Públio Elon Correa da Silva

ThermalEdge: Uma solução em hardware para o reconhecimento de embriaguez em tempo real a partir de imagens térmicas utilizando edge computing

Sorocaba, SP

30 de Setembro de 2022

Públio Elon Correa da Silva

ThermalEdge: Uma solução em hardware para o reconhecimento de embriaguez em tempo real a partir de imagens térmicas utilizando edge computing

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCC-So) da Universidade Federal de São Carlos como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação. Linha de pesquisa: Computação Científica e Inteligência Computacional.

Universidade Federal de São Carlos – UFSCar

Centro de Ciências em Gestão e Tecnologia – CCGT

Programa de Pós-Graduação em Ciência da Computação – PPGCC-So

Orientador: Prof. Dr. Siovani Cintra Felipussi

Coorientador: Prof. Dr. Samuel Lourenço Nogueira

Sorocaba, SP

30 de Setembro de 2022

Correa da Silva, Públio Elon
ThermalEdge: Uma solução em hardware para o reconhecimento de embriaguez em tempo real a partir de imagens térmicas utilizando edge computing/ Públio Elon Correa da Silva. – 2022.
110 f. : 30 cm.

Dissertação (Mestrado) – Universidade Federal de São Carlos – UFSCar
Centro de Ciências em Gestão e Tecnologia – CCGT
Programa de Pós-Graduação em Ciência da Computação – PPGCC-So.
Orientador: Prof. Dr. Siovani Cintra Felipussi
Banca examinadora: Prof. Dr. Siovani Cintra Felipussi, Prof. Dr. Fábio Luciano Verdi, Prof. Dr. Wesley Angelino de Souza
Bibliografia

1. Classificação de Embriaguez. 2. Termografia. 3. Edge Computing. I. Siovani Cintra Felipussi. II. Universidade Federal de São Carlos. III. ThermalEdge: Uma solução em *hardware* para o reconhecimento de embriaguez em tempo real a partir de imagens térmicas utilizando *edge computing*

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências em Gestão e Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Públio Elon Correa da Silva, realizada em 30/09/2022.

Comissão Julgadora:

Prof. Dr. Siovani Cintra Felipussi (UFSCar)

Prof. Dr. Wesley Angelino de Souza (UTFPR)

Prof. Dr. Fabio Luciano Verdi (UFSCar)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

Dedico este trabalho aos meus pais Walter Corrêa da Silva e Edneia Fontolan Corrêa da Silva que me deram suporte durante toda esta jornada.

Agradecimentos

Agradeço,

em primeiro lugar a Deus, por ter cuidado de mim e de minha família durante o mestrado, e também, em meio a pandemia.

aos meus pais Edneia Fontolan Correa da Silva e Walter Correa da Silva pelo suporte, conselhos e sabedoria durante os meus estudos.

Ao Programa de Pós-graduação em Ciência da Computação da UFSCar pela estrutura e pela oportunidade me estudar e ampliar meus conhecimentos.

ao meu orientador Professor Doutor Siovani Cintra Felipussi por me aceitar como aluno de mestrado, pela paciência, pelos conhecimentos, experiências e sabedoria transmitidos durante todo o mestrado.

ao meu co-orientador Professor Doutor Samuel Lourenço Nogueira por me aceitar como co-orientado, pela paciência, pela experiência, pelos conhecimentos e sabedoria passados durante o mestrado.

ao Washington por ter iniciado esta pesquisa de mestrado pela solicitude e parceria ao longo do desenvolvimento deste projeto de mestrado.

ao Alan Pietro pela valiosa ajuda com a reunião de participantes e supervisão da segurança dos mesmos durante os experimentos.

ao Luiz Pedro Barreto de Alcântara pela valiosa ajuda com a reunião de participantes e supervisão da segurança dos mesmos durante os experimentos.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) pelo auxílio financeiro (Código de financiamento 001).

"A sabedoria é a coisa principal: adquire pois a sabedoria; sim, com tudo o que possues adquire o conhecimento.
Exalta-a, e ela te exatalrá; e, abraçando-a tu, ela te honrará;
(Provérbios, 4:7-8)

"It is not because things are difficult that we do not dare; it is because we do not that things are difficult;
(Séneca)

Resumo

O consumo do álcool tem apresentado problemas para a segurança viária e, subsequentemente, a fiscalização de trânsito pode se beneficiar do desenvolvimento dos métodos para classificação do estado do indivíduo. Desta forma, os avanços recentes na termografia associados ao advento do 5G, resultam em novas oportunidades para estudo, como a computação na borda da rede e o uso de algoritmos de redes neurais profundas em microcomputadores. O objetivo deste projeto de mestrado é o desenvolvimento de um *framework* que possibilite a classificação de imagens térmicas, para a rotulação do estado do indivíduo em tempo real, capturadas por uma câmera térmica acoplada a um dispositivo celular que realiza o envio das imagens para um servidor na borda da rede que possui um modelo de rede neural convolucional treinado para reconhecer o estado do indivíduo a partir de imagens térmicas. Em adição, para a detecção de embriaguez, são utilizados aceleradores de *hardware* em dispositivos embarcados. Em adição, um aplicativo para dispositivos móveis foi desenvolvido para possibilitar o envio das imagens em tempo real para classificação, utilizando protocolo UDP. Em vista disso, para que o *framework* proposto possa reconhecer o estado de embriaguez, um modelo de rede neural convolucional foi treinado utilizando *transfer learning* para abstrair características relacionadas a embriaguez. O modelo treinado para o *Edge TPU* obteve a melhor acurácia na classificação de embriaguez, sendo esta de 94,33% no conjunto de teste.

Palavras-chaves: Compressão de Modelos. Edge Computing. Visão Computacional. Redes Neurais Profundas. Imageamento Térmico. Classificação de Imagens em Tempo Real.

Abstract

Alcohol consumption has presented problems for road safety, traffic enforcement can benefit from the development of methods for classifying the individual's status. In this way, recent advances in thermography associated with the advent of 5G, resulted in new opportunities for study, such as computing at the performed at the edge of the network and the use of deep neural network algorithms in microcomputers. The objective of this master's project is the development of a *framework* that allows the classification of thermal images, for the labeling of the individual's state in real-time, captured by a thermal camera coupled to a cellular device that sends the images to a server at the edge of the network that has a convolutional neural network model trained to recognize the individual's state from thermal images. In addition, for detecting drunkenness, *hardware* accelerators are used in embedded devices. In addition, an application for mobile devices was developed to allow the sending of images in real-time for classification, using the UDP protocol. Because of this, for the proposed framework to recognize the drunken state, a convolutional neural network model was trained using *transfer learning* to abstract characteristics related to drunkenness. The model trained for the *Edge TPU* obtained the best accuracy in the classification of inebriation, which was 94.33% in the test set.

Key-words: Model Compression. Edge Computing. Computer Vision. Deep Neural Networks. Thermal Imaging. Real-Time Image Classification

Lista de ilustrações

Figura 1 – Intensidade da radiação de um corpo negro em relação ao comprimento de ondas em diferentes temperaturas. 33
Figura 2 – Comparação entre dispositivos térmicos com diferentes valores de sensibilidade térmica. 36
Figura 3 – Espectro eletromagnético com sub-divisão. 37
Figura 4 – Estrutura de serviços da nuvem. 40
Figura 5 – Edge Computing: Motivações, Desafios e Oportunidades. 43
Figura 6 – Diferentes tipos de compressão de modelos para DNN e métodos de aprendizado de máquina tradicionais. 47
Figura 7 – Diagrama das camadas de uma rede neural convolucional fatorizada. . 48
Figura 8 – Processo genérico de Knowledge-Distillation. 49
Figura 9 – Comparação entre quantização uniforme e quantização não uniforme. Os valores reais em domínio contínuo R são mapeados em valores discretos, os valores de menor precisão são quantizados no domínio Q, que são marcados com os pontos laranjas. 50
Figura 10 – Tela principal do aplicativo densenvolvido. 57
Figura 11 – Menu de alerta utilizado para entrada do endereço de IP do servidor na borda da rede. 58
Figura 12 – Diferentes paletas de pseudo-cores, sendo estas cores: *Iron Rainbow*, *Rainbow*, *Rainbow HC* e *Black Hot* respectivamente. 58
Figura 13 – Antes e após a execução do algoritmo de correção de não uniformidades. 59
Figura 14 – Raspberry Pi 4B utilizado como servidor de Edge Computing. 60
Figura 15 – Modelo de etilômetro utilizado para aferição do BrAC nos participantes. 61
Figura 16 – Imageador Térmico Infiray T3C. 62
Figura 17 – INA219 l2c. 63
Figura 18 – Modelo para classificação de embriaguez treinado a partir da VGG16 por Silva (2021). 64
Figura 19 – Modelo de rede neural convolucional proposto após a extração de atributos e a transferência de aprendizagem. 66
Figura 20 – Fluxo de treinamento do modelo proposto: Transfer Learning e Fine-tuning. 67
Figura 21 – Conversão de modelos para Representação Intermediaria. 68
Figura 22 – Framework proposto. 68
Figura 23 – Google Coral Edge TPU USB. 69
Figura 24 – Intel Neural Compute Stick 2. 70

Figura 25 – Comparação entre um quadro capturado após o dispositivo térmico ser inicializado e 20 minutos após o dispositivo aquecer. 74
Figura 26 – Boxplot com o nível de BrAC de cada participante durante os 4 periodos de aquisição. 79
Figura 27 – Comparação entre 2 quadros do Participante 9: antes e após 30 minutos depois da ingestão de álcool pelo participante. 81
Figura 28 – Antes e 30 minutos após o consumo de álcool - subconjunto do conjunto de imagens capturadas durante o procedimento experimento. 81
Figura 29 – Comparação de quadros onde a quantidade de pixels claros antes e após a ingestão de álcool é maior para o estado de ebriedade. 82
Figura 30 – *Loss* do modelo selecionado durante a *k-fold cross validation.* 85
Figura 31 – Acurácia do modelo selecionado durante a *k-fold cross validation.* . . . 86
Figura 32 – Matriz de confusão do modelo. 86
Figura 33 – Matriz de confusão do modelo Tensorflow Lite acelerado pelo Edge TPU no conjunto de teste. 89
Figura 34 – Matriz de confusão do modelo OpenVino acelerado pelo Intel Neural Compute Stick 2 no conjunto de teste. 90
Figura 35 – Taxa de quadros por classificação (FPS). 91
Figura 36 – Tempo de inferência na borda da rede. 92
Figura 37 – Temperatura durante execução. 93
Figura 38 – Tempo de carregamento do modelo. 94
Figura 39 – Tempo de aquecimento do modelo. 94
Figura 40 – Consumo Energético do Raspberry Pi 4B. 95

Lista de tabelas

Tabela 1 – Comprimento e faixas de onda o espectro infravermelho. 39
Tabela 2 – Características entre diferentes paradigmas - Computação em Nuvem e Edge computing. 42
Tabela 3 – Benchmarking de dispositivos embarcados, redes neurais convolucionais e aceleradores na borda da rede. 53
Tabela 4 – Dados antropométricos de cada participante. 73
Tabela 5 – Estimação de doses a serem consumidas com base na água corporal total de cada participante 74
Tabela 6 – Tabela com as etapas do processo de captura das imagens. 76
Tabela 7 – Resultado da aferição do BrAC para cada participante utilizando o etilômetro 78
Tabela 8 – Resultado da estimação do valor de BrAC para cada participante utilizando a fórmula de Widmark 80
Tabela 9 – Quantidade de pixels acima da limiar estabelecida para os participantes antes e depois da ingestão do álcool 83
Tabela 10 – Diferença entre os valores médios de pixels brilhosos entre os grupos da primeira a segunda aquisição 84
Tabela 11 – Conjunto de hiperparâmetros para a etapa de *transfer learning*. 85
Tabela 12 – Relatório de classificação do modelo após a etapa de treinamento. . . . 87
Tabela 13 – Hiperparâmetros utilizados para o *pruning* do modelo para o Edge TPU. 87
Tabela 14 – Hiperparâmetros utilizados para a quantização do modelo para o *Edge TPU*. 88
Tabela 15 – Comparação entre o resultado final da compressão do modelo. 88

Lista de abreviaturas e siglas

CTB	Código de Trânsito Brasileiro
CoAP	*Constrained Application Protocol*
DDoS	*Distributed Denial of Service*
DNN	*Deep Neural Network*
FPGA	*field-programmable gate array*
FTP	*File Transfer Protocol*
HTTP	*Hypertext Transfer Protocol*
HTTPS	*Hyper Text Transfer Protocol Secure*
IoT	*Internet of Things*
MQTT	*Message Queuing Telemetry Transport*
NCS2	*Neural Compute Stick 2*
NETD	*Noise Equivalent Temperature Difference*
NUC	*Non Uniformity Correction*
PQAT	*Pruning Quantization-Aware Training*
QoS	*Quality of Service*
RNC	Rede Neural Convolucional
STMP	*Simple Mail Transfer Protocol*
TPU	*Tensor Processing Unit*
UDP	*User Datagram Protocol*
VPU	*Visual Processing Unit*

Lista de símbolos

W	Bloco convolucional
$\in$	Pertence a
R	Kernel de um bloco convolucional
w	largura do kernel de um bloco convolucional
h	altura do kernel de um bloco convolucional
c	número de canais de entradas
n	número de canais de saídas
KL	Divergência de Kullback–Leibler.
$\mathcal{L}_{KD}$	Custo total da distilação de conhecimento entre o modelo professor e o modelo aluno.
$Bl_{\mathrm{H_2O}}$	Porcentagem aproximada de água no sangue.

Sumário

1 INTRODUÇÃO 27
1.1 Objetivos 29
1.2 Organização do texto 30

2 REFERENCIAL TEÓRICO 31
2.1 A segurança viária e os efeitos do álcool 31
2.2 Termografia e as mudanças de temperatura na pele 32
2.2.1 Imageamento térmico 36
2.3 Internet of Things (Iot) e Cloud Computing 39
2.4 Edge Computing 41
2.4.0.1 Características da Computação em Nuvem e Edge Computing 42
2.4.1 Edge Computing: Motivações, Desafios e Oportunidades 43
2.5 Deep Neural Networks e compressão de modelos 46
2.6 Trabalhos Relacionados 51

3 IMPLEMENTAÇÃO, MATERIAIS E METODOLOGIA 57
3.1 Implementação 57
3.1.1 Aplicativo móvel 57
3.2 Materiais 60
3.2.1 Raspberry Pi 4B 60
3.2.2 Etilômetro 61
3.2.3 Questionário AUDIT 62
3.2.4 Infiray Xtherm T3C 62
3.2.5 Módulo INA219 63
3.2.6 Modelo de Deep Learning 64
3.2.7 Aceleradores de Hardware 69
3.2.7.1 Google Coral Edge TPU 69
3.2.7.2 Intel Neural Compute Stick 2 69
3.3 Metodologias 70
3.3.1 Estratégia metodológica para estimação de dosagem para bebidas 70
3.3.2 Estratégia metodológica para captura de imagens 73

4 RESULTADOS E DISCUSSÕES 78
4.1 Resultados 78
4.1.1 Considerações Iniciais 78
4.1.2 Resultados experimentais obtidos a partir do etilômetro 78

4.1.3 Resultados das imagens capturadas 80
4.1.4 Resultados do treinamento do modelo de DNN 84
4.1.5 Resultados do servidor na borda da rede 91
4.2 Discussões 96
4.2.1 Estimação do nível de álcool no sangue 96
4.2.2 Sessão de captura das imagens 97
4.2.3 Deep Neural Networks e Edge Computing 98
4.2.4 Trabalhos Futuros 98

Conclusão 99

Referências 101

1 Introdução

Nos últimos anos, a quantidade de acidentes resultantes do estado de embriaguez revelou-se como um problema social e de saúde em muitos países. A correlação entre os sintomas de ebriedade com o valor da alcoolemia é uma tarefa difícil em virtude das manifestações fisiológicas e funcionais diversas. A restrição dos métodos de sintomatologia tradicionais empregados frequentemente contribuem negativamente para este fator. De acordo com a Organização Mundial da Saúde (WHO, 2018a), mais de 3 milhões de mortes ocorridas mundialmente até 2018 são resultantes do uso de bebidas alcoólicas.

Segundo os dados estatísticos da Organização Mundial da Saúde (WHO, 2018b), a tendência à alcoolemia é influenciada por condições sociais, enquanto o tipo de bebida está condicionado à predominância em determinada região. Não obstante, de acordo com relatórios da organização mundial da saúde, houve um acréscimo na quantidade de álcool consumido mundialmente de 5.5 litros per capita em 2005 para 6.4 litros em 2016.

Com a introdução da Lei Seca 11.705 (NETO; SANTOS, 2012), houve uma significativa redução no número de mortes envolvendo acidentes causados por condutores alcoolizados. A atualização na lei penaliza a condução sob influência de álcool caso o motorista disponha de qualquer volume de bebida no sangue. A recusa por parte do condutor a se submeter aos métodos convencionais de fiscalização da alcoolemia pode ser visto como outra variável que impacta nos dados estatísticos em virtude do direito Constitucional de não produzir provas contra si (ROBERTA; RESPLANDES, 2019). Quando os exames toxicológicos são empregados, estes podem ser influenciados pelo tempo de deslocamento do usuário até o local de coleta.

Outro desafio da detecção da alcoolemia é a constatação pelo agente fiscalizador de todos os sinais de embriaguez para efetivamente categorizar o estado de ebriedade do indivíduo (KRUSCINSKI, 2016). O artigo 165 da Lei Federal nº 12.760, determina que o motorista incide em infração penal, uma vez averiguado o estado de intoxicação alcoólica. A dificuldade está na identificação de todas as características indicadas no anexo II do artigo 165 da Lei Federal nº 12.760 do DOU (LEGISLATIVO, 2020), onde devido à unicidade do metabolismo humano e somente após a averiguação de todos os itens, a infração poder ser constatada.

Ademais, os efeitos da ingestão de bebidas alcoólicas foram confirmados repetidamente (KALANT; Lê, 1986) através do uso de técnicas de condução de calor e mediante a amplitude de ondas de pulso digital. Também foi demonstrado consistentemente que o aumento do fluxo sanguíneo provoca gradação de temperatura notadamente no rosto e na superfície da pele devido ao processo de termorregulação, que é regido pelo hipotálamo e

regula o decremento de calor do corpo mediante a dilatação da rede vascular conhecida como termólise e o aumento de temperatura denominada termogênese.

Os sinais supracitados da alcoolemia causam a elevação da temperatura nas regiões faciais onde é localizada a rede vascular, pela qual é possível detectar a diferença de temperatura ao longo do tempo utilizando dispositivos térmicos que realizam a captura de imagens no espectro infravermelho longo e que possuam uma resolução espacial e sensibilidade térmica adequada (BUDDHARAJU; PAVLIDIS; TSIAMYRTZIS, 2006; BUDDHARAJU et al., 2007; BUDDHARAJU; PAVLIDIS, 2012). O aumento de temperatura na face pode ser destacado através do emprego de filtros, que possibilitam a representação estimada dos vasos sanguíneos para observação da vasodilatação.

Em virtude da termografia diversas alterações que indicam enfermidades podem ser detectadas, como demonstra Lahiri, Bagavathiappan e al. (2012). Através de imagens no espectro infravermelho longo, estes indicativos podem ser analisados em decorrência da emissividade de radiação da pele humana ser quase constante. O tecido cutâneo tem valor de emissão de radiação térmica de ± 0.99 e, como comprimento máximo, 9.5 µm de onda. Ainda sob este contexto Khan, Ward e Ingleby (2009), aferiram que para aquisição de informação do tecido cutâneo no espectro infravermelho longo, o intervalo mais adequado é de 2-14 µm.

A partir destes conceitos supracitados e com a finalidade do emprego da termografia, a computação na borda da rede permite combinar aplicações de *Deep Neural Networks* e de IoT (*Internet of Things*). Este conjunto possibilita a captura de dados a partir de sensores em tempo real e redução da latência que advém da aproximação do servidor da entrada de dados. Greco et al. (2020) salientam que a arquitetura IoT evoluiu na área médica nos últimos anos, combinando algoritmos de inteligência artificial para tomada de decisão, reconhecimento e classificação de doenças.

Para a classificação do estado inebriado, Koukiou e Anastassopoulos (2012) apresentaram uma abordagem para seleção de atributos a partir de imagens térmicas que correspondem a estimativa do vaso sanguíneo do rosto. Este trabalho demonstra que a região do nariz tem a maior gradação de temperatura conforme a ingestão do álcool. Posteriormente, Koukiou e Anastassopoulos (2015) empregaram o uso de redes neurais convolucionais para a extração de características locais em toda a face. Todavia, esta abordagem não carrega as características fisiológicas durante a classificação.

Em adição, Koukiou (2021) empregou cadeias de Markov para modelar o comportamento estatístico dos pixels na imagem da testa de uma pessoa para a detecção do estado de intoxicação do indivíduo. O modelo markoviano é utilizado para a obtenção de vetores de características a partir de matrizes de transição de primeira ordem. Este trabalho demonstra o uso de uma arquitetura simples de redes neurais para classificação de embriaguez de acordo com atributos obtidos a partir da região da testa.

Não obstante, Soltuz e Neagoe (2021) apresentaram 3 modelos de aprendizagem profunda para a classificação de embriaguez, utilizando imagens térmicas através da transferência de aprendizagem. Neste trabalho, os autores apresentam o uso da rede GoogLeNet para obter o melhor resultado.

Uma abordagem *data-driven* foi apresentada por (SILVA, 2021), visando induzir uma rede neural convolucional a abstrair características que sejam relevantes para a classificação do estado do indivíduo utilizando transferência de aprendizagem. Em adição, este trabalho apresenta uma explicabilidade maior em relação ao modelo desenvolvido e as características extraídas pelo mapa classificador desenvolvido, considerando os trabalhos em relação a classificação de embriaguez.

Para a implementação de um algoritmo de redes neurais convolucionais para dispositivos móveis na borda rede, (ALIBABAEI et al., 2022) demonstra que é possível realizar a detecção de troncos de videira em tempo real empregando microcomputadores embarcados em conjunto com aceleradores de redes neurais. Em adição, o trabalho realiza o emprego de imagens térmicas para a detecção de troncos de videira na borda da rede.

Desta maneira, a detecção de embriaguez tem se mostrado promissora no emprego de técnicas a partir da termografia. Para isso, este projeto de mestrado apresenta uma abordagem do uso de algoritmo de redes neurais convolucionais para classificação de embriaguez, implantado em um dispositivo embarcado que utiliza a computação em borda para aferir o estado de embriaguez de forma não invasiva e em tempo real, através de imagens capturadas por dispositivos móveis.

1.1 Objetivos

O objetivo principal deste trabalho foi desenvolver um *framework* para o reconhecimento e categorização de embriaguez em tempo real. Este sistema poderá ser utilizado para o monitoramento da temperatura cutânea superficial de um determinado indivíduo de forma não invasiva e de reduzido custo em tempo real, sendo possível correlacionar o estado fisiológico do indivíduo com a classificação da imagem. Desta forma, para se alcançar o objetivo proposto foram delineados os seguintes objetivos secundários:

- Implementação de um algoritmo de redes neurais convolucionais (VGG16), que possui uma considerável quantidade paramétrica e custo computacional, através de técnicas de compressão de modelo para aceleração do tempo de execução do modelo.

- Desenvolvimento de um aplicativo para o envio das imagens em tempo real utilizando o procolo UDP e para a captura dos valores radiométricos de temperatura pixel a pixel em toda a imagem.

- Implementação de coprocessadores como Intel NCS2 e o Coral Edge TPU para a inferência de algoritmos de redes neurais convolucionais na borda da rede, possibilitando a avaliação do desempenho da aceleração do servidor na tarefa de classificação de imagens.

1.2 Organização do texto

Abaixo será apresentado uma breve explicação dos conteúdos dos capítulos seguintes que serão apresentados nesse texto:

- Capítulo 2: Neste capítulo serão apresentados os trabalhos relevantes a variação de temperatura na superficie da face em decorrência da embriaguez. Subsequentemente, a computação na borda da rede e suas oportunidades, desafios e contribuções. Em seguida, será abordado sobre as técnicas de compressão de modelo que possibilitam a implantação de modelos de DNN em dispositivos móveis com capacidade de recursos limitados e, por fim, os trabalhos relacionados que dão base para o *framework* proposto serão detalhados.

- Capítulo 3: Neste capítulo, há a implementação e a estratégia metodológica do *framework* proposto, assim como os materiais empregados para verificar a hipótese de se classificar o estado de embriaguez de um indivíduo em tempo real. Resultados serão abordadas logo em seguida.

- Capítulo 4: Neste capítulo serão apresentados os resultados obtidos durante a estratégia metodológica para captura das imagens, assim como os resultados relacionados ao treinamento do modelo de rede neural convolucional, e subsequentemente, os resultados do desempenho do dispositivo utilizado como *edge server* para a classificação em tempo real das imagens térmicas. Em seguida, será realizada uma discussão em torno dos resultados obtidos, bem como as direções futuras em relação a esta abordagem. Por fim, serão apresentados as considerações finais, bem como as contribuições do presente trabalho.

2 Referencial Teórico

Este capítulo aborda o referencial teórico para o *framework* proposto e descreve a metodologia e *hardware* utilizados para a classificação de embriaguez. Também descreve a relevância da escolha da câmera térmica adequada, os critérios para aquisição das imagens e seleção de quadros, assim como a importância da borda da rede para a classificação e o processamento das imagens em tempo real. Por conseguinte, os trabalhos relevantes para a solução proposta serão apresentados e, subsequentemente, serão discutidos. Ainda sob o contexto do *framework* proposto, são descritas a elaboração e o modelo apresentado de redes neurais convolucionais; é discursada a técnica de otimização do algoritmo de redes neurais e, por fim, os resultados obtidos detalhados.

2.1 A segurança viária e os efeitos do álcool

Os acidentes de trânsito decorrentes de motoristas em estado de embriaguez têm gerado considerável preocupação em relação à segurança viária na última década (HAMMER et al., 2018). De acordo com os dados da Organização Mundial da Saúde (WHO, 2018a), milhões de pessoas estão sujeitas às consequências de acidentes de trânsito. Acrescenta-se ainda que, de acordo com a OMS (WHO, 2018b), houve um aumento na quantidade de álcool consumida per capita de 5,5 litros em 2005 para 6,4 litros em 2016 mundialmente. Em adição, o uso nocivo do álcool é fator causal em mais de 200 doenças e condições de lesões[1]. De acordo com Mellinger (2019), o consumo de cerveja como bebida alcoólica é 34% maior na América do Norte, América do Sul e Europa. Estes dados demonstram que a tendência a alcoolemia é influenciada por condições sociais, enquanto o tipo de bebida está condicionado pela predominância em determinada região.

As causas de acidentes viários têm motivos tríplices, sendo o álcool mais prevalente em comparação a falha do veículo ou ambiente (SILVA et al., 2018). Segundo Abreu, Souza e Mathias (2022), com a implantação da Lei Seca e do CTB[2] houveram impactos positivos na redução dos acidentes até o ano de 2014. Todavia, ainda é necessário que os métodos alternativos de averiguação imediata do estado do usuário sejam desenvolvidos (ROBERTA; RESPLANDES, 2019).

O consumo do álcool reduz a capacidade cognitiva do motorista, contribuindo com o risco de provocar e/ou se tornar vítima de acidentes e, fatores como a idade e a experiência, também são índices que contribuem para a ocorrência destes (MARTIN et al., 2013). Em complemento, o uso do álcool, mesmo que em dose mínima, causa cansaço

[1] https://www.who.int/news-room/fact-sheets/detail/alcohol
[2] https://www.ctbdigital.com.br/artigo/art165

e aumenta a sonolência de quem dirige (BARRET; HOME; REYNER, 2005; HOME; REYNER; BARRETT, 2003). Além disso, mesmo com a eliminação da substância pelo sangue, os efeitos das imparidades ainda são visíveis durante a condução pois é possível perceber uma redução da capacidade motora do motorista (LIU; HO, 2010).

A alcoolemia tem diferentes efeitos no tecido humano e dependendo do teor de álcool presente no sangue em um certo período de tempo, afeta os membros superiores e inferiores do corpo humano (WEAFER; FILLMORE, 2012). O teor de álcool no sangue é determinado pelo tempo necessário para ser absorvido, distribuído, metabolizado e excretado após seu consumo (ZAKHARI, 2006). Segundo Brick (2006), a taxa de eliminação de álcool pelo sangue por indivíduos com o consumo abusivo de álcool tende a ser em média 22mg/dL/h enquanto, para pessoas que realizam o consumo de forma reduzida, esta taxa costuma ser entre 10 a 20mg/dL/h. Não obstante, a taxa de eliminação pelo sangue possui variabilidade de acordo com o sexo, gênero, idade, raça e alimentação (CEDERBAUN, 2012).

Em adição, a taxa de absorção pelo sangue está diretamente correlacionada ao esvaziamento gástrico, sendo o estômago responsável por absorver 70% do álcool consumido e, subsequentemente, o intestino é responsável por metabolizar a substância (HOLT et al., 1980; CORTOT et al., 1986). Um indivíduo em jejum requer um intervalo de 30 minutos a 1 hora para que o nível de álcool no sangue alcance o valor máximo em decorrência da ingestão (LABIANCA, 1992). A metabolização desta substância ocasiona vasodilatação e aumento da temperatura facial nas regiões onde a rede vascular é mais densa, como no nariz, na testa e nos olhos (HERMOSILLA et al., 2018). Após o metabolismo do álcool, o tecido cutâneo sofre um ganho de temperatura através da condução térmica dos vasos sanguíneos e, posteriormente, a vasodilatação permite maior condução térmica e radiação (JONES; PLASSMANN, 2002). Em adição, o processo de termorregulação influencia a temperatura interna do corpo, reduzindo-a através do suor e da vasodilatação (YODA et al., 2005).

2.2 Termografia e as mudanças de temperatura na pele

O imageamento térmico permite a detecção natural da radiação térmica que é emitida pelo tecido cutâneo, como também a interpretação das distribuições de temperaturas que representa mudanças fisiológicas (JONES; PLASSMANN, 2002).

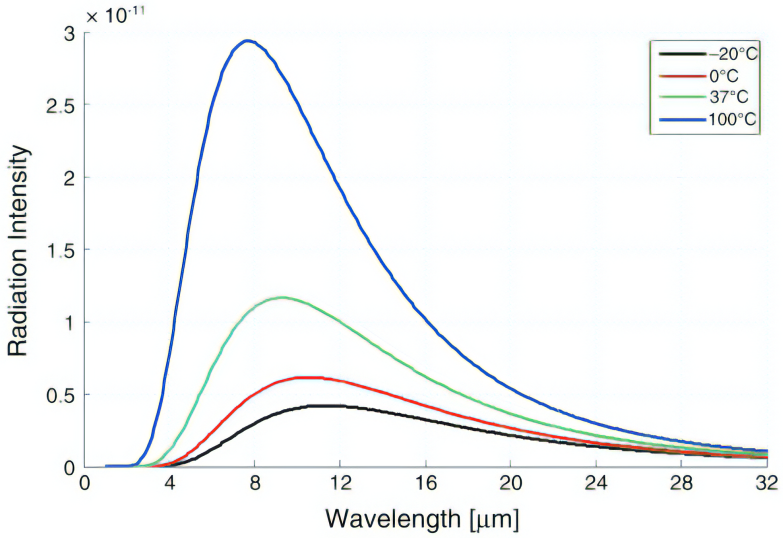
A atmosfera transmite radiação eletromagnética dentro de um comprimento específico de onda devido a absorção de outros comprimentos de ondas da atmosfera. A absorção da radiação eletromagnética no espectro infravermelho é realizada em maior parte pelas moléculas de CO2 e H20 (GADE; MOESLUND, 2014). A radiação decorrente de um determinado objeto de temperatura T é descrita pela função de distribuição de comprimento de onda de Planck (SERWAY; KIRKPATRICK, 1988; GADE; MOESLUND,

2014):

$$I(\lambda, T) = \frac{2\pi hc^2}{\lambda^5 \left(e^{hc/\lambda k_B T} - 1\right)} \tag{2.1}$$

Onde λ é o comprimento de onda, h é a constante de Planck (6,626 x 10^{-34})Js, c é a constante que representa a velocidade da luz (299,792,458 m/s) e k_B é a constante de Stefan–Boltzmann ($1,3806503 \times 10^{-23}$ J/K). Conforme a Figura 1, o pico de intensidade da radiação tende a ter um comprimento de onda menor para temperaturas maiores e a intensidade da radiação aumenta com o incremento da temperatura.

Figura 1 – Intensidade da radiação de um corpo negro em relação ao comprimento de ondas em diferentes temperaturas.



Fonte: Gade e Moeslund (2014).

O pico do comprimento de onda é expresso pela Lei de Wien (SERWAY; KIRKPATRICK, 1988; GADE; MOESLUND, 2014):

$$\lambda_{max} = \frac{2,898 \times 10^{-3}}{T} \tag{2.2}$$

A função de distribuição do comprimento de onda de Planck descreve a radiação de um corpo negro. Segundo Gade e Moeslund (2014), os materiais estudados para aplicações práticas da termografia são denominados corpos cinzas e possuem um valor constante de escala entre 0 e 1. Este fator de escala é chamado de emissividade, onde a pele é considerada um corpo negro perfeito para radiação do espectro eletromagnético e possui valor próximo a 1. (KOUKIOU; ANASTASSOPOULOS, 2012; KOUKIOU; ANASTASSOPOULOS, 2015; GADE; MOESLUND, 2014)

A aplicação da termografia em ambientes controlados para a observação de seres vivos, segundo Tattersall (2016), é crucial para que a temperatura seja capturada através

de vídeos ao longo do tempo para observação de variações de temperatura que são pequenas e ocorridas ao longo deste período. Contudo, diferentes fatores podem afetar a aquisição de imagens térmicas. Fernandez-Cuevas et al. (2015) classificaram-nos em 3 grupos primários:

- Fatores ambientais: está relacionado ao ambiente em que a avaliação de temperatura é realizada.
- Fatores individuais: está relacionado a fatores intrínsecos e extrínsecos às características pessoais do indivíduo.
- Fatores técnicos: está relacionado a fatores relevantes ao dispositivo termográfico empregado para a avaliação da temperatura.

Fatores como o tamanho da sala podem não influenciar diretamente, entretanto, o ambiente deve ter espaço suficiente para acomodar o equipamento e os participantes. Em adição, a temperatura ambiente é relevante para aplicações que se tratam de humanos. A aclimatização é essencial para que a temperatura do corpo do indivíduo se estabilize para não influenciar o termograma posteriormente. Marins et al. (2014) demonstraram estatisticamente que para a aclimatação, o tempo necessário para o balanceamento da temperatura difere conforme o sexo do indivíduo. Neste trabalho, os autores sugerem a espera de no mínimo 10 minutos para que todas as partes corporais sejam aclimatadas. Ainda segundo (FERNANDEZ-CUEVAS et al., 2015), a temperatura ambiente adequada para avaliação de imagens térmicas de seres humanos é entre 18-25^{o}C. Em situações com temperaturas menores, a pessoa pode apresentar arrepios e em temperaturas maiores a transpiração.

Não obstante, fatores externos que podem influenciar o estado psicofisiológico do indivíduo também podem afetar a sua temperatura observada pela imagem térmica. O trabalho de Sonkusare et al. (2019) evidencia a alteração de temperatura nas regiões das bochechas e da ponta do nariz após a indução de ruídos altos com uma latência até a sua detecção de até 10 segundos, através de câmeras térmicas. Entretanto, uma observação com um tempo maior de duração poderia auxiliar na caracterização do tempo de recuperação dessa resposta térmica.

Ademais, (FERNANDEZ-CUEVAS et al., 2015) demonstram que fatores individuais como sexo, idade, altura e peso podem influenciar na temperatura observada em termogramas do indivíduo. Outros estudos da literatura mostram diferentes fatores que afetam a temperatura superficial da pele de uma pessoa relacionados a diferentes locais da face e também outras partes do corpo, como o estado psicofisiológico (ENGERT et al., 2014; IOANNOU; GALLESE; MERLA, 2014) e o ciclo cicardiano (REILLY; WATERHOUSE, 2009; MARINS et al., 2015).

As características técnicas que influenciam a termografia estão relacionadas ao emprego do equipamento e, consequentemente, a qualidade. De acordo com Fernandez-Cuevas et al. (2015), a validade da medição da temperatura refere-se a correspondência da temperatura aferida pelo dispositivo com o valor de temperatura do mundo real de determinada superfície a qual está se medindo, enquanto que a confiabilidade dos valores das medidas obtidas está relacionada à consistência da repetibilidade da aferição. Acrescenta-se que a reprodutibilidade da medida está relacionada a possibilidade de obtenção dos mesmos valores ao longo do tempo.

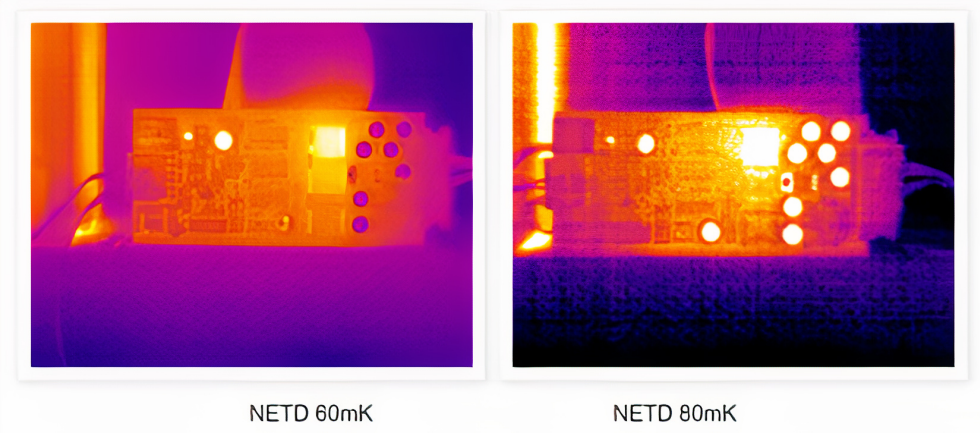
Outro atributo de dispositivos térmicos que afetam a termografia é a resolução espacial da matriz de pixels para a captura de imagens. Fernandez-Cuevas et al. (2015) explicam que, em decorrência da escolha de diversos autores em trabalhar com dispositivos térmicos de resolução espacial de 320x240 (METZMACHER et al., 2018; MACHADO et al., 2021; NKENGNE; PAPILLON; BERTIN, 2013), é possível definir que esta resolução seria o mínimo requerido para trabalhos científicos que utilizam humanos para captura de imagens.

A Diferença de Temperatura Equivalente a Ruído (NEDT) é outra característica dos dispositivos térmicos que influencia na qualidade de uma imagem. Os imageadores térmicos com valor de NEDT reduzidos até 0,05^{o}C, permitem melhor distinguibilidade entre sinal e ruído. O NEDT mede a capacidade do imageador de detectar as menores diferenças de temperaturas em uma superfície. De acordo com Buddharaju (BUDDHARAJU et al., 2007), câmeras térmicas com sensibilidade térmica de até 20mK são suficientes para realizar a segmentação da rede vascular a partir de imagens térmicas. Isto é possível em decorrência dos vasos sanguíneos mais densos que estão em proximidade do tecido cutâneo superficial durante a elevação de temperatura sobre a superfície (JONES, 2010).

O valor do NEDT indica a menor diferença de temperatura que o sensor da câmera consegue detectar. Desta forma, quando o ruído de um sinal de imagem possui o mesmo valor da menor diferença de temperatura a ser medida, significa que o detector de imagens do dispositivo térmico alcançou o seu limite (KIRIMTAT et al., 2020).

A Figura 2 apresenta a imagem de um mesmo objeto a partir de dois dispositivos diferentes com valores de NEDT diferentes.

Figura 2 – Comparação entre dispositivos térmicos com diferentes valores de sensibilidade térmica.



Fonte: MoviTherm [3].

A partir da figura acima, é possível perceber a diferença entre a quantidade de ruído na imagem entre os imageadores térmicos com diferentes valores de sensibilidade térmica. Sob este contexto, os imageadores térmicos que possuem valores maiores de NETD são mais sensíveis ao ruído na imagem.

A importância da termografia é dada pela sua relevância para a extração de atributos de uma imagem térmica que representem a alteração fisiológica em um indivíduo (HERMOSILLA et al., 2018). De acordo com (BUDDHARAJU; PAVLIDIS; TSIAMYRTZIS, 2006), isto é possível em decorrência da fenomenologia dos efeitos da troca de temperatura entre a rede vascular facial e a pele humana.

Em adição, o processo de termorregulação que é regido o hipotálamo realiza a transferência de calor do centro do corpo onde o sangue armazena a temperatura e, através do fluxo sanguíneo, chega às regiões periféricas do corpo como a pele. Consequentemente, a temperatura interna sofre uma redução (LAHIRI; BAGAVATHIAPPAN; AL., 2012). Posteriormente, em decorrência do fenômeno da convecção, existe uma troca de temperatura entre os vasos capilares mais densos na região superficial do rosto e o tecido cutâneo da região superficial do rosto (BUDDHARAJU; PAVLIDIS, 2012; BUDDHARAJU; PAVLIDIS; TSIAMYRTZIS, 2006; BUDDHARAJU et al., 2007).

2.2.1 Imageamento térmico

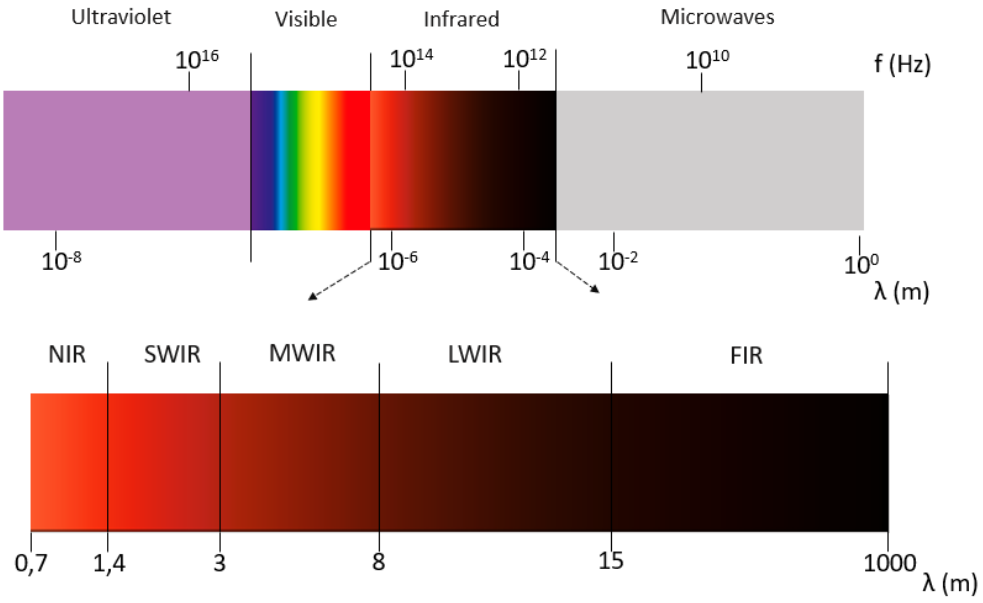
As imagens térmicas capturadas sob o espectro infravermelho têm sido utilizadas ao longo dos anos para observação da distribuição de temperatura da pele humana na área médica. As imagens térmicas permitem que sejam observadas alterações que possam indicar lesões, inflamações, infecções e possíveis malignidades no corpo humano. Essa

análise se torna possível devido os objetos que possuem temperatura acima do zero absoluto emitirem radiação eletromagnética dentro do espectro infravermelho.

O processamento das imagens térmicas permite a obtenção de informações sobre os processos fisiológicos através da análise da distribuição superficial da temperatura da pele que pode ser correlacionada com o estado fisiológico do indivíduo. Desta forma, as câmeras térmicas que possuem suficiente resolução espacial e sensibilidade térmica permitem a observação da variação de temperatura no tecido cutâneo do corpo humano ao longo do tempo (GADE; MOESLUND, 2014).

A radiação de ondas no espectro infravermelho tem emissão a partir de todos os objetos com temperaturas acima do zero absoluto. A Figura 3 apresenta as faixas com sub-divisão do espectromagnético. A radiação do espectro infravermelho está entre as frequências de ondas visíveis e microondas.

Figura 3 – Espectro eletromagnético com sub-divisão.



Fonte: Adaptado de Gade e Moeslund (2014)

Sob o contexto de sensoriamento de temperatura, a obtenção de valores radiométricos de um imageador térmico necessita de repetibilidade e estabilidade da medição do valor de temperatura aferido sobre a região de interesse para cada quadro capturado e sendo necessário o emprego do dispositivo adequado para obtenção de imagens de qualidade e medição correta da temperatura superficial (HOWELL; DUDEK; SOROKO, 2020).

A extração de atributos relacionados a fenomenologia térmica, descrita na seção anterior, utilizando câmeras térmicas, é possível devido ao efeito de convecção de calor produzido pelos vasos sanguíneos (BUDDHARAJU; PAVLIDIS; TSIAMYRTZIS, 2006).

Em(BUDDHARAJU et al., 2007) é demonstrado que, através de técnicas de processamento de imagens, é possível computar a estimativa dos vasos sanguíneos superficiais em imagens térmicas da face através de operações morfológicas para o reconhecimento facial. Não obstante, em (BUDDHARAJU; PAVLIDIS, 2012) é demonstrada a importância do uso de imagens térmicas para extração de características fisiológicas devido a unicidade da rede vascular sanguínea para cada indivíduo. Todavia, o termograma de um indivíduo pode ser afetado por atividades que alteram o metabolismo como o hábito da ingestão de café, atividades físicas e respiração com alta frequência (SOCOLINSKY D. SELINGER, 2004).

Segundo Khan, Ward e Ingleby (2009), o uso de sensores infravermelhos mais sensíveis à radiação eletromagnética pode melhorar a distinção da visualização de atributos faciais. As câmeras térmicas possuem bolômetros como sensores de imagem que determinam o tamanho da matriz de *pixels*. Para realizar a captura da imagem, um detector térmico absorve e converte a radiação em energia térmica e, subsequentemente, o sensor tem um ganho de temperatura onde o sinal é então convertido em imagem (KIMATA, 2018).

O microbolômetro de uma câmera térmica é responsável por converter o sinal elétrico que é detectado pela variação da resistividade do material utilizado. O material utilizado com mais frequência é o óxido de vanádio (VOx) e silício amorfo (a-Si) (GADE; MOESLUND, 2014). As câmeras térmicas são sensíveis a ruídos em decorrência da flutuação da temperatura ambiente conhecida como temperatura diferencial equivalente ao ruído[4]. Este valor é representado na unidade de miliKelvins (mK).

O sensor de uma câmera térmica tem o valor da sua sensibilidade térmica computado pelo desvio padrão de temperatura na abertura de entrada do dispositivo, resultando em uma taxa de razão entre sinal e ruído (MILTON; BARONE; KRUER, 1983). Desta forma, os dispositivos térmicos que possuem baixa sensibilidade térmica podem capturar maiores quantidades de variação de temperatura ao longo do tempo e possuem um melhor filtro de ruído branco. Neste contexto, uma câmera térmica com reduzido valor de NEDT pode fornecer detalhes dos vasos sanguíneos sob o tecido cutâneo (BUDDHARAJU; PAVLIDIS, 2012).

A Tabela 1 mostra a faixa espectral dos comprimentos de onda infravermelho onde o infravermelho longo corresponde ao espectro das câmeras térmicas. Câmeras térmicas podem ser utilizadas como dispositivos de leitura, capturando apenas um ponto único ou uma linha inteira em uma imagem por vez ou pode ser utilizada como uma matriz de observação onde todos os elementos de uma imagem são capturados ao mesmo tempo para cada pixel da matriz de pixels do bolômetro (GADE; MOESLUND, 2014).

Koukiou e Anastassopoulos (2012) demonstraram que, em decorrência da ingestão do álcool, que ocasiona vasodilatação na rede vascular facial do sujeito, é possível realizar

[4] Flir Systems NETD. Disponível em: https://www.flir.com/support-center/Instruments/how-is-nedt-measured/ Acessado em: 05 fev. 2022

Tabela 1 – Comprimento e faixas de onda o espectro infravermelho.

Comprimento de onda	**Faixa**
Infravermelho próximo	0.7 µm - 1.4 µm
Infravermelho de onda curta	1.4 µm - 3 µm
Infravermelho de onda média	3 µm - 8 µm
Infravermelho de onda longa	8 µm - 15 µm

o emprego de câmeras térmicas para observar o aumento de temperatura na face devido à alcoolemia. Este aumento de temperatura ocorre na região superficial do tecido cutâneo facial e se deve ao processo termorregulatório que é regido pelo hipotálamo (HERMOSILLA et al., 2018).

O uso de imageadores térmicos com sensibilidade térmica adequada permite que sejam obtidos termogramas que representem a variação de temperatura de um indivíduo em decorrência da embriaguez. Koukiou e Anastassopoulos (2012) e Hermosilla et al. (2018) realizaram a seleção de pixels de um termograma nas regiões da face em que os vasos sanguíneos se cruzam para verificar que o valor destes aumenta gradualmente após a metabolização do álcool no organismo.

Ainda sob este contexto, Koukiou e Anastassopoulos (2015) investigaram o incremento de temperatura na região da face após o consumo de álcool. Neste trabalho foi observado a elevação da temperatura nas regiões dos olhos, boca e nariz, em consonância, Koukiou e Anastassopoulos (2016) demonstraram que pode ocorrer uma acrescentação de temperatura na região dos olhos de um indivíduo após a ingestão de bebidas alcoólicas. Consequentemente, os trabalhos supramencionados, apresentaram regiões que são relevantes para a identificação da alteração do indivíduo em decorrência da ebriedade.

2.3 Internet of Things (Iot) e Cloud Computing

O conceito de Internet das Coisas é permitir que seja possível computar informações obtidas a partir de sensores sem a intervenção humana, possibilitando que diversos dispositivos produzam grandes quantidades de dados e realizem comunicações através de arquiteturas de redes complexas (SHI; DUSTDAR, 2016). Além disso, a Internet das Coisas pode ser entendida como uma rede compreensiva de objetos inteligentes que possuem a capacidade de auto-organização, compartilhamento de recursos e dados, podendo agir e reagir em situações e mudanças no ambiente (MADAKAM; RAMASWAMY; TRIPATHI, 2015).

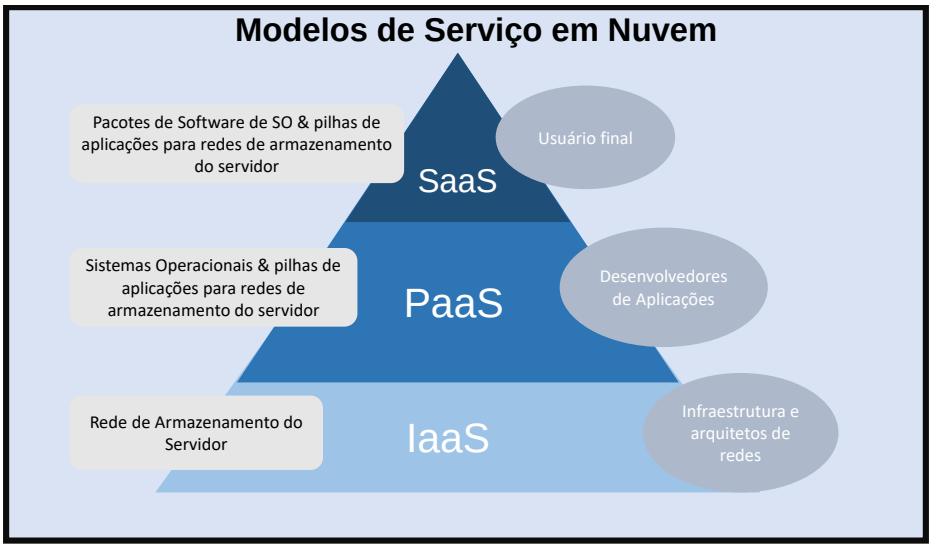
Além disso, a nuvem também tem como vantagem o *pool* de recursos que permite o uso do mesmo *pool* por vários usuários através de multilocação e virtualização de recursos. Todavia, um dos problemas com a computação em nuvem é a segurança e privacidade de

dados (MOHAMMED; ZEEBAREE; AL., 2016). Em adição, a *cloud computing* possui diferentes vulnerabilidades para cada camada de serviço. Em adição, existe a necessidade de enviar os dados a partir de *endpoints* até o servidor *cloud*, além de custos operacionais (PARIKH et al., 2019).

Neste contexto, segundo Varghese (VARGHESE et al., 2016), as aplicações baseadas em nuvem empregam centros de processamento de dados como servidores centrais para realizar o processamento de dados gerados por dispositivos na borda da rede como dispositivos móveis, tablets e dispositivos vestíveis. Este modelo de arquitetura demanda cada vez mais da infraestrutura computacional e da comunicação, impactando de forma adversa a qualidade de serviço.

O trabalho de Mohammed, Zeebaree e al. (2016) evidencia a contribuição da computação em nuvem para o melhor desempenho em aplicações que dependem de tempo, armazenamento e desempenho computacional. Desta forma, apesar da Internet das Coisas diferir da computação em nuvem, ambas as tecnologias se beneficiam mutuamente no desenvolvimento de aplicações devido à convergência de suas propriedades. Acrescenta-se que, a computação em nuvem possui características promissoras como flexbilidade, escalabilidade, disponibilidade e custo reduzido. O crescimento desta tecnologia compromete a segurança de dados, o que leva a complexidade da privacidade e segurança de informações pessoais (SRIKANTH; JAFFRIN, 2022).

Figura 4 – Estrutura de serviços da nuvem.



Fonte: Adaptada de Srikanth e Jaffrin (2022).

Conforme a Figura 4, a computação em nuvem oferece serviços como: infraestrutura como serviço (IaaS), plataforma como serviço (PaaS) e *software* como serviço (SaaS). Todos

estes serviços fornecem serviços de computação sob demanda, como armazenamento e dados processamento (OMETOV et al., 2022). A computação na nuvem é uma solução eficaz para o gerenciamento de serviços de IoT, bem como para implementação de aplicações e serviços que exploram os dados produzidos por estes. Igualmente, a nuvem pode estender o seu escopo para lidar com as coisas do mundo real de forma dinâmica e distribuída (BOTTA et al., 2016).

Embora a nuvem possua os atributos supracitados, ainda possui restrições como a distância para os dispositivos dos usuários e, em alguns cenários, pode não ser viável o envio de dados à nuvem devido a preocupações em relação a privacidade e/ou a localização atual do indivíduo não possibilitar a conexão com o servidor (YOUSEFPOUR et al., 2019). Em decorrência destas fragilidades e com o considerável crescimento da Internet das Coisas, a computação na borda da rede emerge como forma de agilizar o processamento de dados em tempo real sendo uma extensão da computação em nuvem (CAO et al., 2020).

2.4 Edge Computing

A computação em borda aproxima o servidor do local em que ocorre a entrada de dados realizando a troca de informação em tempo real entre sensor e dispositivo na borda da rede. O objetivo é evitar a alta latência e gargalo no tráfego de dados associados ao uso da nuvem, permitindo o gerenciamento de recursos e a segurança das informações pessoais de indivíduos (KHAN et al., 2019). Ainda sob este contexto, o estudo de Yu et al. (2018) ressalta a relevância da computação realizada na borda da rede que permite grandes quantidades de dados possam ser processadas em tempo real próximo a aplicação do usuário, quando comparado ao tradicional uso da nuvem para esta tarefa. Em adição, há uma redução considerável da latência relacionada ao processamento de dados, podendo tratar diretamente de dados que não fazem uso da nuvem. Desta forma, a computação na borda da rede consegue preservar a segurança, controlabilidade e privacidade dos dados do usuário (ZEYU et al., 2020; WANG et al., 2020).

Não obstante, uma das motivações da computação na borda da rede é a preocupação em torno da segurança dos dados quando os sensores são alocados em ambientes públicos (CHEN; RAN, 2019). Em decorrência do uso de dispositivos de computação limitada na borda da rede, estes são mais vulneráveis que a computação em nuvem em relação a ataques como negação de distribuição de serviços (DDoS). Isto se deve à diferença de protocolos empregados na nuvem como HTTP/HTTPS, FTP e STMP, além de possuir mais recursos para mitigação destes ataques, enquanto que a edge emprega protocolos mais leves como CoAP, MQTT e UDP (XIAO; JIA; LIU, 2019).

A computação em nuvem possui características que a torna menos favorável para implementação de aplicações e serviços entre a nuvem e a borda da rede (ZHANG et al.,

2020). Primeiramente, a transmissão de dados se torna difícil em situações onde a conexão com a internet está instável, os dados coletados na *edge* podem conter informações sensíveis e particulares sobre as pessoas e desta forma a transmissão destes dados sem a adjunção de mecanismos de segurança pode acarretar considerável perigo às informações do indivíduo. E finalmente, o crescimento exponencial de nós na borda da rede pode gerar gargalo para o servidor na nuvem, fazendo com que esta abordagem seia inviável ou custo-efetiva para o envio de grandes quantidades de dados.

2.4.0.1 Características da Computação em Nuvem e Edge Computing

A computação na borda da rede evoluiu como um paradigma de computação promissor, alavancado para melhorar a qualidade dos serviços de dispositivos de borda e o desempenho da rede em comparação a computação em nuvem, abordando problemas de latência e *offloading* de computação intensiva realizados em nós da borda da rede (ALSAMHI et al., 2022). De acordo com Ometov et al. (2022), a computação na borda da rede permite a redução do *overhead*[5] entre servidor e cliente em vista da redução na quantidade de saltos em que o pacote enviado deve fazer até chegar ao seu destino. Enquanto que, a nuvem garante que os serviços estejam sempre disponibilizados de forma consistente, a confiabilidade seja mantida e os dados sejam fornecidos conforme a demanda. Contudo, para aplicações móveis que necessitam de tempo real, a computação em nuvem não é adequado em virtude das longas distância para o envio de dados conforme supracitado. Desta forma, uma comparação é realizada abaixo:

Tabela 2 – Características entre diferentes paradigmas - Computação em Nuvem e Edge computing.

Nível de Atributos	**Computação em Nuvem**	**Edge Computing**
Arquitetura	Centralizada	Distribuida
Tipo de serviço	Serviços universais	Redes Móveis
Segurança	Centralizado (garantido pelo provedor da nuvem)	Centralizado (garantido pela operadora móvel)
Mobilidade	Inadequado	Oferecido de forma limitada
Interação em tempo real	Oferecido	Oferecido
Latência	Alta	Baixa
Custo da largura de banda	Alto	Baixo
Capacidade de armazenamento e computação	Alta	Limitada
Escalabilidade	Média	Alta

Fonte: Adaptado de (OMETOV et al., 2022).

A Tabela 2 apresenta as características da computação em nuvem e *Edge computing*. As características da computação realizada na borda da rede faz com que esta seja uma

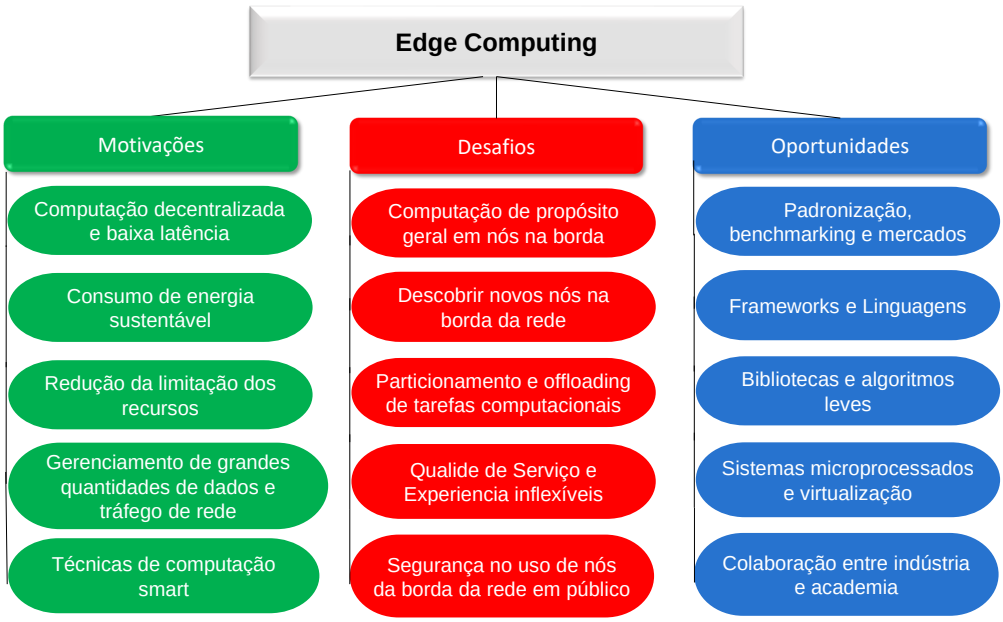
[5] https://www.techtarget.com/whatis/definition/overhead

solução para aplicações em que o tempo de resposta e mobilidade sejam essenciais (SHI; DUSTDAR, 2016). Segundo (OMETOV et al., 2022), a nuvem sendo uma arquitetura centralizada com considerável número de aplicações em IoT, dispõe de inconveniência como alta latência, sensibilidade a localização e tempo de computação. Em vista disso, os paradigmas Edge e Fog tem como objetivo diminuir a carga sobre os sistemas em nuvem e resolver os problemas indicados. Por fim, percebe-se que esses dois paradigmas contribuem para a redução de considerável quantidade de dados enviados para a Nuvem. Por conseguinte, o paradigma da computação em Edge é vantajoso sobre o paradigma da computação em nuvem, em termos de segurança e privacidade.

2.4.1 Edge Computing: Motivações, Desafios e Oportunidades

Esta seção apresenta as características que motivam a computação realizada na borda da rede, assim como os desafios e as oportunidades com o objetivo de detalhar o potencial do uso de dispositivos embarcados.

Figura 5 – Edge Computing: Motivações, Desafios e Oportunidades.



Fonte: Adaptado de Varghese et al. (2016).

A Figura 5 apresenta características relevantes a implementação da computação realizada na ponta da rede. Desta forma, é possível inferir que a computação centralizada na nuvem pode não ser adequada para situações em que a melhor estratégia para as aplicações é a distribuição geográfica. Posteriormente, a computação realizada próxima ao sensor

onde ocorre a entrada de dados pode melhorar o serviço entregue, podendo ser generalizado para outras aplicações. Em vista disso, *Edge Computing* permite a decentralização da nuvem para obter reduzida latência ao aproximar o servidor dos sensores, além de reduzir o consumo de energia em comparação a servidores de nuvem (YANG et al., 2019). Em adição, a computação em nuvem possui um tempo de *round-trip time* maior para a transmissão de dados quando comparado a computação realizada na borda da rede (KITANOV; JANEVSKI, 2017).

Ainda sob este contexto, a computação na borda da rede guarnece, além das motivações, novos desafios e oportunidades que servem como direção para realizar computação na borda da rede. Em adição, a investigação da eficiência de energia na borda da rede tem sido negligenciado em decorrência da considerável dificuldade de interação entre servidores e dispositivos de borda em conjunto com a nuvem (JIANG et al., 2020). Desta forma, técnicas de particionamento de recursos e *offloading* da carga computacional podem ser explorados para otimizar o consumo energético e tempo de execução na borda da rede (XU et al., 2019).

Conforme o supracitado, a borda da rede permite o emprego de ferramentas que possibilitem processar requisições em tempo real em nós das extremidades da rede. Além da compatibilidade com dispositivos leves para realizar tarefas computacionais complexas de forma eficiente (XIAO et al., 2019). Consequentemente, o uso de micro sistemas operacionais, microcomputadores ou microcontroladores pode fornecer opções para lidar com os desafios relacionados à implantação de aplicativos em nós heterogêneos na borda da rede (MANSOURI; BABAR, 2021).

Em adição, a borda da rede permite que algoritmos de redes neurais convolucionais sejam implementados para realizar a inferência em tempo real. O trabalho de Sufian et al. (2020) demonstra que a borda da rede possibilita o treinamento de modelos de *deep learning* através da transferência de aprendizado em um cenário onde há a escassez de exemplos de imagens durante o surgimento do COVID-19, sendo possível agilizar o treinamento de um modelo para um tarefa onde há uma limitada quantidade de amostras.

Os dispositivos embarcados na borda da rede possibilitam otimizar a computação através do uso de redes neurais para torná-la inteligente (CHEN; RAN, 2019). A solução apresentada por Zhang, Wang e Lu (2019) demonstra ser possível otimizar o uso de recursos entre os servidores na borda da rede através do emprego de redes de DNN e, como consequência, possibilita a obtenção de consumo reduzido de banda, mantendo a qualidade de serviço se comparado ao uso da nuvem como servidor para entrada de dados a partir de sensores.

A disponibilidade do servidor na borda da rede é uma das características necessárias para a computação em borda. Monburinon et al. (2019) propuseram uma ferramenta para a detecção de vida selvagem nociva às plantações, onde são implantados diversos

modelos de redes neurais profundas para diferentes dispositivos embarcados que permite a escalabilidade para múltiplos nós e diferentes modelos de DNN. Através do uso de *containers* como o *Docker*, os modelos são atualizados e os dados são enviados para a nuvem. Entretanto, o *hardware* de cada dispositivo tem diferença no tempo de inferência, latência, taxa de quadros por segundo da câmera utilizada e memória consumida, o que é demonstrado nesse trabalho. O trabalho dos autores demonstra a divergência entre as métricas para cada tarefa na borda da rede utilizando redes neurais profundas.

O trabalho de Khan et al. (2019) demonstra que o uso de microcomputadores de placa única na borda da rede como servidor para comunicação com os sensores, ao invés da nuvem, permite obter redução de latência no servidor. O trabalho dos autores apresenta uma arquitetura de computação em borda para reconhecimento facial utilizando redes neurais convolucionais. Para comparação entre a nuvem e borda foram enviadas 5 imagens. O tempo médio de transferência apresentado para o dispositivo na borda foi de 2,4 segundos, enquanto que para a nuvem a transferência foi de 11,8 segundos, em média. Contudo, a latência ainda poderia ser reduzida dado que a classificação é realizada a partir da nuvem.

A execução de modelos de redes neurais convolucionais na borda da rede demanda considerável custo computacional dos dispositivos embarcados de forma eficiente. Segundo Chen e Ran (2019), alguns métodos de otimização do modelo de redes neurais na borda da rede utilizados são: computação no dispositivo, que envolvem técnicas de compressão de modelo, escolha de *hardware* adequado, quantização dos pesos da rede, partição da rede neural e computação distribuída. Ainda é citado que para a computação na borda da rede as métricas de desempenho de maior relevância são a latência, tempo de inferência, consumo de energia, banda disponível e acurácia do classificador.

Para a implantação de modelos de redes neurais profundas na borda da rede é necessária a organização do ciclo de vida de um algoritmo na borda da rede, que consiste na etapa de treinamento através do uso de um conjunto da base de dados para treinamento do modelo. A segunda etapa consiste na predição das classes de novos dados de entrada de um modelo implantado em um servidor onde existe um tempo para a realização da predição, conhecido como tempo de inferência. Este, depende do tamanho de armazenamento do modelo e especifidade técnica do *hardware* utilizado.

A inferência de classificadores na borda da rede possibilita o uso de técnicas de otimização do tempo de inferência das redes neurais na borda da rede. De acordo com Chen e Ran (2019), as técnicas aplicáveis para a otimização se dividem em: (I) Computação realizada no dispositivo que empregam técnicas de design de modelo, compressão e escolha do *Hardware* ideal para a tarefa a ser realizada. (II) Computação na borda da rede que consiste no pré-processamento de dados e gerenciamento de recursos. (III) Computação entre dispositivos da borda da rede, que utilizam técnicas de *offloading*

da carga computacional, particionamento das camadas da rede entre os dispositivos e computação distribuida.

Desta forma, a inferência realizada em tempo real é crucial para diversas aplicações. Como exemplo, os quadros de um veículo autônomo requerem um processamento ágil para detectar e desviar de obstáculos (CHEN; RAN, 2019). Conquanto, o envio de imagens ao servidor da nuvem para inferência pode acarretar um atraso maior, podendo não satisfazer os requisitos de baixa latência ponta-a-ponta necessários para aplicações interativas. Enquanto que, a baixa latência da *edge computing*, permite aos usuários executar aplicações com demanda intensiva de recursos e sensíveis ao atraso na borda da rede (KHAN et al., 2019).

Para a inferência de um classificador na borda da rede, utilizando como servidor um computador de placa única, será empregado uma rede nerural convolucional treinada e ajustada através de técnicas *transfer learning*, com seus parâmetros e ajustes realizados através do *fine-tunning* e *pre-training* que, posteriormente, será otimizada. Este modelo foi desenvolvida com base no trabalho de Silva (2021) dentro do contexto de reconhecimento e classificação de embriaguez e, com isso, será realizada a inferência no dispositivo em borda para a classificação das imagens que serão capturadas pelo dispositivo.

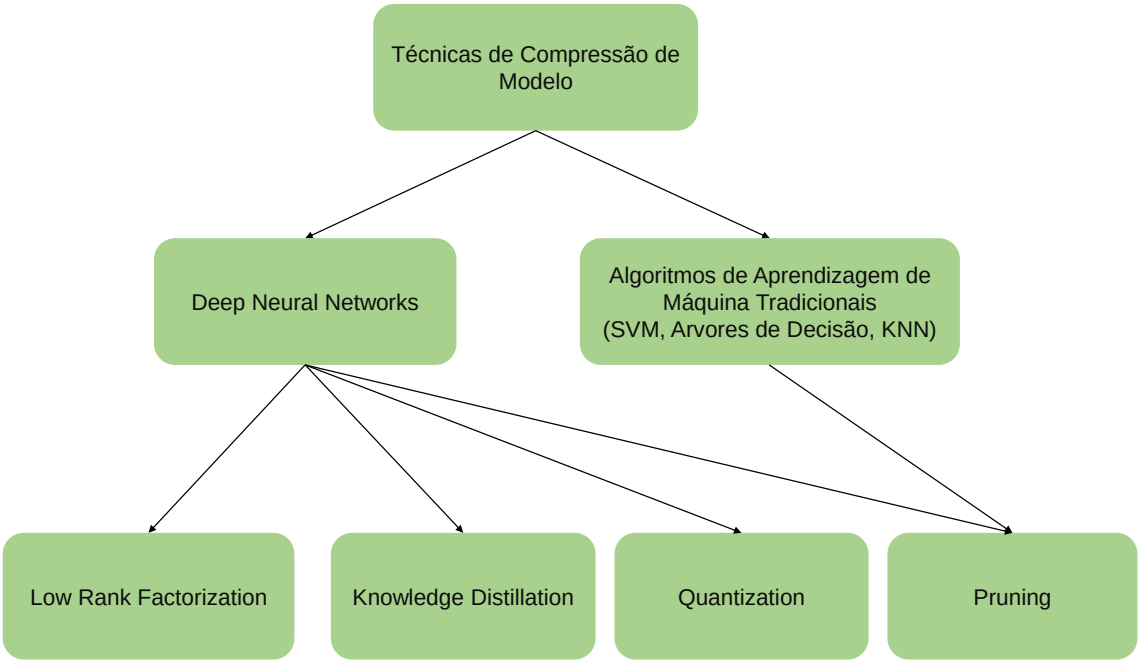
2.5 Deep Neural Networks e compressão de modelos

Ao longo do anos, diversos modelos de redes neurais convolucionais foram propostos com o objetivo de obter a melhor performance possível (CAPRA et al., 2020). Em adição, uma das linhas de IoT tem como objetivo a compressão de modelos de DNN para implementação em dispositivos embarcados de baixa capacidade computacional, como Raspberry Pi [6] ou Arduino[7], tendo benefícios como privacidade, largura de banda e escalabilidade (CHEN; RAN, 2019). Desta forma, uma das preocupações atuais é a implementação de modelos de redes profundas em dispositivos embarcados sem impactar significativamente o desempenho do modelo de RNC (CHENG et al., 2018).

Não obstante, para a implementação de redes neurais profundas na borda da rede, outros autores apresentaram métodos de aceleração de hardware através da otimização de dispositivos lógicos como FPGAs (RODRíGUEZ et al., 2018; KARRAS et al., 2020; XIA et al., 2021). Entretanto, o uso de FPGAs requer que sejam programados especificamente para suas respectivas tarefas e demoram um tempo considerável de programação em nível *software* para seu desenvolvimento. Além disso, outros trabalhos propuseram diferentes métodos para a otimização dos algoritmos de redes neurais profundas para a implementação em dispositivos na borda da rede.

[6] https://www.raspberrypi.com/products/raspberry-pi-4-model-b/
[7] https://www.arduino.cc/

Figura 6 – Diferentes tipos de compressão de modelos para DNN e métodos de aprendizado de máquina tradicionais.



Fonte: Adaptado de Choudhary et al. (2020).

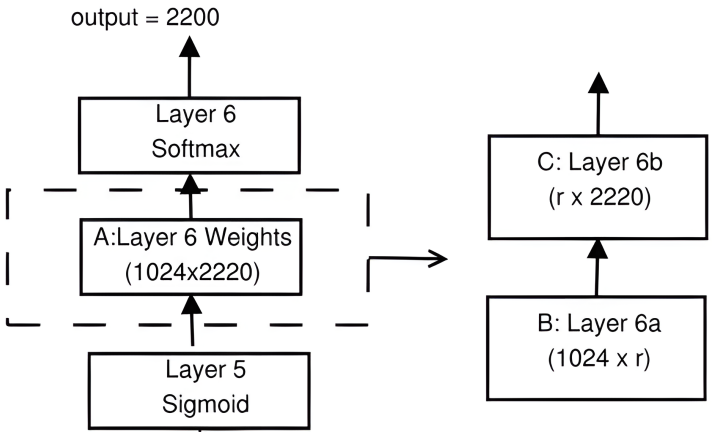
A Figura 6 demonstra as técnicas utilizadas para diferentes algoritmos de aprendizagem de máquina. Sob o contexto de RNC, as técnicas de compressão de modelo mais utilizadas são:

- **Pruning:** esta técnica é utilizada para remover parâmetros redundantes que não contribuem para o treinamento do modelo para reduzir o erro e para melhorar a sua generalização (CHOUDHARY et al., 2020). Esta situação pode ocorrer quando os coeficientes dos pesos possuem valores próximos de zero (LIANG et al., 2021). Subsequentemente, as técnicas de podação podem expandir significativamente a dispersão dos parâmetros e a sua alta dispersão podem beneficiar positivamente o modelo em virtude da redução da memória necessária para armazenamento em disco e redução da complexidade dos parâmetros (CHENG et al., 2018). A podação pode ser aplicada aos filtros, neurônios ou camadas da rede (LIANG et al., 2021). O pruning dos pesos consiste em realizar a remoção dos pesos que não são relevantes, assim os pesos que estão abaixo de um determinado limiar ou são redundantes e tem seus valores zerados (HAN et al., 2015; CHOUDHARY et al., 2020). A apodação de neurônios redundantes pode ser realizada individualmente utilizando métodos de seletividade ou probabilísticos (DENG et al., 2020), ou a partir da regra de conhecimento (CHOUDHARY et al., 2020).

- **Low-rank Decomposition:** O filtro de uma camada de convolução $\boldsymbol{W} \in \mathbb{R}^{w \cdot h \cdot c \cdot n}$ é um tensor 4D, que é uma generalização de matrizes representada por uma matriz em espaço N-dimensional[8] e corresponde às dimensões do filtro como largura, altura, tamanho da entrada e canal de saída. A motivação é encontrar um tensor aproximado $\hat{\boldsymbol{W}}$ que se aproxima o máximo possível de $\boldsymbol{W}$ e torna a computação dos tensores mais eficientes (CHENG et al., 2018; CHOUDHARY et al., 2020).

- **Low-rank Factorization**: A fatorização de matrizes consiste em reduzir os parâmetros através da substituição de uma camada por outras camadas. No exemplo da Figura 7, onde a camada 6 possui pesos A e dimensões m x n, se A possui ordem R, então existe uma fatorização A = B x C (SAINATH et al., 2013) onde B é uma matriz de tamanho m x r e C é uma matriz de tamanho r x n. Desta forma, é possível substituir A pelas matrizes B e C. A redução de parâmetros pode ser realizada desde que o número de parâmetros em B(mr) e C(rn) sejam menores do que em A.

Figura 7 – Diagrama das camadas de uma rede neural convolucional fatorizada.



Fonte: Autor.

- **Knowledge Distillation**: A destilação de conhecimento é um método agnóstico de treinamento de redes neurais que, a partir de um modelo mais complexo, este supervisiona o treinamento de uma rede neural menor auxiliando no treinamento (WANG; YOON, 2022). Desta forma, o conhecimento do modelo professor é transferido para o modelo estudante através da minimização da diferença entre os vetores de probabilidades produzidos pelo modelo professor e estudante de rede neural convolucional. Entretanto, de acordo com Wang e Yoon (2022), em diversas situações os vetores de probabilidades de saída da função softmax do modelo professor possuem classes preditas corretamente com valores de probabilidades altos, enquanto

[8] https://mathworld.wolfram.com/Tensor.html

que as demais classes possuem valores de probabilidades próximos de zero, o que acaba não fornecendo muita informação além dos rótulos informados com a base de dados. Consequentemente, as redes neurais que utilizam softmax como a camada de saída, produzem probabilidades de classes que convertem o logit z_i, computado para cada classe em uma probabilidade q_i, comparando z_i com outros logits (HINTON; VINYALS; DEAN, 2015).

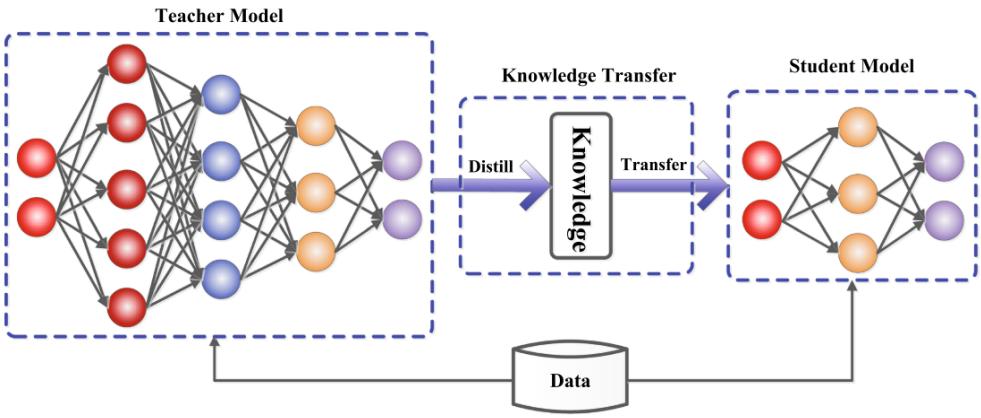
$$q_i = \frac{\exp\left(z_i/T\right)}{\sum_j \exp\left(z_j/T\right)} \tag{2.3}$$

q_i é o parâmetro de temperatura, quando $q_i = 1$, é possível obter a função padrão do softmax. Conforme o valor de q aumenta, a distribuição de probabilidades produzida pela softmax se torna mais suave, fornecendo mais informações sobre a quais classes o modelo professor encontrou maior similaridade em relação ao que foi predito (WANG; YOON, 2022). A equação que computa a destilação do conhecimento composta pelo custo do modelo estudante e da destilação é expressa por (HINTON; VINYALS; DEAN, 2015; WANG; YOON, 2022):

$$\begin{aligned} \mathcal{L}_{KD} &= \alpha * H\left(y, \sigma\left(z_s\right)\right) + \beta * H\left(\sigma\left(z_t; \rho\right), \sigma\left(z_s; \rho\right)\right) \\ &= \alpha * H\left(y, \sigma\left(z_s\right) + \beta * \left[KL\left(\sigma\left(z_t; \rho\right), \sigma\left(z_s, \rho\right)\right) + H\left(\sigma\left(z_t\right)\right)\right] \end{aligned} \tag{2.4}$$

Onde H é a função de custo, y é a classe ground truth, σ é a função de softmax parametrizada pelo valor de temperatura p ($\rho \neq 1$ para o custo de distilação), e α e $\dot{\beta}$ são coeficientes. z_s e z_t são os logits do modelo estudante e professor respectivamente.

Figura 8 – Processo genérico de Knowledge-Distillation.



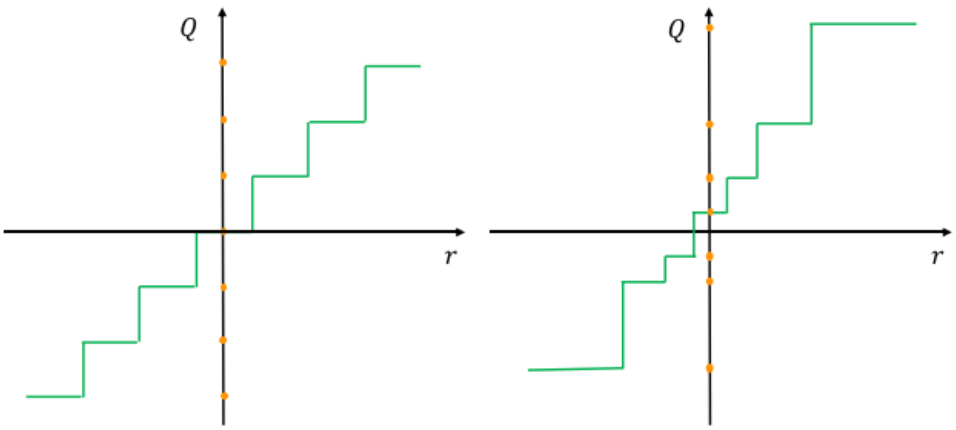
Fonte: Gou et al. (2021).

Na destilação de conhecimento, Gou et al. (2021) explicam que a arquitetura professor-estudante é uma forma genérica de realizar a transferência de conhecimento. Os

autores realizam uma comparação com a forma em que os humanos aprendem. Nesse contexto, espera-se que um aluno encontre o professor compatível para transferir o aprendizado. Analogamente, o design da arquitetura do modelo aluno e professor impactam diretamente do desempenho adquirido pelo modelo estudante. Logo, o modelo estudante tem como objetivo minimizar a combinação entre 2 objetivos (HINTON; VINYALS; DEAN, 2015).

- **Quantization**: A quantização de modelos de redes neurais consiste em comprimir a rede original utilizando redução no número de bits que são necessários para representar os valores dos pesos da rede (CHENG et al., 2017). As tarefas de treinamento e inferência de redes neurais são custosas computacionalmente. Em adição, as redes neurais profundas possuem considerável quantidade de parâmetros. Entretanto, é possível obter uma alta taxa de erro e distância entre um modelo quantizado e sem quantização ao mesmo tempo manter um bom desempenho de generalização do modelo quantizado (GHOLAMI et al., 2022).

Figura 9 – Comparação entre quantização uniforme e quantização não uniforme. Os valores reais em domínio contínuo R são mapeados em valores discretos, os valores de menor precisão são quantizados no domínio Q, que são marcados com os pontos laranjas.



Fonte: (GHOLAMI et al., 2022).

A quantização uniforme pode ser dividida em 2 passos, conforme (WU et al., 2020). Inicialmente, a escolha da faixa dos números reais a serem quantizados, fixando os valores fora desta faixa. Sequencialmente, os valores reais são mapeados para inteiros que podem ser representados pela largura de bits da representação quantizada. A Figura 8 mostra a comparação entre uma quantização uniforme e outra não uniforme.

Desta forma, a quantização de um modelo para int8 pode reduzir a latência durante a inferência de um modelo, tamanho e sem perder significativamente a acurácia

(WU et al., 2020). Os modelos MobileNet, ResNet e Inception são avaliados por (JACOB et al., 2018), onde os autores demonstraram que no caso da MobileNet há um aumento na acurácia após a quantização e para os modelos ResNet e Inception, a queda da acurácia é de no máximo 3%.

2.6 Trabalhos Relacionados

Os trabalhos relacionados a essa pesquisa de mestrado e que dão base para o sistema proposto são descritos na sequencia. Inicialmente, o trabalho de Koukiou e Anastassopoulos (2012) está relacionado à relevância do uso de imagens térmicas para avaliação do estado fisiológico de indivíduos para a detecção do estado de embriaguez. Em relação ao modelo de RNC para a classificação de embriaguez, a estratégia utilizada por Silva (2021) é escolhida como base para o desenvolvimento do classificador desenvolvido. Neste trabalho, um modelo da VGG16 é utilizado somente com os blocos convolucionais para obter *bottleneck features*, e subsequentemente, este vetor de atributos é utilizado para treinar um classificador. Em seguida, a técnica de *fine-tunning* é utilizada para produzir um modelo final. Desta forma, o modelo final é obtido repetindo as etapas de *transfer learning* e *fine-tunning*. O autor obteve 88,5% de acurácia no conjunto de teste utilizando a base de dados da Universidade de Pátras na Grécia. Apesar da acurácia ser menor em comparação aos trabalhos mais recentes da literatura, este trabalho apresenta uma abordagem para indução de características representativas do estado fisiológico do indivíduo, e não obstante, uma explicabilidade melhor do que os demais trabalhos.

Para a implementação de redes neurais convolucionais na borda da rede, Alibabaei et al. (2022) apresentaram as redes neurais convolucionais MobileDet Edge TPU e MobileNet edge TPU para detecção de troncos de videiras utilizando localização de robôs através de imagens térmicas. Em adição, os autores investigaram o impacto da compressão e otimização de modelos de DL na borda da rede.

Uma base de dados pública foi apresentada por Koukiou e Anastassopoulos (2015) utilizando uma câmera térmica do modelo Thermo Vision Micron/A10 que possui uma resolução espacial de 128x160 pixels, e sensibilidade térmica de NEdT $\leq$ 85 mK e uma taxa de quadros de 25Hz em uma temperatura ambiente de até 25ºC operando na faixa de 7,5 a 13um no espectro infravermelho longo. Não obstante, os autores descrevem que a base de dados[9] consiste de 40 participantes dos quais 31 são do sexo masculino e 10 do sexo feminino. Os indivíduos consumiram 120ml de vinho dentro de um período de 1 hora, totalizando 62,4 gramas de álcool no total. O primeiro grupo de imagens foi capturado em um intervalo de 100 milissegundos, resultando em um total de 50 quadros por captura. Este procedimento foi repetido pelos autores antes da ingestão da bebida pelos

[9] http://old.physics.upatras.gr/sober/

participantes e, subsequentemente, a cada 30 minutos após a última dose uma captura foi realizada totalizando 50 quadros de imagens por captura. Todavia, os autores não levam em consideração o tempo de aquecimento do dispositivo antes de realizar a captura das imagens, dado que inicialmente uma câmera térmica possui uma temperatura diferente da temperatura ambiente, o que acaba afetando a sensibilidade térmica do dispositivo. Em adição, a base de dados disponibilizada pelos autores é a única conhecida que permite acesso a pesquisadores.

O uso de imagens térmicas e a borda da rede é estudado por Hegde et al. (2020) onde os autores propuseram um método para a detecção de sintomas que possam indicar o covid-19 em uma pessoa em tempo real. Neste trabalho, os autores utilizaram uma câmera térmica FLIR One Pro que possui uma resolução 120x160 e de 70 mK de sensibilidade térmica. A câmera térmica é utilizada para aferir a temperatura da testa e alteração na temperatura dos lábios para detecção de cianóse, que consiste no azulamento labial que possa indicar COVID-19. Em adição, os autores empregaram o Raspberry Pi como dispositivo na borda da rede e a rede neural convolucional PoseNet. O coprocessador Google Coral USB [10] é utilizado para aceleração da RNC. Contudo, o dispositivo FLIR One Pro possui uma variância no valor de aferição em seus dados radiométricos na faixa de 5ºC, fazendo com que o dispositivo seja ineficiente para a obtenção de repetibilidade e confiabilidade nos valores de temperatura. Em adição, a baixa resolução do dispositivo e a sensibilidade térmica podem não demonstrar na imagem do termograma diferenças significativas no mapa de cores da imagem devido às características supracitadas.

Um sistema para a classificação de dígitos a partir de sinais de mão em imagens térmicas de baixa resolução foi proposto por Breland et al. (2021). Neste trabalho, os autores empregaram a câmera térmica do modelo Omron D6T para capturar imagens em resolução de 32x32 pixels e um *Raspberry Pi*. Subsequentemente, para a classificação das imagens foram coletadas 3200 imagens, sendo 320 quadros para cada classe em um problema de classificação multi-classe com 10 classes. Desta feita, o trabalho apresenta um modelo residual que emprega o uso de uma camada de RNC convolucional. As camadas de *bottleneck* tornam a rede consideravelmente mais leve e também forçam a rede a compactar as representações de características para melhor se adequar ao espaço disponível, fornecendo um resultado aprimorado de treinamento. Desta feita, o modelo proposto neste trabalho apresentou uma acurácia de teste de 99,52%. Conquanto, o presente trabalho não leva em consideração o tempo de inferência do modelo, consumo de energia do *edge server*. Em adição, os autores não levam em consideração o desempenho do servidor na borda da rede em termos de consumo de GPU e temperatura, além de não informar o modelo do servidor na borda da rede utilizado.

A Tabela 3 demonstra trabalhos relacionados que empregaram dispositivos embar-

[10] https://coral.ai/products/accelerator/

cados na borda da rede para tarefas como classificação de imagens e detecção de objetos através do uso de algoritmos de redes neurais profundas e aceleradores de *hardware* para a inferência dos modelos de RNC.

Tabela 3 – Benchmarking de dispositivos embarcados, redes neurais convolucionais e aceleradores na borda da rede.

Autores	Dispositivos	Acelerador	Rede Neural	Métricas	Tarefa
Kristiani, Yang e Huang (2020)	Raspberry Pi 4	Google Coral USB	MobileNet, VGG16, Inception	Acurácia, Tamanho da Rede, Tempo de Carregamento	Classificação de Imagens
Puchtler e Peinl (2020)	Raspberry Pi 4, Jetson NANO	Google Coral USB, Neural Compute Stick 2	MobileNet	Precisão Média, Taxa de Quadros por Segundo	Detecção de Objetos
Baller et al. (2021)	Coral Devboard, Raspberry Pi 4, Jetson NANO, Arduino Nano 33 BLE, ASUS Tinker Edge R	Google Coral USB, Intel Neural Compute Stick 2	MobileNetV2, MobileNetV1	Tempo de Inferência, consumo energético, acurácia	Detecção de Objetos, Classificação de imagens
Lujic et al. (2021)	Raspberry Pi 4, Raspberry Pi 3B+	Google Coral USB	MobileNet SSD V1 e V2	Acurácia, Latência, Tempo de inferência	Detecção de Objetos
Dinh et al. (2021)	Google Coral Devboard, Jetson NANO	–	MobileNet V1, MobileDet, MobileNet V2, YOLOv4, TinyYOLOv4, TinyYOLOv5, Inception	Tamanho de Modelo, Acurácia, Quadros por segundo, uso de memória, uso de CPU	Detecção de Objetos
Petrosino et al. (2021)	Google Coral Devboard, NVIDIA Jetson NANO, Raspberry Pi, MacBook Pro	Movidius NCS, Intel Neural Compute Stick 2, Google Coral USB	MobileNet V1, MobileNet V2	Acurácia, Tempo de Inferência, Temperatura do dispositivo	Classificação de Imagens
Feng et al. (2022)	Raspberry Pi 4B, Jetson NANO, Jetson Xavier NX	Intel Neural Compute Stick 2	YOLOV4, YOLOv4-tiny, YOLOV3, YOLOv3-tiny	Acurácia, FPS, Uso da CPU, Uso de Memória, Consumo energético	Detecção de Objetos

Fonte: Autor.

O uso de acelerador para dispositivos embarcados na borda da rede, também é demonstrado no trabalho de Ward et al. (2021). Neste trabalho, o autor apresenta um sistema para observação de sintomas que possam indicar influenza através da detecção

de tosse utilizando imagens térmicas dados de áudio relacionados a tosse. O *Intel Neural Compute Stick 2* é utilizado como acelerador de um modelo de aprendizagem profunda desenvolvido para detecção de tosse. Os autores utilizam como câmera térmica a *Seek Thermal Compact Pro* que possui uma resolução especial de 320x240 pixels. O *Raspberry Pi* é utilizado como servidor na borda da rede e, subsequentemente, o modelo desenvolvido pelos autores para detectar tosse obteve uma acurácia de 95,59%. Contudo, o trabalho dos autores não leva em consideração as métricas relacionadas a borda da rede como consumo energético, tamanho do modelo e tempo de inferência.

O trabalho de Kristiani, Yang e Huang (2020) utiliza a VPU NCS2[11] para a classificação de imagens, obtendo a melhor acurácia através da rede VGG16[12].

Em concordância com o supracitado, a aceleração de hardware na borda da rede permite o aumento na taxa de quadros dos sensores de imagem, a redução da latência em decorrência da compressão de modelos e redução da carga computacional durante a etapa de classificação. O trabalho de Puchtler e Peinl (2020) evidencia que o Raspberry Pi 4 com o acelerador Coral USB obteve maiores quantidades de quadros por segundo e menor consumo energético para detecção de objetos em comparação a mesmo computador embarcado utilizando o Intel NCS2 e o NVIDIA Jetson Nano.

Em concordância com o trabalho supracitado, Jainuddin et al. (2020) realizaram uma comparação entre os dispositivos Raspberry Pi 3B+, Raspberry Pi 4B e o Intel NUC acelerados pelo coprocessador Edge TPU Coral USB. Os experimentos dos autores demonstram que o Raspberry Pi 4B de 4GB de RAM possui um desempenho equivalente ou superior a CPU do Intel NUC, quando acelerado pelo Edge TPU. Para a rede Inception V4 que possui 43 milhões de parâmetros, o Raspberry Pi 4B obteve um tempo de inferência de 134,33ms enquanto o intel NUC alcançou o valor médio de 95,15ms. Entretanto, o acelerador Coral USB permite a possibilidade de se realizar o overclock para alcançar e maximizar a frequência de processamento do dispositivo. Acrescenta-se ainda que os experimentos realizados foram 2, sendo estes provenientes de imagens obtidas por uma câmera USB e imagens carregadas a partir do disco. Contudo, não foi considerado o tempo de carregamento inicial dos modelos de RNC para se fazer a inferência.

A avaliação entre dispositivos embarcados considerando microcontroladores como o Arduino Nano 33 BLE é realizada no trabalho de Baller et al. (2021). Os autores realizaram uma comparação entre o desempenho dos dispositivos de borda da rede para a tarefa de classificação de imagens e detecção de objetos utilizando a MobileNetV1 e MobileNetV2. Subsequentemente, os autores demonstram que o uso da TPU no microcomputador *Google Coral Devboard* teve o melhor desempenho em tempo de inferência e consumo energético. O trabalho dos autores contribui com o estado-da-arte demonstrando a eficiência energética

[11] https://www.intel.com/content/www/us/en/developer/tools/neural-compute-stick/overview.html
[12] https://keras.io/api/applications/vgg/

do *Devobard*.

O trabalho de (LUJIC et al., 2021) emprega uma abordagem centralizada de servidor para a detecção de trafego no trânsito utilizando o Raspberry Pi como servidor de *edge computing* e os modelos de CNN MobileNet SSD v1 e MobileNet SSDv2, acelerados pelo Coral USB, onde os autores empregam o protocolo MQTT que utiliza o protocolo TCP/IP para transmissão de dados e oferece 3 níveis diferentes de qualidade de serviço, sendo um protocolo leve para aplicações móveis no contexto de IoT. Os autores ainda realizam uma comparação entre a nuvem e o servidor na borda da rede em diferentes camadas de QoS utilizando conexões 3G, 4G e 5G. Neste trabalho é observado que a latência é reduzida em 50% para todos os modelos empregados, obtendo resultados de latência entre 700ms e 1000ms para detecção de tráfego e acurácia aproximada de 89%. Entretanto, a heterogeneidade dos dispositivos móveis e suas respectivas capacidades de conexão podem gerar latências maiores.

Não obstante, a aceleração de hardware também possibilita a redução do tempo de inferência de um modelo de RNC, sendo compatível com aplicações que demandam processamento e classificação em tempo real de imagens. Em (DINH et al., 2021), uma contagem de veículos é apresentada através da detecção de objetos. Neste trabalho, os computadores de placa única como NVIDIA Jetson Nano e o Coral Devboard são utilizados para a inferência de modelos de DNN. Os autores demonstraram que a compressão de modelos e a aceleração de hardware através do uso da TPU do Coral DevBoard permite que seja possível obter uma acurácia considerável e um tempo reduzido para inferência de modelos, obtendo 92,1% de acurácia na detecção de veículos empregando a MobileDet SSD e um tempo médio de inferência de 37,31ms.

O estudo de Feng et al. (2022) realiza um *benchmarking* entre o Raspberry Pi 4B acelerado pelo Intel NCS2, Jetson Nano da NVIDIA e o Jetson Xavier NX para uma tarefa de detecção de objetos empregando as redes YOLOv3, YOLOv3-tiny, YoloV4, YoloV4-tiny. A avaliação é realizada utilizando quadros de 2 videos diferentes, e subsequentemente, o resultado apresentado demonstra que o Raspberry Pi 4B acelerado pelo NCS2 obteve a melhor acurácia para as RNCs empregadas, exceto a YoloV3-tiny onde não houve compatibilidade do acelerador com a rede, e consequentemente, os autores obtiveram 0% de acurácia. Entretanto, o Raspberry Pi obteve os melhores resultados de acurácia consistentemente no trabalho, sendo a maior acurácia obtida pela YoloV4 (94%) e a segunda melhor quantidade de quadros por segundo e menor consumo energético quando comparado aos demais dispositivos considerados neste trabalho.

Por fim, os trabalhos supracitados demonstram que Raspberry Pi 4B tem obtido os melhores resultados quando acelerado pelo Intel NCS2 ou o Edge TPU. Contudo, este microcomputador embarcado pode utilizar até 8GB de memória RAM[13]. Sob este contexto,

[13] https://www.raspberrypi.com/products/raspberry-pi-4-model-b/specifications/

é possível perceber que as avaliações de desempenho do Raspberry Pi utilizando o máximo de memória RAM não é considerada pelos trabalhos apresentados. Em adição, o dispositivo Coral USB (Edge TPU) possui a opção de aumentar a frequência do *clock* do dispositivo, o que também não é considerado durante a comparação entre os dispositivos. Desta forma, este projeto de mestrado também propõe uma avaliação considerando estas características de *hardware* utilizando o modelo de classificação de embriaguez proposto.

3 Implementação, Materiais e Metodologia

3.1 Implementação

Este capítulo apresenta a implementação do aplicativo desenvolvido durante o projeto de mestrado, assim como os materiais e métodos utilizados para o desenvolvimento da solução proposta para a validação da hipótese de se classificar o estado de embriaguez do indivíduo em tempo real na borda da rede. Em vista disso, as funcionalidades do aplicativo serão descritos na subseção 3.1.1, e em seguida, os materiais utilizados para a captura das imagens e utilizados durante os experimentos serão detalhados na seção 3.2. Enquanto que, a estratégia metodológica para a obtenção dos resultados relacionados a captura de imagens e estimação de dosagem de álcool para os participantes é tratado na seção 3.3.

3.1.1 Aplicativo móvel

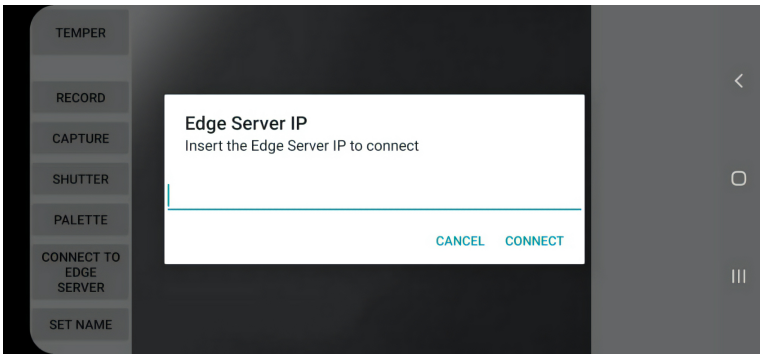
Um aplicativo móvel para o uso da câmera térmica é disponibilizado pela Infiray e pode ser obtido através da Play Store. A versão utilizada é para Android 10 ou acima. O aplicativo é executado após do sistema *android* do celular detectar a conexão da câmera térmica realizada via USB. Em adição, o aplicativo realiza a correção de não uniformidades em dois pontos, assim que o dispositivo é conectado para a remoção de ruídos na imagem. Conquanto, o aplicativo não possui o envio de imagens ou vídeo através de conexão com a internet. Em adição, no aplicativo original disponibilizado pela empresa, é possível analisar somente a temperatura de até 9 pixels em tela e, também, as imagens e vídeos capturados são redimensionados para o tamanho da câmera nativa do dispositivo celular onde o imageador está conectado. Esta interpolação na imagem adiciona ruído e desta forma, um novo aplicativo é desenvolvido para o presente trabalho.

Figura 10 – Tela principal do aplicativo densenvolvido.

Fonte: Autor.

O aplicativo desenvolvido durante o presente trabalho inclui as funções originais do aplicativo disponibilizado em loja. Isto, Todavia, um método de envio dos frames capturados através de conexão de internet utilizando o protocolo UDP é empregado para realizar o envio dos quadros capturados em tempo real para o servidor na borda da rede. Salienta-se ainda que, um dos produtos do presente projeto de mestrado é o aplicativo supracitado. Este aplicativo tem compatibilidade com os dispositivos da linha T3 da Infiray.

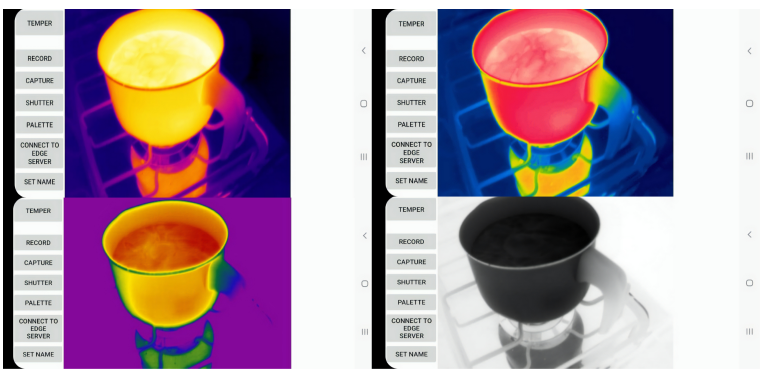
Figura 11 – Menu de alerta utilizado para entrada do endereço de IP do servidor na borda da rede.



Fonte: Autor.

Em adição, a função de alterar o mapa de cores da imagem também é adicionada para avaliar a diferença entre a distribuição de cores conforme mostrado na Figura 12. Contudo, o mapa de cores escolhido para o presente projeto é a *White-hot* que representa as imagens escalas de cinzas, em decorrência de seu reduzido tamanho em memória de armazenamento. Esta cor é escolhida levando em consideração o envio de imagem através do protocolo UDP e o tempo computacional de processamento da rede neural convolucional na borda da rede.

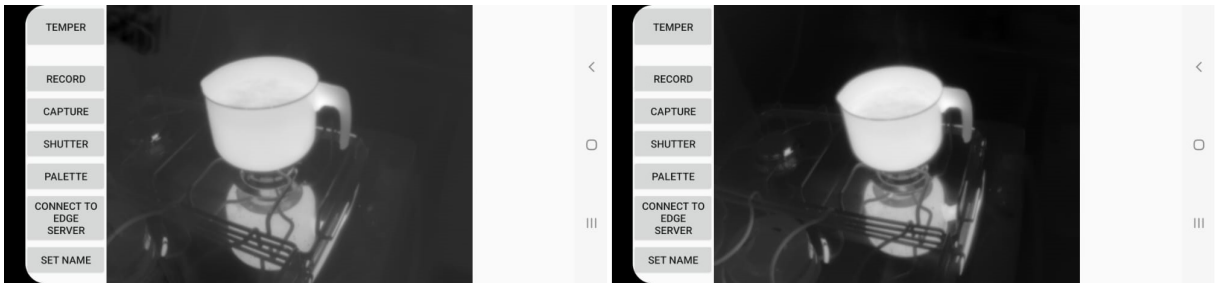
Figura 12 – Diferentes paletas de pseudo-cores, sendo estas cores: *Iron Rainbow*, *Rainbow*, *Rainbow HC* e *Black Hot* respectivamente.



Fonte: Autor.

A Figura 12, apresenta as paletas de cores implementadas no aplicativo e que pode ser alterada ao pressionar o botão *palette* em tela. Não obstante, para a correção de ruídos e ajuste de temperatura pelo imageador, também foi adicionada a função de realizar a correção de não uniformidades. Esta função é disponibilizada através do uso do botão de *shutter*. O imageador automaticamente executa a correção de uniformidades utilizando 2 pontos de temperaturas distintas em tela para calibrar a imagem do sensor. Entretanto, é necessário que essa função seja implementada para que antes de cada captura seja possível realizar a calibração manualmente, evitando ruídos durante a captura das imagens de rosto.

Figura 13 – Antes e após a execução do algoritmo de correção de não uniformidades.



Fonte:Autor.

Conforme mostrado na Figura 13, o botão de correção de não uniformidades (*shutter*) realiza um ajuste que reduz o erro de temperatura em decorrência do aquecimento dos componentes internos do imageador. Este ajuste melhora a representatividade das imagens ajustando o ganho e o *offset* para cada pixel na imagem. Ao analisar a imagem anterior, percebe-se que a da esquerda possui um fundo cinza, enquanto que a da direita possui um fundo preto limpo. A tonalidade cinza ao fundo da imagem da esquerda ocorre em decorrência da diferença de temperatura entre o ambiente e o sensor da imagem, fazendo com que seja necessário a correção não uniforme da imagem para corrigir esta diferença de temperatura.

Nesta seção será descrita a implementação do método de captura da imagem, envio e classificação. Subsequentemente, serão descritos os materiais utilizados para validar a hipótese proposta por este projeto de mestrado, bem como os resultados obtidos. Desta forma, a câmera térmica Infiray Xtherm T3C será conectada a um dispositivo celular móvel que estará conectado via protocolo de transmissão UDP para envio das imagens em tempo real ao Raspberry Pi 4B, que será empregado como servidor na borda da rede.

O *framework* proposto utilizou como linguagem de programação para o treinamento do modelo e inferência na borda da rede a linguagem Python. No ambiente do *Google Colab*, foram utilizados os seguintes pacotes:

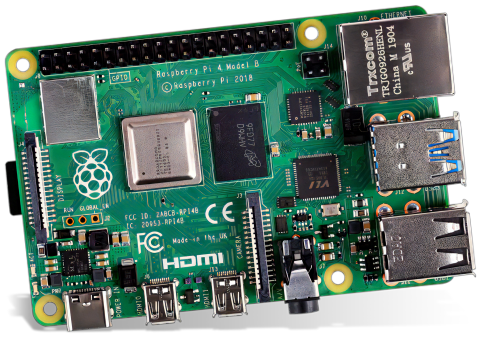
- Keras: biblioteca de fonte aberta que permite utilizar a linguagem de programação Python para o treinamento de redes neurais convolucionais[1].
- Matplotlib: biblioteca utilizada para visualização gráfica de dados e informações para a linguagem de programação Python.[2]
- Numpy: Pacote para computação científica utilizando a linguagem Python. Permite realizar operações vetoriais, matriciais e lógicas[3].
- OpenCV: Biblioteca que permite realizar processamento de imagens e tarefas relacionadas a visãocomputacional[4].
- Scikit-learn: Biblioteca que permite implementar técnicas de aprendizagem de máquina como classificação, regressão, seleção de modelos, redução de dimensionalidade e pré-processamento[5].

3.2 Materiais

3.2.1 Raspberry Pi 4B

No presente trabalho, é utilizado como servidor para computação na borda da rede um Raspberry Pi modelo 4B com 8 GB de memória RAM. O sistema operacional utilizado é o Raspbian Bullseye de 32 e 64 bits[6]. Um dissipador do tipo armadura de alumínio preto na região da superfície possui integrado 2 *coolers* e é acoplado ao servidor para resfriamento.

Figura 14 – Raspberry Pi 4B utilizado como servidor de Edge Computing.



Fonte: Autor.

[1] https://keras.io/
[2] https://matplotlib.org/
[3] https://numpy.org/doc/stable/user/whatisnumpy.html
[4] https://docs.opencv.org/4.x/d0/de3/tutorial$_p y_i ntro.html$
[5] https://scikit-learn.org/stable/
[6] https://www.debian.org/releases/stable/amd64/release-notes/

O modelo do Raspberry Pi apresentado na Figura 14, possui 2 portas USB 2.0 e 2 portas USB 3.0. O processador utilizado é Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit e 1.5GHz de frequência. Em adição, o dispositivo possibilita a conexão via wi-fi nas frequências de 2.4 GHz e 5.0 GHz com protocolo IEEE 802.11ac wireless e Bluetooth 5.0.

3.2.2 Etilômetro

O modelo do alcoolímetro utilizado é o NBAC-03. Este dispositivo utiliza um semicondutor como sensor de gás e um circuito microprocessado para realizar as aferições de níveis da taxa de álcool por ar exalado. O dispositivo adquirido tem 5 bocais em um compartimento traseiro para uso inicial. Entretanto, uma quantidade reserva com 50 bocais também foi adquirida contemplando a avaliação de todos os participantes.

Para realizar aferições é necessário esperar um tempo equivalente a 15 segundos de aquecimento e, subsequentemente, o indivíduo deve assoprar o bocal encaixado ao etilômetro durante 10 segundos. O valor detectado pelo sensor pode ser observado em miligramas por litro, conteúdo de álcool no sangue ou a porcentagem de taxa de álcool no sangue.

Figura 15 – Modelo de etilômetro utilizado para aferição do BrAC nos participantes.

Fonte: Autor.

O instrumento possui uma faixa de medição em porcentagem de CAS de 0,000 a 0,1999% BAC com uma resolução de 0,001%. A sua faixa de medição em miligramas por litro é entre 0,000 a 1,999mg/L. Em adição, o dispositivo possui um certificado de calibração rastreável onde os protocolos de calibração são realizados por uma empresa seguindo as recomendações do INMETRO. A calibração apresentada é dada por 3 aferições onde são calculados os valores médios de cada aferição e uma taxa de erro de calibração de 0,045mg/L por ar alveolar exalado.

3.2.3 Questionário AUDIT

O AUDIT foi desenvolvido para avaliar o consumo excessivo de álcool e, em particular, para ajudar a identificar as pessoas que se beneficiariam com a redução do consumo do mesmo. Os indivíduos que costumam consumir o álcool de forma excessiva, frequentemente apresentam sintomas ou problemas que normalmente não estariam ligados ao seu hábito de beber. (LAWFORD et al., 2012). Desta forma, o AUDIT pode ser empregado em forma de entrevista ou questionário. Neste trabalho, esta ferramenta foi empregada em forma de questionário.

Este instrumento possui 10 perguntas, com 4 itens para cada pergunta, onde cada item perfaz uma pontuação de 0 a 4 pontos, sendo o máximo 40 pontos. A somatória dos pontos para cada pergunta que o indíviduo obtém após responder aos itens do AUDIT permite a classificação do uso da substância da seguinte forma: baixo risco de 0 a 7 pontos; uso de risco de 8 a 15 pontos; uso nocivo de 16 a 19 pontos; provável dependência 20 a 40 pontos (MORETTI-PIRES; CORRADI-WEBSTER, 2011).

Os participantes que obtiveram pontuação acima de 16 pontos foram desconsiderados durante a experimentação em decorrência do risco do indivíduo desenvolver um hábito nocivo no consumo de bebidas álcoolicas. Em adição, foi solicitado aos voluntários que dispusessem de acompanhantes para a condução segura dos mesmos. Da mesma forma, foi disponibilizado um veículo de transporte caso a pessoa não dispusesse de um acompanhante ou condução. Acrescenta-se ainda que os participantes foram questionados sobre seu estado de saúde para verificar se passaram bem, 24 horas após o experimento.

3.2.4 Infiray Xtherm T3C

Para a aquisição das imagens térmicas dos participantes durante os experimentos, é utilizado o imageador térmico da Xtherm, o Infiray T3C[7]. Conforme exibido na Figura 16.

Figura 16 – Imageador Térmico Infiray T3C.

Fonte: Infiray[8].

[7] https://pt.aliexpress.com/i/1005002447765285.html

Este dispositivo pode ser encaixado através de uma entrada USB-C macho/fêmea em um dispositivo móvel com o sistema operacional Android. A versão requirida é Android 10 ou acima. Igualmente, este dispositivo tem compatibilidade com sistemas que possuam *drivers* Video4Linux2, sendo possível obter a partir do imageador imagens no formato YUV420. Em adição, este dispositivo captura imagens com 16 bits, utilizando apenas 14 dos 16 bits de informações. As especificações técnicas de *hardware* deste dispositivo são descritos abaixo:

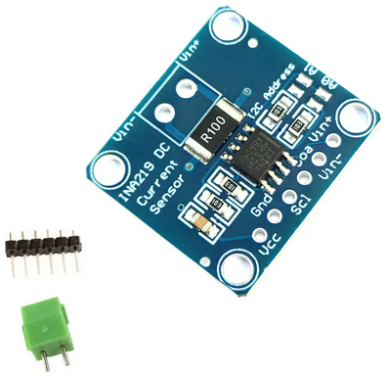
* Resolução espacial: 384x288.
* Taxa de quadros: 25Hz.
* NEDT: $\leq$40mK em temperatura ambiente de 25ºC
* Pixel Pitch: 17um
* Modo de foco: Manual
* Faixa Espectral: 8 14um

3.2.5 Módulo INA219

O módulo INA219 permite medir a tensão e a corrente através da ligação em shunt com um resistor do dispositivo com alguma carga de alimentação. O módulo possui barramento l2c com uma *breakout board*.

A Figura 17 demonstra o componente utilizado. Este dispositivo pode medir a tensão em bus de até 26V e uma corrente de até 3,2 Amperes. Para a medição um resistor em shunt e uma bateria de 5V foi utilizada para medir o consumo energético do Raspberry Pi 4B durante a classificação de imagens.

Figura 17 – INA219 l2c.

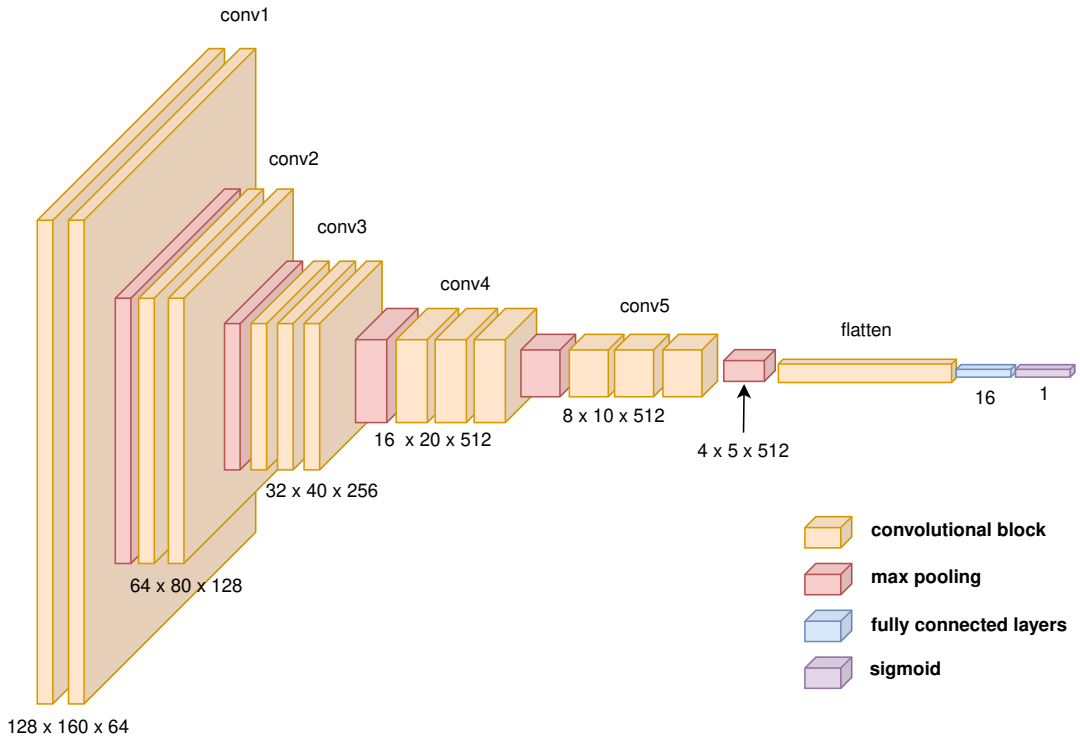


Fonte: Autor.

3.2.6 Modelo de Deep Learning

Para a implementação de um modelo de rede neural convolucional na borda da rede, uma abordagem para o desenvolvimento do classificador apresentado por Silva (2021) é utilizado como base, em virtude da capacidade do algoritmo de identificar respostas fisiológicas com base nas características faciais extraídas. O classificador supracitado foi treinado utilizando como base de dados as imagens capturadas por Koukiou e Anastassopoulos (2015), de resolução espacial de 128x160, conforme apresentado na seção de trabalhos relacionados do Capítulo 2.

Figura 18 – Modelo para classificação de embriaguez treinado a partir da VGG16 por Silva (2021).



Fonte: Autor.

Inicialmente, para a etapa de treinamento do classificador proposto foi conduzida uma seleção de quadros onde as imagens capturadas foram balanceadas em 100 quadros por pessoa. Este conjunto foi subdividido entre 50 imagens para a classe de pessoas em estado de sobriedade e ebriedade, respectivamente. Consequentemente, observou-se experimentalmente uma melhoria no desempenho do classificador ao utilizar o primeiro e segundo conjunto amostral para cada pessoa. Não obstante, as aquisições do estado sóbrio e ébrio de 3 pessoas são subtraídas do conjunto total e particionadas de forma pré-definida para o conjunto de teste.

Em adição, a normalização do conjunto de imagens entre os valores 0 e 1 é necessária em conformidade com os valores de intensidade dos pixels das imagens durante a etapa de treinamento no classificador proposto por Silva (2021). Consequentemente, esta técnica é empregada durante a transferência de aprendizagem, e posteriormente, as imagens do conjunto remanescente de 9 pessoas são particionadas em treino e validação através do emprego da técnica de validação cruzada em k-folds estratificada.

Subsequentemente, são obtidos 5 conjuntos para a técnica de *k-fold cross validation*. Esta estratégia particiona de forma estratificada um número k de subconjuntos onde cada partição de validação possui n/k elementos, em que n representa o volume de dados disponível e K o número de partições (SILVA, 2021).

A técnica de validação cruzada é empregada com o objetivo de estimar a acurácia média entre cada partição e reduzir a variância da acurácia em diferentes conjuntos de treino, validação e teste (WONG; YEH, 2020). A implementação da validação cruzada em k-folds pode reduzir o viés e o erro de predições do classificador conforme maior for o número de K (SAEZ; ROMERO-BéJAR, 2022).

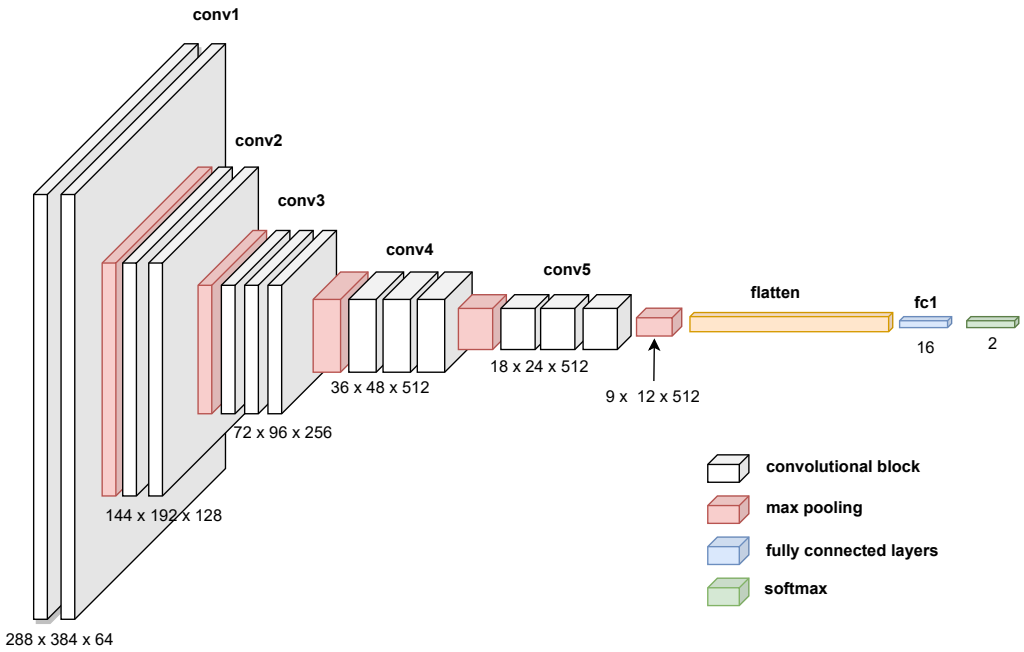
Em seguida, um classificador base é empregado com a técnica de validação cruzada estratificada descrita acima para realizar a extração e obter mapas de atributos, respectivamente. O uso de extração de atributos e transferência de aprendizagem permite uma melhor transferibilidade e generalização do conhecimento adquirido pela rede (WENBO; ZHANG; ROMAGNOLI, 2020).

Acrescenta-se ainda que esta técnica também reduz o desvio padrão, além de melhorar a robustez para o treinamento de um modelo em que o domínio dos dados são diferentes (PAHAR et al., 2022). Sob o contexto do classificador, uma camada de entrada, uma camada totalmente conectada de 16 neurônios e um classificador *softmax* de 2 classes são adicionados ao modelo respectivamente. O classificador *softmax* foi escolhido devido a sua compatibilidade mutual entre ambos os aceleradores (Coral USB e Intel NCS2).

Além disso, a alteração da camada de classificação que utiliza a função *sigmoid* para uma que utiliza a função de *softmax* do modelo proposto foi realizada em decorrência da incompatibilidade do acelerador *Coral USB* com a função de sigmoid. Subsequentemente, o modelo é utilizado para a etapa seguinte de *fine-tuning*. No trabalho de Silva (2021) as camadas convolucionais são descongeladas durante a etapa de *fine-tuning* após a transferência de aprendizagem.

Todavia, em decorrência das características extraídas pelo modelo base e modelo proposto durante a etapa de treinamento, para a etapa de *fine-tuning*, as técnicas de compressão, otimização e conversão são utilizadas sem descongelar as camadas convolucionais. A Figura 19 demonstra o modelo proposto que foi obtido seguindo a abordagem empregada por (SILVA, 2021) para a tarefa de classificação de embriaguez.

Figura 19 – Modelo de rede neural convolucional proposto após a extração de atributos e a transferência de aprendizagem.



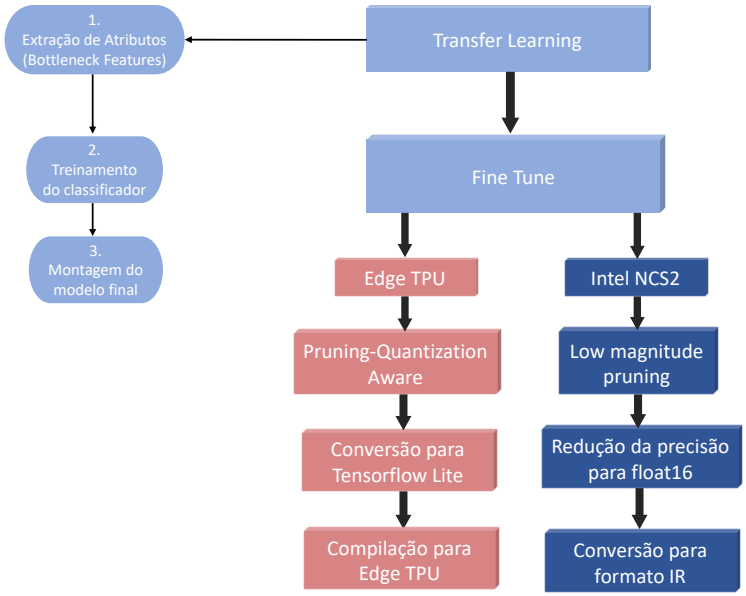
Fonte: Autor.

Nesta etapa, o modelo é retreinado por mais 1 *epoch* utilizando todo o conjunto de imagens de treino para calibrar o modelo e reduzir o *overfitting* durante as etapas de otimização. O retreinamento se deve ao espaço ainda existente para convergência do modelo ao seu ótimo local. Entretanto, para a inferência de um modelo RNC em conjunto com algum acelerador de *hardware*, é necessário que o modelo seja comprimido e otimizado para sua respectiva plataforma.

Conforme a Figura 20, o *pruning* do modelo é realizado durante a etapa de *fine-tuning*, onde a técnica de pruning empregada é baseada em magnitude dos pesos e possibilita a redução de um número significativo de parâmetros das camadas totalmente conectadas e pode reduzir consideravelmente os custos computacionais nas camadas convolucionais devido à esparsidade irregular na topologia da rede podada (LI et al., 2016; LI et al., 2022).

Esta técnica é utilizada similarmente durante a etapa de conversão do modelo para ambos os aceleradores utilizados neste trabalho antes da otimização para seus respectivos aceleradores após a etapa de transferência de aprendizado a partir do modelo base.

Figura 20 – Fluxo de treinamento do modelo proposto: Transfer Learning e Fine-tuning.



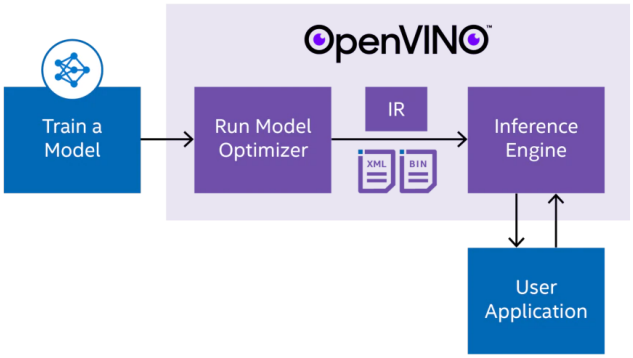
Fonte: Autor.

Contudo, para o Edge TPU, a quantização ciente de treinamento é utilizada para realizar a compressão do modelo para pontos fixos de 8 bits. Salienta-se ainda que o emprego da técnica de quantização do modelo após o *pruning* pode eliminar a esparsidade do modelo durante a etapa de *fine-tuning*. Em vista disso, é necessário empregar a técnica de *pruning quantization aware training* (treinamento consciente de quantização com preservação de esparsidade)[9]. Esta técnica permite que a esparsidade obtida durante a podação dos pesos seja preservada durante a quantização do modelo, possibilitando reduzir consideravelmente o tamanho do modelo e otimizar a sua convergência em decorrência do *fine-tuning* (GIL et al., 2022).

Paralelamente, para o Intel NCS2 a técnica de podação foi empregada sem as quantizações dos pesos, em decorrência da VPU não ter compatibilidade com a técnica de quantização de 8 bits. Desta forma, o modelo em formato keras é utilizado na ferramenta de otimização da Intel para redução da precisão para float16 e, em seguida, o modelo é convertido para formato de representação intermediária (IR) conforme a Figura 21.

[9] https://www.tensorflow.org/model$_o$*ptimization*/*guide*/*combine*/*pqat*$_e$*xample*

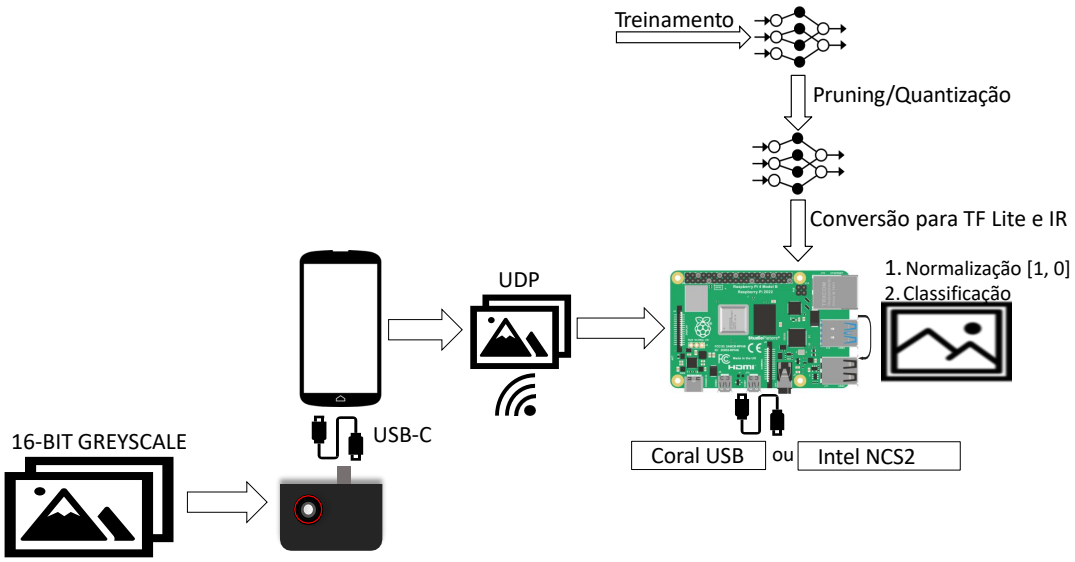
Figura 21 – Conversão de modelos para Representação Intermediaria.



Fonte: Site da Intel[10].

Para o envio das imagens a partir do dispositivo móvel ao servidor Raspberry Pi, é utilizado o protocolo de comunicação UDP/IP. A imagem é codificada para o formato *bytes* e enviada em pacotes onde, para cada pacote recebido, a imagem é montada até que todos os pacotes sejam enviados. Após o envio de todos os pacotes, a imagem é remontada no servidor na borda da rede e classificada pelo algoritmo de RNC, sendo o resultado obtido enviado para o dispositivo móvel conectado ao servidor, conforme mostrado na Figura 22.

Figura 22 – Framework proposto.



Fonte: Autor.

3.2.7 Aceleradores de Hardware

Neste projeto de mestrado são utilizados para aceleração de hardware 2 dispositivos: Google Coral Edge TPU e o Intel Neural Compute Stick 2. Estes dispositivos são utilizados para realizar a classificação de imagens térmicas na borda da rede. O modelo de RNC treinado deve ser convertido para formatos adequados para cada acelerador de hardware, descritos abaixo.

3.2.7.1 Google Coral Edge TPU

O acelerador Google Coral USB é um dispositivo USB que possui unidades de processamento de tensores que têm a capacidade de acelerar as operações de forward passes em modelos de deep learning. Todavia, é necessário que os modelos implementados estejam quantizados em pontos fixos de 8 bits e em formato Tensorflow Lite.

Figura 23 – Google Coral Edge TPU USB.

Fonte: Mouser Electronics[11].

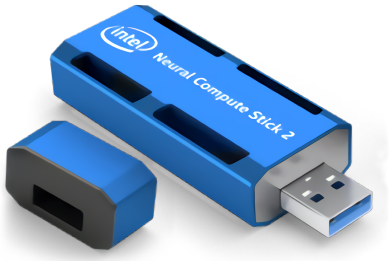
Este acelerador será utilizado no Raspberry Pi para realizar a inferência do modelo implementado. A porta utilizada será a USB 3.0, para incrementar a vazão de dados e reduzir o tempo de inferência.

3.2.7.2 Intel Neural Compute Stick 2

O Neural Compute Stick 2 é uma unidade de processamento visual (VPU) que permite a aceleração de hardware e requer o uso do kit de ferramentas OpenVino da Intel para conversão e otimização do modelo. É utilizado para conversão para uma representação intermediária (IR). Desta forma, para o uso do Intel NCS2 é necessário que as camadas do modelo sejam otimizadas para float16 e convertidas para formato IR, que consiste de 3 arquivos separados. O primeiro arquivo possui a topologia da rede convertida para o formato da VPU e possui o formato .xml, o segundo arquivo possui formato binário e

carrega os pesos da rede, e por fim, o terceiro arquivo é do tipo .mapping e contém a estrutura original do modelo. Logo, os arquivos .xml e .bin são utilizados para a inferência.

Figura 24 – Intel Neural Compute Stick 2.



Fonte: Site da Intel[12].

3.3 Metodologias

Nesta seção, é descrito o protocolo experimental empregado para avaliar a solução proposta. Inicialmente, será descrito o procedimento experimental para a seleção de participantes, o cálculo para dosagem de bebidas e, subsequentemente, o procedimento relacionado ao imageador térmico para captura de imagens dos voluntários. Em seguida, será realizada uma avaliação quantitativa do conjunto de imagens obtidas durante a sessão de captura dos quadros.

Posteriormente, serão detalhados os resultados do modelo de rede neural convolucional proposto para extrair as características das imagens que representam os sintomas de embriaguez e sua capacidade de generalização, bem como sua capacidade de realizar inferências em um dispositivo embarcado em tempo real, acelerado por coprocessadores. Em seguida, serão apresentados os resultados das métricas selecionadas para avaliação da solução proposta, a fim de classificar a embriaguez na borda da rede como consumo energético, tempo de inferência, acurácia e tamanho do modelo.

3.3.1 Estratégia metodológica para estimação de dosagem para bebidas

Esta pesquisa experimental aborda de forma quantitativa os testes e os resultados obtidos. Inicialmente, um procedimento experimental para captura de imagens é realizado seguindo os protocolos de captura empregados nos trabalhos de Koukiou e Anastassopoulos (2012), Hermosilla et al. (2018) onde as imagens são capturadas entre intervalos de 30 minutos com 100 millisegundos entre cada quadro para possibilitar a metabolização do álcool e sequencialmente a visualização das alterações fisiologicas relacionadas a intoxicação pelo álcool.

Para este experimento a cerveja Heineken é selecionada, contendo 5% de teor alcoólico. Inicialmente, é calculado o total de gramas de álcool por mililitro seguindo a equação demonstrada em (BRICK, 2006):

$$\text{mL} \times \%\text{v/v} \times 0,79 \tag{3.1}$$

Na equação acima, mL representa em mililitros o volume da bebida alcóolica, %v/v representa o teor alcóolico da bebida em concentração por volume e o valor 0,79 é um valor utilizado por pesquisadores que representa a gravidade específica do álcool. Desta forma, para uma cerveja Heineken de 350mL e 5% de teor alcoólico, esta equação é resolvida abaixo:

$$\frac{350 \text{ mL} \times 5\% \times 0,79}{100} \tag{3.2}$$

Através da equação acima, obtém-se o valor 13,82g para cada *container* de bebida da marca Heineken de 350mL.

Posteriormente, para avaliar a sintomatologia dos efeitos fisiológicos do etanol no corpo do participante, é necessário realizar o cálculo da água total no corpo. De acordo com Brick (2006), o álcool é uma substância hidrofílica que se distribui primeiramente aos líquidos corporais como o sangue e sua distribuição no corpo de um indivíduo varia de acordo com altura, idade, peso e gênero. Segundo (KALANT, 2005), a diluição do álcool administrado oralmente tem a mesma concentração da água. Ainda segundo Brick (2006), diversos autores asseveram que o nível de conteúdo de álcool no sangue pode ser estimado de forma acurada, levando em consideração variáveis como absorção, metabolismo e gênero. Em concordância com o supracitado, para homens com idade entre 17 e 86 anos de idade, o cálculo de água total no corpo é expresso por:

$$\begin{aligned} TBW =& 2,44 - (0,09516 \times \text{ idade }) + [0,1074 \\ & \times (\text{ altura em polegadas } \times 2,54)] \\ & + [0,3362 \times \text{ peso em libras } /2,2045)] \end{aligned} \tag{3.3}$$

Igualmente, para mulheres de 17 a 84 anos de idade, o mesmo cálculo de água total no corpo pode ser expresso por:

$$\begin{aligned} TBW =& -2,097 + [0,1069 \times (\text{altura em polegadas} \\ & \times 2,54)] + [0,2466 \times (\text{peso em libras}/2,2045)] \end{aligned} \tag{3.4}$$

Os valores constantes da Equação 3.3 e Equação 3.4, inicialmente foram descobertos experimentalmente por Watson, Watson e Batt (1981) e representam valores utilizados em regressão linear para estimar a quantidade de água total no corpo. Utilizando o cálculo abordado pelos autores, o volume total de água corporal pode ser estimado com uma acurácia razoável($\pm$ 9 a 10% de diferença) em relação a regressão linear utilizando valores de dados antropométricos simples. Para mulheres, a idade não teve resultados significativos

quando levada em consideração. Não obstante, conhecendo o valor de água total no corpo de um determinado indivíduo, é possível estimar o nível de conteúdo de álcool no sangue com base na ingestão de álcool.

$$\mathrm{BAC} = \mathrm{g}/\sum V_{\mathrm{d}} \times Bl_{\mathrm{H_2O}} - [(\beta_{1-n} \times (t_{\mathrm{s}} + t_{\mathrm{p}})] \tag{3.5}$$

Onde g é o valor de gramas de álcool, $\sum V_{\mathrm{d}}$ representa o valor de água total no corpo ou o volume de distribuição de água baseado na idade, peso, altura e gênero (WATSON; WATSON; BATT, 1981; BRICK, 2006); $Bl_{\mathrm{H_2O}} = 80,65$ sendo o valor aproximado de porcentagem de água no sangue; β_{1-n} representa a taxa de eliminação de álcool pelo sangue, onde os valores considerados na literatura variam de 10-20mg/dL/h;t_{S} é o tempo passado entre o ínicio do consumo até a última dose consumida em horas. Este valor é representado de forma decimal. t_{p} é o valor da taxa de absorção de álcool pelo sangue desde a última dose consumida e o valor da concentração máxima de álcool no sangue alcança o pico dentro de 30 a 90 minutos, desde o último consumo.

Conquanto, o uso de dispositivos como o etilômetro para estimação do nível de etanol no sangue de um indivíduo requer um fator de calibração. Este fator é chamado de *blood/breath ratio (BBR)* ou relação sangue/respiração (JONES, 2010). Em vista disso, a taxa de relação sangue/respiração é um valor constante e varia entre indivíduos em função do tempo após o consumo do álcool (JONES; ANDERSSON, 2003). De acordo com Ganert e Bowthorpe (2000) esta razão entre sangue/respiração é menor que 2000:1 em um curto prazo após a ingestão do álcool e, subsquentemente, este valor é incrementado para um razão de 2100:1 até 90 minutos após o consumo. Por conseguinte, faz-se necessário que haja uma conversão do valor do conteúdo de álcool no sangue (BAC) para a taxa de conteúdo no sangue (BrAC), através do produto do resultado do CAS.

Utilizando o valor obtido para a estimação do BAC dos participantes, utilizou-se a conversão de BAC para BrAC, conforme supracitado.

A Tabela 5 apresenta os valores antropométricos de cada participante e o seu valor resultante de água corporal, obtidos a partir da Equação 3.3 e Equação 3.4 em concordância com o gênero de cada indivíduo. Desta maneira, a Equação 3.5apresentada por Brick (2006) pode ser rearranjada para estimar a quantidade em gramas necessária para alcançar um valor aproximado de conteúdo de álcool no sangue. O valor considerado para o cálculo é de 0,34mg/L a cada 100mL de sangue por ar expirado, sendo o valor considerado por lei como crime de trânsito de acordo com o CTB[13].

Consequentemente, a Equação 3.5 pode ser rearranjada da seguinte forma para estimar a quantidade em gramas necessária para um sujeito alcançar um valor desejado de CAS (BRICK, 2006):

$$\mathrm{g} = \mathrm{BAC}_{\mathrm{alvo}} + \left[(\beta_{1-n} \times (t_{\mathrm{s}} + t_{\mathrm{p}})] \times \sum V_{\mathrm{d}}/Bl_{\mathrm{H_2O}} \right. \tag{3.6}$$

[13] https://www.ctbdigital.com.br/artigo/art306

Tabela 4 – Dados antropométricos de cada participante.

Índice	Idade	Sexo	Peso (libras)	Altura (polegadas)	Total de Água Corporal (TBW)
Participante 1	23	Masculino	174	70,4	45,99
Participante 2	22	Feminino	172	70	36,10
Participante 3	22	Masculino	136,7	68,9	40
Participante 4	20	Feminino	116,84	64,17	28,3
Participante 5	22	Masculino	165,3	68,47	43,95
Participante 6	21	Masculino	152,11	67,32	42,00
Participante 7	25	Masculino	154,32	71,25	43,50
Participante 8	26	Feminino	173,06	67,71	44,83
Participante 9	26	Masculino	143,3	69,68	40,82
Participante 10	29	Masculino	275,57	73,22	61,68
Participante 11	23	Masculino	158,733	76	45,19
Participante 12	26	Masculino	169,75	69,68	44,86
Participante 13	29	Masculino	202,82	70,86	49,94
Participante 14	62	Masculino	169,75	69,2	31,68
Participante 15	29	Masculino	202,82	67,71	49,93

Fonte: Autor.

O valor de 0,34mg/L para ser expresso em taxa de conteúdo de álcool no sangue (BrAC) deve ser convertido levando a razão de 2100:1 por ar exalado. Consequentemente, obtém-se que o valor do BAC, alvo a ser empregado na Tabela 3.3.1, seria de 71 mg/dL. Em adição, para o valor de T_s considerou-se o tempo de 1 hora para o consumo das bebidas e 30 minutos após a ingestão da última dose para t_p, em concordância com (KOUKIOU; ANASTASSOPOULOS, 2012).

A Tabela 5 mostra os valores de água corporal total obtidos a partir da Equação 3.3 e Equação 3.4. Utilizando estes valores na é possível estimar quantas gramas de álcool são necessárias para cada participante ingerir, obtendo os valores de BrAC escolhido para verificar as alterações fisiológicas nas regiões faciais através do uso do imageador térmico. Não obstante, os valores de dose a serem consumidas são obtidos dividindo o valor da coluna gramas necessárias pelo valor da equação 3.2 de 13,82g para cada pessoa.

3.3.2 Estratégia metodológica para captura de imagens

Os participantes da captura de imagens foram instruídos a não realizar o consumo de bebidas alcóolicas até 48 horas antes da sessão de captura de imagens, como também foram orientados a não realizar a ingestão de alimentos até 6 horas antes do experimento para evitar que o esvaziamento gástrico seja uma variável durante a aquisição dos quadros.

Tabela 5 – Estimação de doses a serem consumidas com base na água corporal total de cada participante.

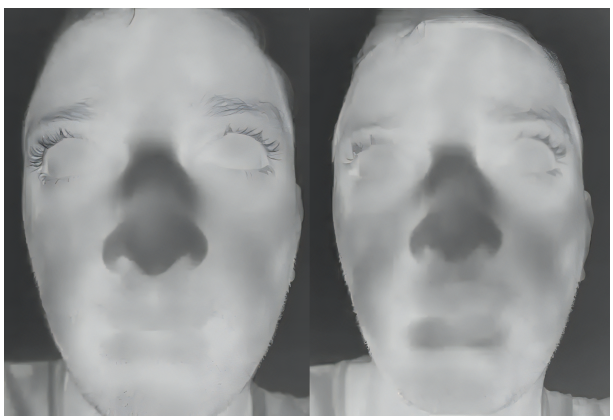
ìndice	**Água Corporal Total**	**Gramas necessárias (g)**	**Doses a serem consumidas**
Participante 1	45,99	53,32	4
Participante 2	36,10	41,85	3
Participante 3	46,35	52,39	3
Participante 4	28,3	47,86	3
Participante 5	32,92	41,85	2
Participante 6	42	50,96	4
Participante 7	42	50,96	4
Participante 8	43,50	48,69	4
Participante 9	44,50	50,44	4
Participante 10	40,82	47,33	3
Participante 11	45,19	71,50	5
Participante 12	44,86	52,39	4
Participante 13	49,94	52,01	4
Participante 14	31,68	51,72	4
Participante 15	49,94	57,89	4

Fonte:Autor.

Acrescenta-se ainda que todos os participantes foram orientados a solicitar acompanhantes ao local de captura e aos que não dispuseram de meios de retorno, foi providenciado transporte de forma segura de volta para seu respectivo domicílio.

A despeito da captura de imagens térmicas discutidas no seção 2.2, para a aquisição de imagens térmicas é necessário levar em consideração a diferença de temperatura ambiente e a temperatura do sensor térmico da câmera durante o seu uso. Em adição, é necessário que haja um tempo de aclimatação dos indivíduos para que a temperatura esteja estabilizada em temperatura ambiente. Em conformidade com a discussão da seção 2.2, experimentalmente é observado que o tempo adequado para que o sensor do imageador se estabilize é de 20 minutos.

Figura 25 – Comparação entre um quadro capturado após o dispositivo térmico ser inicializado e 20 minutos após o dispositivo aquecer.



Fonte: Autor.

Na Figura 25, a imagem a esquerda foi capturada assim que o dispositivo foi inicializado e, posteriormente, a imagem a direita foi capturada 20 minutos após a inicialização do imageador. Sob este contexto, é possível perceber que na primeira imagem, a esquerda, a distribuição do nível de intensidades em escalas de cinza é mais uniforme se comparada a imagem da direita, que possui uma distribuição maior de intensidades de escalas de cinza. Este procedimento é empregado para reduzir a quantidade de ruído em decorrência da diferença de temperatura do ambiente e do sensor (KIRIMTAT et al., 2020).

Ainda sob o mesmo contexto, para evitar o ruído nas imagens, as sessões de capturas foram executadas durante o periodo noturno e em temperatura ambiente de até 25ºC, em virtude da reduzida temperatura com o objetivo de evitar o superaquecimento do sensor ocasionando ruídos adicionais nos quadros capturados. De acordo com Villa e Arteaga-marrero (2020), o microbolômetro das câmeras térmicas depende da temperatura ambiente local. Variações de temperatura em objetos próximos aos sensores, incluindo o case de armazenamento do bolômetro podem modificar o nível de distribuição de temperatura e alteração da irradiação para o plano focal da câmera, influenciando a responsividade do plano focal. Desta forma, a correção não uniforme realizada pelo sensor acaba tendo desempenho melhor para a correção em 2 pontos na imagem. Em adição, seguindo os protocolos de padronização para aferições com imageadores térmicos, a calibração do valor de emissividade é ajustada para 0,98 (JONES; PLASSMANN, 2002). Não obstante, os participantes foram posicionados contra um fundo uniforme com uma temperatura constante de 26,5ºC $\pm 0,5$ de temperatura (FORMENTI et al., 2013).

Logo, a sessão de captura contou com 5 conjuntos de capturas por pessoa e, segundo os protocolos de captura de imagens executados por (KOUKIOU; ANASTASSOPOULOS, 2012), as imagens devem ser capturadas de forma sequencial, com um intervalo de 100 milissegundos para cada quadro. Contudo, neste trabalho as capturas foram realizadas em formato de vídeo (.mp4) com duração de 5 segundos totalizando 125 quadros, e posteriormente, as imagens são extraídas para formato de arquivo TIFF em escalas de cinza. Inclusive, o grupo amostral contou com 15 pessoas e foi dividido em 4 grupos menores.

Tabela 6 – Tabela com as etapas do processo de captura das imagens.

Etapa	Descrição	Periodo
Aclimatação	Tempo para repouso dos participantes e aquecimento do imageador	20 min
Aferição 1	Avaliação do nível de BrAC dos indivíduos	2 min
Captura 1	125 quadros são capturados dos participantes em estado de sobriedade	5 min
Consumo	Os participantes são permitidos realizar a ingestão máxima de 6 latas	1 hora
Repouso	Tempo de espera de metabolização do álcool e para o CAS atingir o nível máximo	20 min
Aferição 2	Avaliação do nível de BrAC do indivíduo	2 min
Captura 2	125 quadros são capturados dos indivíduos em estado de embriaguez	5 mins
Repouso 2	Tempo de espera para queda do CAS de cada indivíduo	30 mins
Aferição 3	Avaliação do nível de BrAC do indivíduo	2 min
Captura 3	125 quadros são capturados conforme a queda do CAS	5 mins
Repouso 3	Tempo de espera para queda do CAS de cada indivíduo	30 mins
Aferição 4	Avaliação do nível de BrAC com o etilômetro	2 min
Captura 4	125 quadros são capturados conforme a queda do CAS	5 mins
Repouso 4	Tempo de espera para queda do CAS cada indivíduo	30 mins
Aferição 5	Avaliação do nível BrAC com o etilômetro	2 min
Captura 5	125 quadros são capturados conforme a queda do CAS	5 mins

Fonte: Autor.

A Tabela 6, contém as etapas da estratégia metodológica empregada para a aquisição das imagens térmicas de cada indivíduo. Inicialmente, o etilômetro é empregado para verificar se há a presença de álcool no ar exalado pelos participantes. Após a avaliação inicial, o primeiro conjunto de imagens é obtido por cada participante antes de qualquer ingestão de álcool e após o período de aclimatação de 20 minutos. Além do conjunto de imagens, também são capturados os valores de temperatura de cada pixel em um quadro.

Os participantes são orientados a permanecer em estado de repouso e distanciados uns dos outros. Consecutivamente ao processo de captura de imagens e análise dos valores de temperatura pixel a pixel, as bebidas são fornecidas aos participantes conforme necessário. Em adição, os participantes são instruídos a consumirem a bebida dentro do período de 1 hora. Desta forma, as bebidas são levadas aos participantes para consumo após o término de cada lata.

Subsequentemente, o tempo de consumo da quantidade total de álcool pelos participantes é considerado de forma individual e, após cada indivíduo ingerir as bebidas,

um intervalo de 30 minutos é acrescentado para o participante em estado de repouso até a próxima avaliação e sessão de captura de imagens. Acrescenta-se ainda que, para todas as sessões de captura de imagens, o dispositivo é aproximado a 15 centímetros da face de cada participante, de forma que a imagem do rosto preencha completamente a tela de captura de acordo com a base de dados disponibilizada por Koukiou e Anastassopoulos (2015).

4 Resultados e Discussões

4.1 Resultados

4.1.1 Considerações Iniciais

Esta seção descreve os resultados experimentais obtidos a partir dos protocolos descritos na seção anterior. Primeiramente serão descritos os resultados obtidos a partir do emprego do etilômetro e serão comparados com os valores estimados a partir da equação de Widmark abordada por Brick (2006). Posteriormente, serão apresentados os resultados do modelo de redes neurais convolucionais obtidos com base no trabalho de Silva (2021). Subsequentemente, são apresentados os resultados em torno das métricas selecionadas para este trabalho relevantes à computação na borda da rede, como acurácia, tempo de inferência, tamanho do modelo, consumo energético.

4.1.2 Resultados experimentais obtidos a partir do etilômetro

A Tabela 7 demonstra os valores aferidos para cada participante através da avaliação no nível de BrAC resultante do emprego do etilômetro.

Tabela 7 – Resultado da aferição do BrAC para cada participante utilizando o etilômetro.

ìndice	**30 minutos**	**60 minutos**	**90 minutos**	**120 minutos**	**BrAC médio**
Participante 1	0,360mg/L	0,290mg/L	0,290mg/L	0,250mg/L	0,300mg/L
Participante 2	0,370mg/L	0,255mg/L	0,155mg/L	0,315mg/L	0,365mg/L
Participante 3	0,430mg/L	0,320mg/L	0,285mg/L	0,320mg/L	0,340mg/L
Participante 4	0,410mg/L	0,390mg/L	0,340mg/L	0,290mg/L	0,360mg/L
Participante 5	0,395mg/L	0,370mg/L	0,385mg/L	0,375mg/L	0,380mg/L
Participante 6	0,380mg/L	0,370mg/L	0,365mg/L	0,350mg/L	0,365mg/L
Participante 7	0,420mg/L	0,210mg/L	0,385mg/L	0,395mg/L	0,350mg/L
Participante 8	0,430mg/L	0,405mg/L	0,335mg/L	0,370mg/L	0,385mg/L
Participante 9	0,345mg/L	0,395mg/L	0,380mg/L	0,370mg/L	0,370mg/L
Participante 10	0,270mg/L	0,260mg/L	0,260mg/L	0,215mg/L	0,250mg/L
Participante 11	0,325mg/L	0,320mg/L	0,410mg/L	0,370mg/L	0,350mg/L
Participante 12	0,390mg/L	0,360mg/L	0,370mg/L	0,370mg/L	0,370mg/L
Participante 13	0,325mg/L	0,295mg/L	0,280mg/L	0,265mg/L	0.350mg/L
Participante 14	0,345mg/L	0,245mg/L	0,240mg/L	0,340mg/L	0,290mg/L
Participante 15	0,395mg/L	0,180mg/L	0,355mg/L	0,345mg/L	0,320mg/L

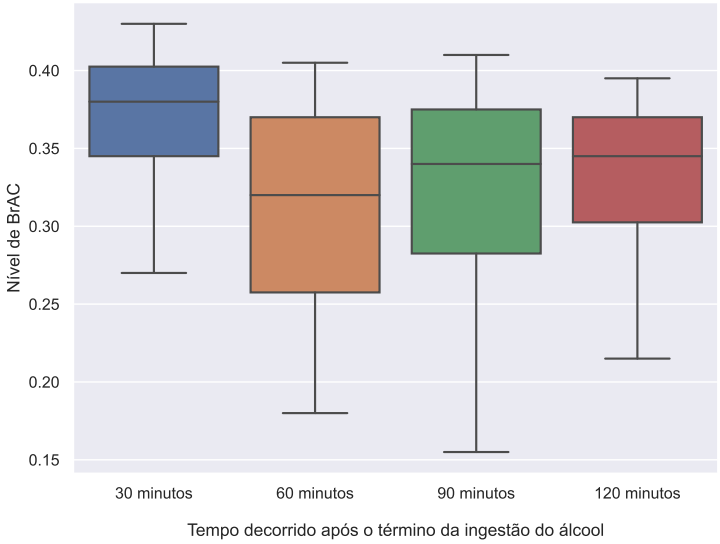
Fonte:Autor.

Os menores valores de BrAC são observados em pessoas com um valor de água total no corpo maior em relação aos demais, como o caso dos participantes do sexo masculinos nos índices 1, 10 e 15. Todavia, percebe-se que os participantes do sexo feminino foram os que obtiveram valores consideravelmente superior de BrAC como o caso dos participantes

2, 4 e 8 em relação aos participantes do sexo masculino considerando o valor médio de BrAC. Salienta-se ainda que, o etilômetro demonstrou ter uma calibração insuficiente para a obtenção dos valores de calibração de forma precisa. Entretanto, percebe-se que os valores de BrAC alcançaram um equilíbrio após 1 hora e 30 minutos e 2 horas que correspondem a quarta e quinta sessão de avaliação respectivamente, quando comparado as primeiras aquisições onde a variância entre as sessões foi maior.

Em adição, é possível perceber a queda do valor de conteúdo de álcool no sangue em todos os participantes exceto o participante número 7. Uma inconsistência na queda do conteúdo de álcool no sangue pode ser observada entre os intervalos de 60 a 90 minutos após o consumo da última dose de bebida. Não obstante, é possível perceber que os valores médios de BrAC para cada participante difere conforme o sexo, peso, altura e idade.

Figura 26 – Boxplot com o nível de BrAC de cada participante durante os 4 periodos de aquisição.



Fonte: Autor.

A Figura 26 apresenta um *boxplot* dos valores de BrAC de todos os participantes, e desta forma, é possível observar que houve uma estabilização na flutuação dos valores de conteúdo de álcool no sangue de cada participante a partir dos 90 minutos.

Logo, o cálculo da fórmula de Widmark leva em consideração que o pico do nível de álcool no sangue ocorrerá em um intervalo de 30 a 90 minutos em virtude da unicidade do metabolismo de cada participante. Consequentemente, não é possível estimar de forma precisa quanto tempo levou para o participante atingir o pico de conteúdo de álcool no sangue.

Tabela 8 – Resultado da estimação do valor de BrAC para cada participante utilizando a fórmula de Widmark.

ìndice	**30 minutos**	**60 minutos**	**90 minutos**	**120 minutos**	**BrAC médio**
Participante 1	0,240mg/L	0,230mg/L	0,220mg/L	0,210mg/L	0,225mg/L
Participante 2	0,345mg/L	0,325mg/L	0,305mg/L	0,280mg/L	0,320mg/L
Participante 3	0,320mg/L	0,310mg/L	0,290mg/L	0,280mg/L	0,300mg/L
Participante 4	0,410mg/L	0,390mg/L	0,370mg/L	0,350mg/L	0,380mg/L
Participante 5	0,290mg/L	0,275mg/L	0,265mg/L	0,255mg/L	0,270mg/L
Participante 6	0,270mg/L	0,260mg/L	0,250mg/L	0,240mg/L	0,255mg/L
Participante 7	0,290mg/L	0,280mg/L	0,270mg/L	0,260mg/L	0,275mg/L
Participante 8	0,350mg/L	0,330mg/L	0,310mg/L	0,290mg/L	0,320mg/L
Participante 9	0,290mg/L	0,275mg/L	0,265mg/L	0,250mg/L	0,270mg/L
Participante 10	0,240mg/L	0,230mg/L	0,220mg/L	0,210mg/L	0,225mg/L
Participante 11	0,280mg/L	0,270mg/L	0,260mg/L	0,250mg/L	0,260mg/L
Participante 12	0,280mg/L	0,270mg/L	0,260mg/L	0,250mg/L	0,270mg/L
Participante 13	0,250mg/L	0,240mg/L	0,230mg/L	0,220mg/L	0,235mg/L
Participante 14	0,320mg/L	0,310mg/L	0,300mg/L	0,290mg/L	0,305mg/L
Participante 15	0,250mg/L	0,240mg/L	0,230mg/L	0,220mg/L	0,235mg/L

Fonte: Autor.

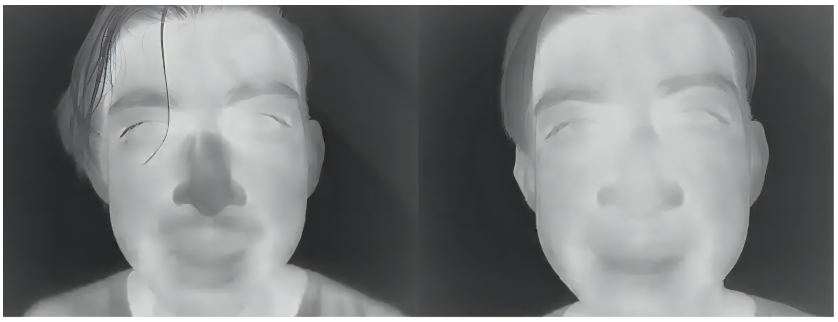
O cálculo de estimação do valor de conteúdo de álcool no sangue, não leva em consideração o metabolismo do invídivuo. Desta forma, seria necessário realizar o mesmo cálculo considerando diferentes taxas de eliminação de álcool e também o tempo de ingestão de álcool na Equação 3.5 por um período maior de observação para avaliar a queda do nível de CAS por participante. Por fim, é possível observar que houve uma correlação entre os valores médios obtidos pela fórmula de Widmark e pelo etilômetro empregado. Neste caso, os valores dos participantes 1, 11, 12 e 13 foram os menores em decorrência dos participantes possuírem peso e altura maiores que os demais.

4.1.3 Resultados das imagens capturadas

As imagens obtidas dos participantes foram decompostas em quadros, sendo 125 quadros por sessão de captura, totalizando 625 quadros por participante.

A Figura 27 demonstra 2 quadros capturados de um dos participantes durante o procedimento experimental descrito na seção anterior enquanto que, a Figura 28, demonstra 6 dos 15 participantes antes e após 30 minutos da última ingestão da dose necessária para alcançar o BrAC alvo. Todas as imagens capturadas durante o experimento foram capturadas em escalas de cinza. Neste caso, para todos os participantes, a primeira imagem foi capturada antes da ingestão de qualquer dose de álcool e 20 minutos após o período de aclimatação.

Figura 27 – Comparação entre 2 quadros do Participante 9: antes e após 30 minutos depois da ingestão de álcool pelo participante.



Fonte: Autor.

Figura 28 – Antes e 30 minutos após o consumo de álcool - subconjunto do conjunto de imagens capturadas durante o procedimento experimento.

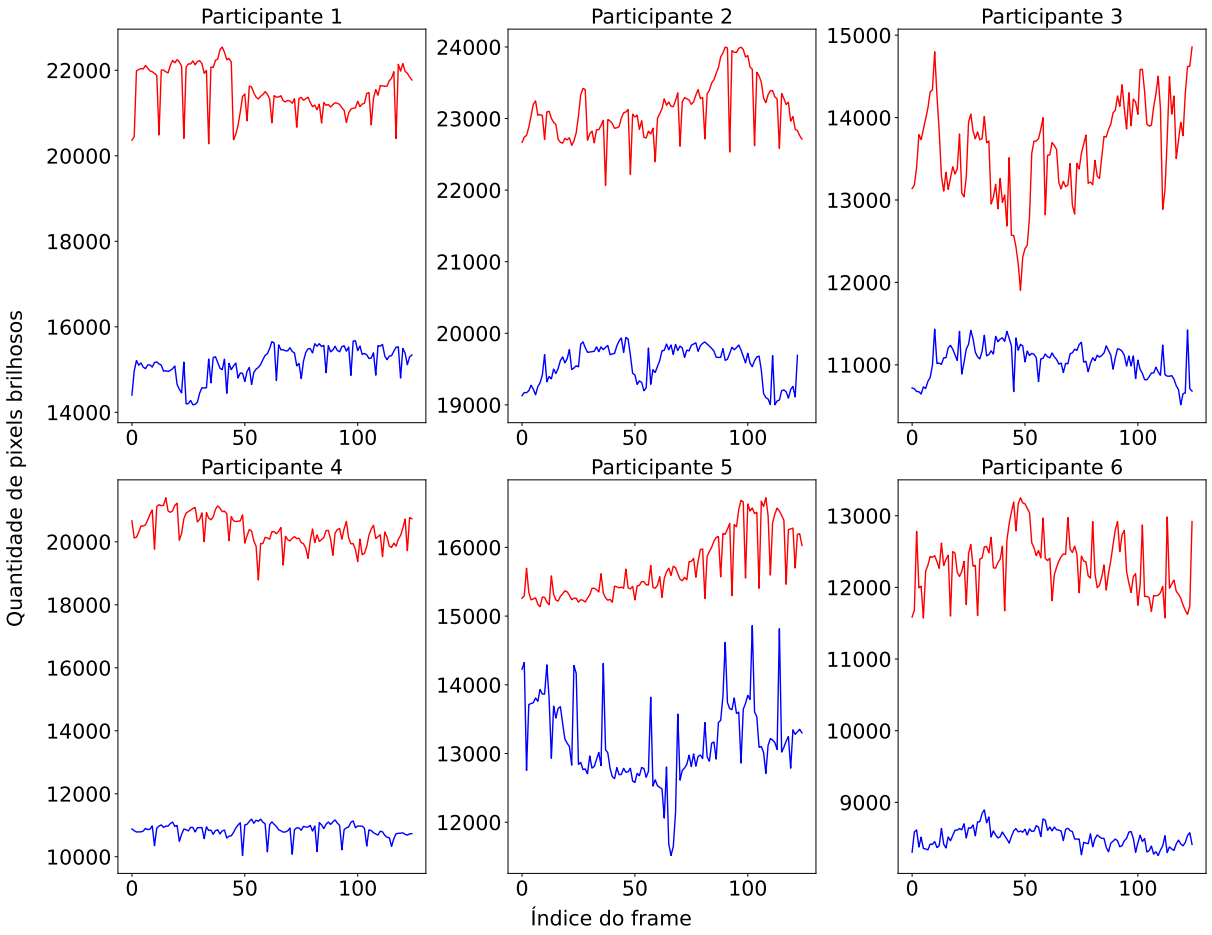
Fonte: Autor.

A segunda imagem para cada participante, apresentada na Figura 28, foi capturada 30 minutos após ingerirem a quantidade necessária para obter um valor próximo de 0.34mg/L de álcool no sangue. Conquanto, para alguns indivíduos a percepção da diferença de temperatura entre as aquisições pode ser subjetiva em decorrência da temperatura na região superficial da pele não ser suficiente para distinguir esta diferença de forma clara.

Uma análise utilizando as imagens capturadas de 6 indivíduos é conduzida com o objetivo de verificar a diferença de intensidade entre os pixels de um sujeito em estado sóbrio e ébrio. Para esta análise experimental foram selecionados 125 quadros da primeira e segunda aquisição respectivamente para os 6 participantes e, posteriormente, são contabilizados todos os pixels em cada quadro onde o valor de intensidade do pixel é acima de 200. O valor de intensidade 200 é considerado de forma arbitrária para representar os pixels mais claros. Desta forma, a Figura 29 demonstra a diferença entre a quantidade de pixels claros nos quadros de uma pessoa antes da ingestão de qualquer dose de álcool e 30 minutos após a ingestão de 4 latas de cervejas.

Figura 29 – Comparação de quadros onde a quantidade de pixels claros antes e após a ingestão de álcool é maior para o estado de ebriedade.

.



Fonte: Autor.

Em adição, na Figura 29, é possível observar que as maiores diferenças de intensidade de pixels está entre os participantes 1, 2 e 4, sendo possível estabelecer que a região superficial da pele destes indivíduos sofreu maiores alterações na temperatura da região superficial cutânea quando comparado aos demais. Durante o procedimento experimental, notou-se que os participantes que se mantiveram de forma comedida durante todo o experimento foram os mesmos que obtiveram a maior discrepância entre os quadros que representam a sobriedade e ebriedade respectivamente. Analogamente, os voluntários que estavam em estado de agitação e euforia após a ingestão do álcool, demonstraram menor diferença entre as imagens térmicas do estado de sobriedade e embriaguez como os participantes 3, 6 e 7.

Acrescenta-se ainda que a variação nos valores dos pixels da Figura 29 se dá em decorrência da variação da posição da face dos indivíduos ao longo da captura dos quadros,

enquanto que, os picos observados ocorrem em decorrência da oclusão dos olhos e o incipiente incremento na distância da face e o imageador térmico durante a sessão de captura de imagens.

Tabela 9 – Quantidade de pixels acima da limiar estabelecida para os participantes antes e depois da ingestão do álcool.

ìndice	**estado**	**valor mínimo**	**valor máximo**	**média**	**desvio padrão**
Participante 1	sóbrio	14176	15674	15148,68	369,19
Participante 1	ébrio	20281	22544	21534,76	532,92
Participante 2	sóbrio	19001	19941	19578,81	256,32
Participante 2	ébrio	22066	23999	23123,77	398,87
Participante 3	sóbrio	15736	18701	17130,52	779,94
Participante 3	ébrio	22200	23568	23114,66	271,05
Participante 4	sóbrio	6345	7337	6811,52	261,01
Participante 4	ébrio	11509	12549	12045,2	234
Participante 5	sóbrio	49982	51219	50492,43	296,99
Participante 5	ébrio	35783	47232	40794,48	2916,57
Participante 6	sóbrio	8158	8980	8479,28	196,33
Participante 6	ébrio	13185	14513	13746,19	328,25
Participante 7	sóbrio	7025	7861	7399,30	167,34
Participante 7	ébrio	7278	8281	7828,06	175,90
Participante 8	sóbrio	7298	8182	7731,76	195,77
Participante 8	ébrio	14683	15804	15205,95	271,11
Participante 9	sóbrio	10515	11433	11046,38	197,71
Participante 9	ébrio	11904	14856	13602,37	574,29
Participante 10	sóbrio	10039	11194	10849,67	221,50
Participante 10	ébrio	18787	21400	20357,18	480,16
Participante 11	sóbrio	11516	14862	13268	609,30
Participante 11	ébrio	15137	16722	15737,51	484,60
Participante 12	sóbrio	31101	32704	31852,24	520,59
Participante 12	ébrio	31819	34683	33615,89	874,07
Participante 13	sóbrio	29710	32172	31074,39	801,8
Participante 13	ébrio	31434	38174	36224,80	2124,99
Participante 14	sóbrio	8258	8769	8493,54	115,52
Participante 14	ébrio	11497	12983	12196,89	364,38
Participante 15	sóbrio	32527	33979	33213,80	293,89
Participante 15	ébrio	39345	41376	40551,40	479,60

Fonte: Autor.

A Tabela 9 demonstra a quantidade de pixels claros do conjunto de imagens térmicas dos participantes em estado sóbrio e ébrio. Desta forma é possível perceber o quinto participante obteve um valor menor de quantidade de pixels brilhosos após a ingestão do álcool em relação a aquisição em que estava sóbrio.

Tabela 10 – Diferença entre os valores médios de pixels brilhosos entre os grupos da primeira a segunda aquisição.

ìndice	**diferença entre as médias**
Participante 1	6386,08
Participante 2	3545,00
Participante 3	5984,13
Participante 4	5233,67
Participante 5	-9697,95
Participante 6	5266,91
Participante 7	428,76
Participante 8	7474,19
Participante 9	2556
Participante 10	9507
Participante 11	2470
Participante 12	1763,64
Participante 13	5150,41
Participante 14	3703,35
Participante 15	7337

Fonte: Autor.

A Tabela 10 apresenta os resultados da diferença entre os valores médios dos pixels considerados brilhosos entre o conjunto de quadros do indivíduo em estado de sobriedade e 30 minutos após a ingestão da última dose de cerveja. Desta forma, é possível analisar que os participantes 1, 8, 10 e 15 obtiveram a maior diferença de temperatura após a ingestão do álcool. Salienta-se ainda que, a diferença entre as médias do participante 5 resultou em um valor negativo em vista da obstrução da face em decorrência da quantidade de cabelo, barba e bigode durante a captura de imagens. Consequentemente, se comparado aos resultados da tabela Tabela 8, percebe-se que os indivíduos que possuem uma estimativa menor de água total no corpo, conforme a Tabela 5 obtiveram as maiores diferenças de temperatura entre o primeiro e segundo conjunto de amostras. Contudo, o valor de BrAC apresentado pelo etilômetro conforme a Tabela 7 demonstra que em um cenário real, o valor do CAS de cada participante pode variar da estimativa inicial em virtude da unicidade do metabolismo humano. A consequência da diferença da metabolização do álcool em cada organismo se dá pela diferença do tempo levado para cada participante atingir o pico de álcool no sangue para então ocorrer a queda do nível de CAS.

4.1.4 Resultados do treinamento do modelo de DNN

Com base na na escolha de hiperparâmetros realizada no trabalho de Silva (2021), os hiperparâmetros selecionados resultaram em uma acurácia de validação e teste próximos de 100%, em vista do tamanho do *batch size*, *epochs* e *learning rate* e uma acurácia de teste de 88,4%. Não obstante, outro fator que contribui com a alta taxa de acurácia no conjunto de treino e validação é a técnica de *feature extraction* empregada. Consequentemente, uma

busca de hiperparâmetros conforme a Tabela 11, demonstra que o a redução da faixa dos hiperparâmetros podem reduzir o *overfitting* e ajudar o modelo a convergir.

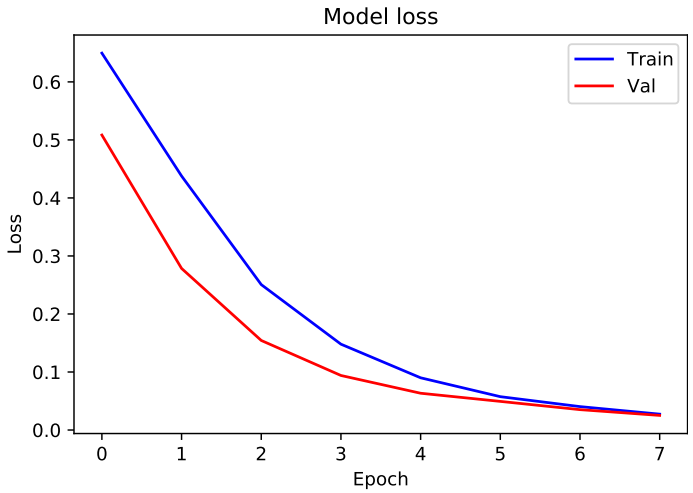
Tabela 11 – Conjunto de hiperparâmetros para a etapa de *transfer learning*.

Epochs	*Batch Size*	*Learning Rate*	*Fold Splits*	*Average Accuracy*	*Standard Deviation*
300	300	1e-4	5	88,33%	0,3
200	300	1e-4	5	88,67%	0,1
100	300	1e-4	5	88,06%	0,3
300	200	1e-4	5	88,13%	0,1
200	100	1e-4	5	88,33%	0,1
100	50	1e-4	5	88,19%	0,1
300	200	1e-5	5	88,06%	0,3
200	100	1e-5	5	88,53%	0,1
100	50	1e-5	5	87,96%	0,3
300	50	1e-3	5	87,93%	0,3

Fonte: Autor.

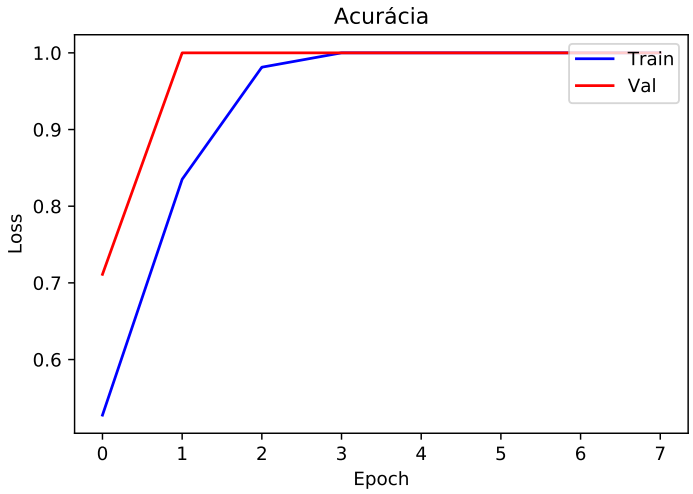
Subsequentemente, os resultados obtidos utilizando como hiperparâmetros os seguintes valores: 8, 300 e 1e-4 como número de *epochs*, tamanho do *batch size* e *learning rate*, respectivamente demonstraram ser os melhores resultados para a seleção de modelos durante a *stratified k-fold cross validation* e, por conseguinte, durante a validação cruzada, foi selecionado o *fold* em que o modelo obteve a melhor convergência, utilizando os hiperparâmetros citados anteriormente.

Figura 30 – *Loss* do modelo selecionado durante a *k-fold cross validation*.



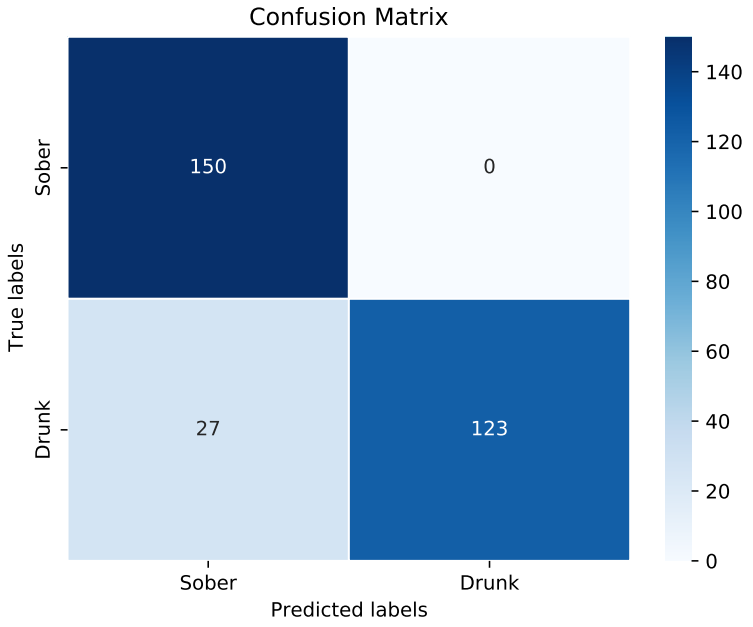
Fonte: Autor.

Figura 31 – Acurácia do modelo selecionado durante a *k-fold cross validation*.



Fonte: Autor.

Figura 32 – Matriz de confusão do modelo.



Fonte: Autor.

O modelo obtido foi selecionado a partir do primeiro *fold* que mostrou a melhor convergência em relação aos demais e suas curvas de acurácia e custo podem ser vistas na Figura 30 e Figura 31. Em adição, o modelo selecionado obteve 91% de acurácia no conjunto de teste. A Tabela 12 apresenta os resultados do relatório de classificação obtidos.

Tabela 12 – Relatório de classificação do modelo após a etapa de treinamento.

Classe	Precisão	Revocação	F-score	Support
Sober	0,85	1,00	0,92	150
Drunk	1,00	0,82	0,90	150

Fonte: Autor.

Após o treinamento do modelo, observando a diferença entre a acurácia no conjunto de teste em relação ao desempenho nos conjuntos de treino e validação, é possível inferir que o modelo ainda pode ser otimizado. Desta forma, durante a condução do *fine-tuning* as técnicas de *pruning* e quantização reduzem o tamanho do modelo conforme descrito anteriormente. Desse modo, melhoram o desempenho no conjunto de teste em vista do espaço para otimização ainda restante. Contudo, o acelerador do *Google* requer que o modelo esteja quantizado em pontos fixos de 8 bits enquanto que o acelerador visual da Intel (NCS2), requer que o modelo esteja em *float16*. Desta forma, para o Google Coral, a técnica de *pruning-quantization aware training* é conduzida para otimizar o modelo, enquanto que, para o NCS2, somente a técnica de *pruning* é conduzida para a otimização do modelo. Dessarte, os parâmetros empregados para a etapa de *PQAT* são apresentados abaixo:

Tabela 13 – Hiperparâmetros utilizados para o *pruning* do modelo para o Edge TPU.

Esparsidade	*Epochs*	*Learning Rate*	*Validation split*	Acurácia (conjunto de teste)
10%	**1**	**1e-5**	**10%**	**91,33%**
10%	2	1e-5	10%	90,33%
10%	3	1e-5	10%	50%
90%	1	1e-5	10%	91,13%
90%	2	1e-5	10%	91,13%
90%	3	1e-5	10%	91,13%

Fonte: Autor.

A Tabela 13 demonstra que para a podação dos pesos da rede neural convolucional, apenas 1 *epoch* de treinamento é necessário para melhorar a acurácia de teste e, consequentemente, como parâmetro, uma esparsidade de 10% da quantidade de pesos da rede como limite da podação. Ademais, a mesma taxa de aprendizagem empregada durante a etapa de treinamento é suficiente para realizar o *fine-tuning* da rede com a técnica de podação e em seguida, a técnica de quantization consciente de treinamento é conduzida para realizar o *fine-tuning* final.

Posteriormente, a quantização do modelo podado é realizada para convertê-lo ao formato *Tensorflow Lite* e compilado para o *Edge TPU*. Os resultados obtidos a partir da

seleção de hiperparâmetro são apresentados na Tabela 14:

Tabela 14 – Hiperparâmetros utilizados para a quantização do modelo para o *Edge TPU*.

Epochs	*Learning Rate*	*Validation split*	Acurácia (conjunto de teste)
1	1e-5	10%	91,33%
2	1e-5	10%	93,33%
3	**1e-5**	**10%**	**94.34%**
1	1e-6	10%	88,99%
2	1e-6	10%	91,00%
3	1e-6	10%	91,00%

Fonte: Autor.

Não obstante, para a podação da rede e uso do NCS2, a mesma busca de hiperparâmetros da Tabela 13 foi utilizada, obtendo os mesmos resultados. Contudo, após a conversão do modelo para um formato de representação intermediária (IR), a acurácia obtida foi de 91% no conjunto de teste.

Desta forma, em conjunto com a acurácia obtida acima, as técnicas de compressão de modelo também possibilitaram a redução do tamanho dos arquivos gerados durante as etapas de otimização.

Tabela 15 – Comparação entre o resultado final da compressão do modelo.

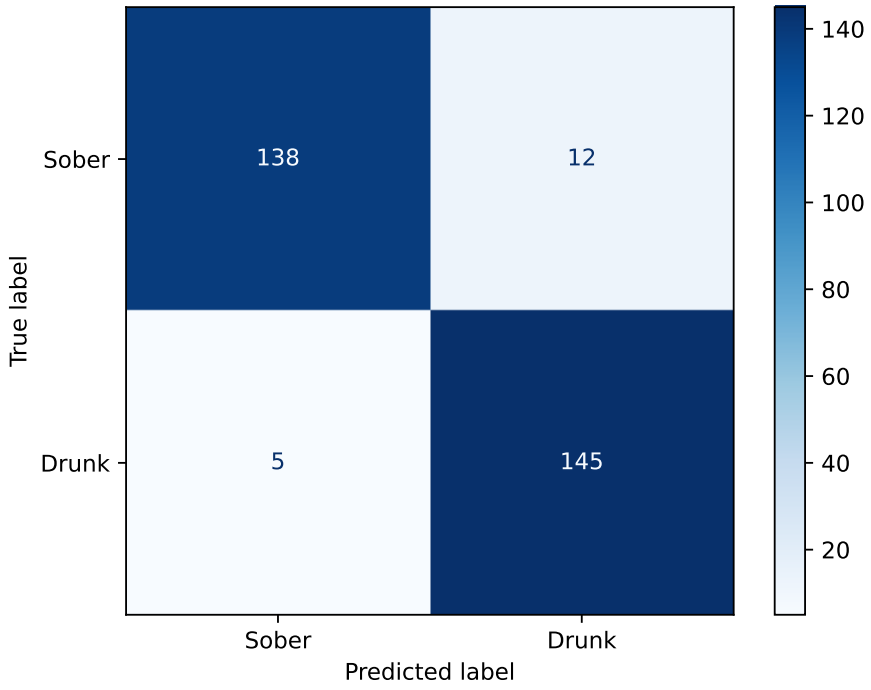
Formato do Modelo	Keras	*Pruning*	Redução da Precisão para float16 (IR)	PQAT (Edge TPU)
Tamanho do Modelo	218 MB	59,5MB	**29.9MB**	**14.9MB**
Acurácia	91,00%	91,33%	**91,00%**	**94,34%**

Fonte: Autor.

A Tabela 15 apresenta o resultado das técnicas de compressão empregadas para a otimização do modelo proposto. Sob este contexto, após a etapa de treino o modelo possuía um tamanho de 218 MB. Após o emprego da técnica de podação, o modelo obteve um incipiente aumento na acurácia e uma redução do tamanho do modelo em aproximadamente 3,6 vezes, sendo reduzido para 59,5MB. Dessa forma, a redução da precisão dos pesos do modelo para o Intel NCS2 através do OpenVino foi possível reduzindo o modelo para 29,9MB após a técnica de podação da rede, preservando a acurácia de 91,00%. Contudo, o melhor desempenho foi obtido pelo modelo utilizado pelo acelerador *Edge TPU*, sendo a redução do modelo de 218 MB para 14,9MB, que representa uma redução de 14,5 vezes o tamanho original, além da maior acurácia obtida entre os demais formatos do modelo. Sob o contexto da acurácia na borda da rede, em concordância com a Tabela 15, é possível

perceber que o modelo em formato Tensorflow Lite obteve uma melhor generalização em relação a tarefa de classificação do estado de indivíduo, conforme mostra a Figura 33. O modelo obteve 145 acertos das 150 amostras do estado de embriaguez do sujeito.

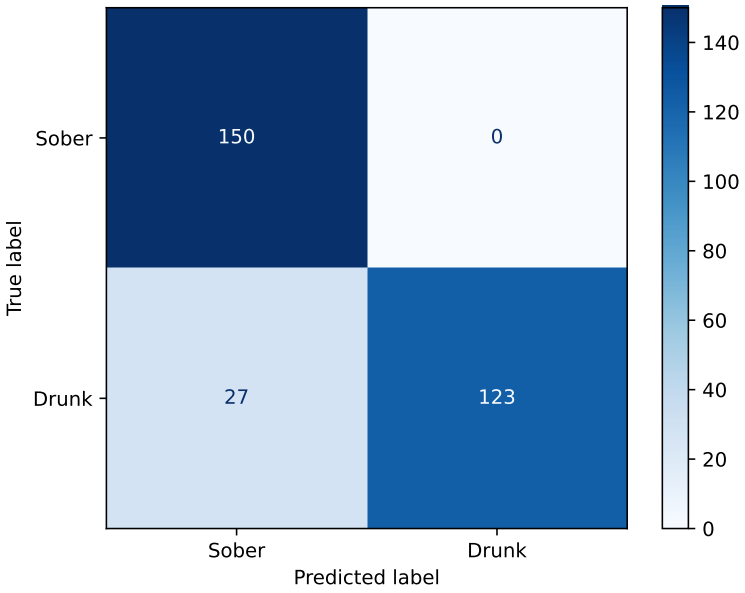
Figura 33 – Matriz de confusão do modelo Tensorflow Lite acelerado pelo Edge TPU no conjunto de teste.



Fonte: Autor.

A Figura 34 apresenta a matriz de confusão obtida durante a inferência na borda da rede utilizando as imagens do conjunto de teste. Não obstante, observando a figura supracitada, é possível perceber que, em decorrência da otimização realizada pela técnica de podação ciente de quantização durante o treinamento, foi possível reduzir consideravelmente o *overfitting* obtido pelo modelo convertido em para o formato de representação intermediária para o acelerador da Intel.

Figura 34 – Matriz de confusão do modelo OpenVino acelerado pelo Intel Neural Compute Stick 2 no conjunto de teste.



Fonte: Autor.

Observando a Figura 34, percebe-se que o modelo otimizado para o Intel NCS2 obteve um viés em relação a classe dos indivíduos em estado sóbrio, considerando a quantidade de predições em relação a classe sóbria onde o modelo conseguiu classificar corretamente todas as amostras do estado sóbrio e classificou incorretamente 27 amostras da classe de embriaguez. Assim sendo, o uso da técnica de podação, e subsequentemente, a redução da precisão do modelo através do uso da ferramenta *OpenVino* da Intel, acarretou a remoção da esparsidade do modelo após a etapa de treinamento.

Desta forma, percebe-se que a ferramenta da Intel não possibilita a podação do modelo cienta da otimização, o que justifica a acurácia obtida. Consequentemente, percebe-se que a acurácia de 91% obtida pelo modelo em formato de representação intermediária é similar a acurácia obtida ao final da etapa de treinamento antes do *fine-tuning*.

Por fim, o melhor resultado obtido foi a partir do modelo treinado e convertido para o formato adequado ao acelerador Coral USB. Neste modelo, foi possível obter uma acurácia no conjunto teste de 94.34%. Em relação a classificação de imagens térmicas para a detecção de embriaguez, (SOLTUZ; NEAGOE, 2021) obtiveram a acurácia máxima de 98.54%. No entanto, a RNC utilizada para a tarefa de classificação do estado de embriaguez foi a GoogLeNet[1], que se trata de um modelo computacionalmente custoso em decorrência

[1] https://www.mathworks.com/help/deeplearning/ref/googlenet.html

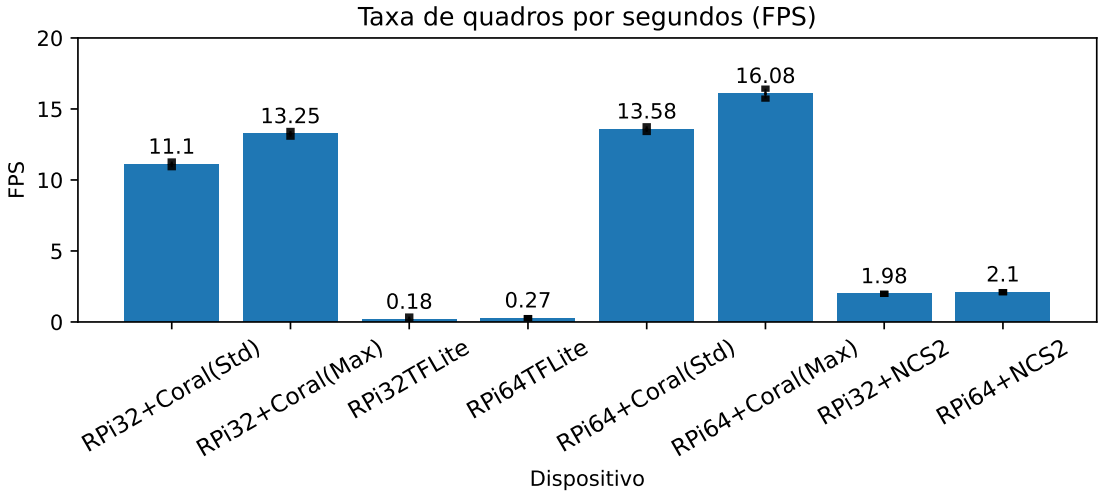
da sua consideravel quantidade paramétrica. Em adição, os autores utilizaram a base de dados disponibilizada por (KOUKIOU; ANASTASSOPOULOS, 2015), no entanto, os autores realizaram uma divisão de 50% dos quadros de cada participante para treino e teste. Logo, considerando a similaridade entre cada quadro por se tratar da mesma pessoa, acredita-se que esta abordagem pode induzir o modelo a *overfitting*. Além disso, os autores não apresentaram as curvas de desempenho do modelo treinado.

4.1.5 Resultados do servidor na borda da rede

Esta seção apresenta os resultados obtidos a partir da inferência do modelo proposto na borda da rede. As métricas utilizadas para avaliar o desempenho do servidor foram: tempo de inferência, o tempo de carregamento do modelo, tempo de aquecimento, a taxa de quadros por segundo (fps), temperatura, consumo energético e uso da memória RAM. Para a avaliação da taxa de quadros, as imagens são enviadas ao servidor na borda da rede sem considerar a acurácia. Para obter uma significância estatística, foram enviadas 300 imagens ao Raspberry Pi e a taxa de quadros é computada a partir de cada inferência realizada.

Ressalta-se ainda que, para a avaliação da acurácia, as imagens foram carregadas a partir do armazenamento em disco. Enquanto que, para avaliação das demais métricas, uma conexão foi estabelecida via UDP a partir do dispositivo móvel, e subsequentemente, as imagens foram enviadas em tempo real para o servidor na borda da rede sem considerar a acurácia.

Figura 35 – Taxa de quadros por classificação (FPS).



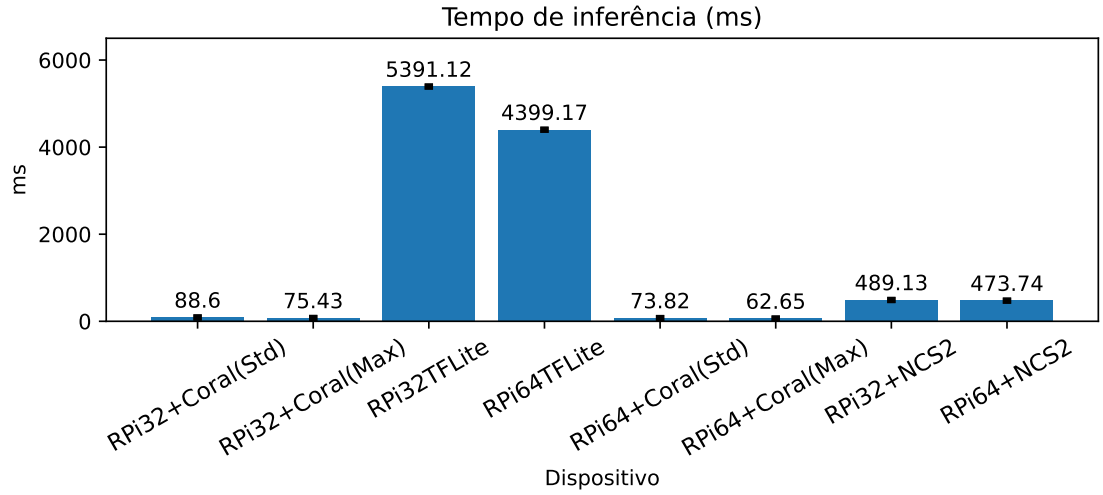
Fonte: Autor

Desta forma, a Figura 36 apresenta os valores médios e o desvio padrão da taxa

de quadros durante a inferência considerando as imagens enviadas a partir do dispositivo móvel e sem considerar a acurácia.

A Figura 36 demonstra o tempo de inferência médio e o desvio padrão obtido utilizando os aceleradores NCS2 e EdgeTPU, sendo o Edge TPU avaliado em frequência padrão e frequência máxima, e também, com as versões do sistema operacional Bullseye 32 e 64 bits.

Figura 36 – Tempo de inferência na borda da rede.



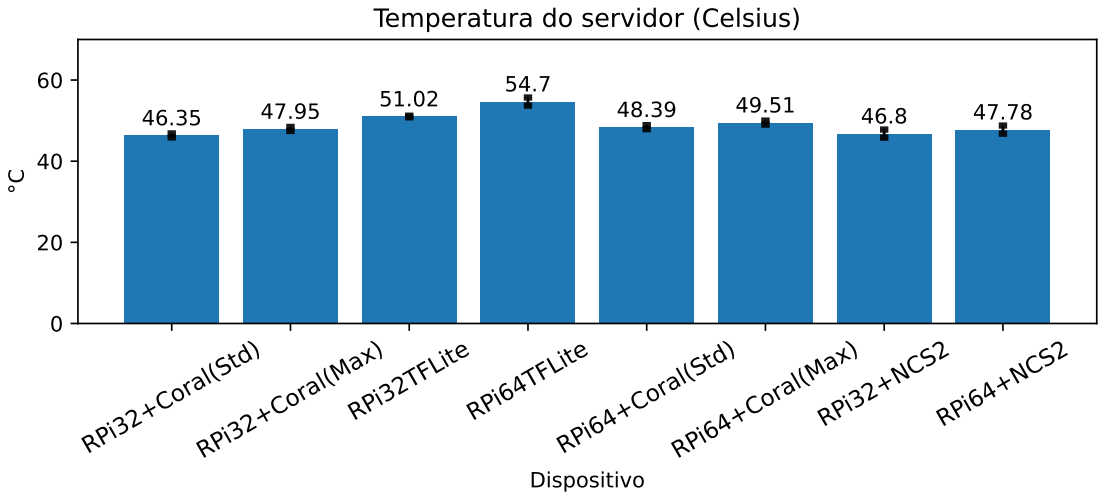
Fonte: Autor

Neste experimento, foram utilizados diferentes aceleradores e formatos do modelo proposto, sendo estes formatos: *Tensorflow Lite* e Representação Intermediária obtido a partir da ferramenta do *OpenVino*. Não obstante, foi possível notar que o dispositivo *Google Coral USB* em frequência máxima do *clock* no Raspberry Pi 4B obteve a maior taxa de quadros.

Observando a figura supracitada, é possível perceber que o Raspberry Pi com o sistema operacional em 64 bits acelerado pelo Edge TPU em frequência máxima obteve uma velocidade aproximadamente 87 vezes menor que o o modelo sem aceleração no sistema operacional de 32 bits. Enquanto que, o NCS2 obteve um ganho na velocidade média de inferência de aproximadamente 11 vezes em relação ao modelo em *Tensorflow Lite*.

A Figura 37 apresenta a temperatura média durante a execução da inferência em cada interação. Sob este contexto, o Raspberry Pi utilizando o sistema operacional de 64 bits em aceleração obteve a maior temperatura média durante o experimento.

Figura 37 – Temperatura durante execução.



Fonte: Autor

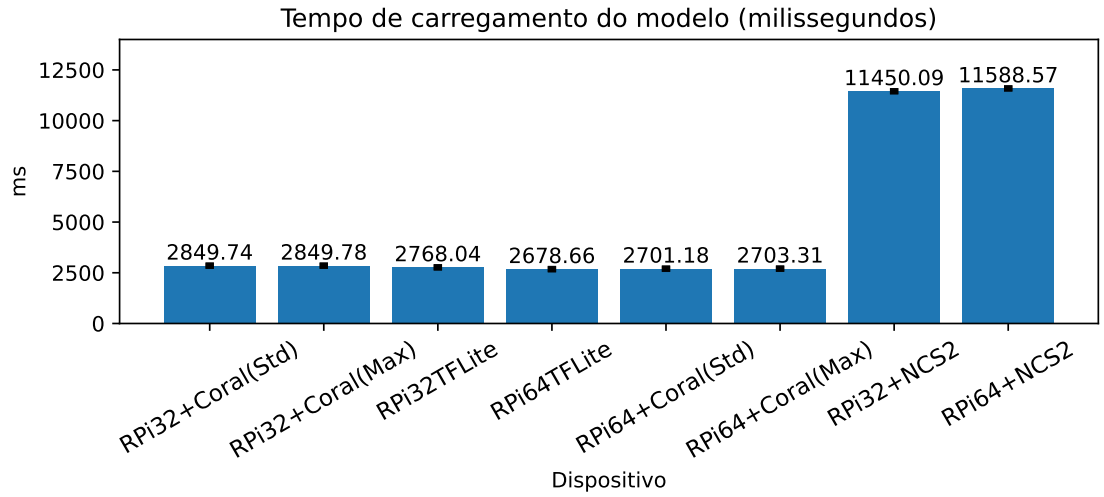
Neste gráfico é possível inferir que o NCS2 teve uma temperatura reduzida se comparada a inferência realizada somente pelo Raspberry Pi e a inferência realizada pelo Raspberry Pi acelerado com o Edge TPU.

Desta forma, para avaliar a temperatura durante a inferência do modelo, um dissipador do tipo case armadura foi utilizado para evitar dano aos componentes, e em adição, as inferências foram realizadas em temperatura ambiente com temperatura máxima de 25ºC. Não obstante, as avaliações ocorreram com um intervalo de 20 minutos a cada inferência para evitar interferência com as avaliações subsequentes.

A Figura 38 demonstra o tempo de carregamento do modelo que está relacionado ao consumo de memória que pode ocasionar a gargalo de recursos do dispositivo em vista da grande quantidade de parâmetros nos modelos de redes neurais convolucionais. O Coral USB possui uma memória interna para alocação de tensores, isto permite reduzir consideravelmente o tempo de inferência em relação ao Intel NCS2 que realiza o carregamento do modelo em sua memória.

Logo, é possível perceber que o Intel NCS2 teve um tempo de carregamento maior que o *Raspberry Pi 4B* sem acelerador e com o *Coral USB*.

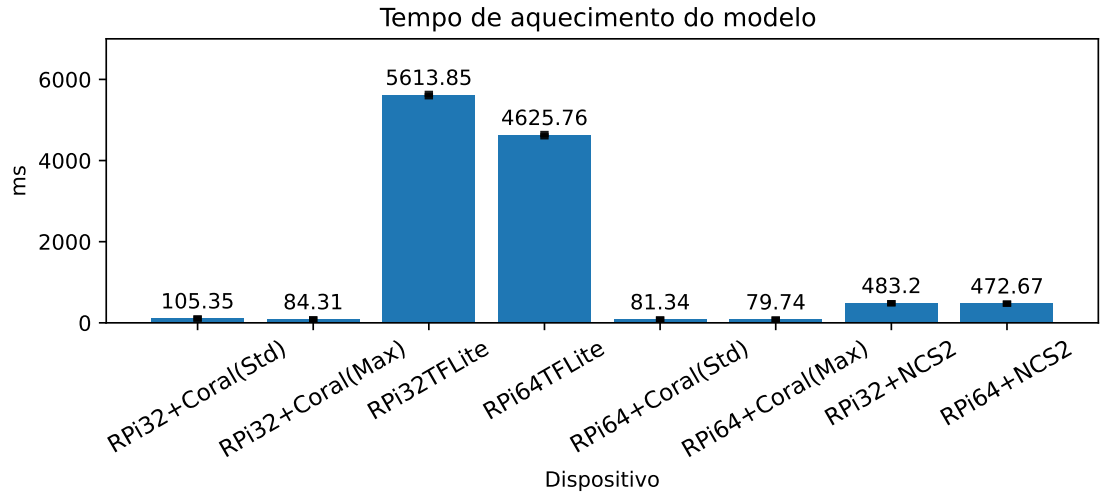
Figura 38 – Tempo de carregamento do modelo.



Fonte: Autor

Isto posto, durante a execução, a memória é gradualmente alocada em três momentos diferentes: (1) durante o carregamento do modelo, (2) na primeira inferência (aquecimento) e (3) durante as inferências subsequentes. Acrescenta-se ainda que, a alocação da memória durante as etapas de carregamento e aquecimento é mantida para todas as inferências, sendo desalocada somente quando o modelo é descarregado.

Figura 39 – Tempo de aquecimento do modelo.



Fonte: Autor

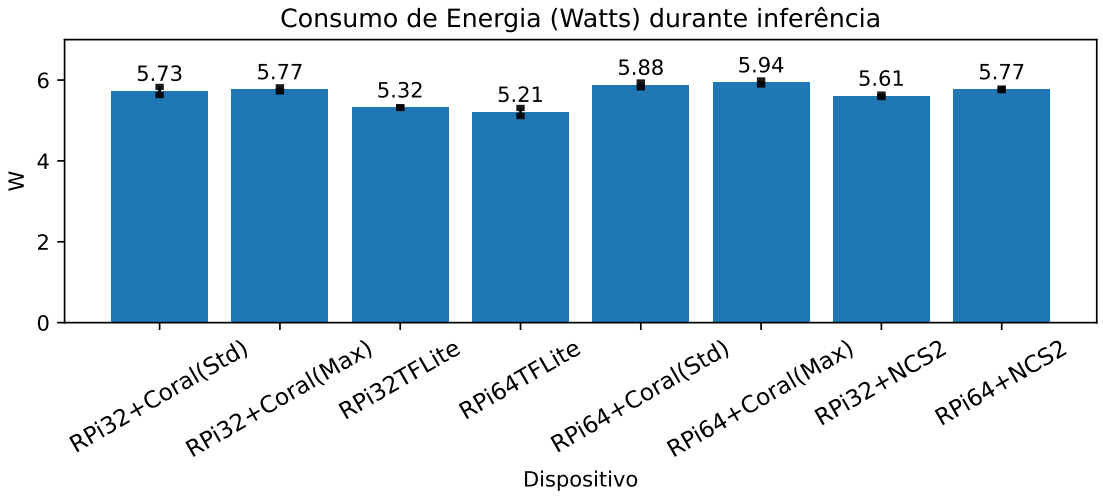
A Figura 39 apresenta os resultados da primeira inferência que correspondem ao tempo de aquecimento do dispositivo para a execução de inferências na borda da rede.

Para a avaliação do tempo de carregamento e o tempo de aquecimento do modelo, foram realizadas 20 avaliações para cada uma dessas métricas, e subsequentemente, observou-se que os desvios padrões se mantiveram dentro dos mesmos intervalos de valores.

Desta forma, a avaliação do tempo de aquecimento do Google Coral USB no sistema operacional de 64 bits na frequência máxima de processamento obteve o menor tempo de aquecimento em comparação com a unidade de processamento visual, o Intel NCS2, o modelo sem aceleração. Igualmente, o Google Coral USB demonstrou ter o menor tempo de carregamento mesmo no sistema operacional de 32 bits.

Por fim, a Figura 40 apresenta o consumo energético reference o processo de inferência. Para avaliar o consumo energético, foram utilizados o Raspberry Pi 4B com o Google Coral USB com a frequência padrão e em frequência máxima. Em adição, também foram avaliados durante os experimentos o consumo energético sem o uso de aceleradores.

Figura 40 – Consumo Energético do Raspberry Pi 4B.



Fonte: Autor

Salienta-se ainda que, para a avaliação do consumo do Raspberry Pi 4B como servidor na borda da rede, o consumo foi avaliado considerando as etapas de carregamento do modelo, tempo de aquecimento e consumo durante a inferência de 300 imagens subsequentes. Ainda sob o contexto do consumo de energia do dispositivo, foi possível inferir que o Intel NCS2 e o EdgeTPU incrementaram o consumo do dispositivo em comparação ao modelo em formato Tensorflow Lite.

Em adição, a Figura 40 demonstra que o tempo de inferência levado no sistema operacional de 32 bits, levou um tempo levemente maior do que no sistema operacional de 64 bits. Contudo, considerando o desvio padrão obtido, é possível perceber que a diferença é negligenciável, possivelmente não havendo diferenças significativas.

4.2 Discussões

4.2.1 Estimação do nível de álcool no sangue

Em conformidade com os resultados obtidos na Capítulo 4, é possível realizar uma correlação com a literatura relacionada ao metabolismo do álcool. De acordo com Cederbaun (2012), o equilíbrio da concentração de álcool em um tecido depende do teor relativo de água neste. Logo, o equilíbrio do álcool dentro de um tecido depende do teor de água, da taxa de fluxo sanguíneo e de sua massa.

Ainda segundo o autor, a mesma dose de álcool por unidade de peso corporal pode produzir diferentes taxas de concentração de álcool no sangue em diferentes indivíduos devido as grandes variações nas proporções de gordura e água em seus corpos, e ao reduzido coeficiente de partição lipídio/água do etanol. As mulheres geralmente têm um volume de distribuição de álcool menor que os homens devido ao maior percentual de gordura corporal. Em adição, as mulheres possuem o pico de álcool no sangue mais alto que os homens quando recebem a mesma dose de álcool em gramas por quilograma de peso corporal, mas não ocorrem diferenças quando recebem a mesma dose por litro de água corporal.

Em vista do supracitado, a equação de Widmark abordada por Brick (2006) foi empregada considerando diferentes doses para a quantidade total de água corporal, e consequentemente, observou-se que houve um incremento considerável no nível de BrAC dos participantes do sexo feminino em relação aos participantes do sexo masculino. Desta forma, apesar da margem de erro do etilômetro empregado de 0,045mg/L de ar alveolar expirado, foi possível perceber que o comportamento observado nos participantes seguiu o esperado conforme a literatura susodita.

Em adição, a tolerância aos efeitos do álcool pelo sujeito afeta consideravelmente as respostas fisiológicas de intoxicação e, em seguida, o valor estimado de nível de álcool no sangue (ASTON; LIGUORI, 2013). Os sintomas de intoxicação podem ser mais leves para os indivíduos com maior tolerância e, portanto, seria mais difícil de associar com um valor de CAS específico.

Não obstante, de acordo com Hustad e Carey (2005), o uso de equações para estimação do nível de CAS possui limitações, mesmo conhecendo os valores de altura, peso e quantidade de álcool a ser consumida, em vista unicidade nas taxas de absorção, distribuição e metabolização do etanol no corpo do indivíduo. Em adição, o uso de equações para estimar o CAS em cenários reais de consumo de álcool introduz mais variabilidade do que é observado em ambientes controlados. Desta forma, o emprego de cálculos de estimação de CAS pode subestimar o valor de álcool no sangue do indivíduo após o consumo de álcool conforme observado na Tabela 7 e Tabela 8 respectivamente.

Ademais, este estudo contribui com uma análise acerca da correlação entre a estimativa do nível de álcool no sangue (CAS) com valores obtidos através do emprego de um etilômetro. Isto posto, percebeu-se que as variáveis experimentais encontradas durante o experimento em que os participantes realizaram a ingestão de álcool, dificultam a correlação do CAS estimado com o valor aferido na prática. Subsequentemente, os resultados obtidos demonstraram que foi possível verificar que os participantes demonstraram um incremento, e subsequentemente, uma queda no nível de CAS de cada indivíduo após 1 hora e 30 minutos.

4.2.2 Sessão de captura das imagens

O protocolo empregado para captura de imagens dos participantes teve como base o trabalho de (KOUKIOU; ANASTASSOPOULOS, 2015) e (HERMOSILLA et al., 2018). Contudo, os trabalhos dos autores não levaram em consideração um protocolo padronizado pela literatura relacionada à termografia. Neste contexto, de acordo com Fernandez-Cuevas et al. (2015), o uso de abordagens padronizadas para o emprego da termografia em seres humanos pode favorecer a literatura e minimizar a influência potencial de fatores técnicos durante o procedimento para aquisição de imagens térmicas.

Desta forma, outra contribuição deste projeto de mestrado é uma abordagem seguindo os padrões discutidos em (FERNANDEZ-CUEVAS et al., 2015). Durante a aquisição de imagens foi considerado o tempo de estabilização do bolometro do dispositivo térmico, assim como a diferença de temperatura do ambiente ao longo do tempo. Em adição, acredita-se que para a representação fisiológica do estado do indivíduo, o emprego de um imageador com uma resolução maior e reduzida sensibilidade térmica pode favorecer a detecção de embriaguez em virtude da quantidade de atributos em uma imagem com uma resolução maior em conformidade com (FERNANDEZ-CUEVAS et al., 2015).

Sob o contexto das repostas fisiológicas dos indivíduos, conforme discutido em (KOUKIOU; ANASTASSOPOULOS, 2015; KOUKIOU; ANASTASSOPOULOS, 2016), as respostas fisiológicas observadas nos participantes incluiram o aumento de temperatura na esclera e queda da temperatura no nariz. Em relação ao incremento de temperatura na esclera em vista do consumo do álcool, os participantes de 1, 3, 5, 6, 10 e 11 demonstraram acrescimento do valor de temperatura, enquanto que, os participantes 2, 4, 7, 8, 9, 14 e 15 não demonstraram alterações de temperatura significantes.

Em vista da quantidade amostral de participantes não foi possível identificar ou correlacionar o motivo deste comportamento. Entretanto, este trabalho contribui com a literatura acerca da termografia apresentando as características fisiológicas de um indivíduo em estado de embriaguez em decorrência do consumo da cerveja. Enquanto que, em (KOUKIOU; ANASTASSOPOULOS, 2015) foram observadas as características fisiológicas do indivíduo após a ingestão do vinho.

4.2.3 Deep Neural Networks e Edge Computing

Na subseção 4.1.4 foi apresentado o resultado do treinamento do modelo proposto para classificação de embriaguez. Isto posto, é possível observar um aumento de acurácia entre os resultados da Tabela 13 e Tabela 14 de aproximadamente 3,01% no conjunto de teste. Desta forma, ressalta-se que este incremento ocorre em decorrência do *fine-tuning* realizado utilizando todo o conjunto de imagens durante o retreinamento nas etapas de podação e quantização, respectivamente. O emprego de técnicas de compressão podem reduzir o tamanho do modelo em memória e, em casos específicos, podem preservar ou degradar a acurácia (LIANG et al., 2021; GHOLAMI et al., 2022).

Em vista da seleção de modelos durante a etapa de *stratified k-fold cross validation*, o modelo ainda não havia alcançado a convergência e, subsequentemente, o treinamento adicional durante a etapa de *fine-tuning* favoreceu a sua generalização em vista do treinamento em conjunto com as técnicas de otimização.

Sob o contexto dos aceleradores de *hardware*, o Google Coral USB demonstrou uma velocidade de inferência superior ao Intel NCS2. Contudo, uma das vantagens do Coral USB é a quantização em pontos fixos de 8 bits que acarreta uma quantização agressiva em relação a redução de precisão do Intel NCS2. Em adição, o Coral USB permite aumentar ou diminuir a frequência de processamento do coprocessador através de uma linha de comando no sistema operacional, enquanto que o Intel NCS2, permite o emprego de outras unidades de processamento visual em *cluster* para maximizar o desempenho.

4.2.4 Trabalhos Futuros

A avaliação de outros modelos e outros dispositivos embarcados pode contribuir com a literatura, acerca da classificação de embriaguez, utilizando imagens térmicas. Não obstante, a aquisição de um conjunto amostral maior de indivíduos em estado de ebriedade pode favorecer a capacidade de generalização do classificador.

Ademais, a avaliação do classificador em um cenário real pode também favorecer o presente trabalho, validando o uso de imagens térmicas para o reconhecimento e categorização do estado do indivíduo onde não há exemplos de termogramas do sujeito em estado de sobriedade.

Por fim, o uso de *federated learning* pode favorecer a capacidade de generalização do modelo, e também, permitir o emprego de mais *edge nodes* para avaliar a escalabilidade da solução proposta neste projeto de mestrado

Conclusão

O presente trabalho de mestrado objetivou propor uma solução para a classificação em tempo real do estado de ebriedade ou sobriedade do indivíduo. Desta forma, um aplicativo de envio de imagens em tempo real através do protocolo UDP foi desenvolvido, além de um modelo computacionalmente complexo para classificação de imagens treinado e comprimido para possibilitar que a classificação de embriaguez fosse realizada na borda da rede.

Sob o contexto da classificação da embriaguez, o modelo desenvolvido utilizou imagens capturadas com uma resolução maior que os demais trabalhos da literatura, o que possibilitou a extração das características fisiológicas representativas do estado do sujeito. Em adição, a técnica de transferência de aprendizagem permitiu ao modelo abstrair mais características relevantes ao reconhecimento da embriaguez a partir das imagens térmicas capturadas. Consequentemente, uma das contribuições deste trabalho é o uso de técnicas de compressão de modelo para diferentes aceleradores de *hardware* para viabilizar a implantação de modelos de aprendizagem profunda da borda da rede.

Em relação a borda da rede, uma comparação foi conduzida para avaliar o desempenho do Raspberry Pi com um sistema operacional de 32 bits, comparado ao desempenho do mesmo dispositivo com um sistema operacional de 64 bits. Dessa forma, foi possível perceber que existe uma relação inversamente proporcional entre o consumo energético e o tempo de inferência. Outra contribuição deste trabalho foi uma avaliação entre o Intel NCS2 e o Google Coral USB durante a execução de um modelo complexo desenvolvido a partir da VGG16. Ambos os aceleradores demonstraram redução considerável no tempo de inferência e no tempo de aquecimento do modelo. No entanto, o NCS2 demonstrou ter o maior tempo de carregamento do modelo em relação às outras opções.

Neste contexto, os trabalhos da literatura que apresentam o uso do Coral USB para o Raspberry Pi não levam em consideração a possibilidade de realizar a maximização da frequência do *clock*. Logo, foi possível observar que o acelerador Google Coral USB possui o menor tempo de inferência e a maior taxa de quadros e, consequentemente, o uso do Raspberry Pi com o sistema operacional de 64 bits acelerado por este coprocessador demonstrou ser o mais adequado para aplicações em que a classificação de imagens é imprescindível.

Em vista do supracitado, a borda da rede contribui evitando a heterogeneidade entre os dispositivos conectados ao servidor na borda da rede, ao mesmo tempo que possibilita a offloading da carga computacional no servidor. Consequentemente, se torna mais viável o envio das imagens para a borda da rede para a classificação do que executar o modelo

diretamente em dispositivos móveis, onde não há a possibilidade do uso de aceleradores. Com base nisso, a compressão do modelo e aceleração do Raspberry Pi através do Google Coral USB e do NCS2, possibilita a classificação em tempo real. Contudo, para uma aplicação que depende consideravelmente da qualidade da imagem para a classificação correta das imagens, o uso do protocolo UDP para envio das imagens via pacotes para a borda da rede pode acarretar a perda de informações relevantes na imagem em situações onde ocorre perda de pacotes em decorrência de problema de conexão.

Por fim, acrescenta-se ainda que o presente projeto de mestrado não tem como objetivo promover ou incentivar o consumo de bebidas alcoólicas e sim contribuir com um método de fiscalização e com o estado-da-arte em relação a classificação de embriaguez em tempo real, utilizando *edge computing*. Assim sendo, todos os participantes foram orientados, após o término da captura de imagens, a não consumir álcool novamente por pelo menos 24 horas. No dia seguinte, também foram contatados para garantir seu estado sóbrio.

Referências

ABREU, D.; SOUZA, d. E.; MATHIAS, T. *Impact of the brazilian traffic code and the law against drinking and driving on mortality from motor vehicle accidents*. 2022. Cadernos de Saude Publica. Disponível em: <https://www.scielo.br/j/csp/a/hMC54dJfRLwnqN9ZcnRFmhC/?format=pdf&lang=pt>. Acesso em: 06 fev. 2022. Citado na página 31.

ALIBABAEI, K. et al. Real-time detection of vine trunk for robot localization using deep learning models developed for edge tpu devices. *Future Internet*, v. 14, n. 7, p. 199, 2022. Citado 2 vezes nas páginas 29 e 51.

ALSAMHI, S. et al. Computing in the sky: A survey on intelligent ubiquitous computing for uav-assisted 6g networks and industry 4.0/5.0. *Drones*, v. 6, n. 7, p. 177, 2022. Citado na página 42.

ASTON, E.; LIGUORI, A. Self-estimation of blood alcohol concentration: A review. *Addictive Behaviors*, v. 38, n. 4, p. 1944–1951, 2013. Citado na página 96.

BALLER, S. et al. Deepedgebench: Benchmarking deep neural networks on edge devices. In: *2021 IEEE International Conference on Cloud Engineering (IC2E)*. San Francisco, CA, USA: IEEE, 2021. p. 20–30. Disponível em: <https://ieeexplore.ieee.org/document/9610432/>. Acesso em: 15 mai. 2022. Citado 2 vezes nas páginas 53 e 54.

BARRET, P.; HOME, J.; REYNER, L. Early evening low alcohol intake also worsens sleepiness-related driving impairment. *Human Psychopharmacology*, v. 20, n. 4, p. 287–290, 2005. Citado na página 32.

BOTTA, A. et al. Integration of cloud computing and internet of things: A survey. *Future Generation Computer Systems*, v. 56, n. 1, p. 684–700, 2016. Citado na página 41.

BRELAND, D. et al. Deep learning-based sign language digits recognition from thermal images with edge computing system. *IEEE Sensors Journal*, v. 21, n. 9, p. 10445–10453, 2021. Citado na página 52.

BRICK, J. Standardization of alcohol calculations in research. *Alcoholism: Clinical and Experimental Research*, v. 8, n. 30, p. 1276–1287, 2006. Citado 5 vezes nas páginas 32, 71, 72, 78 e 96.

BUDDHARAJU, P.; PAVLIDIS, I. Physiology-based face recognition in the thermal infrared spectrum. *Medical Infrared Imaging: Principles and Practices*, v. 29, n. 4, p. 18, 2012. Citado 3 vezes nas páginas 28, 36 e 38.

BUDDHARAJU, P.; PAVLIDIS, I.; TSIAMYRTZIS, P. Pose-invariant physiological face recognition in the thermal infrared spectrum. In: *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. New York, NY, USA: IEEE, 2006. p. 53–53. Disponível em: <http://ieeexplore.ieee.org/document/1640493/>. Citado 3 vezes nas páginas 28, 36 e 37.

BUDDHARAJU, P. et al. Physiology-based face recognition in the thermal infrared. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 29, n. 4, p. 613–626, 2007. Citado 4 vezes nas páginas 28, 35, 36 e 38.

CAO, K. et al. An overview on edge computing research. *IEEE Access*, v. 8, n. 1, p. 85714–85728, 2020. Citado na página 41.

CAPRA, M. et al. Hardware and software optimizations for accelerating deep neural networks: Survey of current trends, challenges, and the road ahead. *IEEE Access*, v. 8, n. 1, p. 225134–225180, 2020. Citado na página 46.

CEDERBAUN, A. Alcohol metabolism. *Clinics in Liver Disease*, v. 16, n. 4, p. 667–685, 2012. Citado 2 vezes nas páginas 32 e 96.

CHEN, J.; RAN, X. Deep learning with edge computing: A review. *Proceedings of the IEEE*, v. 107, n. 8, p. 1655–1674, 2019. Citado 4 vezes nas páginas 41, 44, 45 e 46.

CHENG, J. et al. Recent advances in efficient computation of deep convolutional neural networks. *Frontiers of Information Technology and Electronic Engineering*, v. 19, n. 1, p. 64–77, 2018. Citado 3 vezes nas páginas 46, 47 e 48.

CHENG, Y. et al. A survey of model compression and acceleration for deep neural networks. p. 1–10, 2017. Citado na página 50.

CHOUDHARY, T. et al. A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, v. 53, n. 7, p. 5113–5155, 2020. Citado 2 vezes nas páginas 47 e 48.

CORTOT, A. et al. Gastric emptying and gastrointestinal absorption of alcohol ingested with a meal. *Digestive Diseases and Sciences*, v. 31, n. 4, p. 343–348, 1986. Citado na página 32.

DENG, S. et al. Edge intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet of Things Journal*, v. 7, n. 8, p. 7457–7469, 2020. Citado na página 47.

DINH, D. et al. Towards ai-based traffic counting system with edge computing. *Journal of Advanced Transportation*, v. 2021, n. 1, p. 1–15, 2021. Citado 2 vezes nas páginas 53 e 55.

ENGERT, V. et al. Exploring the use of thermal infrared imaging in human stress research. *PLoS ONE*, v. 9, n. 3, p. 1–11, 2014. Citado na página 34.

FENG, H. et al. Benchmark analysis of yolo performance on edge intelligence devices. *Cryptography*, v. 6, n. 2, p. 1–16, 2022. Citado 2 vezes nas páginas 53 e 55.

FERNANDEZ-CUEVAS, I. et al. Classification of factors influencing the use of infrared thermography in humans: A review. *Infrared Physics Technology*, v. 71, n. 1, p. 28–55, 2015. Citado 3 vezes nas páginas 34, 35 e 97.

FORMENTI, D. et al. Thermal imaging of exercise-associated skin temperature changes in trained and untrained female subjects. *Annals of Biomedical Engineering*, v. 41, n. 4, p. 863–871, 2013. Citado na página 75.

GADE, R.; MOESLUND, T. Thermal cameras and applications: a survey. *Machine Vision and Applications*, v. 25, n. 1, p. 245–262, 2014. Citado 4 vezes nas páginas 32, 33, 37 e 38.

GANERT, P.; BOWTHORPE, W. Evaluation of breath alcohol profiles following a period of social drinking. *Journal of the Canadian Society of Forensic Science*, v. 132, n. 3, p. 137–144, 2000. Citado na página 72.

GHOLAMI, A. et al. *A Survey of Quantization Methods for Efficient Neural Network Inference*. [S.l.]: Chapman and Hall/CRC, 2022. Citado 2 vezes nas páginas 50 e 98.

GIL, Y. et al. Quantization-aware pruning criterion for industrial applications. *IEEE Transactions on Industrial Electronics*, v. 69, n. 3, p. 3203–3213, 2022. Citado na página 67.

GOU, J. et al. Knowledge distillation: A survey. *International Journal of Computer Vision*, v. 129, n. 6, p. 1789–1819, 2021. Citado na página 49.

GRECO, L. et al. Effects of ethanol on thermoregulation. *Pattern Recognition Letters*, v. 135, n. 23, p. 346–353, 2020. Citado na página 28.

HAMMER, J. et al. Global status report on alcohol and health 2018. *Global status report on alcohol*, v. 65, n. 1, p. 74–85, 2018. Citado na página 31.

HAN, S. et al. Learning both weights and connections for efficient neural networks. v. 2015, p. 1135–1143, 2015. Citado na página 47.

HEGDE, C. et al. Autotriage - an open source edge computing raspberry pi-based clinical screening system. p. 1–13, 2020. Disponível em: <https://doi.org/10.1101/2020.04.09.20059840>. Citado na página 52.

HERMOSILLA, G. et al. Face recognition and drunk classification using infrared face images. *Journal of Sensors*, v. 2018, n. 1, p. 1–8, 2018. Citado 5 vezes nas páginas 32, 36, 39, 70 e 97.

HINTON, G.; VINYALS, O.; DEAN, J. Distilling the knowledge in a neural network. p. 1–9, 2015. Citado 2 vezes nas páginas 49 e 50.

HOLT, S. et al. Alcohol absorption, gastric emptying and a breathalyser. *British Journal of Clinical Pharmacology*, v. 9, n. 2, p. 205–208, 1980. Citado na página 32.

HOME, J.; REYNER, L.; BARRETT, P. Driving impairment due to sleepiness is exacerbated by low alcohol intake. *Occupational and Environmental Medicine*, v. 60, n. 9, p. 689–692, 2003. Citado na página 32.

HOWELL, K.; DUDEK, K.; SOROKO, M. Thermal camera performance and image analysis repeatability in equine thermography. *Infrared Physics Technology*, v. 110, n. 4, p. 103447, 2020. Citado na página 37.

HUSTAD, J.; CAREY, K. Using calculations to estimate blood alcohol concentrations for naturally occurring drinking episodes: A validity study. *Journal of Studies on Alcohol*, v. 66, n. 1, p. 130–138, 2005. Citado na página 96.

IOANNOU, S.; GALLESE, V.; MERLA, A. Thermal infrared imaging in psychophysiology: Potentialities and limits. *Psychophysiology*, v. 51, n. 10, p. 951–963, 2014. Citado na página 34.

JACOB, B. et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018. p. 2704–2713. Disponível em: <https://ieeexplore.ieee.org/document/8578384/>. Acesso em: 15 mai. 2022. Citado na página 51.

JAINUDDIN, A. A. et al. Evaluation of deep learning accelerators for object detection at the edge. In: *43rd German Conference on Artificial Intelligence (KI2020)*. Bamberg, Germany: IEEE, 2020. p. 323–328. Disponível em: <https://ieeexplore.ieee.org/document/9243367/>. Acesso em: 1 jun. 2020. Citado na página 54.

JIANG, C. et al. Energy aware edge computing: A survey. *Computer Communications*, v. 151, n. 2018, p. 556–580, 2020. Citado na página 44.

JONES, A. Evidence-based survey of the elimination rates of ethanol from blood with applications in forensic casework. *Forensic Science International*, v. 200, n. 1-3, p. 1–20, 2010. Citado 2 vezes nas páginas 35 e 72.

JONES, A.; ANDERSSON, L. Comparison of ethanol concentrations in venous blood and end-expired breath during a controlled drinking study. *Forensic Science International*, v. 132, n. 1, p. 18–25, 2003. Citado na página 72.

JONES, B.; PLASSMANN, P. imaging of skin structure of human skin. *IEEE Engineering in medicine and biology*, v. 17, n. 2002, p. 41–48, 2002. Citado 2 vezes nas páginas 32 e 75.

KALANT, H. Effects of food and body composition on blood alcohol levels. *Comprehensive Handbook of Alcohol Related Pathology*, v. 1-3, n. 4, p. 87–101, 2005. Citado na página 71.

KALANT, H.; Lê, A. Effects of ethanol on thermoregulation. *Pharmacology and Therapeutics*, v. 1, n. 1, p. 71–97, 1986. Citado na página 27.

KARRAS, K. et al. A hardware acceleration platform for ai-based inference at the edge. *Circuits, Systems, and Signal Processing*, v. 39, n. 2, p. 1059–1070, 2020. Citado na página 46.

KHAN, M. et al. Deep unified model for face recognition based on convolution neural network and edge computing. *IEEE Access*, v. 7, n. 2, p. 72622–72633, 2019. Citado na página 45.

KHAN, M.; WARD, R.; INGLEBY, M. Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature. *ACM Transactions on Applied Perception*, v. 6, n. 2009, p. 1–22, 2009. Citado 2 vezes nas páginas 28 e 38.

KHAN, W. et al. Edge computing: A survey. *Future Generation Computer Systems*, v. 97, n. 1, p. 219–235, 2019. Citado 2 vezes nas páginas 41 e 46.

KIMATA, M. Uncooled infrared focal plane arrays. *IEEJ Transactions on Electrical and Electronic Engineering*, v. 13, n. 1, p. 4–12, 2018. Citado na página 38.

KIRIMTAT, A. et al. Flir vs seek thermal cameras in biomedicine: comparative diagnosis through infrared thermography. *BMC Bioinformatics*, v. 21, n. 21, p. 88, 2020. Citado 2 vezes nas páginas 35 e 75.

KITANOV, S.; JANEVSKI, T. State of the art: Fog computing for 5g networks. In: *24th Telecommunications Forum, TELFOR 2016*. Serbia, Belgrade: IEEE, 2017. p. 3–6. Disponível em: <10.1109/TELFOR.2016.7818728>. Acesso em: 13 fev. 2022. Citado na página 44.

KOUKIOU, G. Intoxicated person identification using markov chains and neural networks. *Neural Computing and Applications*, v. 33, n. 8, p. 3459–3467, 2021. Citado na página 28.

KOUKIOU, G.; ANASTASSOPOULOS, V. Drunk person identification using thermal infrared images. *Forensic Science International*, v. 4, n. 4, p. 229, 2012. Citado 8 vezes nas páginas 28, 33, 38, 39, 51, 70, 73 e 75.

KOUKIOU, G.; ANASTASSOPOULOS, V. Neural networks for identifying drunk persons using thermal infrared imagery. *Forensic Science International*, v. 252, n. 1, p. 69–76, 2015. Citado 8 vezes nas páginas 28, 33, 39, 51, 64, 77, 91 e 97.

KOUKIOU, G.; ANASTASSOPOULOS, V. Drunk person screening using eye thermal signatures. *Journal of Forensic Sciences*, v. 61, n. 1, p. 259–264, 2016. Citado 2 vezes nas páginas 39 e 97.

KRISTIANI, E.; YANG, C.; HUANG, C. isec:an optimized deep learning model for image classification on edge computing. *IEEE Access*, v. 8, n. 1, p. 27267 – 27276, 2020. Citado 2 vezes nas páginas 53 e 54.

KRUSCINSKI, F. *A embriaguez ao volante e a autuação pela recusa na realização do teste do etilômetro sem que o condutor apresente sinais de alteração da capacidade psicomotora*. 2016. Jus. Disponível em: <https://jus.com.br/artigos/45742/a-embriaguez-ao-volante-e-a-autuacao-pela-recusa-na-realizacao-do-teste-do-etilometro-sem-que-o-co
Acesso em: 08 fev. 2022. Citado na página 27.

LABIANCA, D. Estimation of blood-alcohol concentration. *Journal of Chemical Education*, v. 69, n. 8, p. 628–632, 1992. Citado na página 32.

LAHIRI, B.; BAGAVATHIAPPAN, S.; AL. et. Medical applications of infrared thermography: A review. *Infrared Physics and Technology*, v. 55, n. 4, p. 221–235, 2012. Citado 2 vezes nas páginas 28 e 36.

LAWFORD, B. et al. Alcohol use disorders identification test (audit) scores are elevated in antipsychotic-induced hyperprolactinaemia. *Journal of Psychopharmacology*, v. 26, n. 2, p. 324–329, 2012. Citado na página 62.

LEGISLATIVO Órgão: Atos do P. *LEI Nº 14.071, DE 13 DE OUTUBRO DE 2020*. 2020. DOU. Disponível em: <https://www.in.gov.br/web/dou/-/lei-n-14.071-de-13-de-outubro-de-2020-282461197>. Acesso em: 05 fev. 2022. Citado na página 27.

LI, G. et al. Stage-wise magnitude-based pruning for recurrent neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, p. 1–15, 2022. Citado na página 66.

LI, H. et al. Pruning filters for efficient convnets. In: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. IEEE, 2016. p. 1–13. Disponível em: <10.1109/ECAI52376.2021.9515097>. Acesso em: 15 mai. 2022. Citado na página 66.

LIANG, T. et al. Sparknoc: An energy-efficiency fpga-based accelerator using optimized lightweight cnn for edge computing. *Neurocomputing*, v. 461, n. 1, p. 370–403, 2021. Citado 2 vezes nas páginas 47 e 98.

LIU, Y.; HO, C. Effects of different blood alcohol concentrations and post-alcohol impairment on driving behavior and task performance. *Traffic Injury Prevention*, v. 11, n. 4, p. 334–341, 2010. Citado na página 32.

LUJIC, I. et al. Increasing traffic safety with real-time edge analytics and 5g. In: *Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking*. New York, NY, USA: ACM, 2021. p. 19–24. Disponível em: <https://dl.acm.org/doi/10.1145/3434770.3459732>. Acesso em: 1 jun. 2020. Citado 2 vezes nas páginas 53 e 55.

MACHADO, et al. Influence of infrared camera model and evaluator reproducibility in the assessment of skin temperature responses to physical exercise. *Journal of Thermal Biology*, v. 98, n. September 2020, p. 102913, 2021. Citado na página 35.

MADAKAM, S.; RAMASWAMY, R.; TRIPATHI, S. Internet of things (iot): A literature review. *Journal of Computer and Communications*, v. 3, n. 5, p. 164–173, 2015. Citado na página 39.

MANSOURI, Y.; BABAR, M. A review of edge computing: Features and resource virtualization. *Journal of Parallel and Distributed Computing*, v. 150, n. 1, p. 155–183, 2021. Citado na página 44.

MARINS, J. et al. Circadian and gender differences in skin temperature in militaries by thermography. *Infrared Physics Technology*, v. 71, n. 1, p. 322–328, 2015. Citado na página 34.

MARINS, J. et al. Time required to stabilize thermographic images at rest. *Infrared Physics Technology*, v. 65, n. 1, p. 30–35, 2014. Citado na página 34.

MARTIN, T. et al. A review of alcohol-impaired driving: The role of blood alcohol concentration and complexity of the driving task. *Journal of Forensic Sciences*, v. 58, n. 5, p. 1238–1250, 2013. Citado na página 31.

MELLINGER, J. Epidemiology of alcohol use and alcoholic liver disease. *Clinical Liver Disease*, v. 5, n. 13, p. 136–139, 2019. Citado na página 31.

METZMACHER, M. et al. Circadian and gender differences in skin temperature in militaries by thermography. *Energy and Buildings*, v. 158, n. 1, p. 1063–1078, 2018. Citado na página 35.

MILTON, A.; BARONE, F.; KRUER, M. Influence of non-uniformity on ir focal plane array performance. *IEE Conference Publication*, v. 24, n. 228, p. 1–5, 1983. Citado na página 38.

MOHAMMED, M.; ZEEBAREE, S.; AL. et. Iot and cloud computing issues, challenges and opportunities: A review. *Qubahan Academic Journal*, v. 1, n. 2, p. 1–7, 2016. Citado na página 40.

MONBURINON, H. et al. Proceedings of 2019 4th international conference on information technology: Encompassing intelligent technology and innovation towards the new era of human life, incit 2019s the new era of human life, incit 2019. In: *Proceedings - International Conference on Distributed Computing Systems*. Dallas, TX, USA: IEEE, 2019. p. 294–299. Disponível em: <https://ieeexplore.ieee.org/document/8885335/>. Acesso em: 08 fev. 2022. Citado na página 44.

MORETTI-PIRES, R.; CORRADI-WEBSTER, C. Adaptação e validação do alcohol use disorder identification test (audit) para população ribeirinha do interior da amazônia, brasil. *Cadernos de Saude Publica*, v. 27, n. 3, p. 497–509, 2011. Citado na página 62.

NETO, F. S.; SANTOS, E. L. *Lei nº 12.760/2012: a nova Lei Seca. 2012*. 2012. Jus. Disponível em: <https://jus.com.br/artigos/23321/lei-n-12-760-2012-a-nova-lei-seca>. Acesso em: 05 fev. 2022. Citado na página 27.

NKENGNE, A.; PAPILLON, A.; BERTIN, C. Evaluation of the cellulite using a thermal infra-red camera. *Skin Research and Technology*, v. 19, n. 1, p. 1–7, 2013. Citado na página 35.

OMETOV, A. et al. A survey of security in cloud, edge, and fog computing. *Sensors*, v. 22, n. 3, p. 1–27, 2022. Citado 3 vezes nas páginas 41, 42 e 43.

PAHAR, M. et al. Covid-19 detection in cough, breath and speech using deep transfer learning and bottleneck features. *Computers in Biology and Medicine*, v. 141, n. 141, p. 105153, 2022. Citado na página 65.

PARIKH, S. et al. Security and privacy issues in cloud, fog and edge computing. *Procedia Computer Science*, v. 160, n. 1, p. 734–739, 2019. Citado na página 40.

PETROSINO, L. et al. Image sensors and vpu acceleration for data analysis and classification. In: *2021 IEEE International Workshop on Metrology for Industry 4.0 IoT (MetroInd4.0IoT)*. Rome, Italy: IEEE, 2021. p. 392–396. Disponível em: <https://dl.acm.org/doi/10.1145/3434770.3459732>. Acesso em: 1 jun. 2020. Citado na página 53.

PUCHTLER, P.; PEINL, R. Deep learning in the era of edge computing: Challenges and opportunities. In: _____. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, 2020. p. 320–326. Disponível em: <https://doi.org/10.1007/978-3-030-58285-2_29>. Acesso em: 17 mai. 2022. Citado 2 vezes nas páginas 53 e 54.

REILLY, T.; WATERHOUSE, J. Circadian aspects of body temperature regulation in exercise. *Journal of Thermal Biology*, v. 34, n. 4, p. 161–170, 2009. Citado na página 34.

ROBERTA, G.; RESPLANDES, A. *Trânsito , embriaguez e os avanços na fiscalização viária Trânsito , embriaguez e os avanços na fiscalização viária trânsito frente à inexistência de mecanismos de aferição imediata*. 2019. Jus. Disponível em: <https://jus.com.br/artigos/73678/transito-embriaguez-e-os-avancos-na-fiscalizacao-viaria>. Acesso em: 06 fev. 2022. Citado 2 vezes nas páginas 27 e 31.

RODRíGUEZ, A. et al. Fpga-based high-performance embedded systems for adaptive edge computing in cyber-physical systems: The artico³ framework. *Sensors (Basel, Switzerland)*, v. 18, n. 6, p. 1877, 2018. Citado na página 46.

SAEZ, J.; ROMERO-BéJAR, J. Impact of regressand stratification in dataset shift caused by cross-validation. *Mathematics*, v. 10, n. 14, p. 1–14, 2022. Citado na página 65.

SAINATH, T. et al. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC, Canada: IEEE, 2013. p. 6655–6659. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6638949>. Acesso em: 15 fev. 2022. Citado na página 48.

SERWAY, R.; KIRKPATRICK, L. Physics for scientists and engineers with modern physics. *The Physics Teacher*, v. 26, n. 4, p. 254–255, 1988. Citado 2 vezes nas páginas 32 e 33.

SHI, W.; DUSTDAR, S. The promise of edge computing. *Computer*, v. 49, n. 5, p. 78–81, 2016. Citado 2 vezes nas páginas 39 e 43.

SILVA, D. O. et al. Accidentes de circulación y su asociación con el consumo de bebidas alcohólicas. *Enfermería Global*, v. 17, n. 52, p. 365–400, 2018. Citado na página 31.

SILVA, W. Dias da. *Classificação de embriaguez a partir de imagens faciais usando redes neurais convolucionais: Uma abordagem baseada em respostas fisiológicas induzidas pelo álcool*. 95 p. Dissertação (Mestrado) — Universidade Federal de São Carlos, 2021. Citado 8 vezes nas páginas 17, 29, 46, 51, 64, 65, 78 e 84.

SOCOLINSKY D. SELINGER, A. Thermal face recognition over time. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Cambridge, UK: IEEE, 2004. p. 187–190. Disponível em: <https://ieeexplore.ieee.org/document/1333735>. Acesso em: 10 mar. 2020. Citado na página 38.

SOLTUZ, A.; NEAGOE, V. Facial thermal image analysis with deep convolutional neural network architectures for subject dependent drunkenness diagnosis. In: *Proceedings of the 13th International Conference on Electronics, Computers and Artificial Intelligence, ECAI 2021*. Pitesti, Romania: IEEE, 2021. p. 1–4. Disponível em: <10.1109/ECAI52376.2021.9515097>. Acesso em: 1 jun. 2020. Citado 2 vezes nas páginas 29 e 90.

SONKUSARE, S. et al. Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking. *Scientific Reports*, v. 9, n. 1, p. 1–11, 2019. Citado na página 34.

SRIKANTH, G.; JAFFRIN, L. Security issues in cloud and mobile cloud: A comprehensive survey. *Information Security Journal: A Global Perspective*, v. 00, n. 00, p. 1–25, 2022. Citado na página 40.

SUFIAN, A. et al. A survey on deep transfer learning to edge computing for mitigating the covid-19 pandemic: Dtl-ec. *Journal of Systems Architecture*, v. 108, n. May, p. 101830, 2020. Citado na página 44.

TATTERSALL, G. Infrared thermography: A non-invasive window into thermal physiology. *Comparative Biochemistry and Physiology Part A: Molecular Integrative Physiology*, v. 202, n. 1, p. 78–98, 2016. Citado na página 33.

VARGHESE, B. et al. Challenges and opportunities in edge computing. In: *Proceedings - 2016 IEEE International Conference on Smart Cloud, SmartCloud 2016*. New York, NY, USA: IEEE, 2016. p. 20–26. Disponível em: <http://ieeexplore.ieee.org/document/7796149/>. Acesso em: 11 mar. 2022. Citado 2 vezes nas páginas 40 e 43.

VILLA, E.; ARTEAGA-MARRERO, N. Performance assessment of low-cost thermal. *Sensors*, v. 20, n. 1321, p. 1–16, 2020. Citado na página 75.

WANG, L.; YOON, K. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 44, n. 6, p. 1789–1819, 2022. Citado 2 vezes nas páginas 48 e 49.

WANG, X. et al. Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Communications Surveys and Tutorials*, v. 22, n. 2, p. 869–904, 2020. Citado na página 41.

WARD, R. et al. Flunet: An ai-enabled influenza-like warning system. *IEEE Sensors Journal*, v. 21, n. 21, p. 24740–24748, 2021. Citado na página 53.

WATSON, P.; WATSON, I.; BATT, R. Prediction of blood alcohol concentrations in human subjects. updating the widmark equation. *Journal of Studies on Alcohol*, v. 42, n. 7, p. 547–556, 1981. Citado 2 vezes nas páginas 71 e 72.

WEAFER, J.; FILLMORE, M. Acute tolerance to alcohol impairment of behavioral and cognitive mechanisms related to driving: Drinking and driving on the descending limb. *Psychopharmacology*, v. 220, n. 4, p. 697–706, 2012. Citado na página 32.

WENBO, Z.; ZHANG, J.; ROMAGNOLI, J. General feature extraction for process monitoring using transfer learning approaches. *Industrial and Engineering Chemistry Research*, v. 61, n. 15, p. 1586–1594, 2020. Citado na página 65.

WHO. *Alcohol*. 2018. WHO.int. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/alcohol>. Acesso em: 05 fev. 2022. Citado 2 vezes nas páginas 27 e 31.

WHO. *Global status report on alcohol and health*. 2018. WHO.int. Disponível em: <https://apps.who.int/iris/bitstream/handle/10665/274603/9789241565639-eng.pdf?ua=1>. Acesso em: 05 fev. 2022. Citado 2 vezes nas páginas 27 e 31.

WONG, T.; YEH, P. Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, v. 32, n. 8, p. 1586–1594, 2020. Citado na página 65.

WU, H. et al. Integer quantization for deep learning inference: Principles and empirical evaluation. p. 1–20, 2020. Disponível em: <http://arxiv.org/abs/2004.09602>. Citado 2 vezes nas páginas 50 e 51.

XIA, M. et al. Sparknoc: An energy-efficiency fpga-based accelerator using optimized lightweight cnn for edge computing. *Journal of Systems Architecture*, v. 115, n. 1, p. 101991, 2021. Citado na página 46.

XIAO, Y.; JIA, Y.; LIU, C. Edge computing security: State of the art and challenges. *Proceedings of the IEEE*, v. 107, n. 8, p. 1608–1631, 2019. Citado na página 41.

XIAO, Y. et al. Edge computing security: State of the art and challenges. In: *Proceedings of the IEEE*. IEEE, 2019. p. 1608–1631. Disponível em: <https://ieeexplore.ieee.org/document/8741060/>. Acesso em: 08 fev. 2022. Citado na página 44.

XU, X. et al. A computation offloading method over big data for iot-enabled cloud-edge computing. *Future Generation Computer Systems*, v. 95, n. 3, p. 522–533, 2019. Citado na página 44.

YANG, R. et al. Integrated blockchain and edge computing systems: A survey, some research issues and challenges. *IEEE Communications Surveys and Tutorials*, v. 21, n. 2, p. 1508–1532, 2019. Citado na página 44.

YODA, T. et al. Effects of alcohol on thermoregulation during mild heat exposure in humans. *Alcohol*, v. 36, n. 3, p. 195–200, 2005. Citado na página 32.

YOUSEFPOUR, A. et al. All one needs to know about fog computing and related edge computing paradigms: A complete survey. *Journal of Systems Architecture*, v. 98, n. 3, p. 289–330, 2019. Citado na página 41.

YU, W. et al. A survey on the edge computing for the internet of things. *IEEE Access*, v. 6, n. 1, p. 6900–6919, 2018. Citado na página 41.

ZAKHARI, S. Overview: How is alcohol metabolized by the body? *Alcohol Research and Health*, v. 29, n. 4, p. 245–254, 2006. Citado na página 32.

ZEYU, H. et al. Survey on edge computing security. In: *2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*. Fuzhou, China: IEEE, 2020. p. 96–105. Disponível em: <https://ieeexplore.ieee.org/document/9196442/>. Acesso em: 15 mar. 2022. Citado na página 41.

ZHANG, M. et al. Deep learning in the era of edge computing: Challenges and opportunities. In: _____. *Fog Computing*. Wiley Online Library, 2020. p. 67–78. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/9781119551713.ch3>. Acesso em: 05 fev. 2022. Citado na página 42.

ZHANG, X.; WANG, Y.; LU, S. Openei: An open framework for edge intelligence. In: *Proceedings - International Conference on Distributed Computing Systems*. Dallas, TX, USA: IEEE, 2019. p. 1840–1851. Disponível em: <https://ieeexplore.ieee.org/document/8885335/>. Acesso em: 08 fev. 2022. Citado na página 44.