EEL7513 - Tópico Avançado em Processamento Digital de Sinais

CLASSIFICADOR DE GÊNEROS MUSICAIS

João Paulo Vieira[1], Pedro Pordeus Santos[2]
[1-2]Departamento de Engenharia Elétrica e Eletrônica, Universidade Federal de Santa Catarina, UFSC, Florianópolis, Brasil
joaopaulovieirajpv@gmail.com, pedropordeuss@gmail.com

Sumário

Sumário

1. Introdução

- Problemática: classificar gêneros musicais.
- Abordagens trabalhadas:
  - Extrair características sonoras (espectrais, rítmicas, temporais, harmônicas, e etc.) manualmente e aplicá-las em modelos clássicos de aprendizado de máquina supervisionado;
  - Utilizar redes neurais convolucionais para realizar uma extração automática dessas features (utilizando como entrada espectrogramas de Mel).
- Dataset utilizado: GTZAN.
- Literatura consultada:
  - Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches [1];
  - Convolutional Neural Networks Approach for Music Genre Classification [2].
- Objetivo: encontrar uma solução ótima dentre as abordagens propostas, maximizar o desempenho realizando as extrações de atributos, transformações e otimizações necessárias.

2. Dataset GTZAN

2. a. Descrição do conteúdo

- Contém 1000 músicas, dividido em dez gêneros musicais: Blues, Classical, Jazz, Rock, Country, Disco, Hip-hop, Reggae, Pop e Metal;
- Trechos de áudio de aproximadamente 30 segundos, amostrados em 22.05 kHz;
- Fornece planilhas contendo uma seleção de features extraídas das faixas:
    - O arquivo ‘features_30_sec.csv’ tem features referentes aos trechos de 30 segundos;
    - O arquivo ‘features_3_sec.csv’ tem 10x mais informações, pois, para cada faixa é dividida em dez trechos e são extraídas as features;
- Fornece Espectrogramas de Mel;
- Utilizado em diversos trabalhos acadêmicos.

2. Dataset GTZAN

2. b. Dataset GTZAN: EDA inicial

- 10 trechos faltantes;
- Falta de informações sobre o processo de extração de features, e como foram realizadas as divisões dos trechos;
- Inconsistências na variável ‘length’ (problemática na primeira abordagem);
- Arquivo ‘jazz00054.wav’ corrompido;
- Bordas brancas nos espectrogramas;

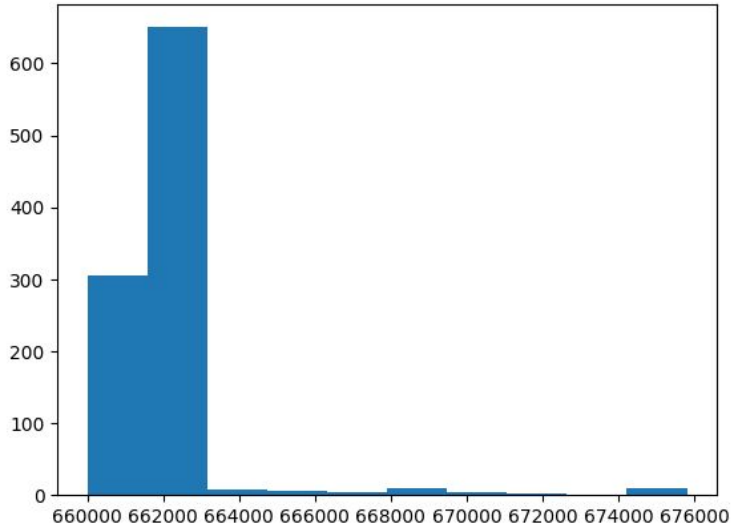


Figura 1: Histograma do comprimento em amostras das faixas de áudio do dataset.

3. Abordagem 1: Extração Manual das Features e Modelos de Classificação

- Features: contêm informações espectrais, rítmicas, temporais e harmônicas das faixas;
- São extraídas em janelas do sinal de entrada, tendo como saída uma matriz Nx1 (ou NxN, no caso dos MFCCs);
- Extrai-se a média e variância associada a cada matriz.

Nome	**Dimensões**
RMS (Média + Variância)	2
Zero-crossing Rate (Média + Variância)	2
Tempo	1
Harmônicas (Média + Variância)	2
Percussive (Média + Variância)	2
Spectral Centroid (Média + Variância)	2
Spectral Rolloff (Média + Variância)	2
Spectral Bandwidth (Média + Variância)	2
MFCCs (Média + Variância)	40
Chromagram (Média + Variância)	2
Nº total de dimensões:	57

Tabela I
FEATURES DO GTZAN E SUAS DIMENSÕES.

3. a. Metodologia

- Entradas de 3 segundos;
- Dataset expandido (10x maior);
- Necessidade de uma divisão agrupada de dados;

3. b. Modelos de Classificação

Modelos	Hiperparâmetros Fixos	Hiperparâmetros Otimizados
Regressão Softmax	max_iter=10000, solver='newton-cg', multi_class = 'multinomial'	C
SVM	max_iter=10000, kernel = 'rbf', random_state = 0	C, gamma
K-Nearest Neighbours	-	n_neighbours, weights, metric
Random Forest	random_state = 0, n_estimators = 250	max_depth, ccp_alpha
Gradient Boosting	random_state = 0, n_iter_no_change = 200	n_estimators, learning_rate
Multi-layer Perceptron	random_state = 0, max_iter = 2000	activation, solver, learning_rate, hidden_layer_sizes, alpha

Tabela II
MODELOS DE CLASSIFICAÇÃO E HIPERPARÂMETROS

- Para os hiperparâmetros não citados, utilizou-se os valores padrões da biblioteca sklearn.

3. c. Divisão dos conjuntos

- Split 60/20/20, em treino, validação e teste;
- Após as otimizações, juntou-se o conjunto de validação ao de treino;

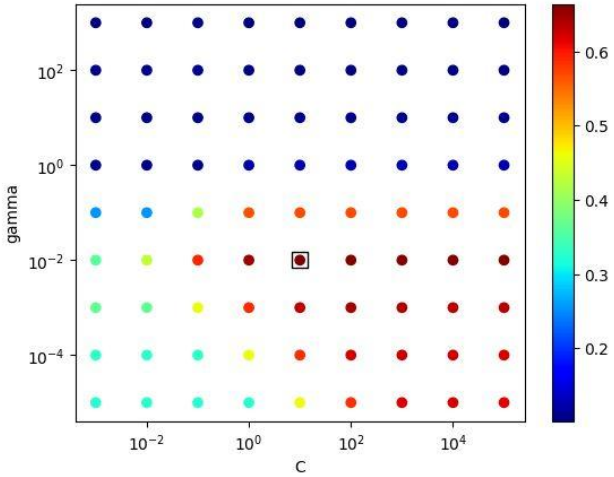
3. d. Métrica

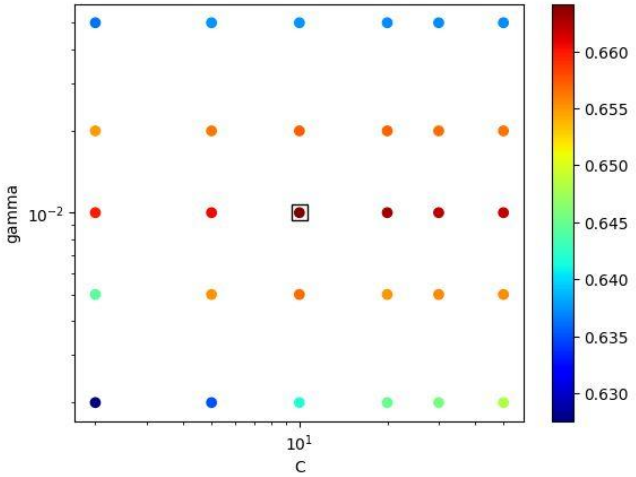
- Considerando um conjunto balanceado, escolheu-se a acurácia;
- Remoção de apenas uma amostra após a EDA;
- Utilizada em ambas as abordagens;

3. e. Otimização de Hiperparâmetros

- Função GridSearchCV, habilitado o parâmetro ‘refit’;
- Transformação realizada: escalonamento dos dados;
- Otimização realizando ajuste fino dos hiperparâmetros.

Figuras 2 e 3: otimização de hiperparâmetros da SVM





3. f. Resultados da Literatura

- O artigo [1] é um estudo dividido em três fases;
- A Fase C utiliza as features do GTZAN;

Classifier	Accuracy	Training Time (s)	Hyperparameters
k-Nearest Neighbours	92.69%	0.0780	nearest neighbours=1
Multilayer Perceptron	81.73%	60.620	activation=ReLu solver lbfgs
Random Forests	80.28%	52.890	number of trees=1000, max depth=10, $\alpha = e^{-5}$, and hidden layer sizes=(5000,10)
Support Vector Machines	74.72%	3.8720	decision function shape=ovo
Logistic Regression	67.52%	3.6720	penaty=12, multi class=multinomial

Figura 4. Resultados de classificação durante a Fase C do artigo [1].

3. f. Resultados da Literatura

- Resultados obtidos ao realizar uma divisão estratificada por classe sem agrupamento dos dados, com entradas de 3 segundos:

Conjunto	SVM	Softmax Regression	KNN	Random Forest	Gradient Boos-ting	MLP
Validação	89,40%	71,80%	91,60%	86,38%	81%	89,30%
Teste	87,50%	71,90%	90,50%	85,50%	78,60%	88,90%
Teste (refit)	89,10%	72,50%	92,10%	87,30%	80,80%	90,70%

Tabela III
DESEMPENHOS EM MODELOS DE CLASSIFICAÇÃO - COM VAZAMENTO DE DADOS

3. g. Divisão agrupada dos dados

- Verificado o vazamento de dados, realizou-se divisão agrupada dos dados, mantendo-se o conjunto balanceado;
- Queda no desempenho;
- Maior confiabilidade nos resultados obtidos;
- Melhores hiperparâmetros encontrados (Tabela V).

Conjunto	SVM	Softmax Regression	KNN	Random Forest	Gradient Boosting	MLP
Validação	66,40%	61,70%	60,69%	64,38%	65,60%	66,70%
Teste (sem refit)	65,20%	61,70%	61,30%	65,40%	66,30%	66,00%
Teste (refit)	74,20%	65,90%	68,00%	72,10%	70,70%	74,00%

Tabela IV
DESEMPENHOS FEATURES INICIAIS - SEM VAZAMENTO DE DADOS - CONJUNTO CSV DATASET

Modelos	Hiperparâmetros
Regressão Softmax	C = 100
SVM	C = 10, alpha = 0.01
K-Nearest Neighbours	metric = 'manhattan', n_neighbours = 3, weights = 'distance'
Random Forest	ccp_alpha = 0.0001, max_depth = 50, n_estimators = 250
Gradient Boosting	learning_rate = 0.001, n_estimators = 100
Multi-layer Perceptron	hidden_layer_sizes = (1000,500,200), activation = 'relu', solver = 'adam', learning_rate = 'adaptive'

Tabela V
MELHORES HIPERPARÂMETROS ENCONTRADOS PARA OS MODELOS DA TABELA IV

3. h. Extração manual de features

- Motivação: aumentar a confiabilidade no procedimento realizado para extrair as features;
- Como deseja-se adicionar novas features, busca-se manter uma equivalência: as features iniciais devem estar associadas ao mesmos trechos que as novas features;
- Procedimento para segmentação: dividir cada faixa em dez partes e eliminar as amostras restantes, caso hajam;
- Baixa perda de informação (máxima perda de 408us de áudio - 9 amostras);
- Erros relativos baixos (~0,1%) para a maioria das features;
- Equivalência nos resultados obtidos.

Conjunto	SVM	Softmax Regression	KNN	Random Forest	Gradient Boosting	MLP
Validação	64,84%	61,12%	59,49%	63,05%	63,50%	65,80%
Teste (sem refit)	64,70%	60,80%	59,10%	63,90%	64,30%	64,40%
Teste (refit)	73,80%	66%	68,10%	72,00%	70,00%	72,50%

Tabela VI
DESEMPENHOS FEATURES INICIAIS - SEM VAZAMENTO DE DADOS - FEATURES EXTRAÍDAS MANUALMENTE

3. i. Adição de features extra

- Foram adicionadas 16 novas features, com base na Fase B do artigo [1];
- Todas possuem a extração de alguma característica espectral (realizou-se a FFT e retirou-se o espectrograma do sinal);

Nome	**Dimensões**
Spectral Contrast (Média + Variância)	2
Spectral Flux (Média + Variância)	2
Spectral Crest (Média + Variância)	2
Spectral Flatness (Média + Variância)	2
Spectral Decrease (Média + Variância)	2
Spectral Slope	1
Spectral Skewness	1
Spectral Spread	1
Spectral Entropy	1
Spectral Variability	1
Peak Smoothness	1
Features GTZAN	57
Nº total de dimensões:	73

Tabela VII
NOVAS FEATURES EXTRAÍDAS E SUAS DIMENSÕES.

3. j. Adição de features extra

- Resultados: aumento de 2,9% na SVM e 3,42% na MLP;
- Entre os desempenhos com e sem ‘refit’, houveram aumentos de 8,6% e 6,7%, evidenciando seu impacto positivo no modelo;

Conjunto	**SVM**	**MLP**
Val	68,20%	69,01%
Teste (sem refit)	68,20%	69%
Teste (refit)	76,70%	75,92%

Tabela VIII
DESEMPENHOS NA SVM E MLP: FEATURES MANUAIS + EXTRAS

Desempenhos [%]	76,88	76,18	75,43	74,67	75,42
Média		75,92%	**Desvio Padrão**		0,46%

Tabela IX
VARIAÇÃO ESTATÍSTICA DO DESEMPENHO DA MLP - FEATURES EXTRA

3. k. Data Augmentation

- Evidenciou-se overfitting em todos os treinamentos;
- Possível causa: poucas amostras de entrada para o modelo;
- Solução proposta: realizar transformações no conjunto de treino;
- Transformações feitas:
  - Normalização de ganho (transformação linear);
  - Pitch Shift (+5%);
  - Speed (+5%);
  - Echo;
  - Reverb;
  - Contrast;
  - Low-pass Filter (8kHz);
- Houve queda no desempenho para os melhores modelos (SVM e MLP);

3. l. Variando o número de MFCCs + Histograma de Batidas

- Proposta: verificar o impacto do número de MFCCs no desempenho do modelo;
- Seria necessário aumentar, diminuir, ou manter o mesmo?

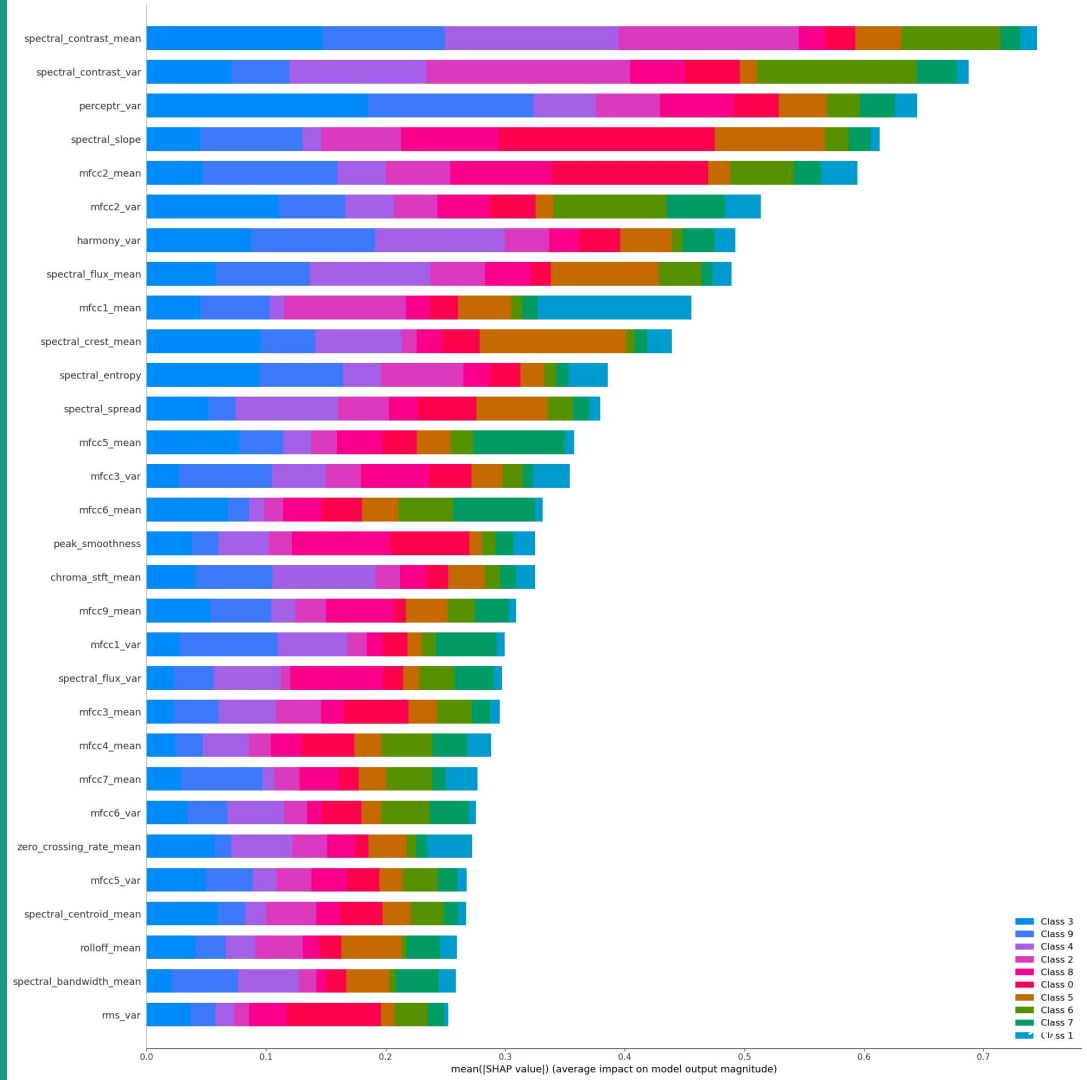
Nº MFCCs	15	20	25	30	40
Desempenho Teste (refit)	73,20%	76,70%	77%	76,00%	75,10%

Tabela X
DESEMPENHOS PARA O CONJUNTO DE FEATURES MANUAIS + EXTRAS - VARIANDO NÚMERO DOS MFCCS (SVM)

- Adição do histograma de batidas, feature que faz uma representação do padrão rítmico da música;
- Ganho baixíssimo no desempenho com um aumento considerável de dimensões (109).

3. m. Features mais importantes para o modelo

- Utilizar os Shap Values para estimar as variáveis mais importantes para o modelo (MLP);
- Deep Explainer;
- Dentre as 30 features mais importantes, 9 delas pertencem ao grupo de features extra adicionados, 13 são médias e variâncias dos MFCCs e as 8 restantes pertencentes ao grupo de inicial.



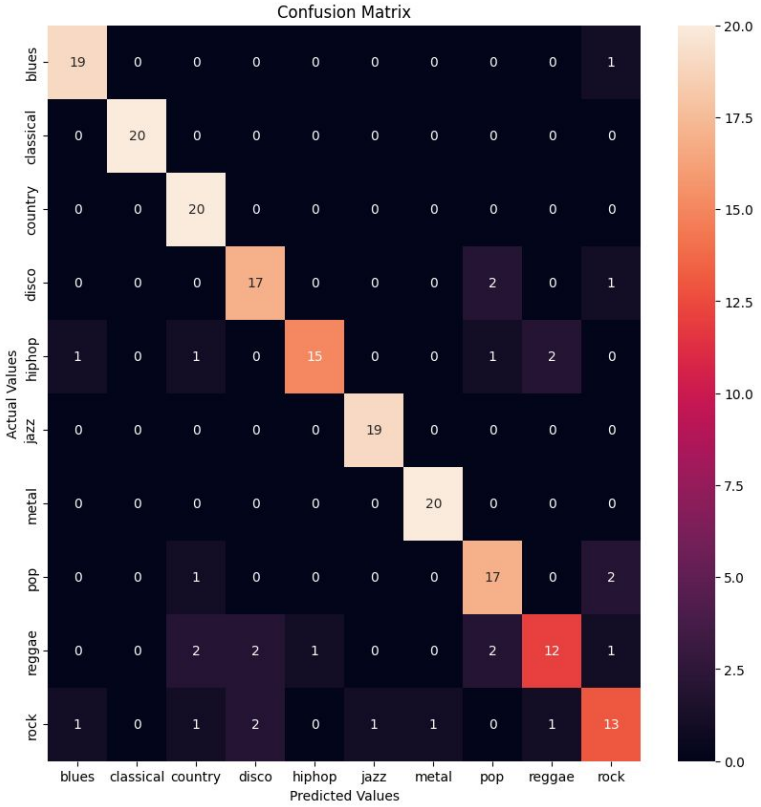
3. n. Redução da dimensionalidade

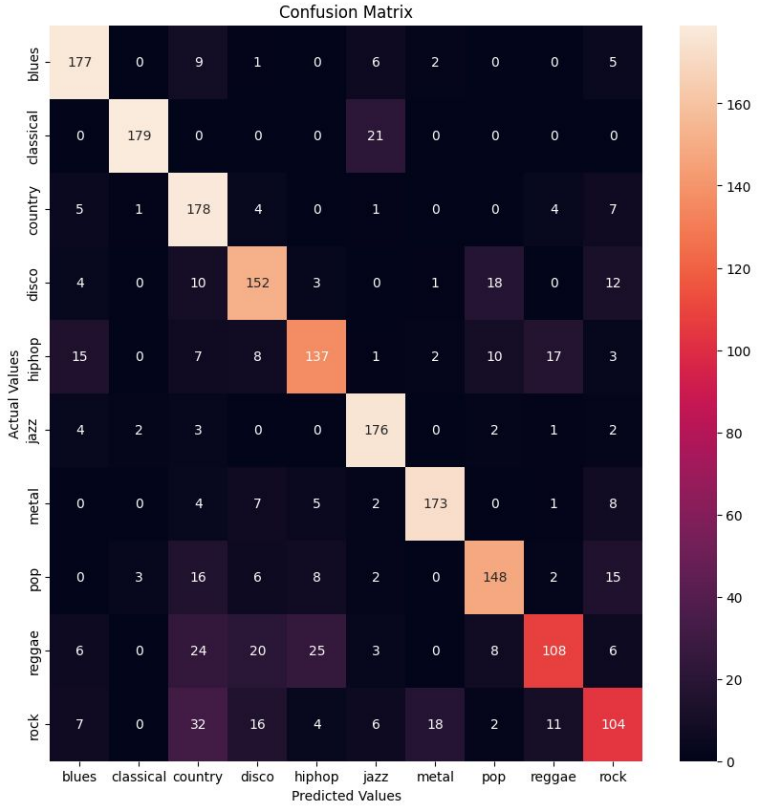
- Testes com as 30, 40 e 50 features mais importantes;
- Queda considerável no desempenho para os dois primeiros casos;
- Ganho de 0,1% no desempenho com 50 features -> reduziu-se de 83 para 50 o número de dimensões;

3. o. Ensemble

- O modelo prevê 10 saídas para cada trecho -> é possível extrair a média aritmética entre essas saídas para obter a previsão para uma única faixa.
- Desempenho máximo obtido: 86,4%.

3. o. Ensemble: matriz de confusão





3. p. Últimos testes realizados e conclusões

- Por que não utilizar features de 3 segundos ao mesmo tempo que features de 6, 15 e 30 segundos?
- Aumentar o número de features sem aumentar o número de linhas no dataset;
- Realizado esse teste, obtiveram-se resultados de 79,5% (MLP) e 81,5% (SVM) -> 1275 features (999 entradas);
- Para features extraídas de apenas 30s (com 25 MFCCs e as features extra), obteve-se 80,5% na SVM);
- Outra possibilidade, a ser testada, seria dividir as faixas, também, em 5 trechos de 6 segundos e 2 trechos de 15s, o que aumentaria em 70% o conjunto de dados.

4. Abordagem 2: redes neurais convolucionais

- Aplicação de redes neurais convolucionais em espectrogramas de escala Mel e MFCCs
- Testes de duas arquiteturas encontradas na literatura
- Testes de modelos pré treinados disponíveis na biblioteca Torchvision
- Busca pelo melhor modelo e aprimoramentos

4. a. Implementação de modelos da literatura

- Modelo de Ndou, Ajoodha e Jadhav(2021)
- 5 blocos convolucionais compostos por
  - Uma camada convolucional de kernel (3,3), stride(1,1) e padding = “same”
  - Função de Ativação ReLU
  - Max Pooling com stride e janela (2,2)
  - Regularização Dropout de probabilidade 0.2
- Camada FC com função de ativação Softmax
- Otimizador SGD (Não houve convergência com o Adam)
- Testes com 2 Splits: 60-20-20 e 80-10-10

Split 60-20-20

Momentum	Weight_decay	Épocas	Acurácia de Validação
0.9	0.00001	200	52%
0.9	0.0001	200	47%
0.9	0.001	200	58%
0.8	0.00001	200	52.5%
0.8	0.0001	200	49.5%
0.8	0.001	200	52.5%
0.7	0.00001	300	49%
0.7	0.0001	300	53.5%
0.7	0.001	300	45.5%
0.6	0.00001	500	46.5%

Tabela XI
OTIMIZAÇÃO DE HIPERPARÂMETROS DO MODELO DE NDOU, AJOODHA E JADHAV(2021) PARA O SPLIT 60-20-20

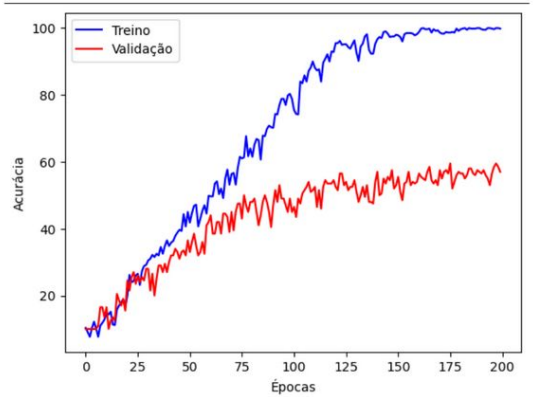


Figura 7. Melhor resultado de treinamento do modelo de Ndou, Ajoodha e Jadhav para o split 60-20-20.

Split 80-10-10

Momentum	Weight_decay	Épocas	Acurácia de Validação
0.8	0.001	250	62%
0.8	0.0001	250	61%
0.7	0.01	500	60%
0.9	0.01	200	63.1%

Tabela XII
OTIMIZAÇÃO DE HIPERPARÂMETROS PARA O MODELO DE NDOU, ADOODHA E JADHAV(2021) PARA O SPLIT 80-10-10

4. a. Implementação de modelos da literatura

- Modelo de Cheng, Chang e Kuo (2020)
- Mesma arquitetura com algumas modificações
  - Regularização Dropout com 0.5 de probabilidade
  - Otimizador Adam
  - Batch Size 32
- Teste com hiperparâmetros do artigo vs Encontrados anteriormente

4. a. Implementação de modelos da literatura

Acurácia de Validação : 41.5%

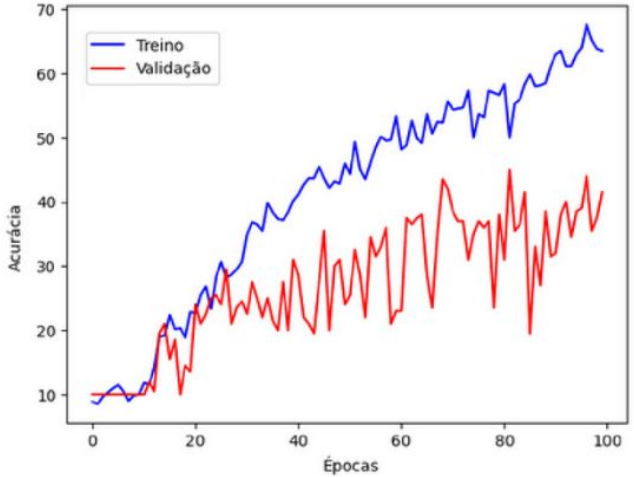


Figura 8. Treinamento do modelo de Cheng, Chang e Kuo (2020) para os hiperparâmetros descritos no artigo

Acurácia de Validação : 58.5%

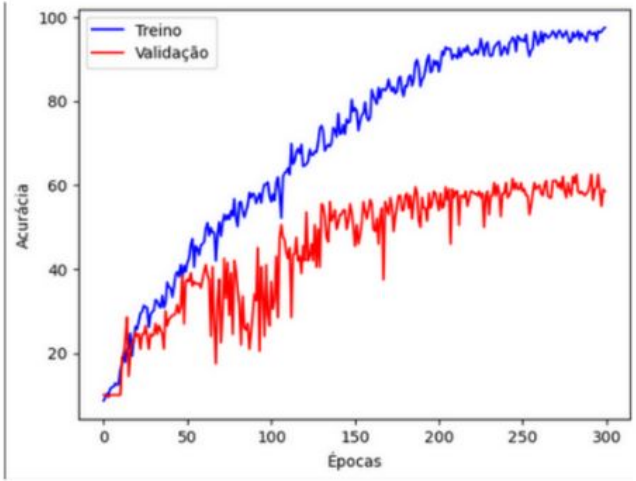


Figura 9. Gráfico de treinamento do modelo de Cheng, Chang e Kuo (2020) para os hiperparâmetros otimizados

4. b. Avaliação de Modelos Pré Treinados

- Implementação de modelos pré treinados disponíveis na biblioteca torchvision
- Modificação da camada de saída para 10 saídas de classificação
- Testes feitos com mesmos hiperparâmetros com melhor resultado dos modelos anteriores, com 100 épocas:
    - SGD: Learning Rate = 0.01, Weight Decay = 0.001, momentum = 0.9, Épocas = 100

4. b. Avaliação de Modelos Pré Treinados

Modelo	Acurácia de Validação	Número de Parâmetros
Resnet18	68,0%	11.7M
AlexNet	63,0%	61.1M
VGG16	70,5%	138.4M
DenseNet161	71,0%	28.7M
GoogleNet	65,0%	6.6M
ShuffleNet_v2_x1_0	64,5%	1.4M
MobileNet_v2	66,0%	3.5M
ResNext50_32x4d	63,0%	25.0M
Wide_resnet50_2	63,5%	68.9M

Tabela XIII
ACURÁCIA DE VALIDAÇÃO E NÚMERO DE PARÂMETROS PARA MODELOS PRÉ TREINADOS DA BIBLIOTECA TORCHVISION

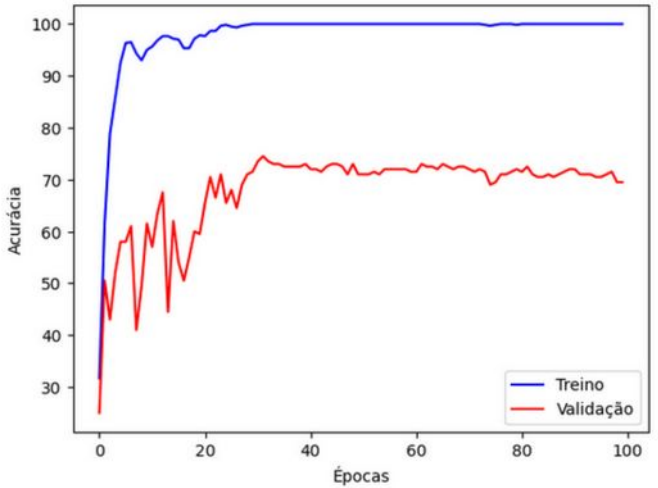
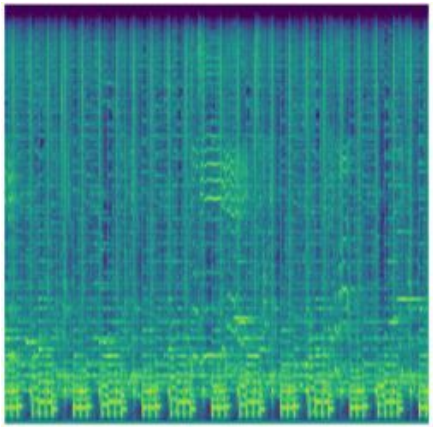


Figura 10. Gráfico de treinamento do modelo DenseNet161

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
I - Extração própria de Espectrogramas

- Obter melhor resolução e informação das músicas
  - Falta de informação do processo de extração original do dataset
- Extração pela biblioteca Torchaudio
  - FFT de tamanho 2048
  - 256 bancos de filtro de escala Mel
  - Imagem final redimensionada para 128x128



4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
I - Extração própria de Espectrogramas

- Aumento da acurácia de validação para 75.5%
- Redução do tempo de treinamento para 70 épocas

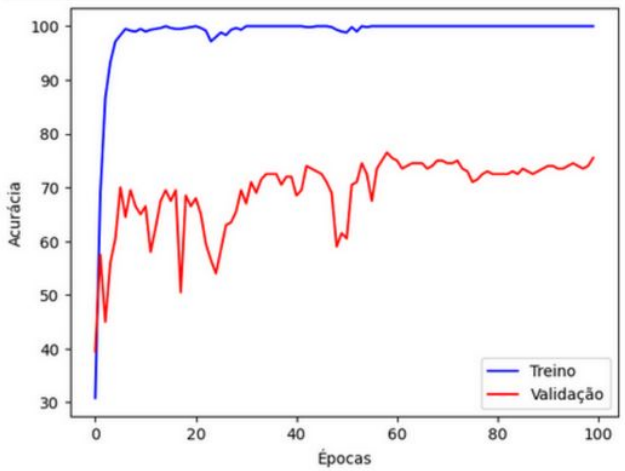


Figura 11. Gráfico de treinamento do modelo DenseNet161 para os espectrogramas extraídos

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
II - Implementação de Early Stopping

- Modelo salva os pesos com melhor acurácia de validação ao decorrer do treinamento
- Interrompe o treinamento caso não haja melhora na acurácia de validação
- Aumento da acurácia de validação para 79.5%
- Todos os testes seguintes aplicados Early Stopping

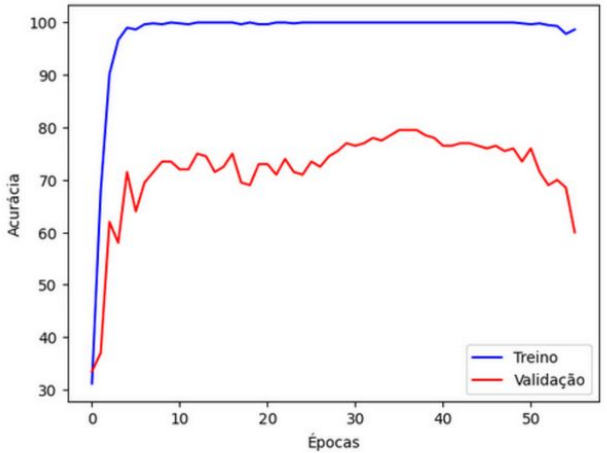


Figura 12. Gráfico de treinamento do modelo DenseNet161 aplicado EarlyStopping

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
III - Otimização de Hiperparâmetros (espec. de 30 segundos)

Momentum	Learning Rate	Weight Decay	Época Limite	Época de Max. Acc	Paciência	Acurácia de Validação
0.9	0.01	0.001	70	35	20	79.5%
0.9	0.01	0.0001	70	35	30	76.0%
0.9	0.01	0.00001	70	47	30	76.0%
0.9	0.001	0.001	100	6	40	68%%
0.9	0.001	0.0001	100	14	40	70.0%
0.9	0.001	0.00001	100	70	40	70.5%
0.9	0.0001	0.001	120	109	50	65.0%
0.9	0.0001	0.0001	120	80	50	67.5%
0.9	0.0001	0.00001	120	88	50	67.5%
0.8	0.01	0.001	70	11	30	72.0%
0.8	0.01	0.0001	70	13	30	73.5%
0.8	0.01	0.00001	70	22	30	72.5%
0.8	0.001	0.001	100	57	40	70.5%
0.8	0.001	0.0001	100	20	40	70.0%
0.8	0.001	0.00001	100	51	40	69.5%
0.8	0.0001	0.001	120	116	50	65.5%
0.8	0.0001	0.0001	120	73	50	66.5%
0.8	0.0001	0.00001	120	118	50	67%

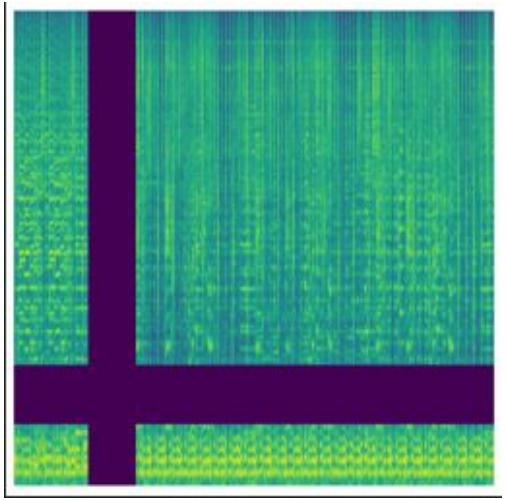
Tabela XIV
TABELA DE OTIMIZAÇÃO DE HIPERPARÂMETROS PARA O MODELO DENSENET161 COM OTIMIZADOR SGD

Learning Rate	Weight Decay	Época Limite	Época de Max. Acc	Paciência	Acurácia de Validação
0.01	0.001	70	26	30	52.5%
0.0005	0.0005	100	49	30	79.0%
0.0005	0.0001	100	42	30	78.0%
0.0001	0.01	70	42	30	75.5%
0.0001	0.001	70	14	30	75.5%
0.0001	0.0001	100	16	30	76.5%
0.0001	0.0005	100	60	30	79.0%
0.00001	0.001	70	51	30	73.5%

Tabela XV
TABELA DE OTIMIZAÇÃO DE HIPERPARÂMETROS PARA O MODELO DENSENET161 COM OTIMIZADOR ADAM

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
IV - Data Augmentation

- Buscando reduzir overfitting aumentando o conjunto de treino, aplicou-se Data Augmentation nos espectrogramas
- Funções de Data Augmentation aplicadas no conjunto treino+validação:
  - Time Masking
  - Frequency Masking
- Piora drástica no desempenho : acurácia de validação de 35.5%
- Funções aplicadas não são adequadas para a aplicação em questão



4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
IV - Data Augmentation

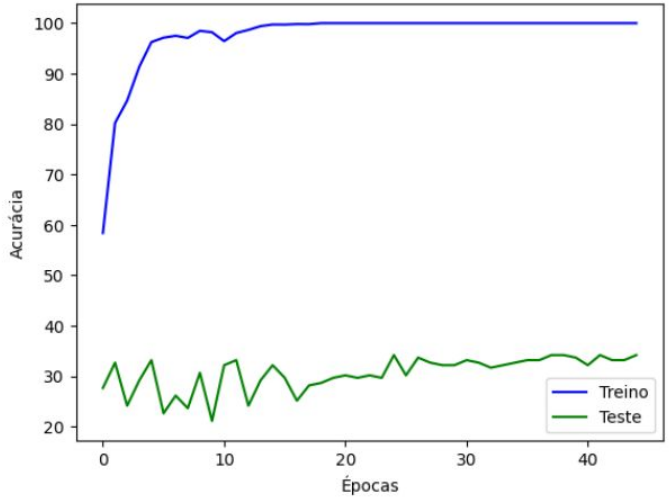


Figura 13. Gráfico de treinamento do modelo DenseNet161 aplicado Data Augmentation

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
V - Extração de espectrogramas para 3 segundos de áudio

- Aumento do conjunto de treino particionando os áudios
  - Equivalência entre as abordagens
- Mesmos parâmetros de extração de espectrogramas
- Redimensionamento das imagens para 64x64 para reduzir o tempo de treinamento
- Leve ganho na acurácia de validação: 80.85%

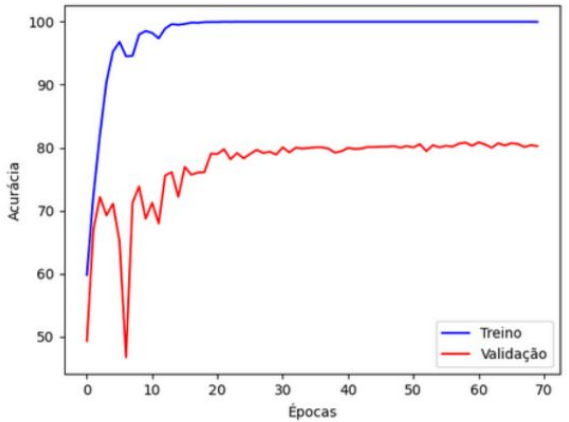


Figura 14. Gráfico de treinamento do modelo DenseNet161 para espectrogramas de 3 segundos

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
V - Otimização de hiperparâmetros (espec. de 3 segundos)

Momentum	Learning Rate	Weight Decay	Época Limite	Época de Max. Acc	Paciência	Acurácia de Validação
0.9	0.1	0.001	70	42	30	61.35%
0.9	0.01	0.01	70	15	30	72.35%
0.9	0.01	0.001	70	60	30	80.85%
0.9	0.01	0.0001	70	43	30	76.6%
0.9	0.01	0.00001	70	22	30	79.0%
0.9	0.001	0.001	100	64	40	73.75%

Tabela XVI
OTIMIZAÇÃO DE HIPERPARÂMETROS PARA O MODELO DENSENET161
COM OTIMIZADOR SGD PARA ESPECTROGRAMAS DE 3 SEGUNDOS

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
VI - Otimização de hiperparâmetros (espec. de 3 segundos)

Learning Rate	Weight Decay	Época Limite	Época de Max. Acc	Paciência	Acurácia de Validação
0.0001	0.0001	70	59	30	80.15%
0.0001	0.0005	70	30	30	78.50%
0.0005	0.0001	70	28	30	80.65%
0.0005	0.0005	70	40	30	80.05%

Tabela XVII
OTIMIZAÇÃO DE HIPERPARÂMETROS PARA O MODELO DENSENET161 COM OTIMIZADOR ADAM PARA ESPECTROGRAMAS DE 3 SEGUNDOS

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
VII - Inferência no conjunto de testes com ensemble

- Implementação do ensemble para a inferência no conjunto de testes da CNN
- Classifica pela média das probabilidades das previsões para trechos de 3 segundos da música
- Ganho considerável na acurácia de treino: 83.9%

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
VIII - Ensemble no treinamento do modelo

- Outra maneira de implementar o ensemble é durante o treinamento do modelo
- O cálculo da perda, e consequentemente dos gradientes, é feita pela média dos outputs dos trechos de 3 segundos de uma dada música
- Inferência também feita pela média
- Acurácia de Treino : 82.9%
- Resultado muito próximo da abordagem do ensemble apenas na inferência de teste sendo inconclusivo qual é superior.

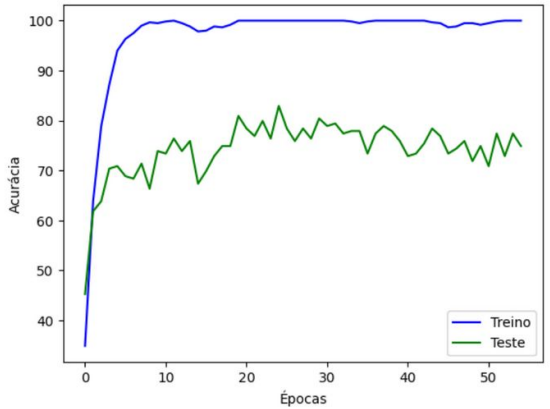
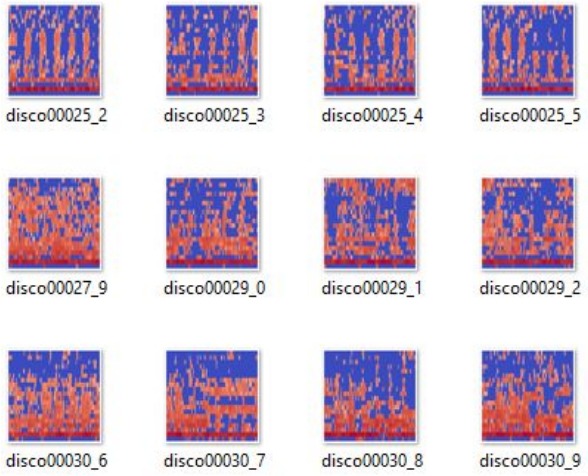


Figura 15. Gráfico de treinamento do modelo DenseNet161 com ensemble no treino

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
IX - Extração própria de MFCCs

- Outra modo de representar os áudios por imagem é por MFCCs
- Realizou-se a extração dos MFCCs dos trechos de 3 segundos com hiperparâmetros com 20 coeficientes de Mel.
- Imagem redimensionada para 64x64
- Otimização dos hiperparâmetros do otimizador
- Inferência no teste com ensemble



4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
IX - Extração própria de MFCCs

Momentum	Learning Rate	Weight Decay	Época Limite	Época de Max. Acc	Paciência	Acurácia de Validação
0.9	0.01	0.001	70	67	35	64.7%
0.9	0.001	0.001	70	2	35	56.8%
0.9	0.0005	0.001	701	17	35	55.0%
0.9	0.01	0.01	70	17	35	57.1%
0.9	0.01	0.0005	70	50	35	63.0%
0.9	0.01	0.0001	70	48	35	60.2%
0.8	0.01	0.001	70	51	35	62.25%
0.8	0.001	0.001	120	95	60	55.1%

Tabela XVIII
OTIMIZAÇÃO DE HIPERPARÂMETROS PARA O MODELO DENSENET161 COM O OTIMIZADOR SGD PARA DADOS DE ENTRADA NO FORMATO MFCC DE 3 SEGUNDOS

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
IX - Extração própria de MFCCs

Learning Rate	Weight Decay	Época Limite	Época de Max. Acc	Paciência	Acurácia de Validação
0.0001	0.0001	100	98	50	62.7%
0.0001	0.0005	100	72	50	62.35%
0.0005	0.0001	100	85	50	65.70%
0.0005	0.0005	100	28	50	63.25%

Tabela XIX
OTIMIZAÇÃO DE HIPERPARÂMETROS PARA O MODELO DENSENET161 COM OTIMIZADOR ADAM PARA DADOS DE ENTRADA NO FORMATO MFCC DE 3 SEGUNDOS

4. c. Aprimoramentos do Modelo Escolhido : (Densenet161)
IX - Extração própria de MFCCs

- Resultado final com inferência de teste com ensemble: 76.88%
- Ganho muito considerável em relação à acurácia de validação de 65.7%, porém com desempenho abaixo do modelo treinado com espectrogramas de Mel.

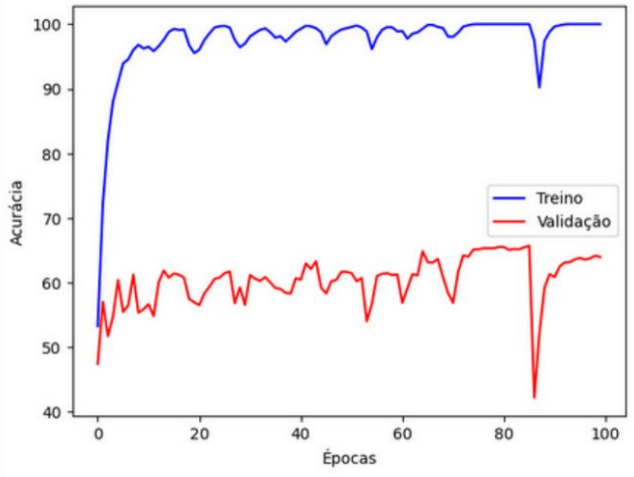
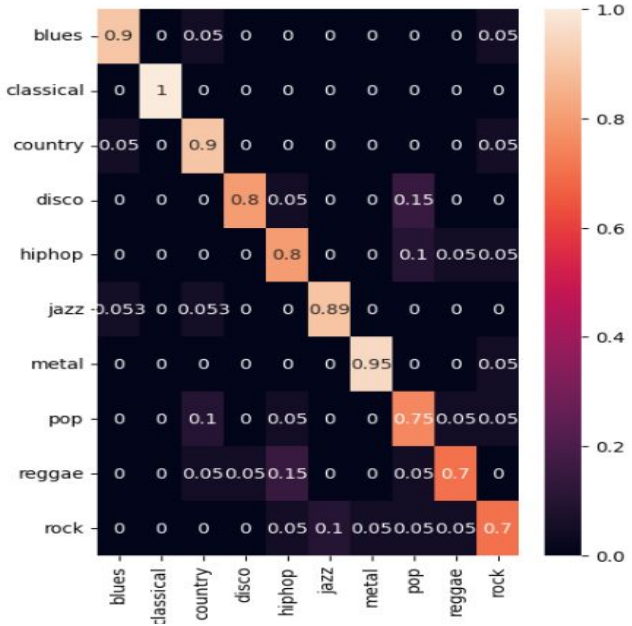


Figura 16. Gráfico de treinamento do modelo DenseNet161 para dados de entrada no formato MFCC de 3 segundos

4. d. Análise de Resultados e Possíveis melhorias

- Modelos pré treinados para classificação de imagem se mostraram muito efetivos para a aplicação
- Early Stopping e Ensemble foram as estratégias mais importantes no ganho de acurácia
- Ensemble no treino ou somente no teste ainda é inconclusivo qual seria o melhor, sendo necessário mais testes para adquirir resultados mais isentos da aleatoriedade do treinamento
- Pela matriz de confusão as previsões erradas são condizentes com as semelhanças em timbre, ritmo, instrumental, etc. Exemplo: Disco e Pop

4. d. Análise de Resultados e Possíveis melhorias



4. d. Análise de Resultados e Possíveis melhorias

- Possíveis Melhorias para a abordagem 2:
  - Aquisição de mais testes de combinação de hiperparâmetros para os diferentes otimizadores, estratégias e tipos de dados de entrada.
  - Otimização de dimensões da imagem de entrada, tanto para espectrogramas quanto MFCCs para melhor acurácia de validação.
  - Otimização dos hiperparâmetros de extração dos espectrogramas e MFCCs para melhor acurácia de validação.
  - Testes com diferentes tempos para as partições dos áudios a serem aplicados no ensemble.
  - Testes com alterações nas camadas de saída do modelo

5. Comparações entre abordagens e conclusões finais

- Ambas abordagens apresentam ganho de desempenho para trechos menores (3 segundos), havendo duas hipóteses possíveis:
  - Aumento considerável no conjunto de treino;
  - Extração de features mais localizada;
- Modelo com melhor desempenho final: SVM com 50 melhores features e ensemble na inferência de teste

Modelo	Desempenho
SVM - Features extra (3s) + Ensemble	86,40%
SVM - Features extra (30 segundos)	80,50%
Densenet161 - Espectrogramas de 30 segundos	79,50%
Densenet161 - Espectrogramas de 3s + Ensemble no teste	83,90%

Tabela XX
COMPARAÇÃO ENTRE OS RESULTADOS DO ENSEMBLE X RESULTADOS PARA ENTRADAS DE 30S

Modelo	Desempenho
SVM - Features extra (3s)	77,1%
CNN - Espectrogramas (3s)	80,85%

Tabela XXI
COMPARAÇÃO ENTRE OS MODELOS COM ENTRADAS DE 3 SEGUNDOS (FEATURES X ESPECTROGRAMAS

6. Referências

- N. Ndou, R. Ajoodha and A. Jadhav, ”Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches,”2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 2021, pp. 1-6, doi: 10.1109/IEMTRONICS52119.2021.9422487.
- Y. -H. Cheng, P. -C. Chang and C. -N. Kuo, ”Convolutional Neural Networks Approach for Music Genre Classification,”2020 International Symposium on Computer, Consumer and Control (IS3C), Taichung City, Taiwan, 2020, pp. 399-403, doi: 10.1109/IS3C50286.2020.00109

Obrigado!

Dúvidas?