Modelos MLP e CNN — UrbanSound8K

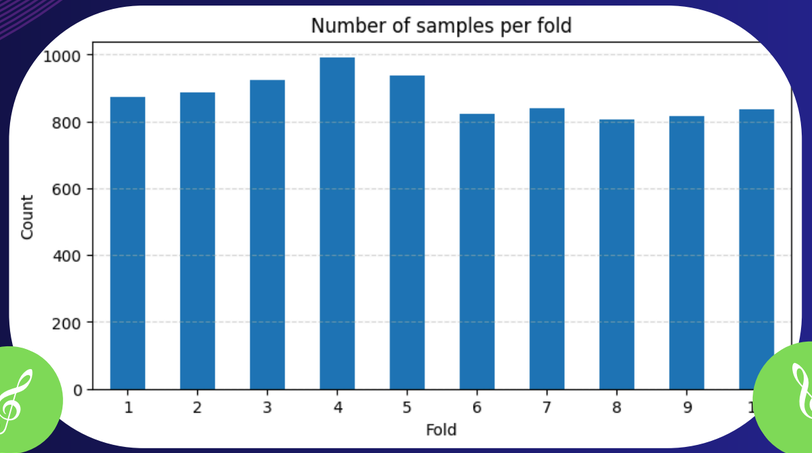
Classificação de Áudio com Deep Learning

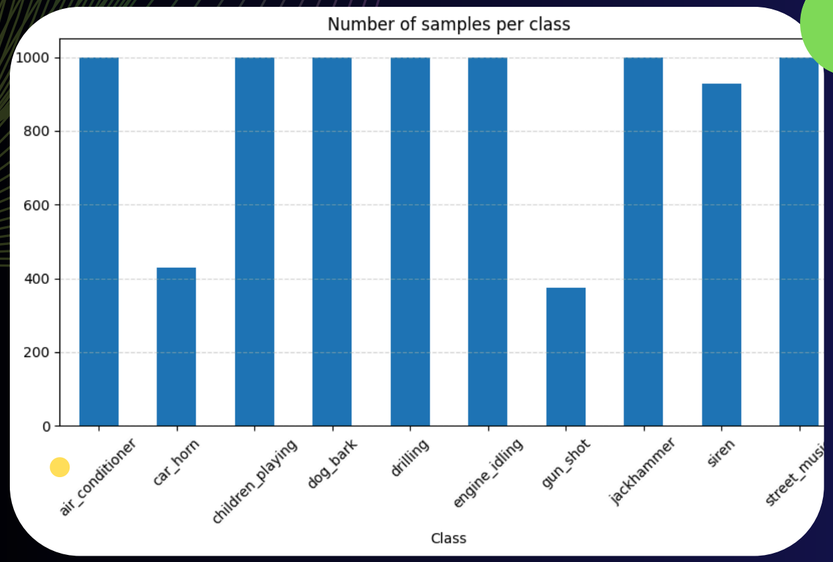
Aprendizagem Computacional II – LIACD 2025/26

Autores: Alexandre Furriel, Daniel Gomes e Liliana Silva

Características do Dataset

- 8732 excertos de áudio (≤ 4s)
- 10 classes de sons urbanos
- Divisão oficial em 10 folds
- Variabilidade elevada entre samples (ambiente, volume, ruído)

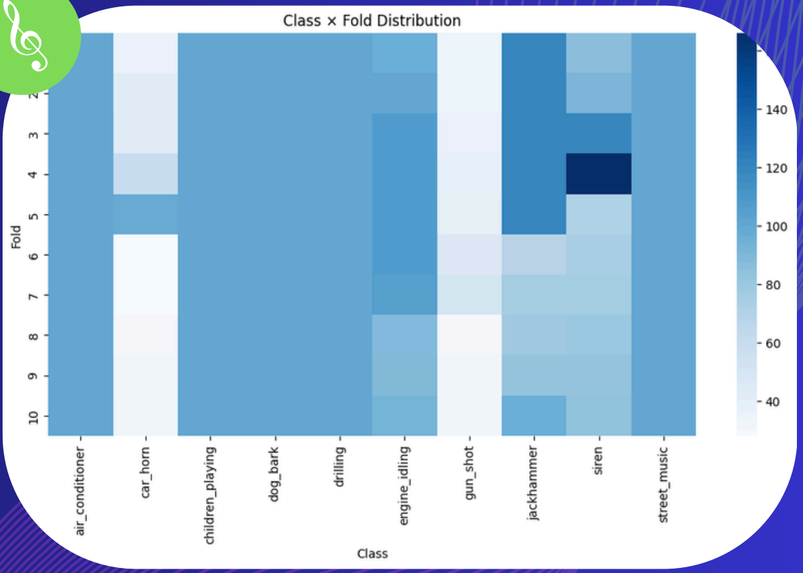




Dataset

Padrões observados na análise

- Desbalanceamento claro em classes como car_horn e gun_shot
- Tamanho dos folds moderadamente desigual
- Distribuição classe × fold bastante estável



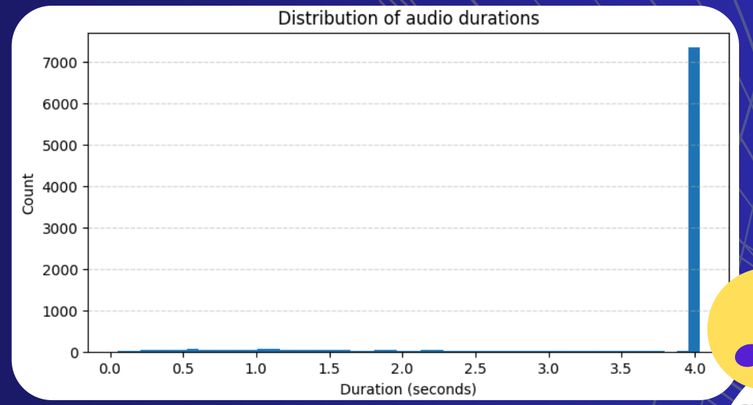
Pré-Processamento

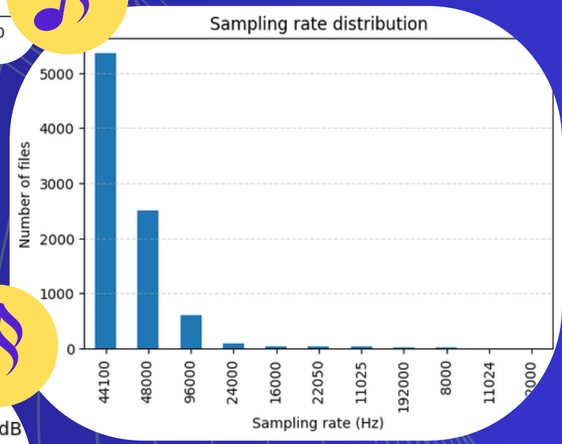
Desafios

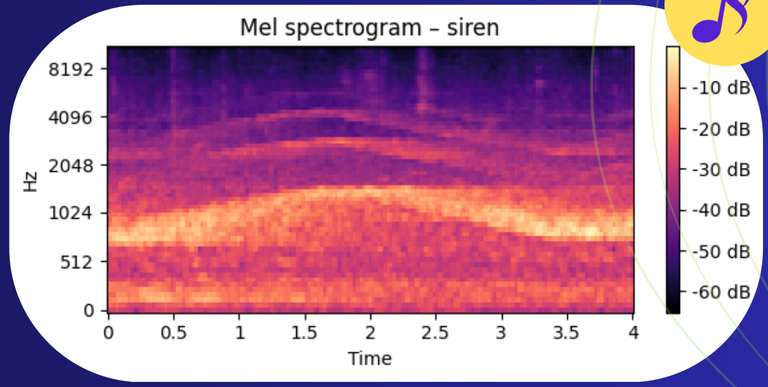
- Durações inconsistentes
- Sampling rates diferentes
- Variabilidade de ruído e volume

Decisões

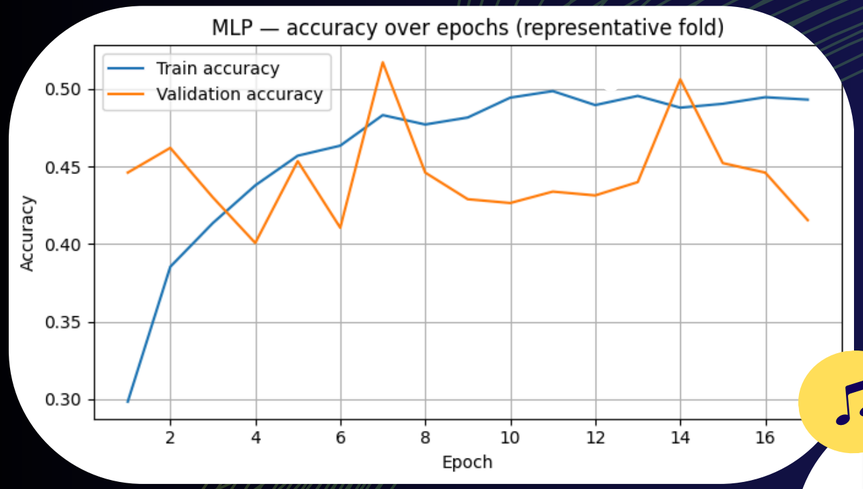
- Uniformização da sampling rate
- Padding para duração fixa
- Normalização dos sinais
- Extração de Mel Spectrograms
- Preparação de tensores para MLP e CNN





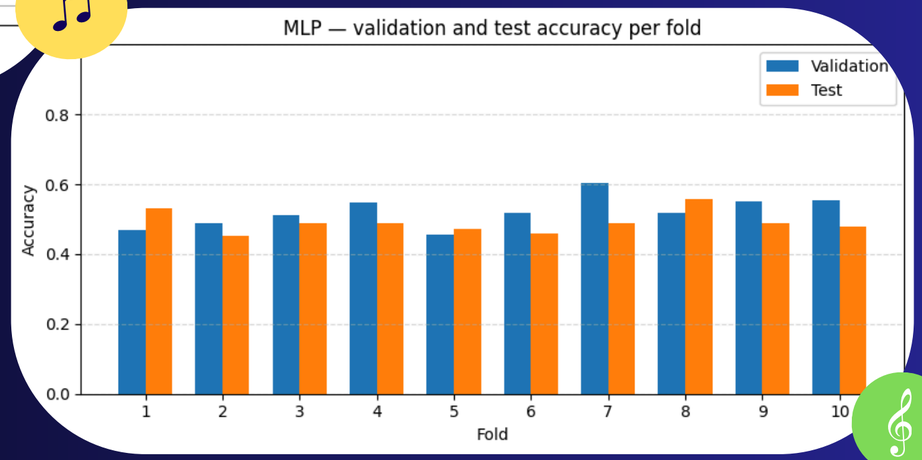


MLP — Modelo Baseline



Input

- MelMel spectrogram (64 × 173) flatten → vetor 1D (~11k features)



Arquitetura

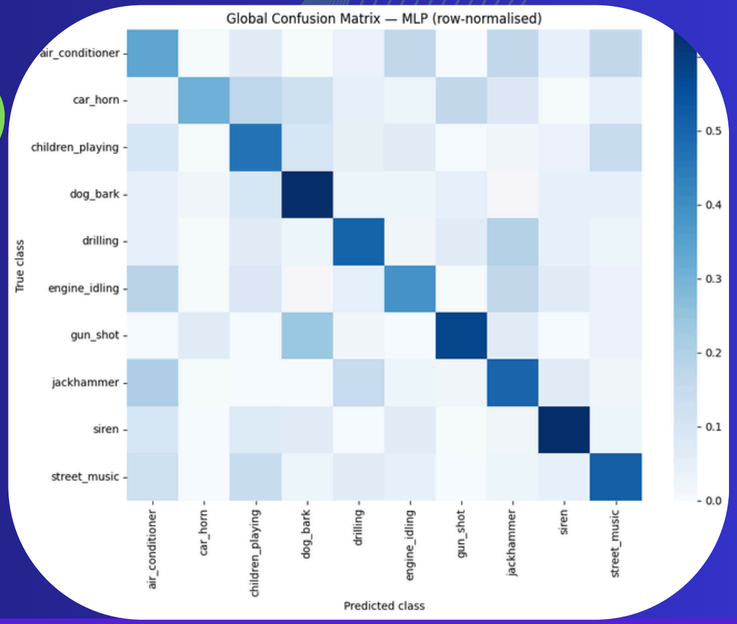
- 2 camadas Dense (512 → 256) + ReLU
- Batch Normalization
- Dropout (0.5)
- Softmax para 10 classes

Comportamento

- Convergência rápida
- Overfitting moderado (train > val)
- Validação instável (flutuações elevadas)

Desempenho (10-fold)

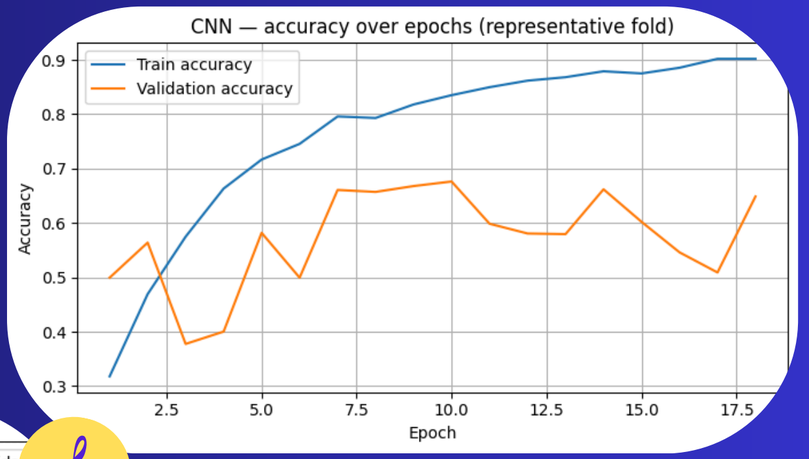
- Val acc ≈ 53% ± 5%
- Test acc ≈ 48% ± 5%
- Erros sistemáticos em classes contínuas (air_conditioner, engine_idling)



CNN (64 Mel Bands)

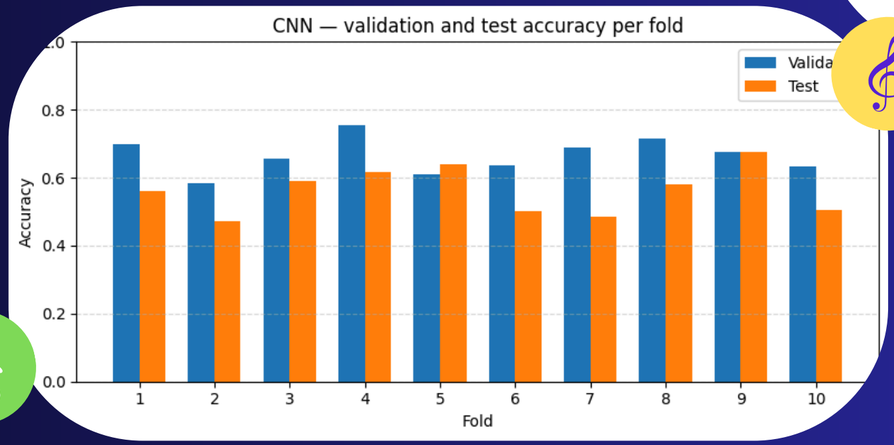
Input

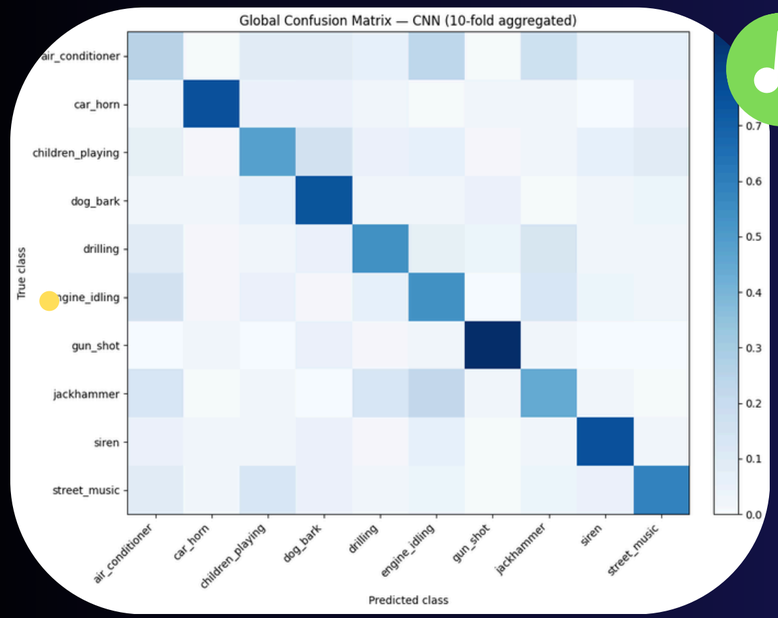
- Mel spectrogram 2D (64 × 173 × 1)



Arquitetura

- Conv2D + BatchNorm + ReLU
- MaxPooling
- Dropout
- Dense final (10 classes)





Comportamento

- Melhor generalização vs MLP
- Menor overfitting
- Validação estável mas com flutuações iniciais
- Captura padrões locais (transientes e texturas)

Desempenho (10-fold)

- Val acc ≈ 0.65 ± 0.05
- Test acc ≈ 0.58 ± 0.06
- Grandes melhorias em classes transitórias (car_horn, dog_bark)
- Confusão ainda presente entre classes contínuas (air_conditioner, engine_idling)

Refinamento Espectral: 64 → 96 Mel Bands

Porque ajustar:

• Baixa resolução nos 64 bins limitava a separação em regiões críticas de baixa frequência.
• Harmónicos mecânicos ficavam sobrepostos e difíceis de distinguir.
• Cenas amplas exigiam maior granularidade espectral.

Decisão:

• Aumentar a resolução para 96 Mel bands, mantendo todo o resto constante (modelo, treino, folds).

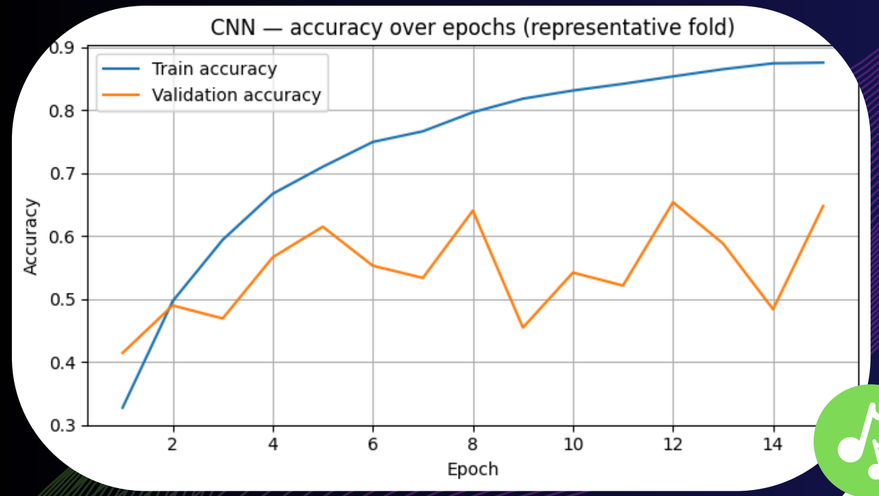
Objetivo:

• Verificar se o aumento de detalhe espectral reduz a confusão e melhora a generalização.

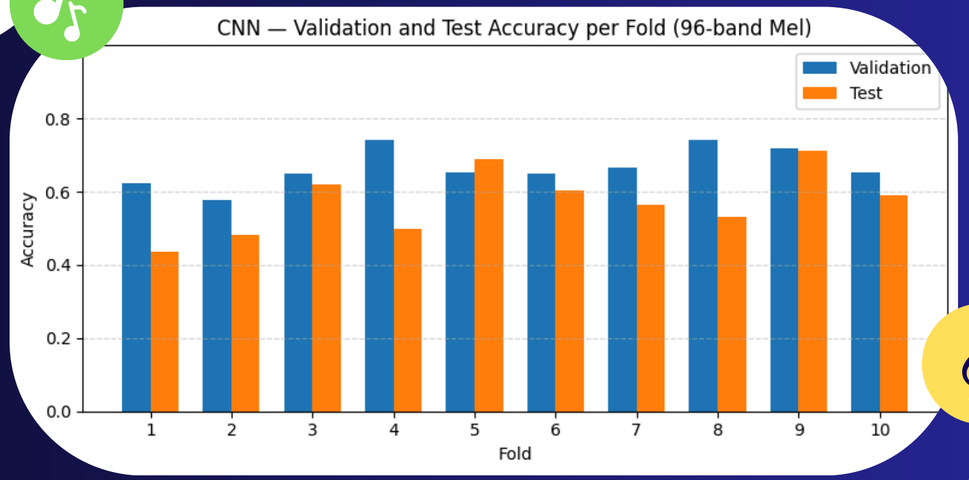
Problema Detetado no CNN-64

Baixa frequência	Confusão entre *air_conditioner* e *engine_idling* devido a espectros quase idénticos
Sons mecánicos periódicos	*drilling* ↔ *jackhammer* apresentam harmónicos sobrepostos
Cenas complexas	*children_playing* ↔ *street_music* apresentam mistura de fontes e variabilidade elevada

CNN (96 Mel Bands)



- Mel spectrogram 2D (96 × 173 × 1)
- Mesma arquitetura do CNN-64

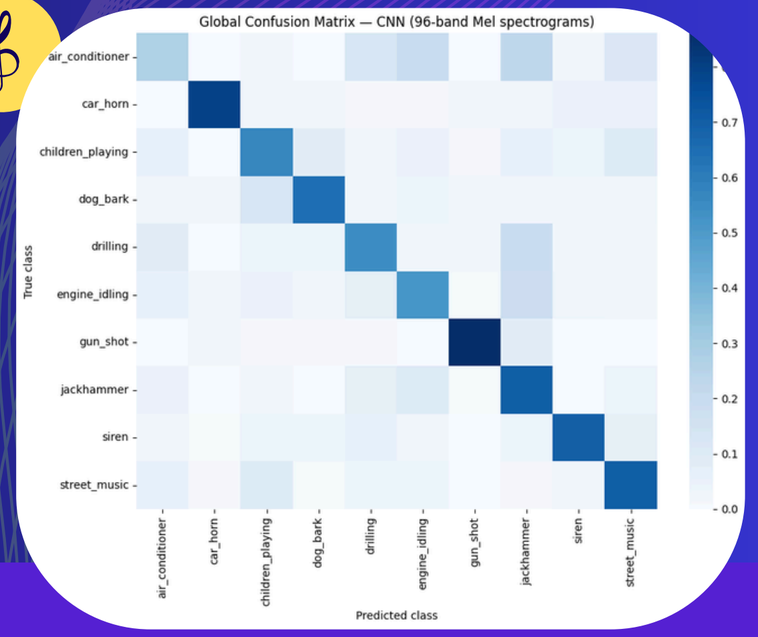


Comportamento

- Gap val→test reduzido (modelo mais estável)
- Melhoria clara em classes contínuas e estacionárias
- Curvas de validação mais regulares vs. CNN-64
- Menor dependência do fold; comportamento consistente

Desempenho (10-fold)

- Val acc ≈ 0.687 ± 0.045
- Test acc ≈ 0.607 ± 0.049
- Variância do teste ≈ 0.049
- Ganhos particularmente fortes em classes mecânicas



Insights da Matriz de Confusão

- Diagonais mais fortes em classes de baixa frequência
- Melhor separação em padrões periódicos (drilling, jackhammer)
- Menor leakage children_playing ↔ street_music
- Confusão persistente apenas em pares acusticamente semelhantes

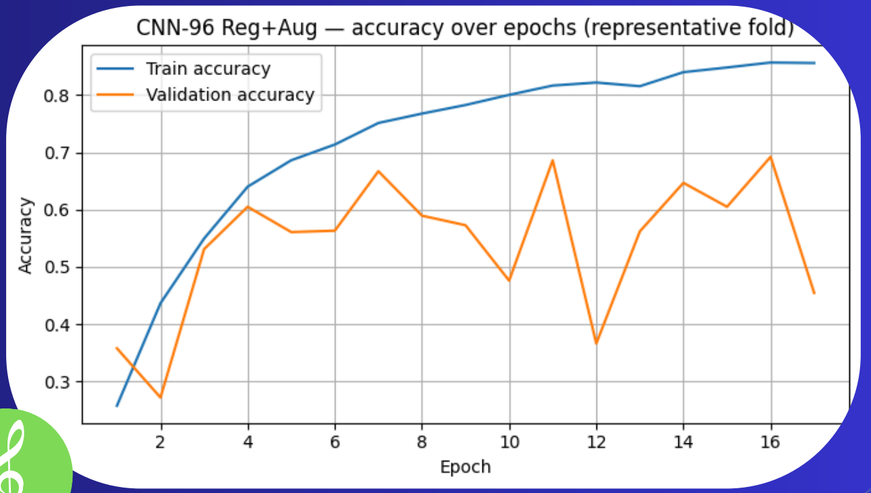
CNN-96 + Reg/Aug

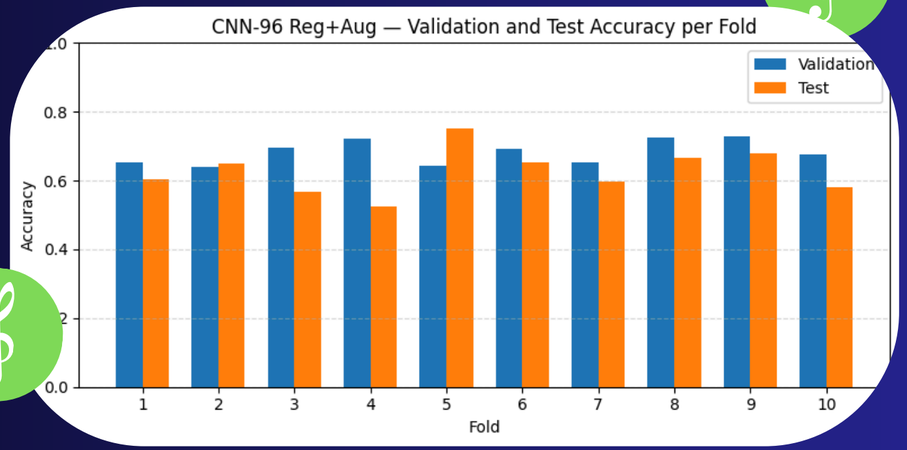
Comportamento

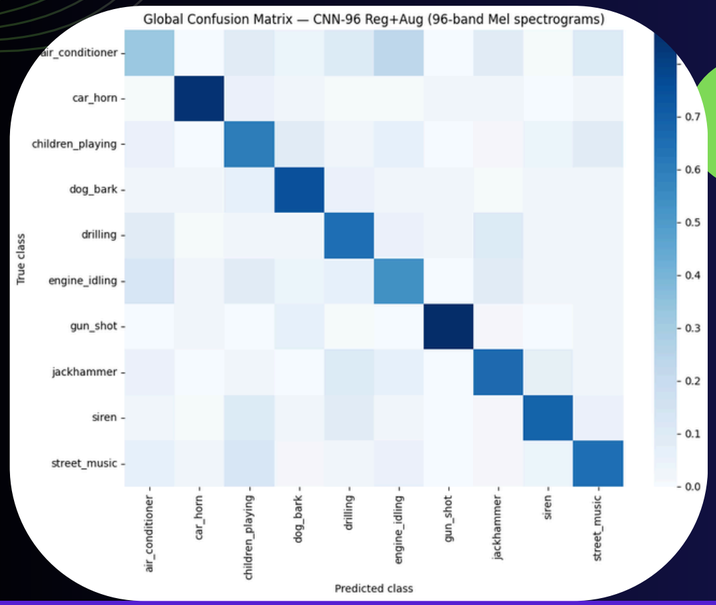
- Diferença val→test mais pequena → overfitting reduzido
- Curvas de treino/validação mais estáveis
- Maior consistência entre folds → modelo mais robusto

Refinamentos aplicados

- Dropout & L2 mais fortes
- Light SpecAugment (time + freq masking) apenas no treino
- Arquitetura idêntica → alteração apenas na variabilidade do treino







Desempenho (10-fold)

- Val acc ≈ 0.679 ± 0.028
- Test acc ≈ 0.620 ± 0.047
- Variância do teste mais baixa (≈ 0.047)
- Pequeno trade-off: ligeira queda em validação, mas melhor generalização

Insights da Matriz de Confusão

- Melhoria adicional em classes mecânicas (drilling, jackhammer)
- Menor leakage residual em children_playing ↔ street_music
- Confusão quase apenas nos pares acusticamente semelhantes

Comparação Final dos Modelos

Modelo	Val acc (média ± std)	Test acc (média ± std)	Var. do teste	Características principais
MLP	0,5304 ± 0,0511	0,4778 ± 0,0529	~ 0,053	Sem estrutura 2D; forte overfitting; má separação entre classes.
CNN 64 bandas	0,6688 ± 0,0385	0,5387 ± 0,1069	~ 0,107	Explora o espectrograma 2D; boa em eventos curtos; variância alta entre folds.
CNN 96 bandas	0,6872 ± 0,0452	0,6072 ± 0,0490	~ 0,049	Melhor resolução em baixas frequências; ganhos em classes contínuas; mais estável.
CNN 96 + Reg/Aug	0,6788 ± 0,0276	0,6197 ± 0,0466	~ 0,047	Regularização + masking; melhor generalização global; menor sensibilidade ao fold.

- O aumento da resolução Mel (64 → 96) foi a melhoria mais significativa, resolvendo o principal bottleneck representacional.
- A regularização + masking reduziu o overfitting residual e estabilizou os resultados sem modificar a arquitetura.
- O modelo final (CNN-96 + Reg/Aug) apresenta o melhor equilíbrio entre desempenho, generalização e robustez.

Bónus e Desafios

Bónus – DeepFool

- Aplicámos o ataque DeepFool à CNN-96 Reg+Aug no último fold de validação.
- Avaliámos 100 amostras de teste.
- O ataque enganou o modelo em 87% dos casos.
- A perturbação média tinha norma $L_2 \approx 1.38 \times 10^6$, praticamente impercetível no Mel-spectrograma.
- Mostra que, apesar da boa accuracy, o modelo mantém baixa robustez adversarial.

Desafios Identificados

- O modelo é sensível a pequenas perturbações → baixa robustez adversarial.
- Classes com padrões espectrais semelhantes continuam difíceis (ex.: air_conditioner vs engine_idling).
- Variância entre folds mostra dependência do conjunto de treino.
- Limitações computacionais impediram explorar arquiteturas maiores ou augmentations avançadas.