UNIVERSIDADE FEDERAL DE ITAJUBÁ

Detecção de Sirenes de Emergência no Trânsito com TinyML

Gustavo Kenji Silva Taniwaki — 2020017770
Pedro Augusto G. O. Rosa — 2020003040
Matheus Máximo Quatio — 2023001880

Itajubá
2025

Sumário

1 INTRODUÇÃO

A detecção de sons de veículos de emergência no trânsito é uma tarefa crítica em diversos contextos, como sistemas de transporte inteligentes, veículos autônomos e soluções de mobilidade urbana. Identificar com precisão sons de sirenes — como os de ambulâncias e caminhões de bombeiros — em meio ao ruído urbano é essencial para garantir respostas rápidas e seguras em situações de emergência.

Métodos tradicionais de reconhecimento sonoro baseados em filtros acústicos ou análise manual de sinais apresentam limitações quanto à robustez e adaptabilidade em ambientes ruidosos e dinâmicos. Com os avanços recentes em aprendizado de máquina e processamento de áudio, modelos treinados a partir de representações visuais do som, como espectrogramas, surgem como alternativas promissoras para esse tipo de detecção. Este projeto propõe o uso de um conjunto de dados contendo áudios urbanos classificados em três categorias — sirene de ambulância, sirene de caminhão dos bombeiros e ruído de trânsito — visando treinar modelos capazes de realizar a classificação automática desses sons.

O foco está na viabilidade de implementação desses modelos em dispositivos de borda, como sistemas embarcados em veículos ou semáforos inteligentes, possibilitando uma solução portátil, eficiente e de baixo custo para reconhecimento sonoro em tempo real no trânsito urbano.

2 DESENVOLVIMENTO

A execução do projeto se deu por meio da plataforma **Edge Impulse**, utilizada para a criação de projetos em *TinyML*. As etapas necessárias para a conclusão do projeto são as seguintes:

- Aquisição de dados;
- Processamento dos dados;
- Criação e treinamento do modelo;
- *Deployment* do modelo.

Dessa forma, cada uma dessas etapas será abordada nesta seção, iniciando pela coleta dos dados necessários. Além disso, foi utilizada a proporção de 80% dos dados para treinamento e 20% para testes.

2.1 AQUISIÇÃO DOS DADOS

Os dados utilizados para treinamento e testes foram obtidos pelo Emergency Vehicle Siren Sounds Dataset, que está disponível no site Kaggle (Dataset link). O conjunto de dados é composto por arquivos de áudio no formato .wav, com duração de 3 segundos cada. Esses áudios contêm sons de sirenes de veículos de emergência, especificamente de ambulâncias e caminhões dos bombeiros. Há ainda uma terceira categoria denominada ”Traffic”, que inclui sons comuns do tráfego urbano, sem a presença de sirenes. Cada uma das três categorias possui 200 arquivos de áudio, acompanhados de 200 imagens de espectrogramas correspondentes, geradas a partir dos áudios.

2.2 PROCESSAMENTO DOS DADOS

O passo inicial, após a escolha e tratamento dos dados, foi a criação da estrutura do modelo, realizada na seção de Impulse Design da plataforma Edge Impulse. Nessa etapa, os blocos responsáveis por processar e treinar o modelo foram definidos de forma modular e intuitiva.

Primeiramente, configurou-se o bloco de entrada como Time series data, com a frequência de amostragem ajustada para 16.000 Hz, uma janela de análise de 1000 ms, e um deslocamento (stride) de 500 ms, garantindo uma boa cobertura temporal dos sinais de áudio. A entrada utilizada foi o eixo ”audio”, com preenchimento automático de zeros para dados incompletos.

Em seguida, adicionou-se o bloco de processamento de áudio MFCC (Mel-frequency cepstral coefficients), que é amplamente utilizado na extração de características acústicas relevantes para tarefas de reconhecimento de som. Esse bloco transforma o áudio bruto em uma representação mais adequada para aprendizado de máquina.

Posteriormente, foi definido o bloco de aprendizado Classification, onde foi selecionado o uso das features extraídas via MFCC. O objetivo é classificar os áudios em três categorias: ambulância, caminhão dos bombeiros e trânsito comum. Esse bloco é responsável por treinar o modelo utilizando as amostras rotuladas, com base nas características extraídas.

Por fim, o bloco de Output Features resume as saídas previstas pelo modelo, que neste caso correspondem às três classes mencionadas. A Figura abaixo apresenta as configurações de entrada e os blocos do modelo configurados no Edge Impulse.

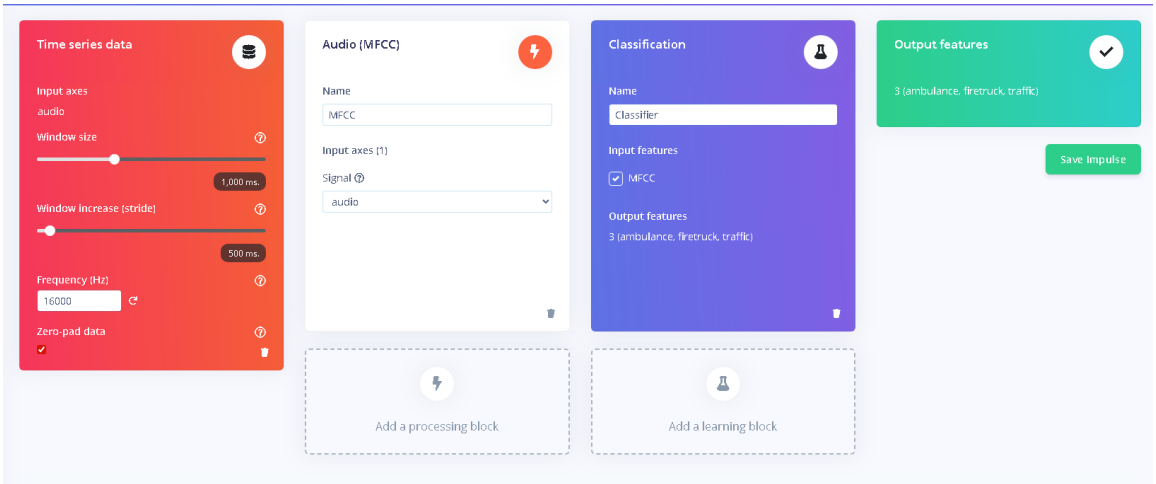


Figura 1: Configuração da estrutura do modelo no Edge Impulse, com entrada de dados, extração de MFCC, bloco de classificação e definição das saídas.

2.3 CRIAÇÃO E TREINAMENTO DO MODELO

Após a definição da estrutura do *Impulse*, foi realizada a configuração dos parâmetros de extração de características e do modelo de aprendizado. Para isso, utilizou-se o método MFCC (*Mel-Frequency Cepstral Coefficients*), amplamente reconhecido pela sua eficiência em tarefas de reconhecimento de padrões em sinais de áudio.

Os parâmetros definidos para a extração das características MFCC estão apresentados na Figura 2. Foram utilizados 13 coeficientes, com *frame length* e *stride* de 0.02 segundos, 32 filtros, tamanho de FFT igual a 256 e janela de normalização com tamanho 101. O coeficiente de *pre-emphasis* foi mantido em 0.98.

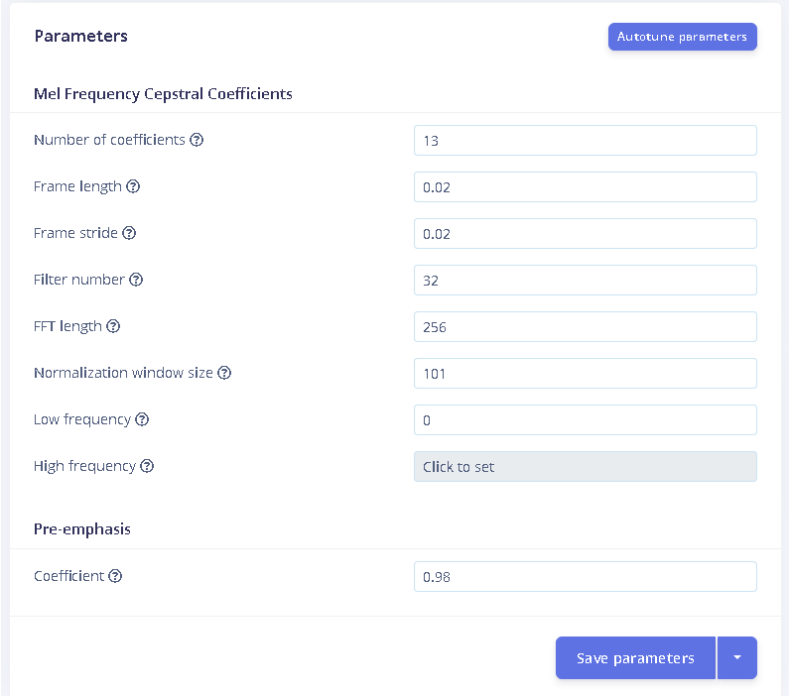


Figura 2: Parâmetros utilizados para a extração dos coeficientes MFCC.

A arquitetura do modelo adotada foi uma rede neural convolucional 1D, definida conforme ilustrado na Figura 3. A estrutura conta com uma camada de entrada de 650 *features*, seguida por uma camada de reshaping (13 colunas), duas camadas convolucionais com 8 e 16 filtros, respectivamente, ambas com tamanho de kernel 3. Cada camada convolucional é seguida por uma camada de *dropout* com taxa de 25%. A camada final é *fully-connected*(totalmente conectada) com três saídas correspondentes às classes: ambulância, caminhão dos bombeiros e trânsito.

Os principais parâmetros utilizados no treinamento foram:

- **Ciclos de treinamento:** 100
- **Taxa de aprendizado:** 0.005
- **Processador:** CPU
- **Otimizador (*learned optimizer*):** desabilitado

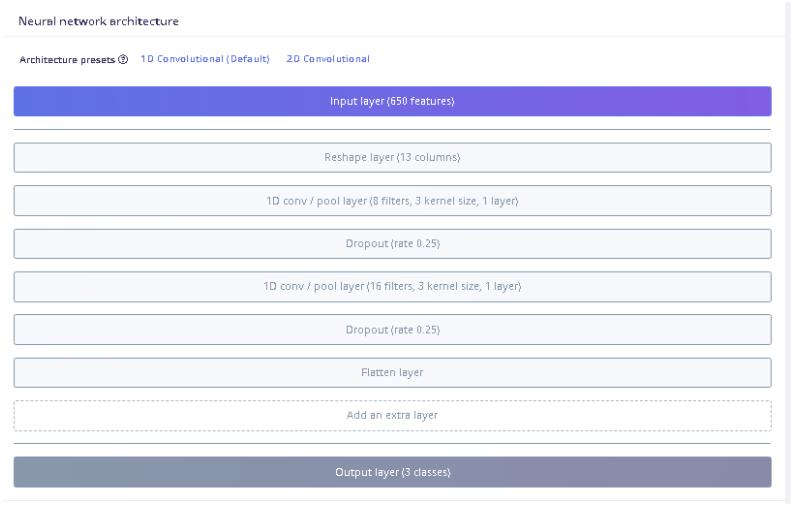


Figura 3: Arquitetura do modelo

Após o treinamento, o modelo obteve uma acurácia de **90.0%** no conjunto de validação, conforme apresentado na Figura 4. A matriz de confusão indica um desempenho elevado na classificação de sons de trânsito (99.6%) e bom desempenho nas classes de ambulância e caminhão dos bombeiros, embora com alguma confusão entre essas duas categorias. As métricas de precisão, recall e F1-score ponderadas ficaram todas em **0.90**, e a curva ROC teve uma área de **0.97**, demonstrando a capacidade geral do modelo.

Além disso, o desempenho embarcado previsto também foi satisfatório, com tempo médio de inferência de 7 ms, uso de RAM de 12.5 KB e uso de memória flash de 45.7 KB.

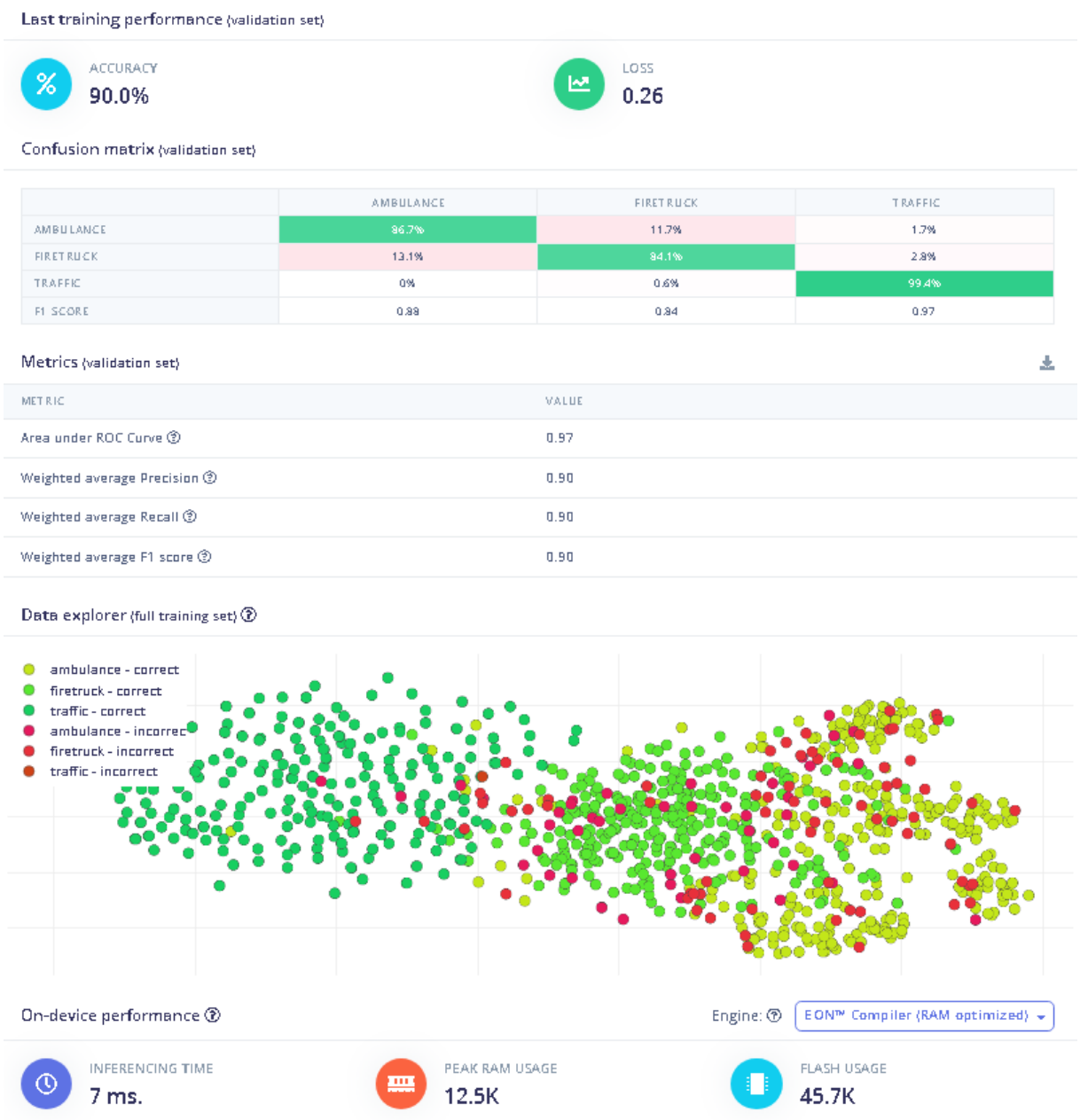


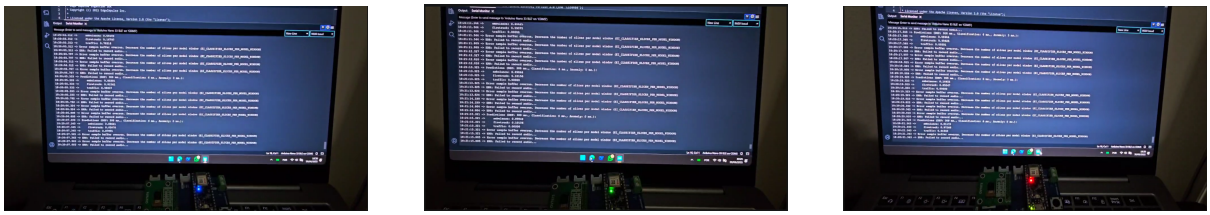
Figura 4: Arquitetura da rede neural e resultados do treinamento do modelo.

2.4 DEPLOYMENT DO MODELO

Para validar a aplicação do modelo em um cenário realista, foi realizada a inferência em tempo real utilizando um dispositivo embarcado. Os testes foram conduzidos a partir da reprodução de vídeos no YouTube contendo sons reais de sirenes de ambulâncias, caminhões dos bombeiros e tráfego urbano comum. Esses áudios foram utilizados como entrada para o modelo, simulando situações práticas de uso, como a detecção embarcada em veículos inteligentes ou em sensores de infraestrutura urbana.

A resposta do sistema foi mapeada para um conjunto de *LEDs* conectados ao dispositivo, permitindo uma visualização imediata da categoria identificada pelo modelo. A sinalização foi definida da seguinte forma:

- **Verde:** som de ambulância;
- **Vermelho:** som de caminhão dos bombeiros;
- **Azul:** som de tráfego comum (sem sirenes).



(a) Tráfego comum (b) Ambulância (c) Caminhão dos bombeiros

Figura 5: Sistema embarcado implementado

Essa abordagem permitiu avaliar, além da acurácia do modelo, seu desempenho embarcado em termos de tempo de resposta e uso de recursos computacionais, especialmente em dispositivos de baixa potência. O comportamento do sistema mostrou-se, em geral, estável e funcional, com reconhecimento satisfatório das categorias em tempo real.

No entanto, observou-se uma limitação importante: o modelo ainda apresenta certa dificuldade em diferenciar sons de tráfego com ruídos intensos — como buzinas, freadas bruscas ou sirenes ao fundo — dos sons reais de emergência. Em alguns casos, esse tipo de ruído urbano foi erroneamente classificado como ambulância ou caminhão de bombeiros, o que indica a necessidade de um refinamento futuro do modelo e, possivelmente, o uso de mais dados ou técnicas de pré-processamento mais robustas.

3 RESULTADOS

O desempenho do modelo foi avaliado utilizando duas versões distintas: a versão original em `float32` (não otimizada) e a versão quantizada em `int8` (otimizada para execução embarcada). Ambas foram submetidas ao mesmo conjunto de validação, permitindo uma comparação direta de suas métricas.

3.1 Versão Não Otimizada (Float32)

A versão não otimizada do modelo apresentou uma acurácia de **84,87%**, com valores equilibrados nas métricas de precisão, revocação e F1-score, todas com média ponderada de **0,89**. A área sob a curva ROC (AUC) manteve-se alta, atingindo **0,97**, demonstrando a capacidade do modelo em distinguir entre as classes mesmo em situações ambíguas.

A matriz de confusão (Figura 6) revela que a classe de tráfego foi identificada corretamente em **99,5%** dos casos, enquanto o caminhão dos bombeiros teve um desempenho de **75,2%**. A classe ambulância teve um desempenho inferior, com **81,5%** de acertos e confusão notável com a classe do caminhão dos bombeiros (**12,4%** dos casos).

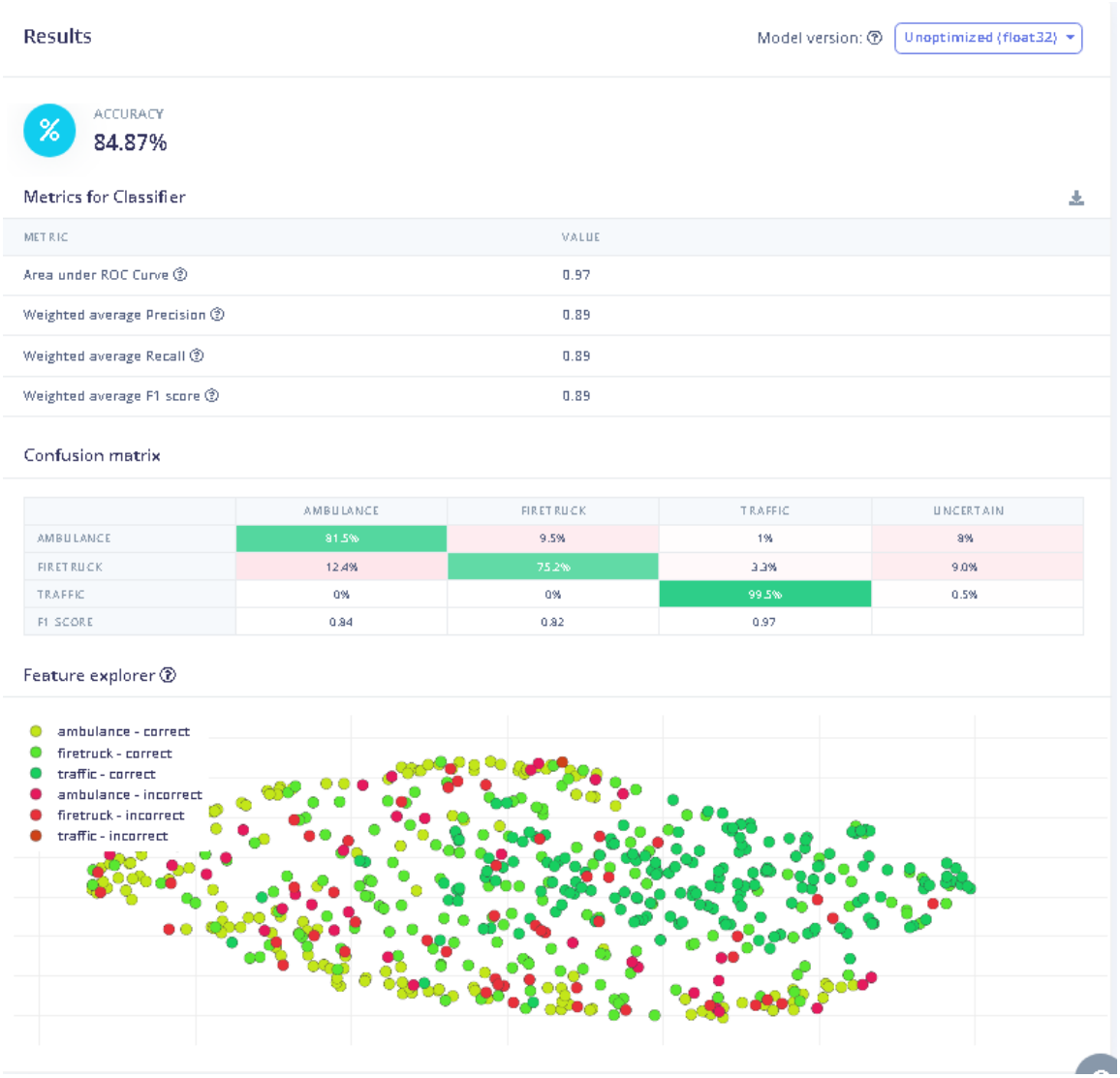


Figura 6: Resultado do teste versão não otimizada (Float32)

3.2 Versão Otimizada (Quantized Int8)

A versão quantizada do modelo, voltada para execução em dispositivos embarcados, apresentou um desempenho praticamente equivalente ao modelo em `float32`, com acurácia de **84,71%** e métricas médias ponderadas de **0,89** para precisão, revocação e F1-score. A área sob a curva ROC permaneceu em **0,97**, indicando estabilidade mesmo com a conversão para formato de menor precisão numérica.

Conforme ilustrado na Figura 7, o desempenho por classe também se manteve muito próximo ao do modelo original: ambulância com **81%**, caminhão dos bombeiros com **75,2%** e tráfego com **99,5%**. A porcentagem de classificações incertas aumentou ligeiramente para **10%** na classe de bombeiro e **8,5%** na de ambulância, o que é esperado em modelos quantizados devido à perda de precisão numérica.

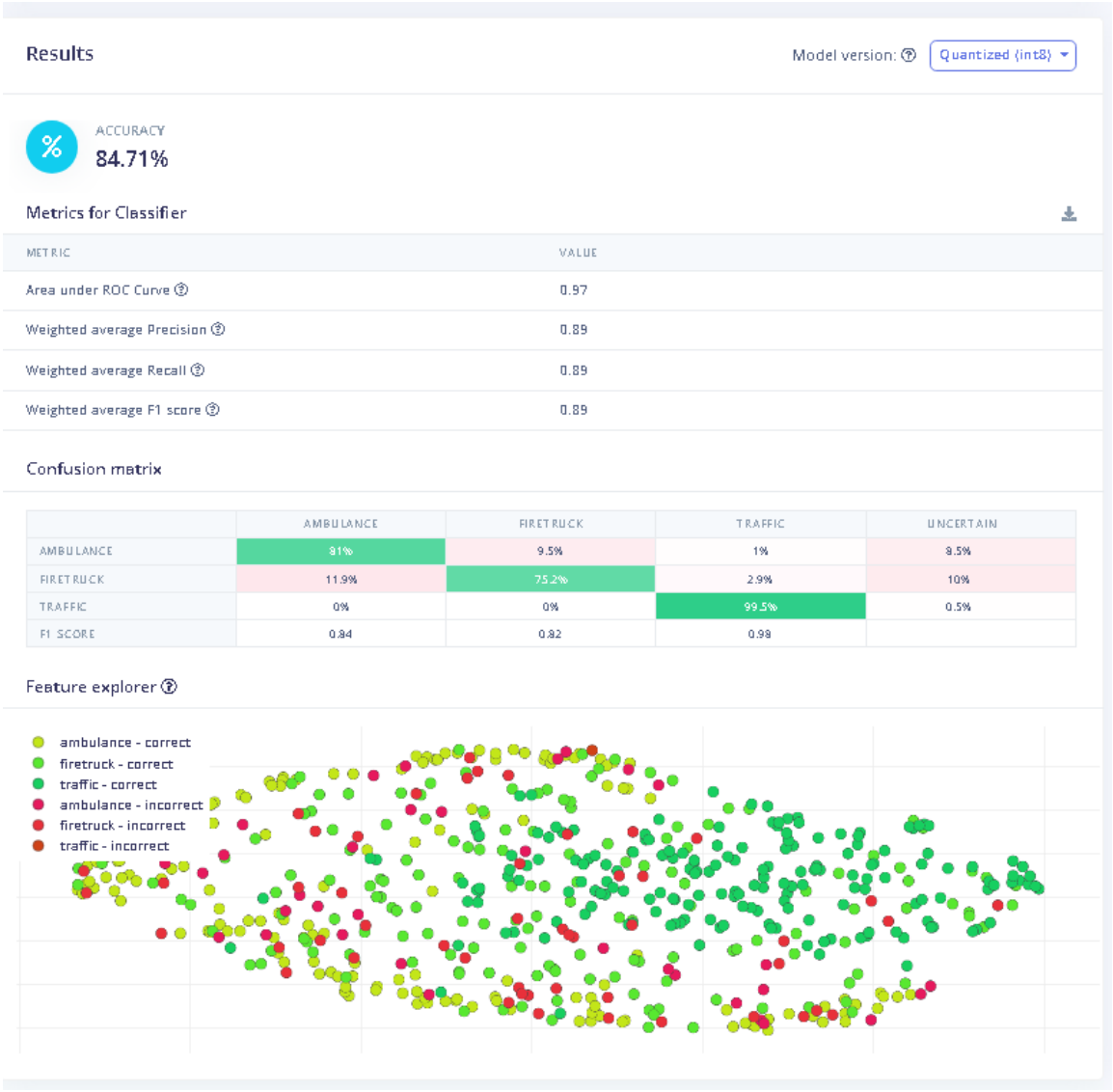


Figura 7: Resultado do teste versão otimizada (Quantized Int8)

4 CONCLUSÃO

O presente projeto demonstrou a viabilidade da aplicação de modelos de aprendizado de máquina embarcados (TinyML) para a detecção de sons de emergência no trânsito. Através da plataforma Edge Impulse, foi possível realizar todo o pipeline de desenvolvimento — desde a aquisição dos dados até a inferência embarcada — de forma prática e eficiente.

Apesar das limitações observadas na distinção entre sirenes semelhantes, o desempenho geral do modelo foi bastante satisfatório, tanto em termos de acurácia quanto de consumo de recursos. A implementação prática com LEDs também contribuiu para validar a aplicabilidade do sistema em cenários reais.

Como trabalhos futuros, sugere-se a ampliação da base de dados com sons variados, além da aplicação de técnicas de *data augmentation* e otimizações nos parâmetros de extração MFCC, visando aumentar a robustez do modelo. A integração com sistemas embarcados mais complexos, como assistentes veiculares ou redes de sensores urbanos, também representa um caminho promissor para evoluções desse projeto.

5 ANEXOS E LINKS

- Dataset utilizado;
- Projeto - Edge Impulse;