,3267
rmsSTD	0,0420	0,0274	0,0001	0,1801
chromaMean1	0,3897	0,1436	0,0087	0,9406
chromaSTD1	0,2707	0,0651	0,0234	0,4521
chromaMean2	0,3759	0,1324	0,0271	0,9155
chromaSTD2	0,2522	0,0595	0,0670	0,4402
chromaMean3	0,3996	0,1451	0,0162	0,9019
chromaSTD3	0,2745	0,0634	0,0393	0,4510
chromaMean4	0,3790	0,1366	0,0278	0,8985
chromaSTD4	0,2511	0,0590	0,0406	0,4290
chromaMean5	0,4040	0,1469	0,0177	0,9215
chromaSTD5	0,2755	0,0619	0,0270	0,4486
chromaMean6	0,3828	0,1381	0,0180	0,9500
chromaSTD6	0,2647	0,0629	0,0263	0,4633
chromaMean7	0,3825	0,1338	0,0287	0,8952
chromaSTD7	0,2585	0,0584	0,0478	0,4727
chromaMean8	0,4083	0,1473	0,0182	0,9530
chromaSTD8	0,2778	0,0621	0,0405	0,4569
chromaMean9	0,4017	0,1391	0,0203	0,8918
chromaSTD9	0,2591	0,0580	0,0352	0,4412
chromaMean10	0,4210	0,1480	0,0355	0,9301
chromaSTD10	0,2826	0,0596	0,0536	0,4620
chromaMean11	0,3879	0,1395	0,0246	0,9766
chromaSTD11	0,2548	0,0605	0,0576	0,4486
chromaMean12	0,3941	0,1416	0,0036	0,8881
chromaSTD12	0,2649	0,0603	0,0067	0,4453
pulseMean	0,2226	0,0436	0,0625	0,3047
pulseSTD	0,2901	0,0411	0,1204	0,3689
mfcc1Mean	-248,6545	125,4504	-880,0508	20,6493
mfcc1STD	72,0147	28,7498	9,7999	228,8510
mfcc2Mean	121,0136	39,2193	-35,9870	252,3913
mfcc2STD	33,2092	10,9616	6,9715	92,9399
mfcc3Mean	11,3775	30,2680	-136,2832	107,3422
mfcc3STD	27,5511	8,7714	6,4757	69,0640
mfcc4Mean	32,3349	19,0016	-78,8123	99,0024
mfcc4STD	20,1999	5,5315	5,3795	51,8807
mfcc5Mean	10,8006	13,3056	-44,8453	105,6812
mfcc5STD	17,2439	4,6397	5,7295	38,6193
mfcc6Mean	15,4576	11,8148	-43,5520	64,3013
mfcc6STD	14,4106	3,7599	5,3779	33,2084
mfcc7Mean	0,3809	8,3940	-34,2453	41,3776
mfcc7STD	13,3677	3,1554	5,3248	29,9809
mfcc8Mean	9,1149	9,2565	-34,1483	49,9314
mfcc8STD	11,7518	2,7589	5,0628	26,1492

mfcc9Mean	-2,3998	6,7988	-32,4414	27,0466
mfcc9STD	11,2064	2,4954	4,9139	23,8553
mfcc10Mean	4,3471	8,1646	-29,9178	33,7426
mfcc10STD	10,6403	2,4244	4,8367	27,2977
mfcc11Mean	-5,5346	6,1142	-39,6084	19,8067
mfcc11STD	10,1965	2,2877	4,5156	24,8124
mfcc12Mean	4,8254	7,1046	-23,5843	29,7812
mfcc12STD	9,7358	2,1727	4,4238	21,0785
mfcc13Mean	-5,7105	5,4389	-38,4339	21,0591
mfcc13STD	9,4246	2,0859	4,4416	19,8432
centroidMean	1626,4693	760,2594	224,3094	4703,4233
centroidSTD	640,1194	332,7742	54,2922	2300,4915
rollofMean	3395,1706	1703,5680	291,5393	8112,0149
rollofSTD	1452,8408	684,6423	62,8573	3699,2867
zcrMean	0,0726	0,0409	0,0081	0,4131
zcrSTD	0,0434	0,0304	0,0026	0,2289
loudness	131,2187	97,2609	0,0119	787,5535
salience	0,2022	0,1418	0,0114	0,8398
melodiaMean	102,3414	66,6051	3,7358	897,1302
melodiaSTD	104,3230	74,1369	19,9922	746,1099
bpm	114,7313	25,4693	57,1100	191,4064
danceability	1,3004	0,3522	0,6553	3,7589
flux	36,2372	20,2029	0,0367	145,0117

FONTE: O autor (2023).

LETRAS				
Característica	**Mínimo**	**Média**	**Desvio Padrão**	**Máximo**
charCount	243,9512	106,1290	25,0000	1020,0000
wordCount	53,1208	23,2074	5,0000	221,0000
avgWord	3,6391	0,4000	2,3902	5,9773
uniqueWords	36,7699	14,5160	5,0000	132,0000
uniqueWordsProp	0,7160	0,1301	0,0602	1,0000
neg	0,0921	0,0892	0,0000	0,5700
neu	0,7797	0,1212	0,1410	1,0000
pos	0,1282	0,1043	0,0000	0,7410
compound	0,1455	0,6249	-0,9964	0,9975

FONTE: O autor (2023).

SIMBÓLICAS				
Característica	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**
averageNoteDuration	0,4510	0,1642	0,1607	1,5093
maximumNoteDuration	2,2870	1,1237	0,3750	14,6250
minimumNoteDuration	0,0816	0,0597	0,0000	0,2500
variabilityOfNoteDuration	0,3589	0,1611	0,0801	1,7119
noteDensity	5,9156	1,5772	1,2899	11,9420
importanceOfBassRegister	0,6077	0,2437	0,0000	1,0000
importanceOfMiddleRegister	0,2744	0,2043	0,0000	0,9524
importanceOfHighRegister	0,1179	0,1300	0,0000	0,9600
mostCommonPitchClassPrevalence	0,2877	0,0870	0,1170	0,7952
primaryRegister	51,5937	8,5867	31,0000	86,7100
pitchClassVariety	10,5320	1,4059	4,0000	12,0000
range	62,3718	13,5963	13,0000	86,0000

FONTE: O autor (2023).

VOZ				
Característica	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**
f1Mean	435,0590	735,0838	145,4982	1318,3267
f1STD	0,1509	0,4300	0,0783	0,7377
f1BandMean	548,0128	1058,4279	139,5690	1403,0151
f1BandSTD	0,1448	0,3101	0,0760	0,6924
f2Mean	1369,4052	1777,4914	151,2974	2294,5850
f2STD	0,1005	0,1961	0,0284	0,3185
f2BandMean	508,2661	833,7917	95,5611	1139,0696
f2BandSTD	0,2232	0,4521	0,0796	0,7680
f3Mean	2401,7532	2831,4320	135,9205	3280,6187
f3STD	0,0637	0,1201	0,0167	0,1918
f3BandMean	427,8066	754,7055	82,9762	1056,1384
f3BandSTD	0,2666	0,5362	0,1095	0,9937
voicedMean	0,1178	0,8668	0,7308	28,5200
voiceSTD	0,0000	0,9621	0,6086	9,4006
unvoicedMean	0,0353	0,2886	0,2800	5,9350
unvoicedSTD	0,0000	0,4484	0,4334	5,9150
f0Mean	22,0075	40,0282	4,7095	55,8514
f0STD	0,0418	0,1524	0,0434	0,4115
hnrMean	-3,8280	6,7531	2,9240	14,6080
hnrSTD	-1302,4741	0,8498	21,4053	279,6650
jitterMean	0,0053	0,0289	0,0136	0,1008
jitterSTD	0,8227	1,9123	0,4618	4,9235
shitterMean	0,4461	1,0505	0,2643	2,0258
shitterSTD	0,3333	0,6855	0,1679	1,3786
mfcc1MeanVoice	-794,6223	-340,5773	80,9140	-91,8420
mfcc1STDVoice	45,4177	123,0368	29,6674	205,7719
mfcc2MeanVoice	-79,8906	32,1848	28,1687	155,5727

mfcc2STDVoice	25,4004	50,4017	8,9586	97,0370
mfcc3MeanVoice	-142,5249	-27,2686	25,6150	60,5220
mfcc3STDVoice	15,3084	40,3989	9,4370	78,2539
mfcc4MeanVoice	-94,6529	0,5306	17,5967	65,9728
mfcc4STDVoice	15,3823	33,3485	7,1771	71,3892
mfcc5MeanVoice	-68,6901	-18,1242	13,3831	39,8565
mfcc5STDVoice	8,6849	28,3011	5,7075	53,5671
mfcc6MeanVoice	-59,9610	-13,9916	12,0638	32,6793
mfcc6STDVoice	7,9572	22,1244	4,6310	44,7199
mfcc7MeanVoice	-59,3508	-19,2734	10,2475	15,3534
mfcc7STDVoice	9,4574	21,8992	4,3186	43,1379
mfcc8MeanVoice	-48,1863	-10,1761	8,7269	29,1585
mfcc8STDVoice	7,0949	17,4064	3,4133	31,1285
mfcc9MeanVoice	-54,4491	-20,5343	9,2908	14,3570
mfcc9STDVoice	9,7393	18,8253	3,2159	31,8577
mfcc10MeanVoice	-41,6568	-3,4594	7,0883	33,4618
mfcc10STDVoice	5,6811	16,0475	2,8458	29,0531
mfcc11MeanVoice	-49,7843	-15,6672	8,2191	34,4364
mfcc11STDVoice	6,7243	16,0968	2,7255	30,5143
mfcc12MeanVoice	-33,3068	-6,7224	5,8966	12,3087
mfcc12STDVoice	5,9696	13,4590	2,5555	27,5247
mfcc13MeanVoice	-40,6988	-13,9983	6,5960	7,6137
mfcc13STDVoice	5,8990	14,2179	2,4461	31,4356
mfcc14MeanVoice	-27,2170	-3,7616	5,8548	29,1720
mfcc14STDVoice	5,5639	13,1151	2,8991	32,1520
mfcc15MeanVoice	-31,7973	-9,8528	6,0576	20,1508
mfcc15STDVoice	4,4780	13,3945	3,1347	29,9652
mfcc16MeanVoice	-26,2119	-3,3098	6,3619	41,5778
mfcc16STDVoice	4,6504	13,1552	3,9613	32,7294
gfccsMean	-31,0630	17,6888	16,6333	56,2674
gfccsSTD	98,4346	209,7258	17,7502	251,0072

FONTE: O autor (2023).

APÊNDICE K – HIPERPARÂMETROS DAS ARQUITETURAS DE *DEEP LEARNING* SELECIONADOS PARA AS MELHORES ACURÁCIAS NO CENÁRIO DE CLASSIFICAÇÃO A

Arquitetura	Hiperparâmetros (base dois)	Hiperparâmetros (concordância)
NODE	{batch_size: 64; max_epochs: 100; num_layers: 3; num_trees: 64; optimizer: 'adam'}	{batch_size: 64; max_epochs: 100; num_layers: 3; num_trees: 128; optimizer: 'adam'}
TabNet	{n steps: 5; batch size: 64; epochs: 100; optimizer: adam}	{n steps: 5; batch size: 64; epochs: 100; optimizer: adam}
MLP	{num_epochs: 100; batch_size: 64; optimizer: 'adam'; dropout: [0,30; 0,25; 0,25]}	{num_epochs: 50; batch_size: 64; optimizer: 'adam'; dropout: [0,25; 0,25; 0,25]}
TabTransform	{num_epochs: 100; batch_size: 64; optimizer: 'adam'}	{num_epochs: 100; batch_size: 64; optimizer: 'adam'}
1D-CNN	{num_epochs: 100; batch_size: 32; optimizer: 'adam'; dropout: [0,30; 0,25; 0,00]}	{num_epochs: 50; batch_size: 64; optimizer: 'adam'; dropout: [0,30; 0,30; 0,00]
FT-Transformer	{num_epochs: 50; batch_size: 64; optimizer: 'adam'}	{num_epochs: 50; batch_size: 32; optimizer: 'adam'}

FONTE: O autor (2023).

APÊNDICE L - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (CENÁRIO A – TODOS OS ATRIBUTOS)

Base dois														
Áudio			**Acordes**			**Letras**			**Simbólicas**			**Voz**		
Modelo	**Acurácia**	**Tempo (s)**	**Modelo**	**Acurácia**	**Tempo (s)**	**Modelo**	**Acurácia**	**Tempo (s)**	**Modelo**	**Acurácia**	**Tempo (s)**	**Modelo**	**Acurácia**	**Tempo (s)**
CatBoost	**62,91%**	**63,53**	**XGBoost**	**40,99%**	**10,61**	**XGBoost**	**45,26%**	**0,59**	**SVM**	**55,48%**	**0,92**	**CatBoost**	**60,40%**	**54,2**
SVM	61,80%	1,85	CatBoost	40,89%	7,85	SVM	44,05%	0,83	CatBoost	55,01%	10,76	SVM	59,75%	1,37
XGBoost	61,33%	8,21	SVM	39,96%	1,39	CatBoost	43,95%	9,82	XGBoost	55,01%	0,42	XGBoost	58,27%	3,51
RF	59,75%	2,78	LightGBM	39,31%	0,65	RF	43,49%	0,96	RF	54,55%	0,85	RF	57,89%	2,16
LightGBM	59,67%	10,18	RF	38,38%	0,61	LightGBM	42,28%	0,69	LightGBM	53,62%	1,16	LightGBM	56,78%	6,72
K-NN	56,22%	0,41	K-NN	36,43%	0,01	K-NN	41,07%	0,07	K-NN	46,84%	0,01	K-NN	54,36%	0,42
Base um														
Áudio			**Acordes**			**Letras**			**Simbólicas**			**Voz**		
Modelo	**Acurácia**	**Tempo (s)**	**Modelo**	**Acurácia**	**Tempo (s)**	**Modelo**	**Acurácia**	**Tempo (s)**	**Modelo**	**Acurácia**	**Tempo (s)**	**Modelo**	**Acurácia**	**Tempo (s)**
SVM	**77,05%**	**0,01**	**CatBoost**	**47,30%**	**5,22**	**SVM**	**53,82%**	**0,01**	**XGBoost**	**65,43%**	**0,20**	**SVM**	**71,95%**	**0,01**
CatBoost	76,48%	61,0	XGBoost	47,30%	0,02	CatBoost	53,25%	8,63	CatBoost	65,15%	9,50	CatBoost	69,97%	51,21
LightGBM	75,63%	4,42	SVM	46,74%	0,01	XGBoost	52,97%	0,02	SVM	64,87%	0,01	LightGBM	69,97%	3,80
RF	72,52%	0,79	RF	46,45%	0,40	RF	50,14%	0,39	RF	64,02%	0,42	RF	68,27%	0,40
XGBoost	72,51%	0,73	LightGBM	46,45%	0,38	LightGBM	48,44%	0,62	LightGBM	59,20%	0,55	XGBoost	67,98%	0,57
K-NN	70,53%	0,01	K-NN	37,39%	0,01	K-NN	44,47%	0,01	K-NN	58,92%	0,01	K-NN	67,13%	0,01

FONTE: O autor (2023).

APÊNDICE M - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (CENÁRIO A – HIPERPARÂMETROS)

Base dois													
Áudio			Acordes			Letras			Simbólicas			Voz		
Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)
SVM	**61,80%**	**1,54**	**SVM**	**42,10%**	**1,50**	**XGBoost**	**45,72%**	**0,38**	**CatBoost**	**56,13%**	**1,51**	**SVM**	**59,75%**	**2,83**
CatBoost	61,71%	5,52	XGBoost	41,63%	0,40	LightGBM	45,07%	0,01	SVM	56,04%	13,13	CatBoost	57,89%	5,44
LightGBM	61,06%	2,95	CatBoost	41,54%	1,66	CatBoost	44,33%	0,97	LightGBM	54,93%	3,32	XGBoost	57,80%	3,56
XGBoost	60,69%	11,00	LightGBM	41,45%	1,33	SVM	44,33%	5,80	RF	54,08%	1,83	LightGBM	57,62%	1,13
RF	60,22%	4,00	RF	39,68%	0,68	RF	44,05%	2,00	KNN	51,95%	0,01	RF	57,52%	4,45
KNN	57,52%	1,54	KNN	37,73%	0,01	KNN	43,49%	0,01	XGBoost	55,01%	0,64	KNN	57,34%	0,63

Base um													
Áudio			Acordes			Letras			Simbólicas			Voz		
Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)
CatBoost	**77,33%**	**19**	**SVM**	**50,42%**	**0,83**	**SVM**	**54,67%**	**1,72**	**SVM**	**68,27%**	**2,27**	**CatBoost**	**72,23%**	**3,81**
SVM	77,05%	0,00	XGBoost	46,74%	0,38	CatBoost	54,10%	3,83	XGBoost	64,31%	0,91	SVM	71,95%	0,38
XGBoost	76,77%	4,15	RF	45,60%	1,20	RF	54,10%	0,44	CatBoost	64,02%	2,06	RF	69,79%	1,98
LightGBM	75,35%	8,13	LightGBM	45,04%	0,01	XGBoost	53,82%	0,01	RF	63,17%	0,38	XGBoost	68,27%	0,83
RF	73,08%	3,00	CatBoost	45,02%	1,62	LightGBM	52,41%	3,57	LightGBM	62,89%	0,01	KNN	66,85%	0,01
KNN	72,80%	0,01	KNN	44,19%	0,01	KNN	50,14%	0,01	KNN	62,32%	0,01	LightGBM	62,89%	0,75

APÊNDICE N - HIPERPARÂMETROS DOS ALGORITMOS DE APRENDIZADO DE MÁQUINA SELECIONADOS

Base dois		Base um	
Modelo	**Melhores hiperparâmetros**	**Modelo**	**Melhores hiperparâmetros**
K-NN (Áudio)	(metric= 'manhattan', n_neighbors= 15, weights= 'distance')	K-NN (Áudio)	(metric= 'manhattan', n_neighbors= 15, weights= 'distance')
K-NN (Acordes)	(metric= 'manhattan', n_neighbors= 15, weights= 'uniform')	K-NN (Acordes)	(metric= 'minkowski', n_neighbors= 15, weights= 'uniform')
K-NN (Letras)	(metric= 'minkowski', n_neighbors= 15, weights= 'uniform')	K-NN (Letras)	(metric= 'minkowski', n_neighbors= 15, weights= 'uniform')
K-NN (Simbólicas)	(metric= 'manhattan', n_neighbors= 15, weights= 'uniform')	K-NN (Simbólicas)	(metric= 'manhattan', n_neighbors= 15, weights= 'uniform')
K-NN (Voz)	(metric= 'manhattan', n_neighbors= 15, weights= 'distance')	K-NN (Voz)	(metric= 'manhattan', n_neighbors= 13, weights= 'distance')
RF (Áudio)	(max_depth= None, max_features= 'auto', min_samples_leaf= 3, n_estimators= 200)	RF (Áudio)	(max_depth= 30, max_features= 'sqrt', min_samples_leaf= 2, n_estimators= 300)
RF (Acordes)	(max_depth= 40, max_features= 'sqrt', min_samples_leaf= 4, n_estimators= 150)	RF (Acordes)	(max_depth= None, max_features= 'sqrt', min_samples_leaf= 5, n_estimators= 150)
RF (Letras)	(max_depth= 20, max_features= 'sqrt', min_samples_leaf= 5, n_estimators= 300)	RF (Letras)	(max_depth= 30, max_features= 'auto', min_samples_leaf= 5, n_estimators= 150)
RF (Simbólicas)	(max_depth= 40, max_features= 'sqrt', min_samples_leaf= 4, n_estimators= 150)	RF (Simbólicas)	(max_depth= 20, max_features= 'sqrt', min_samples_leaf= 5, n_estimators= 200)
RF (Voz)	(max_depth= None, max_features= 'auto', min_samples_leaf= 2, n_estimators= 200)	RF (Voz)	(max_depth= 30, max_features= 'auto', min_samples_leaf= 3, n_estimators= 300)
SVM (Áudio)	(kernel= 'rbf', C= 1)	SVM (Áudio)	(kernel= 'rbf', C= 1)
SVM (Acordes)	(kernel= 'linear', C= 0.1)	SVM (Acordes)	(kernel= 'linear', C= 10)
SVM (Letras)	(kernel= 'linear', C= 10)	SVM (Letras)	(kernel= 'linear', C= 100)
SVM (Simbólicas)	(kernel= 'linear', C= 100)	SVM (Simbólicas)	(kernel= 'linear', C= 100)
SVM (Voz)	(kernel= 'rbf', C= 1)	SVM (Voz)	(kernel= 'rbf', C= 1)

LightGBM (Áudio)	(learning_rate= 0.1, max_depth= 6, min_data_in_leaf= 200, num_leaves= 20)	LightGBM (Áudio)	(learning_rate= 0.1, max_depth= 12, min_data_in_leaf= 10, num_leaves= 100)
LightGBM (Acordes)	(learning_rate= 0.1, max_depth= 3, min_data_in_leaf= 100, num_leaves= 20)	LightGBM (Acordes)	(learning_rate= 0.1, max_depth= 3, min_data_in_leaf= 100, num_leaves= 20)
LightGBM (Letras)	(learning_rate= 0.1, max_depth= 3, min_data_in_leaf= 10, num_leaves= 20)	LightGBM (Letras)	(learning_rate= 0.1, max_depth= 3, min_data_in_leaf= 200, num_leaves= 20)
LightGBM (Simbólicas)	(learning_rate= 0.1, max_depth= 3, min_data_in_leaf= 100, num_leaves= 20)	LightGBM (Simbólicas)	(learning_rate= 0.1, max_depth= 3, min_data_in_leaf= 100, num_leaves= 20)
LightGBM (Voz)	(learning_rate= 0.3, max_depth= 3, min_data_in_leaf= 200, num_leaves= 20)	LightGBM (Voz)	(learning_rate= 0.1, max_depth= 3, min_data_in_leaf= 100, num_leaves= 20)
CatBoost (Áudio)	(depth= 3, iterations= 250, learning_rate= 0.1)	CatBoost (Áudio)	(depth= 5, iterations= 500, learning_rate= 0.1)
CatBoost (Acordes)	(depth= 5, iterations= 250, learning_rate= 0.1)	CatBoost (Acordes)	(depth= 4, iterations= 250, learning_rate= 0.01)
CastBoost (Letras)	(depth= 2, iterations= 250, learning_rate= 0.1)	CastBoost (Letras)	(depth= 5, iterations= 500, learning_rate= 0.01)
CastBoost (Simbólicas)	(depth= 2, iterations= 500, learning_rate= 0.1)	CastBoost (Simbólicas)	(depth= 4, iterations= 500, learning_rate= 0.01)
CastBoost (Voz)	(depth=4, iterations= 250, learning_rate= 0.1)	CastBoost (Voz)	(depth=4, iterations= 100, learning_rate= 0.1)
XGBoost (Áudio)	(colsample_bytree= 0.7, learning_rate= 0.1, max_depth= 12, n_estimators= 300)	XGBoost (Áudio)	(colsample_bytree= 0.4, learning_rate= 0.3, max_depth= 12, n_estimators= 300)
XGBoost (Acordes)	(colsample_bytree= 0.4, learning_rate= 0.1, max_depth= 3, n_estimators= 200)	XGBoost (Acordes)	(colsample_bytree= 0.4, learning_rate= 0.1, max_depth= 3, n_estimators= 100)
XGBoost (Letras)	(colsample_bytree= 0.4, learning_rate= 0.1, max_depth= 3, n_estimators= 100)	XGBoost (Letras)	(colsample_bytree= 0.4, learning_rate= 0.1, max_depth= 3, n_estimators= 100)
XGBoost (Simbólicas)	(colsample_bytree= 1.0, learning_rate= 0.1, max_depth= 3, n_estimators= 100)	XGBoost (Simbólicas)	(colsample_bytree= 0.7, learning_rate= 0.1, max_depth= 3, n_estimators= 100)
XGBoost (Voz)	(colsample_bytree= 0.4, learning_rate= 0.3, max_depth= 12, n_estimators= 200)	XGBoost (Voz)	(colsample_bytree= 0.4, learning_rate= 0.1, max_depth= 6, n_estimators= 100)

FONTE: O autor (2023).

APÊNDICE O - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (PCA – CENÁRIO B)

Base dois													
Áudio			Acordes			Letras			Simbólicas			Voz		
Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)
LightGBM	**62,45%**	**0,91**	***XGBoost***	**40,14%**	**0,71**	***LightGBM***	**43,77%**	**0,48**	***RF***	**51,67%**	**1,83**	***SVM***	**59,10%**	**2,69**
SVM	62,17%	1,69	*CatBoost*	39,86%	2,39	*CatBoost*	43,68%	0,88	***XGBoost***	**51,67%**	**0,40**	*CatBoost*	58,36%	2,86
CatBoost	61,43%	1,80	*LightGBM*	39,68%	0,40	*SVM*	43,21%	1,98	*SVM*	51,57%	84,76	*LightGBM*	57,87%	0,40
XGBoost	59,75%	3,63	*SVM*	38,94%	0,97	*XGBoost*	43,21%	0,02	*CatBoost*	51,02%	2,32	*RF*	57,80%	4,38
RF	59,75%	3,00	*RF*	38,28%	2,40	*K-NN*	42,47%	0,00	*LightGBM*	50,74%	0,41	*K-NN*	57,06%	0,01
K-NN	57,15%	0,01	*K-NN*	37,73%	0,39	*RF*	42,37%	5,01	*K-NN*	49,25%	0,01	*XGBoost*	55,20%	2,47

Base um													
Áudio			Acordes			Letras			Simbólicas			Voz		
Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)
SVM	**76,48%**	**0,01**	***CatBoost***	**47,59%**	**2,14**	***K-NN***	**53,54%**	**0,01**	***LightGBM***	**60,91%**	**0,01**	***SVM***	**72,52%**	**0,01**
CatBoost	74,50%	7,06	*XGBoost*	47,05%	0,01	*XGBoost*	53,26%	0,02	*K-NN*	60,62%	0,01	*K-NN*	70,25%	0,01
LightGBM	74,50%	5,20	*LightGBM*	44,76%	0,43	*CatBoost*	52,69%	1,63	*RF*	60,33%	0,41	*LightGBM*	69,69%	0,40
RF	74,22%	2,15	*SVM*	44,75%	0,64	*SVM*	52,50%	1,67	*SVM*	60,05%	1,28	*XGBoost*	69,41%	0,67
K-NN	73,65%	0,01	*RF*	44,19%	0,93	*RF*	50,70%	0,40	*XGBoost*	59,77%	0,40	*RF*	69,12%	2,02
XGBoost	71,39%	1,47	*K-NN*	44,19%	0,01	*LightGBM*	50,14%	0,01	*CatBoost*	59,20%	1,62	*CatBoost*	62,88%	1,20

FONTE: O autor (2023).

APÊNDICE P - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (SELEÇÃO DE ATRIBUTOS – CENÁRIO B)

Base dois														
Áudio			Acordes			Letras			Simbólicas			Voz		
Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)
CatBoost	**62,45%**	**3,23**	***LightGBM***	**33,36%**	**0,01**	***XGBoost***	**50,28%**	**0,01**	***CatBoost***	**56,04%**	**1,38**	***SVM***	**59,85%**	**1,12**
SVM	60,68%	1,09	*XGBoost*	32,71%	0,01	*CatBoost*	45,44%	0,70	*RF*	55,29%	1,15	*CatBoost*	58,73%	3,71
LightGBM	60,59%	1,41	*CatBoost*	32,06%	0,78	*RF*	45,35%	1,74	*LightGBM*	55,11%	0,41	*LightGBM*	57,53%	0,66
RF	60,13%	3,85	*RF*	31,41%	0,46	*LightGBM*	45,07%	0,39	*XGBoost*	54,92%	0,41	*K-NN*	57,34%	0,39
XGBoost	59,85%	3,99	*SVM*	29,18%	0,41	*SVM*	44,61%	1,34	*SVM*	54,83%	23,36	*XGBoost*	56,97%	4,10
K-NN	57,80%	0,26	*K-NN*	27,78%	0,00	*K-NN*	42,23%	0,01	*K-NN*	52,23%	0,01	*RF*	56,50%	3,99

Base um														
Áudio			Acordes			Letras			Simbólicas			Voz		
Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)
CatBoost	**77,90%**	**17,02**	**XGBoost**	**47,03%**	**0,01**	**XGBoost**	**54,11%**	**0,61**	**SVM**	**67,42%**	**0,75**	**SVM**	**72,80%**	**0,01**
SVM	75,07%	0,01	CatBoost	46,45%	0,98	CatBoost	53,82%	1,78	LightGBM	65,72%	0,01	XGBoost	71,95%	0,40
XGBoost	74,05%	1,17	LightGBM	45,61%	0,01	RF	53,54%	0,35	K-NN	63,45%	0,01	RF	70,82%	1,76
RF	73,65%	1,96	RF	41,36%	0,31	LightGBM	52,41%	0,01	RF	62,32%	1,03	CatBoost	70,82%	1,25
K-NN	73,37%	0,01	K-NN	41,07%	0,01	K-NN	50,14%	0,01	XGBoost	62,32%	0,01	LightGBM	69,41%	0,46
LightGBM	73,09%	7,80	SVM	39,66%	0,01	SVM	49,85%	0,01	CatBoost	62,03%	1,63	K-NN	67,99%	0,01

FONTE: O autor (2023).

APÊNDICE Q - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (*UNDERSAMPLING* – CENÁRIO C)

Base dois														
Áudio			Acordes			Letras			Simbólicas			Voz		
Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)
XGBoost	**60,96%**	**8,36**	**RF**	**40,33%**	**0,70**	**XGBoost**	**43,95%**	**0,22**	**XGBoost**	**55,20%**	**1,12**	**SVM**	**56,97%**	**2,07**
CatBoost	60,22%	5,69	XGBoost	40,24%	0,33	LightGBM	43,86%	1,22	CatBoost	54,46%	1,59	CatBoost	56,41%	4,50
LightGBM	59,38%	2,95	CatBoost	40,05%	2,46	CatBoost	43,49%	1,16	SVM	54,46%	8,81	XGBoost	54,83%	3,29
SVM	59,20%	1,90	SVM	38,38%	0,98	RF	43,49%	2,30	RF	52,88%	0,93	RF	54,73%	3,77
RF	59,01%	4,39	LightGBM	38,01%	0,07	SVM	41,91%	2,50	LightGBM	52,50%	0,45	LightGBM	54,08%	1,33
K-NN	58,92%	0,01	K-NN	35,78%	0,14	K-NN	41,82%	0,06	K-NN	51,85%	0,32	K-NN	53,43%	0,58
Base um														
Áudio			Acordes			Letras			Simbólicas			Voz		
Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)
CatBoost	**76,77%**	**16,7**	**XGBoost**	**52,12%**	**0,18**	**CatBoost**	**53,82%**	**3,83**	**RF**	**66,28%**	**1,14**	**XGBoost**	**72,23%**	**2,45**
SVM	76,48%	0,20	CatBoost	50,70%	1,05	XGBoost	52,40%	0,14	LightGBM	64,87%	0,17	RF	71,95%	3,50
RF	75,07%	2,02	LightGBM	50,42%	0,25	SVM	51,55%	1,77	XGBoost	64,58%	0,10	SVM	71,38%	0,30
XGBoost	74,78%	1,95	RF	49,85%	0,91	RF	50,14%	0,84	SVM	64,30%	1,14	LightGBM	70,25%	3,36
LightGBM	74,78%	9,71	SVM	49,57%	2,55	K-NN	49,00%	0,09	CatBoost	64,30%	4,24	CatBoost	68,55%	3,26
K-NN	71,67%	0,08	K-NN	43,90%	0,04	LightGBM	47,87%	0,12	K-NN	61,47%	0,03	K-NN	67,98%	0,05

FONTE: O autor (2023).

APÊNDICE R - RESULTADOS OBTIDOS PARA O CENÁRIO UM CONSIDERANDO OS ALGORITMOS DE APRENDIZADO DE MÁQUINA (*OVERSAMPLING* – CENÁRIO C)

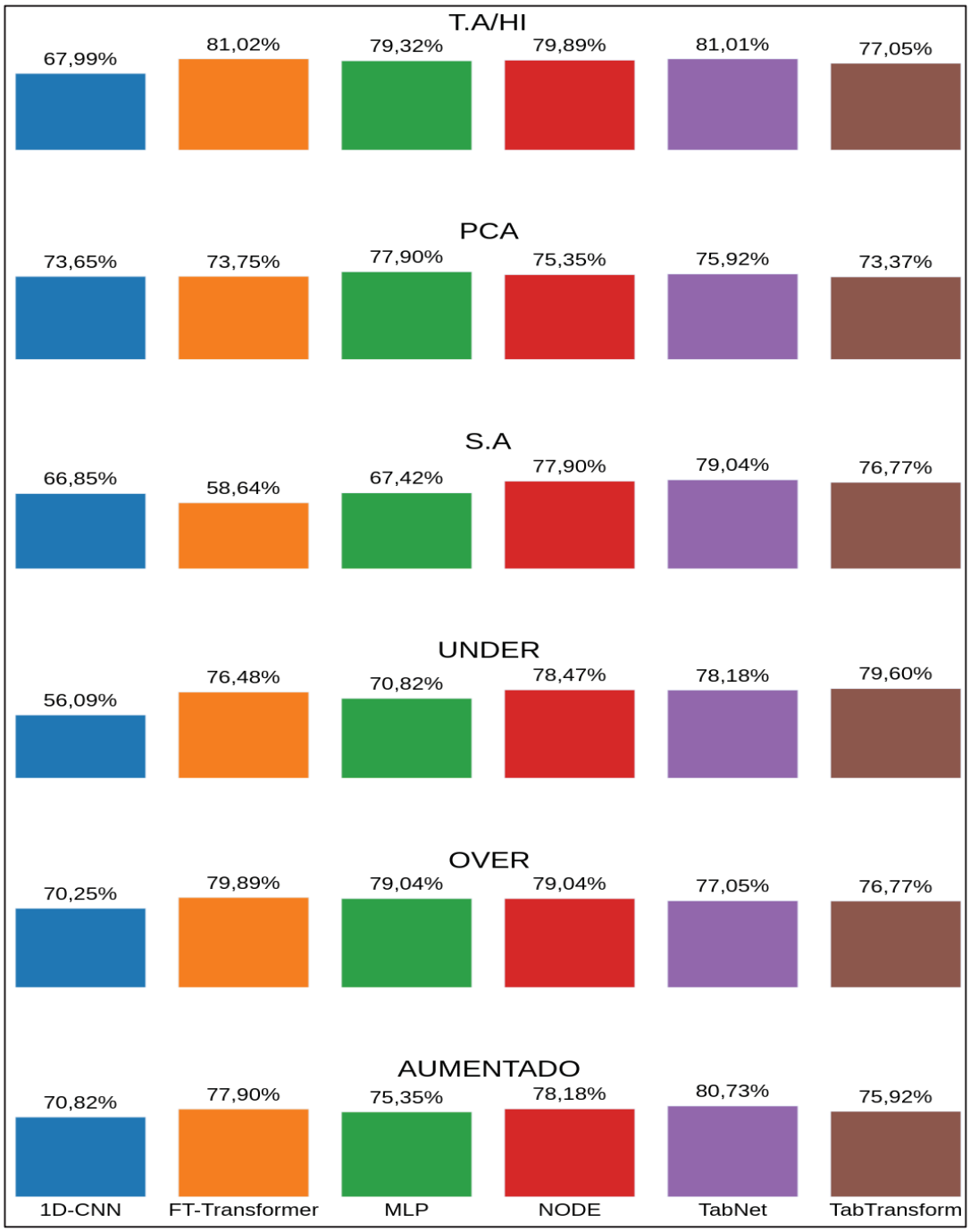
Base dois													
Áudio			Acordes			Letras			Simbólicas			Voz		
Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)
CatBoost	**60,96%**	**3,69**	**CatBoost**	**40,98%**	**4,47**	**CatBoost**	**43,12%**	**1,52**	**CatBoost**	**52,69%**	**5,69**	**CatBoost**	**55,20%**	**7,27**
LightGBM	60,50%	2,13	LightGBM	40,14%	1,04	XGBoost	42,84%	0,63	**XGBoost**	**52,69%**	**2,72**	**XGBoost**	**55,20%**	**6,16**
XGBoost	59,85%	10,71	XGBoost	40,14%	0,60	LightGBM	42,65%	0,49	LightGBM	52,13%	0,53	SVM	55,11%	4,20
RF	59,85%	5,61	RF	39,77%	0,97	SVM	41,63%	4,11	RF	51,95%	1,75	RF	54,83%	5,32
SVM	59,57%	1,50	SVM	39,31%	1,86	RF	41,54%	3,18	SVM	51,39%	15,1	LightGBM	53,25%	0,95
K-NN	53,25%	0,45	K-NN	38,19%	0,38	K-NN	38,66%	0,15	K-NN	50,00%	0,27	K-NN	52,69%	1,45
Base um														
Áudio			Acordes			Letras			Simbólicas			Voz		
Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)	Modelo	Acurácia	Tempo (s)
CatBoost	**80,17%**	**18,95**	**RF**	**52,97%**	**0,71**	**XGBoost**	**52,12%**	**0,30**	**SVM**	**65,72%**	**2,73**	**XGBoost**	**70,53%**	**1,95**
XGBoost	77,05%	1,46	LightGBM	52,12%	0,34	RF	51,27%	1,05	**LightGBM**	**65,72%**	**0,35**	**RF**	**70,53%**	**2,63**
SVM	76,77%	0,32	SVM	51,27%	2,36	CatBoost	50,14%	5,14	XGBoost	64,30%	0,23	**LightGBM**	**70,53%**	**1,58**
LightGBM	76,48%	11,62	XGBoost	48,44%	0,19	SVM	49,00%	2,34	CatBoost	64,30%	3,69	SVM	69,97%	0,27
RF	73,93%	2,76	CatBoost	48,44%	2,47	LightGBM	47,02%	0,69	RF	63,17%	1,83	CatBoost	68,83%	4,08
K-NN	70,25%	0,11	K-NN	41,07%	0,06	K-NN	46,17%	0,04	K-NN	57,22%	0,03	K-NN	63,17%	0,11

FONTE: O autor (2023).

APÊNDICE S – COMPARAÇÃO DO PERCENTUAL DE ACERTO DAS ARQUITETURAS DE DEEP LEARNING (BASE DOIS)

FONTE: O autor (2023).

APÊNDICE T – COMPARAÇÃO DO PERCENTUAL DE ACERTO DAS ARQUITETURAS DE DEEP LEARNING (BASE UM)



FONTE: O autor (2023).

APÊNDICE U – MATRIZES DE CONFUSÃO PARA OS MELHORES RESULTADOS OBTIDOS COM ALGORITMOS DE APRENDIZADO DE MÁQUINA (BASE DOIS)

Acordes				
Emoção	**Alegria**	**Raiva/medo**	**Serenidade**	**Tristeza**
Alegria	**140**	13	44	112
Raiva/medo	63	**27**	21	39
Serenidade	78	2	**57**	166
Tristeza	43	5	37	**229**
Alegria	**45,31%**	4,21%	14,24%	36,25%
Raiva/medo	42,00%	**18,00%**	14,00%	26,00%
Serenidade	25,74%	0,66%	**18,81%**	54,79%
Tristeza	13,69%	1,59%	11,78%	**72,93%**

FONTE: O autor (2023).

Áudio				
Emoção	**Alegria**	**Raiva/medo**	**Serenidade**	**Tristeza**
Alegria	**237**	24	36	12
Raiva/medo	63	**67**	14	6
Serenidade	52	0	**166**	85
Tristeza	23	8	76	**207**
Alegria	**76,70%**	7,77%	11,65%	3,88%
Raiva/medo	42,00%	**44,67%**	9,33%	4,00%
Serenidade	17,16%	0,00%	**54,79%**	28,05%
Tristeza	7,32%	2,55%	24,20%	**65,92%**

FONTE: O autor (2023).

Letras				
Emoção	**Alegria**	**Raiva/medo**	**Serenidade**	**Tristeza**
Alegria	**198**	24	38	49
Raiva/medo	59	**52**	12	27
Serenidade	75	2	**121**	105
Tristeza	54	9	81	**170**
Alegria	**64,08%**	7,77%	12,30%	15,86%
Raiva/medo	39,33%	**34,67%**	8,00%	18,00%
Serenidade	24,75%	0,66%	**39,93%**	34,65%
Tristeza	17,20%	2,87%	25,80%	**54,14%**

FONTE: O autor (2023).

Simbólicas				
Emoção	**Alegria**	**Raiva/medo**	**Serenidade**	**Tristeza**
Alegria	**233**	16	44	16
Raiva/medo	97	**40**	9	4
Serenidade	91	4	**111**	97
Tristeza	25	4	65	**220**
Alegria	**75,40%**	5,18%	14,24%	5,18%
Raiva/medo	64,67%	**26,67%**	6,00%	2,67%
Serenidade	30,03%	1,32%	**36,63%**	32,01%
Tristeza	7,96%	1,27%	20,70%	**70,06%**

FONTE: O autor (2023).

Voz				
Emoção	**Alegria**	**Raiva/medo**	**Serenidade**	**Tristeza**
Alegria	**230**	23	42	14
Raiva/medo	50	**83**	10	7
Serenidade	63	0	**145**	95
Tristeza	33	3	86	**192**
Alegria	**74,43%**	7,44%	13,59%	4,53%
Raiva/medo	33,33%	**55,33%**	6,67%	4,67%
Serenidade	20,79%	0,00%	**47,85%**	31,35%
Tristeza	22,00%	0,96%	27,39%	**61,15%**

FONTE: O autor (2023).

APÊNDICE V – MATRIZES DE CONFUSÃO PARA OS MELHORES RESULTADOS OBTIDOS COM ALGORITMOS DE APRENDIZADO DE MÁQUINA (BASE UM)

Acordes				
Emoção	**Alegria**	**Raiva/medo**	**Serenidade**	**Tristeza**
Alegria	**63**	23	5	20
Raiva/medo	18	**34**	3	6
Serenidade	25	11	**10**	31
Tristeza	15	6	3	**80**
Alegria	**56,76%**	20,72%	4,50%	18,02%
Raiva/medo	29,51%	**55,74%**	4,92%	9,84%
Serenidade	32,47%	14,29%	**12,99%**	40,26%
Tristeza	14,42%	5,77%	2,88%	**76,92%**

FONTE: O autor (2023).

Áudio				
Emoção	**Alegria**	**Raiva/medo**	**Serenidade**	**Tristeza**
Alegria	**97**	9	4	1
Raiva/medo	10	**49**	2	0
Serenidade	12	1	**50**	14
Tristeza	4	0	13	**87**
Alegria	**87,39%**	8,11%	3,60%	0,90%
Raiva/medo	16,39%	**80,33%**	3,28%	0,00%
Serenidade	15,58%	1,30%	**64,94%**	18,18%
Tristeza	3,85%	0,00%	12,50%	**83,65%**

FONTE: O autor (2023).

Letras				
Emoção	**Alegria**	**Raiva/medo**	**Serenidade**	**Tristeza**
Alegria	**80**	8	1	22
Raiva/medo	20	**25**	0	14
Serenidade	20	3	**1**	52
Tristeza	17	3	0	**87**
Alegria	**72,07%**	7,21%	0,90%	19,82%
Raiva/medo	33,90%	**42,37%**	0,00%	23,73%
Serenidade	26,32%	3,95%	**1,32%**	68,42%
Tristeza	15,89%	2,80%	0,00%	**81,31%**

FONTE: O autor (2023).

Simbólicas				
Emoção	**Alegria**	**Raiva/medo**	**Serenidade**	**Tristeza**
Alegria	**89**	11	11	0
Raiva/medo	27	**29**	1	2
Serenidade	16	3	**38**	19
Tristeza	3	2	12	**90**
Alegria	**80,18%**	9,91%	9,91%	0,00%
Raiva/medo	45,76%	**49,15%**	1,69%	3,39%
Serenidade	21,05%	3,95%	**50,00%**	25,00%
Tristeza	2,80%	1,87%	11,21%	**84,11%**

FONTE: O autor (2023).

Voz				
Emoção	**Alegria**	**Raiva/medo**	**Serenidade**	**Tristeza**
Alegria	**97**	5	6	3
Raiva/medo	12	**47**	0	0
Serenidade	15	0	**37**	24
Tristeza	10	1	10	**86**
Alegria	**87,39%**	4,50%	5,41%	2,70%
Raiva/medo	20,34%	**79,66%**	0,00%	0,00%
Serenidade	19,74%	0,00%	**48,68%**	31,58%
Tristeza	9,35%	0,93%	9,35%	**80,37%**

FONTE: O autor (2023).