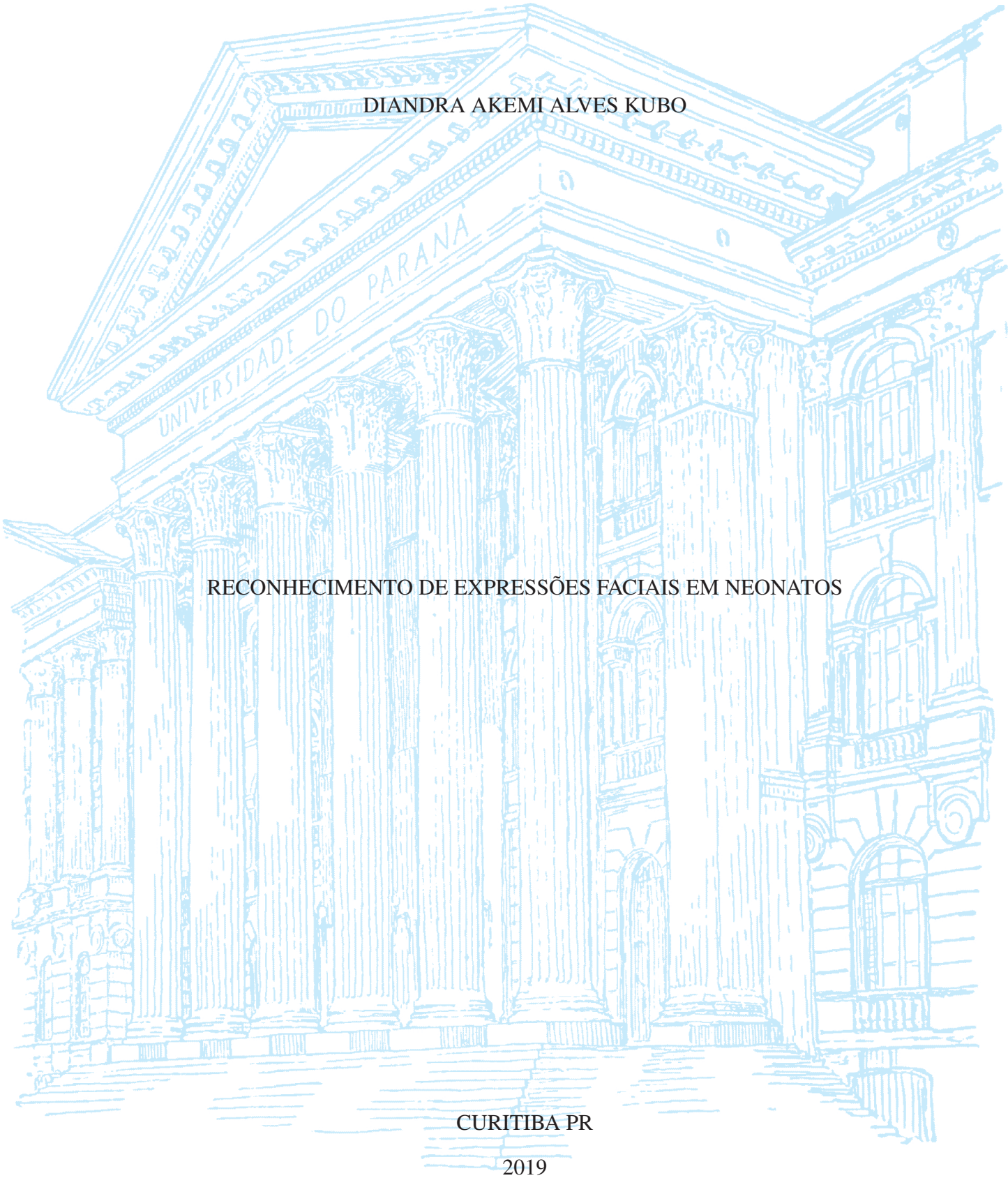
UNIVERSIDADE FEDERAL DO PARANÁ

DIANDRA AKEMI ALVES KUBO



RECONHECIMENTO DE EXPRESSÕES FACIAIS EM NEONATOS

CURITIBA PR

2019

DIANDRA AKEMI ALVES KUBO

RECONHECIMENTO DE EXPRESSÕES FACIAIS EM NEONATOS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática, no Programa de Pós-Graduação em Informática, setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Profª Drª Olga R. P. Bellon.

CURITIBA PR

2019

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

K95r Kubo, Diandra Akemi Alves
Reconhecimento de expressões faciais em neonatos [recurso eletrônico] / Diandra Akemi Alves Kubo. – Curitiba, 2019.

Dissertação – Universidade Federal do Paraná - Setor de Ciências Exatas - Programa de Pós-Graduação em Informática, 2019.

Orientadora: Olga Regina Pereira Bellon.

1. Software. 2. Recém-nascidos. 3. Aprendizado do computador. 4. Neonatologia. I. Universidade Federal do Paraná. II. Bellon. Olga Regina Pereira. III. Título.

CDD - 006.42

Bibliotecária: Vanusa Maciel CRB- 9/1928

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIENCIAS EXATAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO INFORMÁTICA - 40001016034P5

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **DIANDRA AKEMI ALVES KUBO** intitulada: **Reconhecimento de Expressões Faciais em Neonatos**, sob orientação do Prof. Dr. OLGA REGINA PEREIRA BELLON, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

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

CURITIBA, 30 de Outubro de 2019.

OLGA REGINA PEREIRA BELLON
Presidente da Banca Examinadora (UNIVERSIDADE FEDERAL DO PARANÁ)

LUCIANO SILVA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

HENRIQUE SERGIO GUTIERREZ DA COSTA
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

MINISTÉRIO DA EDUCAÇÃO - UNIVERSIDADE FEDERAL DO PARANÁ - Setor de Ciências Exatas - Coordenação de Pós-Graduação em Informática

Aos meus pais, Stela e Milton.

AGRADECIMENTOS

A Deus.

A professora Olga, orientadora, ouvinte, conselheira e amiga desde a graduação, pela paciência, compreensão e sabedoria passada.

A minha família, principalmente meus pais, que sempre incentivaram minha carreira acadêmica e garantiram que eu pudesse ter as melhores oportunidades de estudo.

Aos meus amigos de Curitiba, Rafael, Gabriel, Manuela, Talita e Paulo, que estão comigo pelo que parece ser sempre, me apoiando e incentivando desde o começo. Aos amigos de informática biomédica, Thiago e Lucas, melhores companheiros de curso fora do curso que eu tive. Aos amigos do IMAGO que me ajudaram sempre que precisei. Aos amigos do SIMEPAR, Camila, Jorge, Renan, Tiago, Mariana e Alana que foram essenciais no incentivo e apoio no começo do mestrado.

Ao Artur, pela sinceridade e honestidade que sempre me incentivaram a dar meu melhor.

As minhas amigas Priscila, Lara e Rhaissa, que me acolheram e são minha família em São Paulo, me incentivando, me ouvindo e me motivando sempre que preciso. Ao Felipe, Karina, e Adelmo pelos empurrões pra que esse trabalho saísse e por também serem minha família. Ao Tiago, quem esse trabalho não teria saído não fosse sua ajuda. Ao Helio, pela compreensão e parceria.

RESUMO

A avaliação de dor é uma tarefa difícil e complexa, que é particularmente importante para recém-nascidos, que não conseguem verbaliza-la de maneira adequada e são vulneráveis a danos cerebrais decorrentes do não tratamento da dor. As ferramentas utilizadas no ambiente clínico para auxiliar na avaliação de dor requerem treinamento dos profissionais de saúde que irão utilizá-las, e seu uso é afetado pelo viés no reconhecimento da dor de cada indivíduo. Por essa razão, esforços tem sido colocados em automatizar essa tarefa, e uma das maneiras de fazê-lo é analisando a expressão facial do neonato, uma vez que esta é comprovadamente correlacionada à dor. Nessa dissertação, as diferenças entre os principais trabalhos em reconhecimento automático de expressão facial de neonatos são apresentadas, examinando os métodos utilizados, bases de dados e performances dos sistemas. Com isso em mente, testamos os principais métodos utilizados com objetivo de comparar suas performances mais a fundo. Esse estudo também avança o entendimento da base de dados COPE, a única base de dados de expressão facial de neonatos publicamente disponível. Conduzimos testes com métodos off the shelf para detecção de face, e em 54% das imagens nenhuma face foi detectada, reforçando a necessidade do desenvolvimento de sistemas específicos para recém-nascidos ou mais robustos à mudanças de público. Desde a publicação da base COPE em 2005, avanços significativos foram alcançados na área de processamento de imagens, e por essa razão comparamos métodos clássicos de extração de características em processamento de imagens com características provenientes de redes neurais convolucionais (CNNs), que são consideradas estado da arte para a maioria das aplicações de visão computacional. Um delta de 19% foi observado entre os filtros de gabor (melhor dos métodos clássicos) e características da ResNet50 (melhor das CNNs). Também testamos a robustez dos métodos a ruído, um fator importante em problemas de visão computacional onde devem ser considerados cenários da vida real. Para os métodos clássicos, foi observado um delta menor na performance entre cenários limpos e ruidosos, mas de maneira geral a performance foi pior que das CNNs. Em adição, estressando a performance das CNNs, testamos quais camadas produziriam melhor performance, na tentativa de verificar se camadas mais rasas poderiam ter desempenho igual ou melhor que camadas mais profundas, o que significaria menor custo computacional. Os resultados mostraram melhores resultados utilizando as camadas mais profundas. De maneira geral, estudando a literatura da área notamos uma tendência na utilização de métricas enviesadas, como acurácia, em um campo onde uma visão mais completa de performance de modelos deveria ser utilizada, por se tratar de um público tão vulnerável. Por fim, também observamos uma dificuldade no acesso as bases da literatura. Nossos esforços reforçam o potencial da utilização de métodos de visão computacional, porém fora limitados à base de dados utilizada.

Palavras-chave: Expressões faciais, avaliação de dor, visão computacional

ABSTRACT

Pain evaluation is a difficult and complex task, that is particularly important for newborns, who cannot verbalize it properly and are vulnerable to cerebral damage due to untreated pain. The current pain assessment tools used in clinical settings require extensive training for the caregivers and can be affected by each individual's bias towards pain recognition. For this reason, efforts have been made to automate this task, and one of the ways to do so is analyzing the newborn's facial expression, that has been proved to correlate with pain. In this dissertation, the differences among the most prominent works in automatic neonatal facial expression recognition were outlined, examining methods used, databases and final performance. With this in mind, we tested main methods used to compare their performances more in depth. This study also advances the understanding of the COPE database, the only publicly available newborn facial expression database. We conducted a test with off the shelf methods for face detection, and found that in 54% of the images, no face was found, reinforcing the need to develop either tailored applications or more robust ones. Since the COPE database was published, in 2005, significant advances in image processing have been made, and for this reason, we compared classical image processing feature extraction methods with Convolutional Neural Networks (CNNs), that are considered to be state of the art for most computer vision problems. We saw a difference of 19% in recall when using gabor filters (best of classical methods) and then the ResNet50 features (best of CNNs). We also tested the methods in regards to robustness to image noise, an important factor for computer vision problems when real world scenarios are considered. We found that image processing methods had a smaller delta in performance from clean to noisy scenarios, but had overall poor performance. In addition, stressing the CNNs performance, we also studied which layers yielded best performance in order to verify if shallow layers could produce the same results as deeper ones for this application, meaning less computational cost, but our test showed superior performance in deeper layers. Overall, studying the literature we noticed a tendency to use biased metrics, such as accuracy, in a field where a more complete view of model performance should be used. Moreover, we also found it very difficult to access data for this field. Our findings reinforce the potential of more complex computer vision methods, but are limited to the dataset that was used.

Keywords: Facial expression, pain assessment, computer vision

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A dor pode ser definida como uma experiência sensorial e emocional associada a algum tipo de dano tecidual (Maxwell et al., 2013). A tarefa de avaliá-la de forma correta e segura é um grande desafio para profissionais de saúde e responsáveis pelo cuidado de outras pessoas (Turk e Melzack, 2011). Em neonatos, (Schechter et al., 2002) afirma que a avaliação de dor pode ser mais complexa e ainda necessita muita pesquisa.

Até aproximadamente 1940 acreditava-se que neonatos não tinham a capacidade de experienciar dor, e portanto não recebiam tratamento algum para a mesma. Procedimentos dolorosos eram feitos sem analgesia, e em alguns casos nem mesmo anestesia. Apenas em 1987 a Academia Americana de Pediatria publicou um trabalho estabelecendo o novo consenso de que bebês deveriam receber anestesia e um tratamento adequado para a dor, reconhecendo que o sistema nervoso de neonatos é capaz de sentir dor, e portanto, demonstrá-la (Anand et al., 2007).

1.1 DESAFIOS

(Maxwell et al., 2013) enfatiza que: "A inabilidade de comunicar verbalmente a dor não nega a possibilidade do indivíduo estar experienciando-a e precisando de tratamento". Assim como afirma (Hummel e van Dijk, 2006), avaliação de dor em pacientes de qualquer idade que não conseguem se comunicar continua sendo uma tarefa desafiadora, mesmo com os avanços recentes em técnicas de detecção de dor.

Utilizar indicadores comportamentais de dor é a prática mais acessível num ambiente clínico. Porém, depende da capacidade do neonato em expressar a dor, o que pode não ser possível dependendo do quadro clínico sendo experienciado (Hummel e van Dijk, 2006).

Além disso, essa classe de indicadores também depende da pessoa que observa o paciente, pois sua mensuração pode ser extremamente subjetiva (Guinsburg e Cuenca, 2010).

1.2 MOTIVAÇÕES

Apesar de existirem poucos dados empíricos focando nos efeitos a longo prazo da dor física precoce, estudos mostram que neonatos são muito vulneráveis aos possíveis danos cerebrais que podem ocorrer em decorrência da mesma (Maxwell et al., 2013).

Assim como afirma (Guinsburg e Cuenca, 2010), existe uma linguagem definida de dor em neonatos. Ela é expressada majoritariamente por meio de expressões faciais, movimentos corporais, frequência cardíaca, saturação de oxigênio, pressão arterial e frequência respiratória.

As ferramentas utilizadas para avaliação de dor em sua maioria têm alto grau de subjetividade, por tratar de indicadores comportamentais e por serem influenciadas pelo fator de julgamento humano. Assim como afirma (Hummel e van Dijk, 2006), historicamente o cuidado de neonatos focou em sobrevivência, não reconhecimento e tratamento da dor. Atualmente, avaliação de dor baseia–se grandemente em indicadores comportamentais.

(Prkachin, 2009) afirma que o caminho que a análise de expressões faciais deve seguir é o da análise automática, justamente na tentativa de diminuir a subjetividade inerente ao processo.

1.3 PROPOSTA

Este trabalho tem como objetivo estudar uma solução para monitoramento de expressões faciais de neonatos, para que a expressão de dor possa ser reconhecida automaticamente. Para tanto, métodos de extração de características de duas grandes modalidades são testados: de processamento de imagens tradicional e de visão computacional.

1.4 ORGANIZAÇÃO

No Capítulo 2, os conceitos necessários de avaliação de dor, reconhecimento de dor e sistemas automáticos são apresentados. Já no Capítulo 3, os trabalhos na área de reconhecimento automático de expressões faciais em neonatos são estudados. Os materiais utilizados nesse estudo, bem como os métodos testados são apresentados no Capítulo 4. Por fim, resultados são exibidos no Capítulo 5, e a conclusão no Capítulo 6.

2 FUNDAMENTAÇÃO TEÓRICA

O presente capítulo apresenta os conceitos necessários para o desenvolvimento de um sistema que automaticamente reconhece expressões faciais em neonatos. A Seção 2.1 apresenta os fundamentos de avaliação de dor, base para Seção 2.2, onde algumas escalas de dor são apresentadas. Na Seção 2.3, os fundamentos de sistemas de reconhecimento automático de expressões faciais de dor são apresentados.

2.1 AVALIAÇÃO DE DOR

Assim como afirma (Devlin et al., 2018), padrões de dor são altamente individuais e podem ter origem nas mais diversas fontes, sendo assim um tópico abrangente e de alta complexidade. Essa complexidade aumenta quando os pacientes sendo tratados estão sob alguma condição especial que os impedem de comunicar-se, que alteram o estado mental, que implicam no uso de ventilação mecânica ou uso de dispositivos invasivos, que atrapalham o sono ou ainda sua mobilidade.

Todas essas condições agravantes tem ocorrência de alta frequência em ambientes como UTIs neonatais, principalmente com pacientes prematuros, que fazem parte de um corte populacional com alto número de procedimentos dolorosos realizados regularmente (Cremillieux et al., 2018).

Em outro ambiente de tratamento de dor, como nos departamentos de atendimento de emergência (pronto-socorro), (Walters, 2018) aponta que os níveis de dor são frequentemente subestimados quando se tratam de atendimentos pediátricos, e que apenas 50% desses pacientes recebem qualquer forma de analgesia no pronto-socorro. No mesmo estudo ainda é apontado que para mesmas causas de dor reportadas entre crianças e adultos, para crianças, quando analgesia é administrada, geralmente é administrada em dose proporcionalmente menor.

Exposições prolongadas a dor e também à administração de opioides em recém nascidos foram evidenciadas como danosas e tendo diversos efeitos no desenvolvimento neurológico tanto no curto quanto no longo prazo (Boyle et al., 2018). Tal afirmação reforça ainda mais a importância do cuidado dessa população vulnerável.

Como dependem completamente do cuidado de terceiros, quando hospitalizados, o cuidado de recém nascidos geralmente é designado à enfermeiros. Existem vários estudos que verificam a concordância entre os níveis de intensidade de dor reportados por pacientes e os níveis de dor reportados por enfermeiros, com a maioria dos resultados indicando correlação entre ambos. Esses estudos são apenas possíveis de serem realizados com pacientes adultos ou com capacidade de reportar dor por si mesmos (Dequeker et al., 2018), enquanto nos casos de recém-nascidos, a comparação da avaliação geralmente é feita comparando-se com escalas de dor ou indicadores de dor. Para auxiliar na quantificação e evidência da dor, alguns indicadores mostram-se úteis e serão discutidos. De forma geral, existem três tipos de indicadores de dor que podem ser considerados nessa avaliação: fisiológicos, biomarcadores e comportamentais (Maxwell et al., 2013).

2.1.1 Indicadores fisiológicos

Algumas respostas automáticas produzidas pelo corpo humano podem ser analisadas como indicadores de eventos sendo observados, como a dor. (Santos et al., 2012) afirma que a dor ativa mecanismos do sistema nervoso autônomo como forma de compensação da mesma.

A frequência cardíaca é uma resposta comumente analisada no processo de avaliação de dor. Observa-se um aumento na frequência quando recém nascidos a termo passam por procedimentos dolorosos, como por exemplo punção do calcanhar. Além de estar relacionada com a presença de dor, estudos mostram também que a magnitude da mudança na frequência pode ser associada a intensidade da dor e duração desse estímulo doloroso (Anand et al., 1987). Mesmo para recém nascidos prematuros essas mudanças acontecem (Fitzgerald e McIntosh, 1989).

Outro indicador muito utilizado é a frequência respiratória. Esta também tem correlação diretamente proporcional à dor assim como a frequência cardíaca, de maneira que uma frequência respiratória elevada está associada a presença de dor (Guinsburg e Cuenca, 2010).

A saturação de oxigênio também é um indicador fisiológico, que tem uma correlação inversamente proporcional à presença de dor: sua redução indica possível presença de dor (Fitzgerald e McIntosh, 1989). Em compensação, o aumento na pressão arterial também pode indicar presença de dor (Hummel e van Dijk, 2006).

Contudo, apesar de termos esses indicadores, existem inúmeros fatores, principalmente no contexto de neonatos, que podem induzir esses mesmos comportamentos fisiológicos sem a presença de dor. Por exemplo, o fato do neonato estar sobre ventilação mecânica ou intervenção farmacológica (Maxwell et al., 2013) pode afetar sua frequência cardíaca e respiratória. (Presbytero et al., 2010) afirma: “Os parâmetros fisiológicos parecem úteis para avaliar a dor na prática clínica, mas, em geral, não podem ser usados de forma isolada para decidir se o recém nascido apresenta dor”. Esses indicadores são utilizados para verificar a presença ou ausência de dor, e quando utilizados em conjunto podem ser ferramentas auxiliadoras no processo de tomada de decisão da avaliação de dor, levando apenas em consideração que se a estimativa de intensidade de dor é essencial, outros tipos de indicadores devem ser utilizados em conjunto.

2.1.2 Biomarcadores

(Marchi et al., 2009) afirma que biomarcadores são: “ bioquímicos específicos no corpo com uma característica molecular bem definida que os fazem úteis no diagnóstico e no acompanhamento de doenças, e na determinação de progresso de tratamentos”. Devido a essa natureza, biomarcadores também são utilizados no diagnóstico de dor.

O nível de cortisol é um biomarcador hormonal, que é geralmente utilizado em estudos clínicos pois é comprovadamente relacionado ao nível de stress (Maxwell et al., 2013). Outro biomarcador utilizado, o ácido úrico, aparece no organismo humano como resultado da degradação de ATP em subprodutos de purina em decorrência de um aumento no consumo de oxigênio. Estudo mostram o aumento na concentração de ácido úrico em procedimentos de punção do calcanhar em recém nascidos (Slater et al., 2012).

Outro biomarcador discriminante de dor é a medida de condutância da pele. A coleta dessa informação é dada por meio de eletrodos colocados na superfície palmar ou plantar do neonato, que são capazes de captar a variação na condutância elétrica causada pela variação de suor, produzida pela liberação de certas substâncias quando a dor ocorre (Jesus, 2011).

Esses indicadores biológicos apresentam resultados altamente confiáveis na avaliação de existência de dor. Contudo, biomarcadores em geral tem um tempo de obtenção prolongado, sendo mais indicados para estudos clínicos ou investigações pós fato com uma janela de tempo

de trabalho maior. Em adição, por demandarem coleta de materiais, e análise laboratorial, a maioria desses métodos tem um custo elevado (Guinsburg e Cuenca, 2010; Hummel e van Dijk, 2006), e, principalmente, são invasivos, podendo causar ainda mais dor aos pacientes.

2.1.3 Indicadores comportamentais

Indicadores comportamentais apresentam respostas mais específicas à estímulos dolorosos, tendo assim grande importância. Assim como (Guinsburg e Cuenca, 2010) afirma, os mais estudados são a expressão facial, o choro e o movimento corporal. O movimento corporal, apesar de ser diminuído em neonatos prematuros em relação a neonato a termo (Craig et al., 1993), é um componente importante na avaliação de do r. A rigidez do tórax e alguns movimentos das extremidades podem caracterizar presença de dor, bem como a presença da mão espalmada seguida de um abrupto fechamento das mãos. Reações corporais produzidas por procedimentos táteis são mais significativas que reações faciais e fisiológicas, e também duram mais, tendo uma probabilidade maior de serem reconhecidas por um profissional de saúde responsável pelo cuidado de um neonato (Holsti et al., 2005).

(Pereira et al., 1999) afirma que o choro é a principal forma de comunicação de neonatos, pois serve como mensagem de angústia para a pessoa responsável pelo cuidado dele. Apesar do fato de que cerca de 50% dos recém-nascidos não choram durante procedimento doloroso, (Guinsburg e Cuenca, 2010) enfatiza a existência do choro específico da dor. Seria este caracterizado por uma fase expiratória prolongada, uma tonalidade aguda, perda do padrão melódico característico e maior duração, ao contrário do choro normal definido por uma frequência de 80 dB, e a combinação de uma fase expiratória, inspiração, período de descanso, e por fim, uma fase expiratória. Contudo, essa diferenciação pode ser complexa, e o choro pode ser provocado por outros estímulos não dolorosos (Guinsburg, 1999).

A expressão facial é considerada o indicador mais sensível à dor em neonatos, como afirma (Hummel e van Dijk, 2006). Atividade completa na face e o conjunto de alguns movimentos particulares são associados à dor. Em especial (Guinsburg e Cuenca, 2010) enfatiza a fronte saliente, a fenda palpebral estreitada, o sulco naso-labial marcado, lábios entreabertos, boca estirada, língua tensa e tremor de queixo.

Apesar de serem consolidados, esses indicadores comportamentais sofrem influência direta da interpretação de quem os observa (Guinsburg e Cuenca, 2010). Em um estudo com 405 participantes, onde cada um recebeu a tarefa de classificar qual foto de neonato apresentava uma expressão de dor, (Balda et al., 2009) relata que:

> "Constatou-se um menor número de acertos para os entrevistados sem parceiro fixo, com maior número de filhos, renda per capita elevada, atuação profissional na área da saúde e escolaridade inferior a 16 anos ou com atuação profissional em outras áreas que não a da saúde e escolaridade superior a 16 anos. Ou seja, os entrevistados detentores dessas características tiveram maior dificuldade para reconhecer a expressão facial de dor do recém-nascido."

Pais tem um papel essencial no processo de avaliação de dor, uma vez que conhecem seus filhos melhor que profissionais de saúde temporários (Maxwell et al., 2013). Sem treinamento médico, pais se mostram capazes de detectar a presença de dor em concordância com médicos e auxiliares de enfermagem (Balda et al., 2000), porém diferem de opinião em relação a intensidade da dor detectada (Elias et al., 2008). Essa diferença na avaliação feita por adultos reforça a dificuldade no processo decisório da avaliação de dor.

(Prkachin et al., 2001) conduziram dois estudos com quatro grupos de pessoas, buscando a resposta dos participantes à um vídeo de pacientes com dor no ombro passando por avaliações de fisioterapia. No primeiro estudo, participantes foram divididos entre aqueles com história familiar de doença crônica, e aqueles sem. No segundo estudo, a divisão foi feita entre participante sem experiência em problemas de dor e terapeutas com vasta experiência em problemas de dor. No primeiro estudo, ficou evidenciado que os participantes com história de doença crônica na família atribuíram maior nível de dor aos vídeos, e no segundo estudo, os terapeutas profissionais atribuíram níveis de dor menores que os dos participantes controle. Esse estudo mostra o que também afirma (Brahnam et al., 2005): profissionais de saúde criam uma certa insensitividade à expressão de dor por serem expostos frequentemente à ocasiões de sofrimento.

De forma a auxiliar na diminuição da subjetividade na avaliação de dor, muitas escalas de dor são utilizadas no ambiente clínico, tentando quantificar ou detectar indicadores comportamentais e fisiológicos de dor.

2.2 ESCALAS DE DOR

Em um estudo feito em 2003, (Chermont et al., 2003) mostrou que apenas um terço dos participantes, médicos pediatras, conheciam alguma escala para avaliar dor em neonatos. Mesmo que um número baixo de pediatras as conheçam, nas últimas décadas várias ferramentas têm sido desenvolvidas (Hummel e van Dijk, 2006), sendo essenciais para um diagnóstico preciso e um tratamento adequado (Arias e Guinsburg, 2012).

Especificamente para neonatos, as escalas mais utilizadas para auxiliar a avaliação de dor por profissionais de saúde são: *Neonatal Infant Pain Scale (NIPS)*, *Face, Legs, Activity, Cry, Consolability Scale (FLACC)* e *Neonatal Facial Coding System (NFCS)*. Além de serem altamente utilizadas, essas escalas também tem extensa literatura verificando sua eficácia e usabilidade (Grunau et al., 1998; Manworren e Hynan, 2003; Breaus et al., 2001; Peters et al., 2002).

2.2.1 *FLACC*

Assim como afirma (Voepel-Lewis et al., 1997), com simplicidade em mente, a escala *FLACC* foi criada para ranquear a situação de cinco variações comportamentais em relação a:

1. face;
2. pernas;
3. atividade;
4. choro;
5. consolabilidade.

Na Figura 2.1, os detalhes da pontuação para cada componente sendo analisado podem ser melhor entendidos.

2.2.2 *NIPS*

A *Neonatal Infant Pain Scale* foi criada baseada na opinião de 43 profissionais de saúde experientes em avaliação de dor (Hudson-Barr et al., 2002), que responderam a uma pesquisa listando os principais pontos que eles próprios utilizavam em suas avaliações. Após estudo de

Categorias	Pontuação		
	0	1	2
F Face	Sem expressão particular ou sorriso	Presença ocasional de careta ou sobrancelhas salientes, introspecção, desinteresse. Parece triste ou preocupado	Sobrancelhas esporadicamente ou constantemente salientes, mandíbulas cerradas, queixo trêmulo. Face aparentando estresse: expressão assustada ou de pânico
P Pernas	Posição normal ou relaxada	Desconforto, inquietação, tensão. Tremores ocasionais	Chutes ou pernas soltas. Aumento considerável da espasticidade, tremores constantes ou sacudidelas
A Atividade	Em silêncio, posição normal, *movimentando-se* facilmente	Contorcendo-se, movimentando o corpo para frente e para trás, tensão. Moderadamente agitado (*por exemplo*, movimento da cabeça para a frente e para trás, comportamento agressivo); respiração rápida, superficial, suspiros intermitentes	Corpo arqueado, rígido ou trêmulo. Agitação intensa, cabeça chacoalhando (não vigorosamente), tremores, *respiração presa em* gaspingou inspiração profunda, intensificação da respiração rápida e superficial
C Choro	Sem choro (acordado ou dormindo)	Gemidos ou lamúrias, reclamações ocasionais. Impulsos verbais ou grunhidos ocasionais	Choro regular, gritos ou soluços, reclamações frequentes. Repetidos impulsos verbais, grunhidos constantes
C Consolabilidade	Contente, relaxado	Tranquilizado por toques ocasionais, abraços ou conversa e distração	Difícil de consolar ou confortar. Rejeita o cuidador, resiste ao cuidado ou a medidas de conforto

Figura 2.1: Esquema de pontuação da escala *FLACC*. Fonte: (Bussotti et al., 2015)

gravações de procedimentos dolorosos (Lawrence et al., 1993), os parâmetros decididos como discriminantes foram:

1. expressão facial;
2. choro;
3. respiração;
4. posição das pernas;
5. posição dos braços;
6. estado de sono ou vigília.

A presença diferenciada desses parâmetros denota um ponto à nota final, exceto o item choro, que pode denotar até dois pontos. O resultado final é a soma de todos os pontos, e um resultados maior que três caracteriza dor (Guinsburg et al., 1997).

Diferente da *NIPS*, focando apenas na parte da expressão facial, temos a *Neonatal Facial Coding System*.

2.2.3 *NFCS*

A *Neonatal Facial Coding System* é uma das principais escalas de avaliação de dor em neonatos. Ela foi criada a partir de gravações de vídeos de recém-nascidos, por meio de uma análise em câmera lenta buscando padrões durante procedimentos previamente conhecidos por causarem dor (Grunau e Craig, 1987).

Nessa escala, dez ações faciais são monitoradas:

1. fenda palpebral estreitada;
2. olhos bem fechados;
3. aprofundamento do sulco naso-labial;

4. lábios abertos (qualquer separação dos lábios);
5. boca esticada verticalmente;
6. boca esticada horizontalmente;
7. língua tensa;
8. tremor de queixo;
9. lábios franzidos;
10. protusão da língua (“não dor” em neonatos a termo).

A presença de cada ação enumerada representa um ponto na nota final, exceto no caso do item 10, em sua ausência caracteriza um ponto. O resultado final é a soma de todos os pontos presentes, e uma convenção utilizada é a de que um resultado maior que 3 caracteriza dor. Na Figura 2.2 podemos observar um exemplo de expressão facial com os itens 1, 2, 3, 4, 5, e 6 da *NFCS*.

Figura 2.2: Expressão facial de dor em um neonato. Fonte: (Wong e Hockenberry, 2000)

A validação e confiança da *NFCS* foi estudada e comprovada em estudos de (Grunau et al., 1998), (Peters et al., 2002), (Schiavenato et al., 2008) e (Sweet e McGrath, 1998).

Apesar de bem definidas, métodos que dependem de tomadas de decisões de humanos tem uma subjetividade inerente. Em um estudo feito por (Prkachin, 2009), vários potenciais fatores de confusão em relação ao uso de escalas com indicadores comportamentais foram explicados. Utilizar de forma correta essas escalas é uma tarefa complexa que requer intenso treinamento e experiência de profissionais de saúde, sendo iminentemente cansativa. (Prkachin, 2009) ainda afirma que mesmo profissionais treinados podem sentir dificuldade na realização de tais tarefas.

O tratamento da dor em neonatos está altamente relacionado ao método escolhido para sua avaliação, bem como das interpretações de quem faz sua avaliação (Guinsburg, 1999). Como afirma (Arias e Guinsburg, 2012), as ações faciais se correlacionam mais com a atividade

cortical durante um estímulo de dor em comparação a indicadores fisiológicos, reforçando a confiabilidade de tal abordagem. Desta forma, uma tentativa de minimizar a subjetividade na aplicação de escalas de dor para avaliação é utilizar sistemas de reconhecimento automático da expressão de dor por meio de imagens ou vídeos.

2.3 RECONHECIMENTO AUTOMÁTICO

Assim como afirma (Prkachin, 2009), devido aos diversos vieses que ocorrem na avaliação de dor, essa avaliação deve seguir o caminho da análise e reconhecimento automático.

Presente em várias escalas comportamentais de dor, a expressão facial pode ser avaliada em imagens ou vídeos de forma não invasiva e em tempo real. Nas últimas décadas, a análise automática de expressões faciais se tornou alvo de constante pesquisas acadêmicas, não somente em adultos mas também em neonatos. Um sistema genérico dessa área é composto por três grandes componentes estruturantes: detecção de face, extração de características e classificação (Yuan et al., 2008).

A detecção de face nada mais é do que a tarefa de encontrar, em uma imagem, a posição e área de uma ou mais faces. Essa etapa é crítica pois é necessário que se obtenha corretamente as regiões com faces para que as características das mesmas não sejam confundidas com o fundo da imagem e para que a extração de características seja feita de maneira correta. (Hjelmås e Low, 2001). Essa tarefa tem sido estudada extensamente nas últimas décadas, sendo considerada um tipo especial detecção de objetos em visão computacional devido as particularidades e variabilidades da face humana (Sun et al., 2018).

Uma extensão da detecção de faces é a detecção de pontos fiduciais (*landmarks*) da face. No lugar de apenas determinar a localização e área de faces em uma imagem, a detecção de pontos fiduciais busca pontos específicos pré-determinados. Esse tipo de tarefa pode agregar mais informação do que apenas a detecção da face pois uma vez que os pontos são detectados corretamente, pode-se extrair informação de pose e dos movimentos faciais visíveis (Zhu e Ramanan, 2012). A Figura 2.3 mostra um exemplo de pontos fiduciais (em azul) detectados em uma imagem.

Após o passo de detecção de face e/ou detecção de pontos fiduciais, pode-se fazer a extração de características de cada face encontrada. As características podem ser definidas como propriedades mensuráveis extraídas de um objeto. Segundo (Zamzmi et al., 2017), existem cinco principais abordagens para extração de características faciais no problema de reconhecimento de dor em neonatos: Métodos de redução de características, Variações de *Local Binary Patterns (LBP)*, métodos baseados em movimento, métodos baseados em modelos e sistemas de *Facial Action Coding System (FACS)*.

Nos métodos de redução, os pixels das imagens são tratados como características e métodos de redução da dimensionalidade processam as características, como o *Sequential Floating Forward Selection (SFFS)* e o *Principal Component Analysis (PCA)*.

Nos métodos baseados no *LBP*, que é um descritor de textura em imagens, várias modificações de diferentes aspectos do seu funcionamento já foram testadas na literatura, e por isso (Zamzmi et al., 2017) cria uma categoria para métodos baseados nele.

Já nos métodos baseados em movimento, o objetivo é detectar ocorrência de movimentos faciais entre sequências de imagens. Por outro lado, métodos baseados em modelos buscam otimizar a correspondência entre um modelo com vários parâmetros e uma imagem de entrada, como, por exemplo, o *Active Appearance Model (AAM*. Por fim, métodos baseados em *FACS* buscam detectar as *Action Units (AUs)* que definem esse sistema em uma face (Zamzmi et al., 2017).

Figura 2.3: Exemplo de pontos fiduciais (em verde) em uma imagem.

Após termos obtido as características extraídas da imagem da face, esses dados são passados a métodos de aprendizado de máquina para classificação.

2.4 APRENDIZADO DE MÁQUINA

Na área de aprendizado de máquina, existem dois grandes grupos de conhecimento: o de aprendizado supervisionado e não supervisionado. No aprendizado não supervisionado, não temos uma variável resposta definida que nos explique os exemplos sendo analisados, muitas vezes o objetivo é justamente descobrir uma separação que explique os dados disponíveis. No aprendizado não supervisionado, temos os algoritmos de agrupamento, que podem ser construídos de maneira *Top Down* ou *Bottom Up*, que visam encontram um número ótimo de grupos dentro dos dados (James et al., 2013).

No aprendizado supervisionado, existe um rótulo pra cada exemplo da base sendo analisada. Esse rótulo pode ser uma categoria, o que implica em um problema de classificação, ou pode ser um valor contínuo, o que levaria a um problema de regressão. Ambas áreas tem métodos tradicionais de modelagem mas também tem métodos de redes neurais, que dependendo da complexidade podem ser categorizados como métodos de *Deep Learning*. Esse esquema mencionado pode ser visualizado na figura 2.4.

O uso de classificação é essencial para o melhor entendimento, sumarização e análise de situações (Chandrasekaran e Keuneke, 1987). De modo geral, podemos dizer que a tarefa de classificação envolve qualquer tomada de decisão ou previsão em cima de informações sobre um determinado estado atual aprendido. O processo de classificação consiste então na formalização de um método que possa tomar essas decisões repetidamente (Michie et al., 1994).

Assim como (Michie et al., 1994) afirma, para que a classificação de um novo contexto possa ser feita, é necessário que, previamente, algum tipo de processo ou aprendizado de contextos

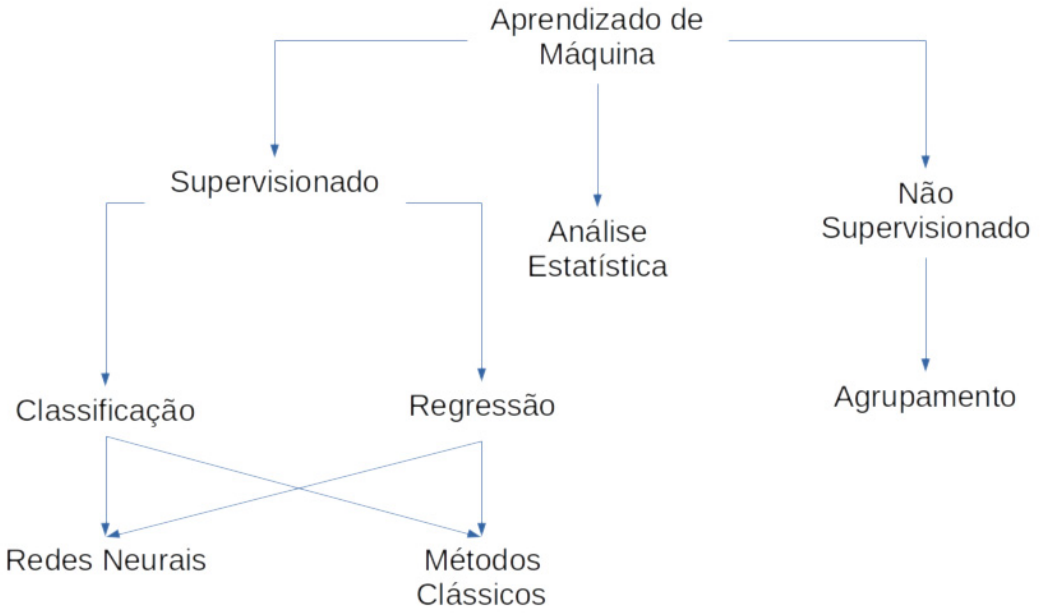


Figura 2.4: Visão geral de métodos de aprendizado de máquina

conhecidos seja realizada. Por isso, podemos observar a existência de três grandes áreas de estudo que dividem tipos de classificação: classificação estatística, clássica e redes neurais.

Na classificação estatística, tem-se conhecimento de que o problema que define os dados que precisam ser classificados seguem um modelo probabilístico. Então, ao invés de obter-se apenas à qual classe o novo contexto pertence, obtêm-se a probabilidade de pertencer à cada classe. No classificação clássica, o processo de classificação depende do processo de aprendizagem e obtenção de conhecimento dos contextos previamente conhecidos para tomar decisões sobre eventuais contextos novos, partindo-se do pressuposto de que instâncias similares pertencem à mesma classe (Quinlan, 2014).

Em redes neurais, a intenção é mimetizar o processo classificatório do cérebro humano, simulando neurônios. De modo geral, temos camadas de nós interconectados, em que cada nó produz uma função não linear de seus dados de entrada (Quinlan, 2014; Michie et al., 1994). Devido a essas conexões, um alto grau de não linearidade pode ser exigida da rede, dependendo da arquitetura escolhida para resolução do problema. Após o dado passar pela rede inteira, eventualmente ele chega a um nó de saída, em que a classe final é calculada.

Para o problema de classificação de expressões faciais em neonatos, as classes a serem analisadas geralmente são *dor* e *não dor*, uma vez que *não dor* pode ainda ser subdividida em classes como *repouso*, *reação a fricção*, *choro normal*, entre outras.

Para construir um classificador, uma separação deve ser feita entre a base de treino e a base de teste. A base de treino será utilizada na construção dos classificadores, enquanto a base de teste será utilizada para avaliar a performance dos mesmos (Dangeti, 2017). Existem várias técnicas para fazer essa divisão dos dados, sendo a mais comum a técnica *Hold Out*, ilustrada na figura 2.6, em que 70% dos dados ficam separados para o treino, e 30% para teste.

Assim como (Dangeti, 2017) afirma, essa divisão dos dados pode ser feita várias vezes, de modo a garantir que a base inteira seja utilizada para construir o modelo pelo menos uma vezes, sempre mudando qual porção constituirá a base de teste. Essa técnica é chama de **K-fold**, um método de *cross validation*. O K é o número de partições em que a base será dividida, consequentemente sendo também o número de treinamentos que acontecerá, cada um utilizando

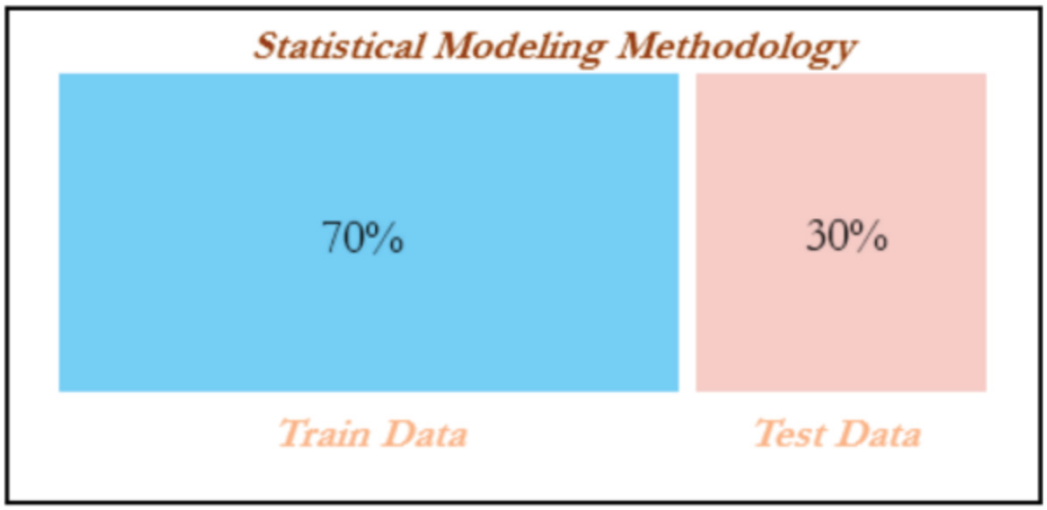


Figura 2.5: Exemplo de esquema *Hold Out* Fonte: (Dangeti, 2017)

uma porção para teste. Uma ilustração de um *K-fold* com K igual a 5 pode ser observado na figura 5.1(a).

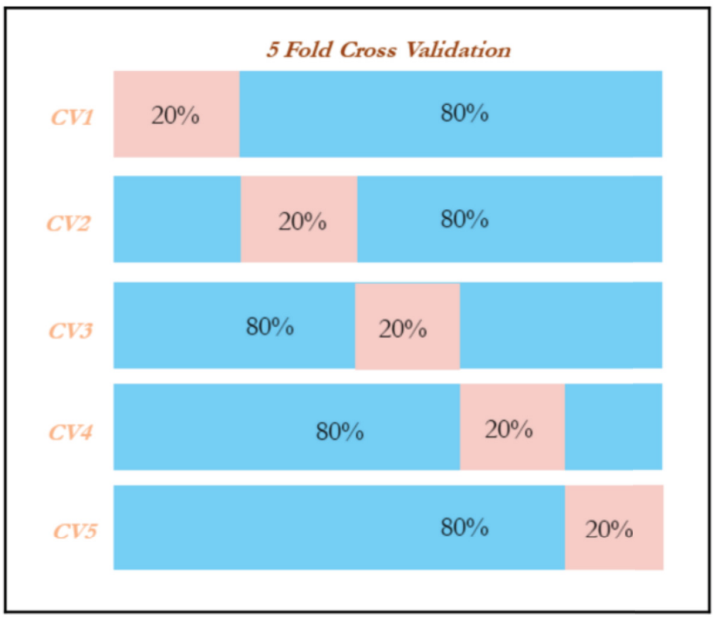


Figura 2.6: Exemplo de esquema *K fold* Fonte: (Dangeti, 2017)

Criado o classificador, para avaliar o quanto uma classificação feita foi boa, vários fatores devem ser levados em consideração, e podemos citar 3 perguntas chave:

1. Qual é o balanceamento entre as classes do meu problema?
2. Qual métrica eu deveria estar avaliando de acordo com o contexto do meu problema?
3. Meus dados representam uma situação real?

Essas perguntas estão relacionadas ao uso final do modelo de classificação criado. É essencial que se saiba a priori o balanceamento entra as classes nos dados, ou seja, qual é a

proporção de cada classe na base. Tal fator é importante pois em casos de desbalanceamento extremo, em que uma das classes é um evento raro, técnicas de balanceamento de dados precisam ser aplicadas para uma melhor construção do modelo e obtenção de resultados verdadeiros.

A segunda pergunta também é extremamente importante ao contexto da aplicação. De acordo com o conjunto de dados que se tem e com o objetivo final, a métrica sendo avaliada deve estar adequada (Hossin e Sulaiman, 2015). Deve ser considerada qual será a aplicação final do modelo, o que será feito com o seu resultado. Se pensamos em um modelo que prevê chuva no dia seguinte, e um sistema se alimenta dessa previsão para emitir um alerta a população, um falso negativo parece ser mais grave que um falso positivo, pois seria pior que a população estivesse despreparada para um possível desastre. Nesse caso, queremos maximizar o recall do nosso modelo, para que ele tenha uma sensibilidade maior ao nosso evento alvo. Em compensação, se imaginarmos um sistema de recomendação de filmes, a intenção não é descobrir todos os filmes que um usuário gosta, mas garantir que ele goste dos que são recomendados. Nesse caso, queremos maximizar o precision do nosso modelo, que mostra quantos casos eram positivos dentro de todos que foram preditos.

Analisar um métrica que não representa o esperado na aplicação do modelo pode levar a resultados enviesados e que são diferentes da realidade.

3 TRABALHOS RELACIONADOS

Nas últimas décadas, vários trabalhos têm sido desenvolvidos na área de reconhecimento automático de expressões faciais de neonatos. Neste capítulo, revisaremos os principais destes trabalhos disponíveis na literatura.

Um dos primeiros trabalhos na área, o trabalho de (Brahnam et al., 2005) não só desenvolveu um *pipeline* para reconhecimento de expressões faciais em neonatos, como também foi um dos poucos estudos da área em que a base de dados coletada é publicamente disponível a partir de um pedido oficial aos donos da base. A detecção de face foi feita de forma manual, com humanos selecionando uma elipse que contemplasse boa parte da face dos neonatos. Os métodos de extração de características testados foram da abordagem de redução de dimensionalidade, o *Principal Component Analysis (PCA)* e o *Linear Discriminant Analysis (LDA)*. Para classificar as expressões, o algoritmo *Support Vector Machines (SVM)* foi utilizado. Na Figura 3.1, esse pipeline de processamento é exemplificado.

Figura 3.1: Pipeline de reconhecimento de expressões. Fonte: (Brahnam et al., 2005)

No trabalho de (Lu et al., 2008a), uma base de dados de imagens também foi coletada. Assim como em (Brahnam et al., 2005), a detecção da face e pre-processamento foram feitos manualmente. Para extração de características, filtros de Gabor 2D foram utilizados. Devido ao alto número de características obtidas, o algoritmo AdaBoost foi utilizado para seleção das mais discriminantes. Ao fim, 900 características foram selecionadas, das 412160 originais. Por último, o método SVM é utilizado para classificar as expressões faciais.

O trabalho de (Lu et al., 2008a) é continuado em (Yuan et al., 2008) e (Lu et al., 2008b). Um método de *Hybrid Boost* é proposto para fazer a seleção de características, que também são filtros de Gabor 2D como no trabalho anterior. O *Hybrid Boost* também é utilizado para selecionar classificadores, objetivando uma solução final com uma hierarquia de classificadores. O esquema do sistema proposto pode ser observado na Figura 3.2.

Em (Zhong e Liu, 2011), uma nova versão do trabalho é empregada, sendo aplicada em um sistema de monitoramento em tempo real. A estrutura do sistema propõe uma rede de terminais sendo utilizados em um ambiente clínico, que apenas transmitiriam os vídeos de monitoramento para um servidor remoto onde o reconhecimento de dor seria feito.

Ainda dos mesmo autores, o trabalho de (Lu et al., 2016) propõe um novo método, e para tanto mais imagens foram coletadas de recém-nascidos nos estados de repouso, choro, dor moderada e dor intensa. Exemplos de dor intensa presentes na base podem ser vistas na Figura 3.3. As imagens foram rotuladas por especialistas em NFCS, e uma série de testes foi realizada para comparar características e classificadores. A etapa de extração de características testou o algoritmo *Local Gradient Code (LGC), Local Directional Pattern (LDP), Local Directional Texture Pattern (LDTP)*, e também variantes de *LBP*. Para classificação de expressões faciais,

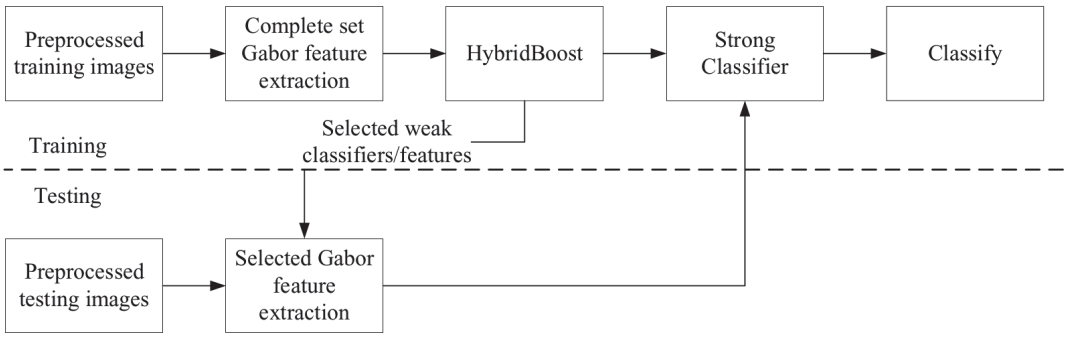


Figura 3.2: Pipeline de reconhecimento de expressões. Fonte: (Yuan et al., 2008)

um modelo de classificação baseado em representação esparsa é testado em comparação a um SVM, e de acordo com os autores, o primeiro modelo se mostrou melhor.

Figura 3.3: Expressões de dor intensa. Fonte: (Lu et al., 2016)

Em (Schiavenato et al., 2008), o trabalho desenvolvido baseia-se em ações descritas no NFCS. De acordo com as ações descritas na escala, 7 pares de pontos fiduciais foram escolhidos para serem analisados. Na Figura 3.4(a) podemos ver em amarelo os pares de pontos definidos. O objetivo do trabalho era quantificar as diferenças observadas nesses pontos, verificando se explicariam ações da NFCS. Vídeos foram gravados de neonatos imediatamente antes, durante, e depois de um procedimento doloroso. Um par de imagens de cada neonato em repouso (antes do procedimento) e no auge da dor (durante o procedimento) foram escolhidas por um especialista, exemplificadas na Figura 3.4 com os pontos em vermelho, e as distâncias euclidianas dos pontos em azul.

(a) (b)

Figura 3.4: Pontos fiduciais (a) e distância euclidiana em diferentes expressões (b). Fonte: (Schiavenato et al., 2008)

O trabalho de (Schiavenato et al., 2008) tem um caráter analítico e exploratório, e por isso os dados das distâncias são utilizados em uma análise estatística da correlação direta com a dor e também em uma análise das diferenças entre os participantes. Um sistema de classificação automática não foi desenvolvido.

Ao invés de tentar classificar dor e não-dor, o sistema de (Hazelhoff et al., 2009) busca classificar se o recém-nascido está dormindo, acordado ou chorando. A detecção da região da face é realizada por meio de um modelo gaussiano que detecta a face de acordo com tamanho e cor da pele. A partir da face, o contorno dos olhos, sobrancelhas e boca são detectados, e distâncias em cada componente são definidas para serem utilizadas. Na figura 3.5 pode-se observar a face detectada com o fundo em preto, os pontos fiduciais das sobrancelhas, olhos e boca e também a linha do nariz.

(a) (b) (c)

Figura 3.5: Exemplos das características detectadas em estados acordado (a), dormindo (b) e choro (c). Fonte: (Hazelhoff et al., 2009)

O esquema de classificação utilizado é definido por uma hierarquia. Para definição dos estados de cada olho, da boca, e de cada sobrancelha, o método *k - Nearest Neighbors (kNN)* é utilizado. Num segundo nível mais alto, o estado de componentes simétricos é combinado, olhos e sobrancelhas, pois de acordo com especialista o movimento facial nos neonatos é simétrico. Por fim, um último classificador baseado em regras define o estado final da face. O esquema hierárquico utilizado nesse trabalho segue a ideia do funcionamento das escalas de dor utilizadas no ambiente clínico (Hazelhoff et al., 2009).

Em (Gholami et al., 2010) os autores tratam o problema de estimar a intensidade de dor, não somente sua presença ou ausência. A base de dados utilizada foi a de (Brahnam et al., 2005), em que especialistas e não–especialistas rotularam a intensidade de dor nas imagens selecionadas. As características utilizadas foram os próprios valores de intensidade dos pixels de cada imagem, e uma comparação foi feita entre o classificador SVM e RVM – *Relevance Vector Machine*. O método RVM se mostrou superior nos experimentos conduzidos.

(Mansor et al., 2012a) aborda o problema de reconhecimento de expressões faciais em recém-nascidos de forma semelhante à abordagem de (Brahnam et al., 2005), utilizando o método PCA para extrair características, e *Support Vector Machines (SVM)* para classificar as expressões. O trabalho se destaca por utilizar, diferente de (Brahnam et al., 2005), um método para detecção automática das faces. O método escolhido foi o Viola-Jones, e exemplos da detecção de face no trablho podem ser visualizados na Figura 3.6.

No trabalho de (Mansor et al., 2012c), a detecção de face é feita por meio de uma análise pixel a pixel da probabilidade de o mesmo pertencer à uma região de pele, e posteriormente uma análise de componentes conexas. As características extraídas são coeficientes de um modelo auto regressivo, que são entrada para um classificador *Fuzzy k-Nearest Neighbor*. Os mesmos autores

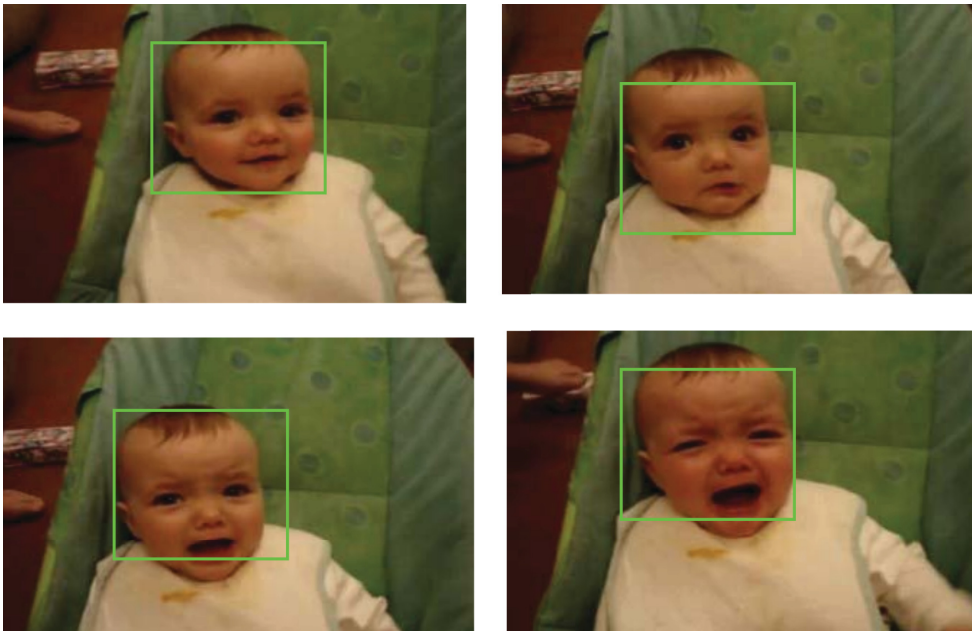


Figura 3.6: Exemplos de detecção de face. Fonte: (Mansor et al., 2012a)

continuaram o trabalho em (Mansor et al., 2012b), onde o esquema de detecção de face continuou o mesmo, porém a extração de características foi dada por meio do uso da transformada rápida de Fourier e a classificação por um *k-Nearest Neighbor* simples.

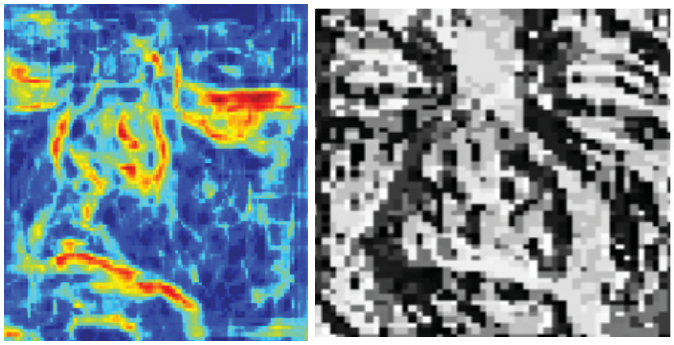
(Mansor e Rejab, 2013a) apresentam uma abordagem utilizando a base de (Brahnam et al., 2005) já com as faces detectadas. A extração de características se dá por meio do método de uma matriz de co–ocorrência dos valores dos pixels em escala de cinza (GLCM), uma abordagem consolidada para análise de textura em imagens. Os classificadores testados foram o *Linear Discriminant Analysis (LDA)* e *Hybrid Genetic Algorithm Neural Network (GANN)*. Os experimentos conduzidos também testaram diferentes condições artificiais de iluminação na base de dados. Apesar de ser consideravelmente mais lento, o classificador GANN mostrou um desempenho superior ao LDA na maioria dos testes realizados.

Os autores de (Mansor e Rejab, 2013b) continuaram o trabalho citado anteriormente e novamente focaram em experimentos com diferentes extratores de características e classificadores. Nesse trabalho, as características testadas foram congruência de fase, um extrator baseado na frequência do modo como visualmente processamos componentes visuais, e o método LBP. Exemplos desses extratores na base utilizada podem ser observados na Figura 3.7. Os classificadores experimentados foram *Naive Bayes* e uma classificador esparso. Os melhores resultados foram obtidos com uma combinação das duas características e com o classificador esparso.

Em (Mansor et al., 2014), novamente a base de (Brahnam et al., 2005) é utilizada e mais experimentos com diferentes características e classificadores são realizados. As características analisadas são a SSQ (*Single Scale Self Quotient Image*), utilizada numa tentativa de restaurar as imagens corrompidas em adição à DCT (Transformada Discreta do Cosseno). Uma rede neural probabilística foi utilizada como classificador.

Assim como afirma (Lu et al., 2016), a direção da pesquisa na área de reconhecimento de expressões faciais em neonatos segue para o uso de métodos baseados em *deep learning*, como redes neurais convolucionais (CNNS).

Com isso em mente, em (Zamzmi et al., 2018) um trabalho é apresentado em que os autores utilizaram CNNs pré-treinadas para extrair características das faces de neonatos. As



(a) (b)

Figura 3.7: Exemplos das características de congruência de fase (a) e LBP (b). Fonte: (Mansor e Rejab, 2013b)

redes escolhidas foram: VGG–F, VGG–M, VGG–S (utilizadas em classificação de imagens) e VGG-Face (utilizada em reconhecimento de faces). Nos experimentos, o melhor extrator de características foi a rede VGG–Face, de uma camada alta, em combinação com o classificador kNN. Os autores ainda testaram a combinação das características da rede com uma abordagem tradicional de processamento de imagens, um método baseado em *optical flow*, utilizado para calcular a tensão em regiões da face. A combinação mais efetiva reportada no estudo foi utilizar características de tensão em conjunto com características extraídas da última camada convolucional da VGG–Face, em adição ao classificador *Naive Bayes*.

Na Tabela 3.1, um resumo das bases e labels pode ser observado. Dos trabalhos revisados nesse capítulo, observa-se que a maioria utiliza a base criada por (Brahnam et al., 2005), explicada em detalhes no Capítulo 4. Nota-se também que não existe um padrão quanto aos labels, com algumas bases lidando apenas com classes de dor e não dor, e algumas especificando em sub-classes a expressão de não dor.

Já na Tabela 3.2, características e classificadores utilizados nos trabalhos discutidos pode ser observada. Com um olhar cronológico, nota–se que apesar de ser uma tendência crescente nos últimos anos, (Zamzmi et al., 2018) é o único analisado que utiliza técnicas de *deep learning*, enquanto os demais trabalhos focam em abordagens tradicionais de processamento de imagens.

Na Tabela 3.3, observamos os resultados de performance do trabalhos analisados. É importante notar, que mesmo tratando-se de uma área de pesquisa médica, quase todos trabalhos focam em apresentar resultados na métrica de acurácia, ou seja, de casos certos no total. Tal prática pode ser tendenciosa por dois motivos: o desbalanceamento das classes, e o peso do tipo de erro. Em um problema de detecção de dor, em um contexto da vida real e que também acontece na maioria das bases (em todas em que tal informação foi reportada), a maioria dos casos são casos em que a dor não está presente. Desta forma, se imaginamos casos cotidianos em que hipoteticamente temos 90% do tempo um estado normal, e criamos um classificador que apenas prevê a classe *não dor*, a acurácia desse classificador seria de 90%. Ainda analisando a metodologia de aprendizado de máquina dos métodos, apenas dois trabalhos ((Mansor e Rejab, 2013b) e (Mansor et al., 2014)) explicitam qual foi a estratégia de treino dos classificadores, porém um utiliza a técnica *Hold Out* num esquema 80%/20% e outro um *K-fold* com 10 splits. Tais práticas também são afetadas por um desbalanceamento de classes e podem apresentar resultados tendenciosos.

Ainda um outro ponto que deve ser notado em relação a apresentação de resultados dos trabalhos citados, é a questão do tipo de erro dado o contexto de avaliação de dor. Nesse contexto, em que a classe sendo predita é *dor*, um erro do tipo I (falso positivo) seria prever ocorrência de dor quando ela não existe, e o erro do tipo II (falso negativo) seria não prever dor

Tabela 3.1: Resumo das bases e labels utilizados nos trabalhos revisados.

Referência	**Base**	**Labels**
(Brahnam et al., 2005)	Própria	Repouso, Choro, Estímulo de Ar, Fricção, Dor
(Lu et al., 2008a)	Própria	Repouso, Choro, Dor
(Yuan et al., 2008)	(Lu et al., 2008a)	Repouso, Choro, Dor
(Lu et al., 2008b)	(Lu et al., 2008a)	Repouso, Choro, Dor
(Lu et al., 2016)	Própria	Repouso, Choro, Dor Moderada, Dor Intensa
(Schiavenato et al., 2008)	Própria	Dor, não dor
(Hazelhoff et al., 2009)	Própria	Acordado, Dormindo, Desconforto
(Gholami et al., 2010)	(Brahnam et al., 2005)	Repouso, Choro, Estímulo de Ar, Fricção, Dor
(Mansor et al., 2012a)	Própria	Dor, não dor
(Mansor et al., 2012c)	Própria	Dor, não dor
(Mansor et al., 2012b)	Própria	Dor, não dor
(Mansor e Rejab, 2013a)	(Brahnam et al., 2005)	Repouso, Choro, Estímulo de Ar, Fricção, Dor
(Mansor e Rejab, 2013b)	(Brahnam et al., 2005)	Repouso, Choro, Estímulo de Ar, Fricção, Dor
(Mansor et al., 2014)	(Brahnam et al., 2005)	Repouso, Choro, Estímulo de Ar, Fricção, Dor
(Zamzmi et al., 2018)	Própria	Dor, não dor

quando ela existe. Para um contexto de monitoramento e emissão de alertas, um erro do tipo II é muito mais grave que um erro do tipo I. Dessa forma, analisar métricas que refletem esses pesos é importante.

Todos os autores dos trabalhos citados nas tabelas 3.1 e 3.2 que coletaram dados foram contatados na tentativa de conseguir acesso as respectivas bases de dados, porém o único retorno positivo veio de (Brahnam et al., 2005), que concederam aos autores deste trabalho acesso a base de dados COPE.

Tabela 3.2: Resumo das características e classificadores utilizados nos trabalhos revisados.

Referência	**Característica**	**Classificador**
(Brahnam et al., 2005)	Pixel + PCA	SVM
(Lu et al., 2008a)	Gabor + AdaBoost	SVM
(Yuan et al., 2008)	Gabor 2D + HybdridBoost	Hierárquico
(Lu et al., 2008b)	Gabor 2D + HybdridBoost	Hierárquico
(Lu et al., 2016)	LGC, LDP, LDTP E LBP	SVM
(Schiavenato et al., 2008)	Distância de pares de pontos	–
(Hazelhoff et al., 2009)	Estado de cada componente	Hierárquico
(Gholami et al., 2010)	Pixel	SVM e RVM
(Mansor et al., 2012a)	Pixel + PCA	SVM
(Mansor et al., 2012c)	Auto regressivo	Fuzzy kNN
(Mansor et al., 2012b)	Transformada rápida de Fourier	kNN
(Mansor e Rejab, 2013a)	GLCM	LDA, GANN
(Mansor e Rejab, 2013b)	Congruência de fase, LBP	Esparso, Naive Bayes
(Mansor et al., 2014)	SSQ + DCT	Rede neural probabilística
(Zamzmi et al., 2018)	VGG–F, VGG–S, VGG–M, VGG–Face, Optical Flow	Naive Bayes

Tabela 3.3: Resumo das performances dos trabalhos revisados.

Referência	**Métrica**	**Valor**
(Brahnam et al., 2005)	Acurácia	88%
(Lu et al., 2008a)	Acurácia	85.29%
(Yuan et al., 2008)	Acurácia	88%
(Lu et al., 2008b)	Acurácia	85.29%
(Lu et al., 2016)	Acurácia	85.50%
(Schiavenato et al., 2008)	Correlação de coeficientes intraclasse	0.37 a 0.95
(Hazelhoff et al., 2009)	Acurácia	95%
(Gholami et al., 2010)	Acurácia	91%
(Mansor et al., 2012a)	Acurácia	93%
(Mansor et al., 2012c)	Acurácia	90.7%
(Mansor et al., 2012b)	Acurácia	90.12%
(Mansor e Rejab, 2013a)	Acurácia	91%
(Mansor e Rejab, 2013b)	Acurácia	88%
(Mansor et al., 2014)	Acurácia	96.2%
(Zamzmi et al., 2018)	Acurácia	92.71%

4 MATERIAIS E MÉTODOS

Seguindo o fluxo de reconhecimento automático da expressão de dor em neonatos, são necessários dados e também algoritmos para cada etapa de detecção de face, extração e características e classificação. Tais detalhes serão explicados neste capítulo.

Todos os autores das bases de dados citadas na seção 3 foram contatados para tentativa de acesso e uso das bases nesse trabalho, porém apenas os autores de (Brahnam et al., 2005) responderam aos pedidos.

4.1 COPE

A base de dados utilizada para o estudo desse trabalho foi a primeira base de dados criada para o avanço do reconhecimento automático de dor em neonatos. Os dados foram coletados no Hospital St. Johns em Missouri, Estados Unidos. O público-alvo da pesquisa foram bebês cujas mães tiveram parto sem complicações (Brahnam et al., 2005).

Dentro do público-alvo, 26 neonatos puderam participar da coleta de dados, 13 meninos e 13 meninas com idades entre 13 horas e 3 dias. 200 imagens foram obtidas, em quatro diferentes cenários de estímulo (Gholami et al., 2010):

1. Transporte de um berço para outro (imagens de choro e repouso);
2. Estímulo de ar: sopro de ar no nariz do bebê;
3. Fricção: calcanhar esfregado levemente com algodão ensopado em álcool;
4. Dor: procedimento de coleta de sangue;

Os três primeiros itens entram na categoria genérica de não dor, e o último, de dor. Na Figura 4.1 exemplos de imagens da base podem ser observados. Nota-se que dentro da primeira categoria existem imagens de choro não causadas por estímulo doloroso, o que na prática clínica pode ser uma tarefa dificilmente discriminável.

A base COPE atualmente está disponível em duas modalidades: A primeira com todas as imagens originais, e a segunda com as faces dos neonatos já detectadas e com a região de interesse cortada.

(a) (b)

(c) (d)

Figura 4.1: Exemplos de imagens da base COPE. Expressões de (a) descanso, (b) estímulo de sopro de ar, (c) de fricção e (d) dor. Fonte: (Brahnam et al., 2005)

5 EXPERIMENTAÇÃO E RESULTADOS

Os experimentos se dividem em duas partes: o primeiro foca na detecção de face em imagens, e o segundo, na extração de características da face. Todos os experimentos a seguir foram desenvolvidos na linguagem Python.

5.1 DETECÇÃO DE FACE

O primeiro teste realizado foi o de detecção de face. Dois métodos foram avaliados, o primeiro sendo o de (Kazemi e Sullivan, 2014) disponível na biblioteca Dlib (King, 2009). O segundo teste avaliou o método de (Viola e Jones, 2004) disponível na biblioteca (Bradski, 2000). O objetivo do teste foi avaliar a performance de dois métodos de *prateleira* muito bem consolidados para detecção de faces de adultos, em neonatos.

Ao rodar os dois métodos na base COPE com as imagens originais, os seguintes resultados foram obtidos:

- O método da biblioteca Dlib foi capaz de detectar faces em 32% das imagens;
- O método da biblioteca OpenCV detectou apenas 21% das faces;
- 54% das imagens não teve face alguma detectada pelos dois métodos;
- Apenas 8% das imagens teve uma face detectada pelos dois métodos.

Esses resultados evidenciam a dificuldade da tarefa de detecção de face em neonatos, uma vez que a performance para adultos é exponencialmente maior.

Das imagens que não tiveram face alguma detectada:

- 15% eram de cenários onde ar foi soprado no neonato;
- 7% eram de cenários de choro;
- 15% eram de cenários onde algum tipo de fricção foi aplicada ao neonato;
- 22% eram de cenários de dor;
- 32% eram de cenários de descanso;
- 5% eram de cenários após analgesia com sucrose;
- 4% eram de bocejo.

Mesmo em um cenário onde as imagens são predominantemente de repouso, o nível de faces de dor que não foram detectadas foi extremamente alto. Na Figura 5.1, três exemplos podem ser observados. Na primeira figura, nota-se que os dois métodos detectaram corretamente a região da face. Em 5.1(b), temos um exemplo onde apenas o método Viola-Jones da biblioteca OpenCV foi capaz de detectar uma face na imagem, porém a região retornada errou na escala e localização da face. Já em 5.1(c), o mesmo tipo de erro acontece com o método de (Kazemi e Sullivan, 2014) da biblioteca Dlib.

Ambos os métodos avaliados são consolidados na literatura, porém falharam em detectar as faces de neonatos em imagens com boa luminosidade, e face semi frontal. Isso evidencia a clara diferença que deve existir em métodos direcionados a imagens de neonatos, necessitando, assim, de métodos mais complexos e robustos para tal tarefa.

(a) (b) (c)

Figura 5.1: Exemplos de detecção de face em imagens da base COPE. Exemplos de (a) detecção pelo biblioteca Dlib e OpenCV, (b) apenas detecção (errônea) do OpenCV, e (c) apenas detecção errônea do Dlib.

5.2 EXTRAÇÃO DE CARACTERÍSTICAS

Focando apenas na extração de características, alguns métodos foram testados na base COPE processada, já com a região de interesse da face recortada. O objetivo desse teste era avaliar a performance de métodos clássicos de extração de características em processamento de imagens. Nessa seção, usaremos *características* e *features* de maneira permutável.

Como as imagens utilizadas já haviam sido processadas, a detecção de face não foi necessária. Sendo assim, os métodos de extração de características foram implementados, e logo após todos foram testados pelo mesmo método classificador, um *Random Forest*.

Para conduzir o treinamento e teste do classificador, uma versão modificada de *Leave One Out* foi seguida. Isso se deveu ao fato de que a base COPE tem apenas 204 imagens, e apenas 26 participantes. Na tentativa de evitar *overfitting*, 26 rodadas e treinamento e teste foram realizadas, sendo que em casa rodada as imagens de um dos participantes foi deixada para teste, e as demais para treinamento.

Primeiramente, assim como alguns trabalhos da literatura, um teste foi conduzido com os dados originais sem nenhuma característica extraída. A acurácia de teste foi de 73%. Logo após, o método *Principal Component Analysis - PCA* foi empregado nas imagens originais, outra metodologia recorrente na literatura. Esse método também resultou em uma acurácia de 73%. Em seguida, um teste foi realizado com filtros de gabor. Esse filtro propõe analisar sinais de tempo e frequência simultaneamente (Liu et al., 2012), modelando os aspectos espaciais do córtex visual (Kamel e Campilho, 2009). Após os testes, a acurácia atingida foi de 76%. O teste seguinte utilizou *Histogram of Oriented Gradients*. Esse método divide a imagem em sub-regiões e calcula histogramas das direções dos gradientes em cada uma (Lowe, 2004). A acurácia obtida foi de 78%. Esses resultados podem ser visualizados na tabela 5.1 e na Figura 5.2 exemplos de cada característica testada podem ser visualizados.

Tabela 5.1: Resumo das performances dos métodos testados

Método	**Precision**	**Recall**	**F1**	**Acurácia**	**FP**	**FN**	**TP**	**TN**	**AUC**
Gabor	0.628	**0.45**	0.524	0.76	16	33	27	128	0.794
HOG	0.722	0.433	0.542	**0.784**	10	34	26	134	0.822
Original	0.6	0.3	0.4	0.735	12	42	18	132	0.784
Daisy	0.567	0.283	0.378	0.725	13	43	17	131	0.758
PCA	0.667	0.2	0.308	0.735	6	48	12	138	0.675
LBP	0.429	0.15	0.222	0.691	12	51	9	132	0.631

Analisando os resultados, temos que a melhor *feature* foi a HOG, quando olhamos apenas a acurácia. Contudo, sabendo que a métrica recall (sensibilidade) nos diz quantos casos de dor foram acertados de todos presentes na base, o que devido a nossa aplicação, é mais importante. Desta forma, o melhor método nesse sentido foi o filtro de gabor, descobrindo 45% dos casos, que consequentemente também teve o menor número de falso negativo (FN).

Na tabela ainda podemos observar o precision dos métodos, que nos diz quantos casos eram da classe *dor* de todos que foram previstos como tal. Nesse caso, o método HOG foi melhor, porém essa métrica deve ser otimizada quando o custo de um falso positivo (FP) é maior do que o custo de um falso negativo (FN), que não é o caso desse trabalho. As métricas F1, e AUC também são apresentadas por motivos de comparação, sendo que F1 é uma média harmônica entre o pricision e o recall, e o AUC é uma relação entre o *false positive rate* e o *true positive rate* .

Diferente da literatura, aqui mostramos os resultados de forma mais tangível por utilizar o esquema de treinamento *Leave One Subjet Out*, pois nesse esquema todos os elementos da base fazem parte da base de teste em algum momento, o que implica em termos todas as imagens classificadas. Isso aumenta a confiabilidade nos resultados, pois deve ser lembrado que a base COPE é uma base de poucas imagens, por isso a importância de sempre mostrar a quantidade de falsos negativos, falsos positivos, verdadeiros positivos e falsos positivos.

Com o objetivo de testar os limites desses métodos e de identificar qual desses métodos é o mais robusto à ruídos, outra rodada de testes foi realizada.

Considerando que o estudo desse trabalho visa auxiliar futuras aplicações para monitoramento de recém nascidos em tempo real, devemos levar em consideração que em contextos da vida real, as condições de captura de imagens podem não ser as melhores. Para replicar esse efeito, reproduzimos na base de dados COPE três tipos de ruídos em 20 intensidades diferentes.

O primeiro ruído testado foi o de Sal e Pimenta, que substitui pixels da imagem aleatoriamente com pixels brancos ou pretos. A proporção de pixels que são substituídos é definida a priori e chamada de *Amount*, e portanto foi o parâmetro que variamos para testar a robustez dos métodos. Na figura 5.3 observamos que após o valor 0.2 a imagem fica extremamente deturpada, e por isso esse valor foi escolhido empiricamente como limiar do teste.

Outro ruído avaliado foi o ruído gaussiano, em que ao invés de substituir os valores dos pixels com pixels brancos ou pretos, os valores seguem uma distribuição gaussiana. Por ter uma função de densidade de probabilidade, aumentando a variância, os valores de pixels sendo substituídos são mais extremos, ou seja, mais ruidosos. Na figura 5.4 observamos que após o valor 0.05 a imagem fica altamente ruidosa, e por isso esse valor foi escolhido empiricamente como limiar do teste.

Assim como comentado anteriormente, a métrica de interesse a ser avaliada é o recall. Para cada tipo de ruído e cada valor de parâmetro do ruído, um esquema de avaliação *Leave One Out* foi empregada, e os resultados para o ruído sal e pimenta podem ser observados na figura 5.5.

De forma geral, observamos que a performance cai para os métodos. Considerando a performance inicial com $amount = 0$ e a performance final com $amount = 0.19$, podemos observar na tabela 5.2 a diferença entre essas performances para cada método. De acordo com a tabela, o método mais resistente ao ruído sal e pimenta foi com a imagem original inteira como feature, porém com um delta muito próximo ao método LBP. O método menos resistente foi o filtro de gabor.

Analisando os resultados do ruído gaussiano, também observamos que de forma geral, a performance cai para os métodos (Figura 5.6). Considerando a performance inicial com $var = 0$ e a performance final com $var = 0.05$, e podemos observar na tabela 5.3 a diferença entre essas performances para cada método. De acordo com a tabela, o método mais resistente ao ruído

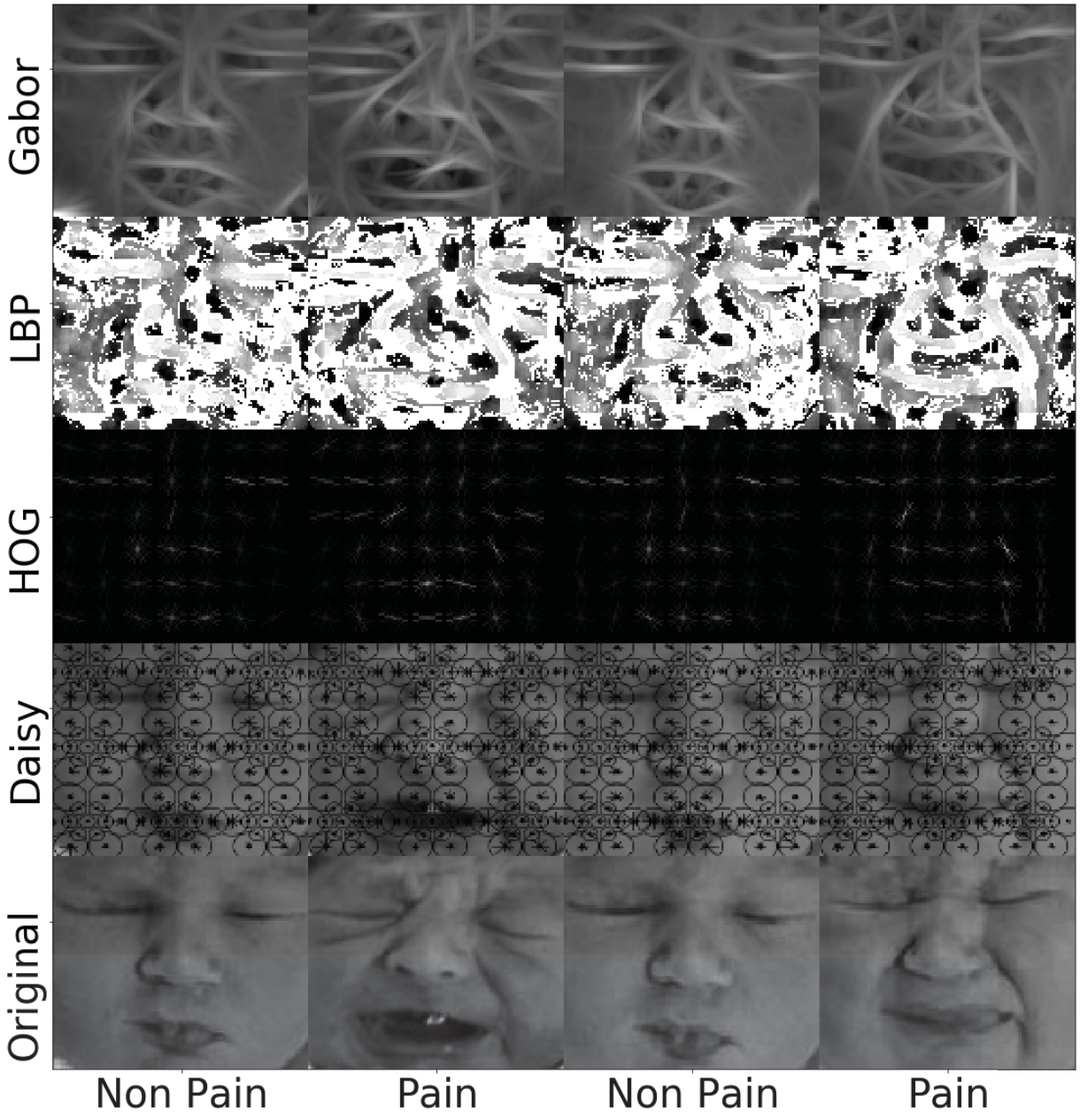


Figura 5.2: Exemplos das características testadas na base COPE.

gaussiano foi o LBP e o menos resistente, foi o HOG, similar aos resultados do ruído sal e pimenta.

Após os testes nos métodos convencionais de processamento de imagem, queríamos verificar o quanto métodos mais avançados de extração de características se compariam, i.e. características de Redes Neurais Convolucionais (CNNs). Mais ainda, assim como (Zamzmi et al., 2018), o objetivo era analisar níveis diferentes de extração de features, comparando diferentes camadas de uma mesma CNN. Em adição, também visava-se avaliar a diferença entre os métodos de *max pooling* e *avg pooling*.Tudo isso, sendo possível extrair características dos 3 canais de cores das imagens.

As redes escolhidas para teste foram a Resnet50 (He et al., 2016) e a MobileNet (Howard et al., 2017), duas redes que performam muito bem no challenge da ImageNet (Russakovsky et al., 2015). Utilizando técnicas de transfer learning, utilizamos as redes já treinadas com os

Figura 5.3: Exemplo do aumento de ruído sal e pimenta na base COPE

pesos finais da ImageNet para extrair características das imagens. Os resultados desse primeiro teste podem ser vistos na tabela 5.4

Na tabela 5.4, podemos observar que as camadas mais altas ou mais profundas tiveram um desempenho melhor, e também que o average pooling teve desempenho melhor que o maximum pooling, de forma geral.

Um ponto interessante de se notar é que o melhor método tradicional, o filtro de Gabor, atingiu um recall de 45%, enquanto a melhor feature de rede, a ResNet50 na camada 49 com average pooling atingiu um recall de 68%, mostrando uma considerável melhora alterando apenas a extracão de feature.

Os mesmos testes de robustez à ruídos foram feitos para as features de redes neurais. Os exemplos de ruídos sal e pimenta testados podem ser vistos na figura 5.7, e do ruído gaussiano na figura 5.8.

Os resultados da performance com o aumento o ruído sal e pimenta pode ser visto na figura 5.9 e do ruído gaussiano na figura 5.10. As tabelas com os resultados em detalhe dos

Figura 5.4: Exemplo do aumento de gaussiano na base COPE

deltas nas performances podem ser vistas na tabela 5.5 para sal e pimenta e tabela 5.6 para o ruído gaussiano.

Para o ruído de sal e pimenta, a feature menos resistente ao ruído foi a MobileNet, na camada 11 com average pooling. A com o delta menor, e portanto teoricamente a mais resistente foi a MobileNet, na camada 3 com maximum pooling, porém ao analisar o desempenho final, este é zerado. Esse comportamento da mais resistente se manteve ara o ruído gaussiano, porém a menos resistente foi a Resnet50, na camada 49, com maximum pooling.

De forma geral, o que observamos é que as features de redes neurais resultaram em um desempenho melhor do que as features de processamento de imagens tradicional. Em relação à robustez ao ruído, as redes neurais apresentaram maiores deltas entre o desempenho inicial, sem ruído, e o desempenho final, com a imagem altamente ruidosa.

Analisando apenas as redes neurais, houve um certo nível de separação de desempenho entre as camadas que foram reduzidas por average pooling e maximum pooling, com as camadas

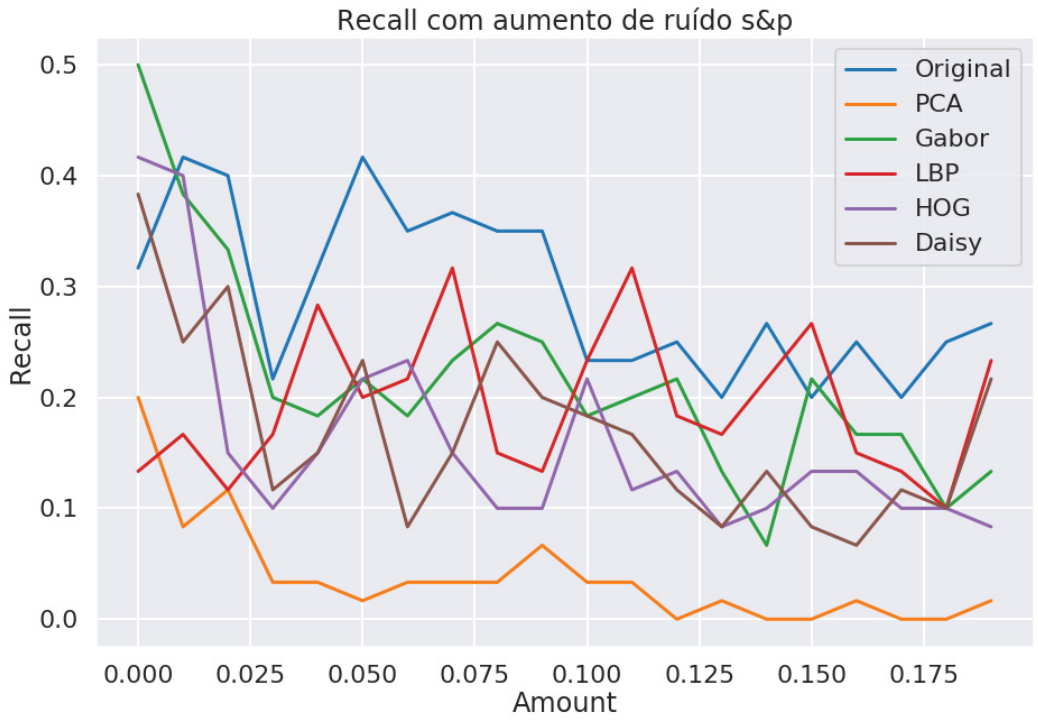


Figura 5.5: Performance dos métodos com o parâmetro *amount* aumentando

Tabela 5.2: Performances nos dois cenários extremos sem e com ruído sal e pimenta

Método	**Inicial**	**Final**	**Delta**
Original	0.316667	0.266667	**0.050000**
PCA	0.200000	0.016667	0.183333
Gabor	0.500000	0.133333	**0.366667**
LBP	0.133333	0.233333	-0.100000
HOG	0.416667	0.083333	0.333333
Daisy	0.383333	0.216667	0.166667

de average pooling ficando no geral, com melhor desempenho que as de maximum pooling. Já na robustez ao ruído, nenhuma ordenação significante foi observada.

Outro aspecto importante do desempenho das redes neurais é que as que tiveram melhor desempenho foram as camadas mais altas (profundas) da rede, porém foram as mais rasas e intermediárias que foram mais robustas ao ruído.

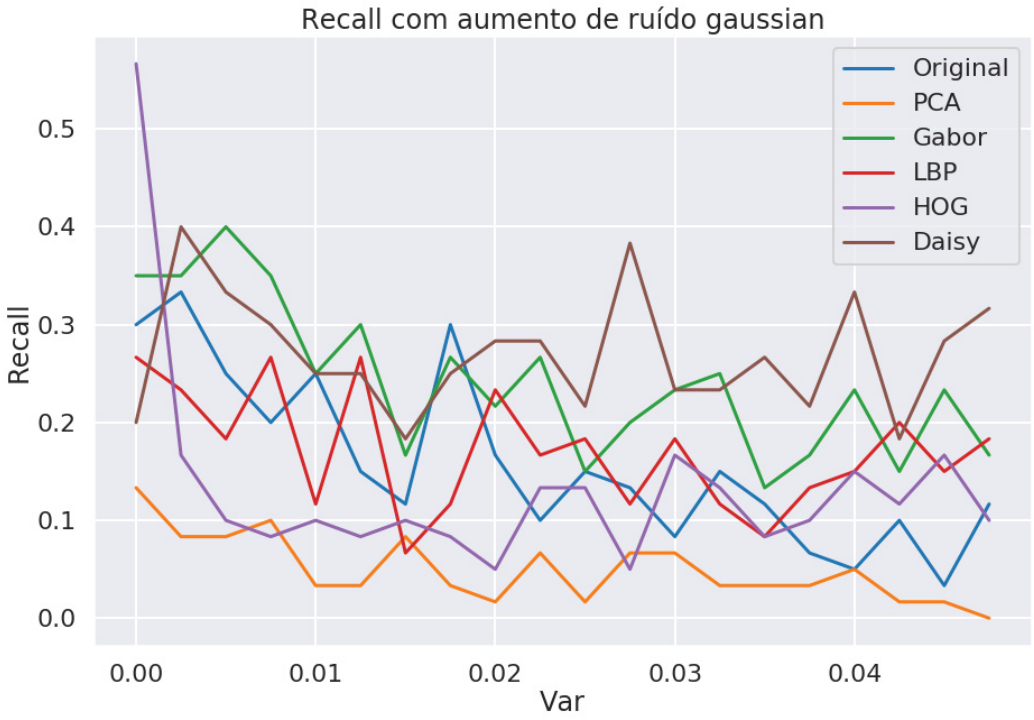


Figura 5.6: Performance dos métodos com o parâmetro *var* aumentando

Tabela 5.3: Performances nos dois cenários extremos sem e com gaussiano

Método	**Inicial**	**Final**	**Delta**
Original	0.300000	0.116667	0.183333
PCA	0.133333	0.000000	0.133333
Gabor	0.350000	0.166667	0.183333
LBP	0.266667	0.183333	**0.083333**
HOG	0.566667	0.100000	**0.466667**
Daisy	0.200000	0.316667	-0.116667

Tabela 5.4: Performances com features de diferentes camadas da ResNet50 e MobileNet

Modelo	**Camada**	**Pooling**	**Precision**	**Recall**	**F1**	**Acurácia**	**FP**	**TN**	**TP**	**TN**	**AUC**
Resnet50	49	AVG	0,707	**0,683**	0,695	0,824	17	19	41	127	0,857
Resnet50	37	AVG	0,791	0,567	0,66	**0,828**	9	26	34	135	0,841
Resnet50	49	MAX	0,708	0,567	0,63	0,804	14	26	34	130	0,832
Resnet50	25	AVG	0,733	0,55	0,629	0,809	12	27	33	132	0,833
MobileNet	11	AVG	0,608	0,517	0,559	0,76	20	29	31	124	0,792
MobileNet	13	MAX	0,705	0,517	0,596	0,794	13	29	31	131	0,815
MobileNet	3	AVG	0,62	0,517	0,564	0,765	19	29	31	125	0,803
MobileNet	13	AVG	0,667	0,5	0,571	0,779	15	30	30	129	0,812
Resnet50	13	AVG	0,638	0,5	0,561	0,77	17	30	30	127	0,782
MobileNet	5	AVG	0,612	0,5	0,55	0,76	19	30	30	125	0,763
Resnet50	37	MAX	0,794	0,45	0,574	0,804	7	33	27	137	0,853
Resnet50	13	MAX	0,727	0,4	0,516	0,779	9	36	24	135	0,827
Resnet50	25	MAX	0,656	0,35	0,457	0,755	11	39	21	133	0,794
MobileNet	11	MAX	0,63	0,283	0,391	0,74	10	43	17	134	0,724
MobileNet	5	MAX	0,577	0,25	0,349	0,725	11	45	15	133	0,677
MobileNet	3	MAX	0,407	0,183	0,253	0,681	16	49	11	128	0,649

Figura 5.7: Exemplos de ruídos de sal e pimenta utilizados

Figura 5.8: Exemplos de ruídos gaussianos utilizados

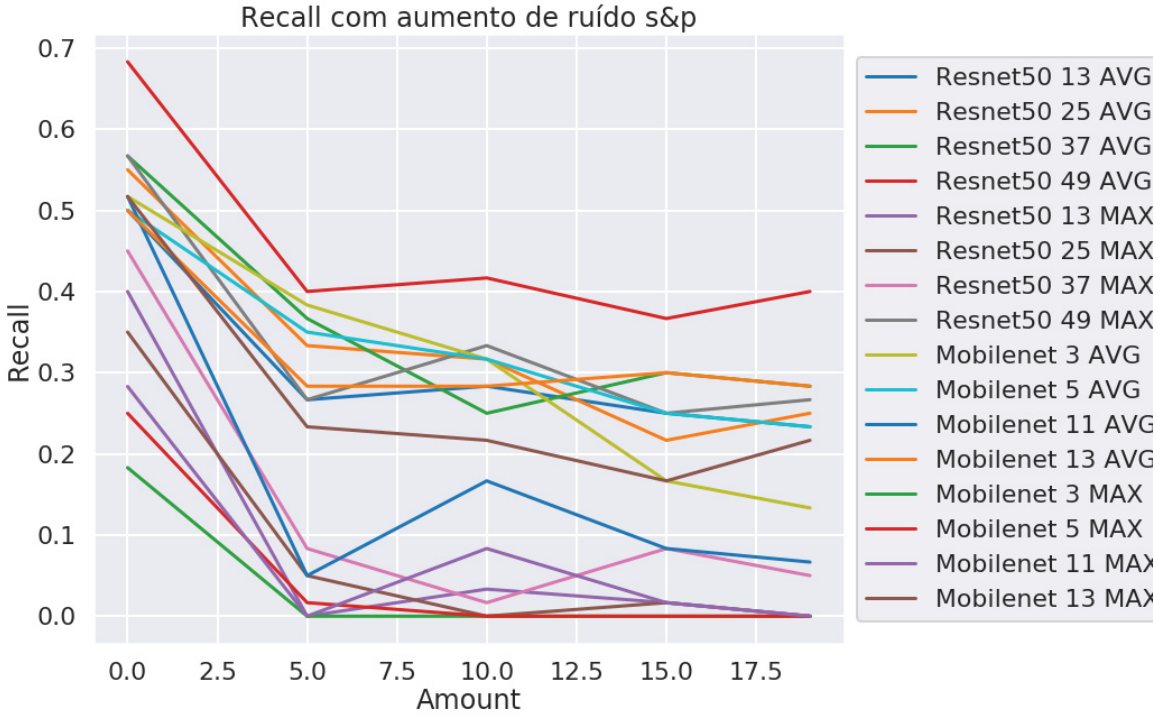


Figura 5.9: Performance dos métodos com o parâmetro *amount* aumentando

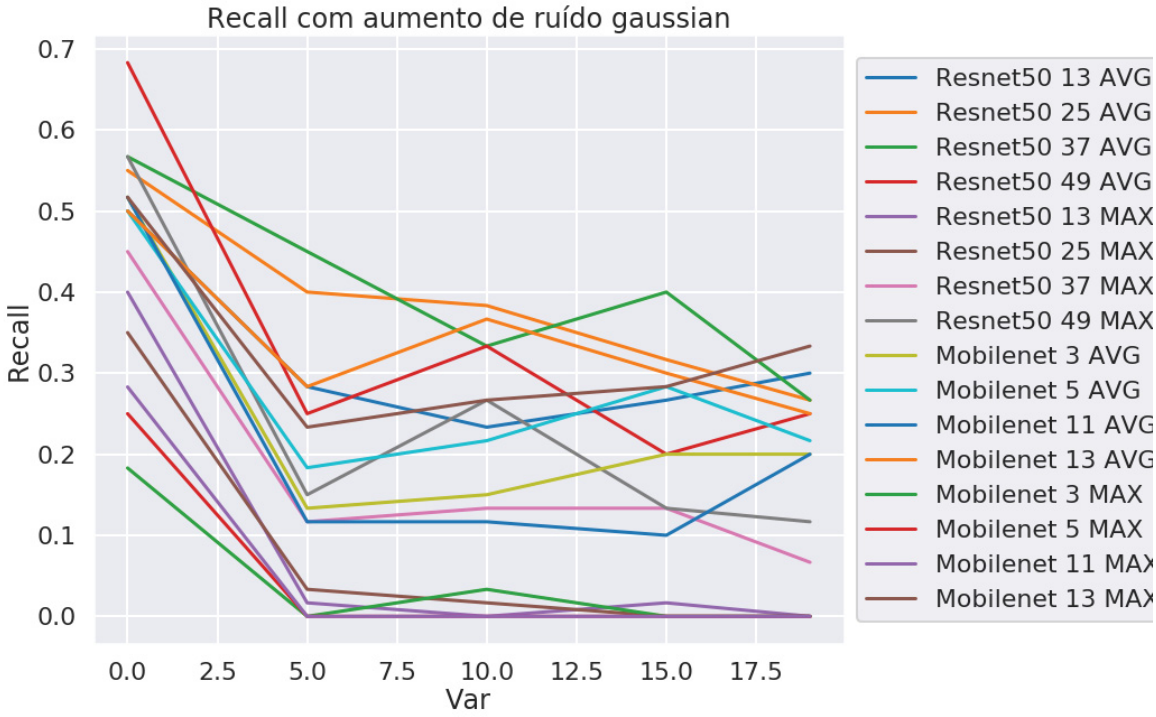


Figura 5.10: Performance dos métodos com o parâmetro *var* aumentando

Tabela 5.5: Performances nos dois cenários extremos sem e com sal e pimenta

Modelo	Camada	Pooling	Inicial	Final	Delta
Resnet50	13	AVG	0.500	0.233	0.267
Resnet50	25	AVG	0.550	0.250	0.300
Resnet50	37	AVG	0.567	0.283	0.284
Resnet50	49	AVG	0.683	0.400	0.283
Resnet50	13	MAX	0.400	0.000	0.400
Resnet50	25	MAX	0.350	0.000	0.350
Resnet50	37	MAX	0.450	0.050	0.400
Resnet50	49	MAX	0.567	0.267	0.300
Mobilenet	3	AVG	0.517	0.133	0.384
Mobilenet	5	AVG	0.500	0.233	0.267
Mobilenet	11	AVG	0.517	0.067	0.450
Mobilenet	13	AVG	0.500	0.283	0.217
Mobilenet	3	MAX	0.183	0.000	0.183
Mobilenet	5	MAX	0.250	0.000	0.250
Mobilenet	11	MAX	0.283	0.000	0.283
Mobilenet	13	MAX	0.517	0.217	0.300

Tabela 5.6: Performances nos dois cenários extremos sem e com ruído gaussiano

Modelo	Camada	Pooling	Inicial	Final	Delta
Resnet50	13	AVG	0.500	0.300	0.200
Resnet50	25	AVG	0.550	0.267	0.283
Resnet50	37	AVG	0.567	0.267	0.300
Resnet50	49	AVG	0.683	0.250	0.433
Resnet50	13	MAX	0.400	0.000	0.400
Resnet50	25	MAX	0.350	0.000	0.350
Resnet50	37	MAX	0.450	0.067	0.383
Resnet50	49	MAX	0.567	0.117	0.450
Mobilenet	3	AVG	0.517	0.200	0.317
Mobilenet	5	AVG	0.500	0.217	0.283
Mobilenet	11	AVG	0.517	0.200	0.317
Mobilenet	13	AVG	0.500	0.250	0.250
Mobilenet	3	MAX	0.183	0.000	0.183
Mobilenet	5	MAX	0.250	0.000	0.250
Mobilenet	11	MAX	0.283	0.000	0.283
Mobilenet	13	MAX	0.517	0.333	0.184

6 CONCLUSÃO

O problema de avaliação de dor em neonatos foi estudado nesse trabalho. Sabe-se que apesar de ser uma área com grande progresso nos últimos anos, avaliar dor em neonatos está longe de ser uma tarefa fácil. Uma das principais maneiras que profissionais de saúde realizam essa tarefa é analisando indicadores comportamentais. Essa tarefa, na maioria dos casos, tem seu desempenho afetado pela subjetividade humana inerente ao processo.

Desta maneira, métodos para avaliação automática de dor são necessários. Um dos aspectos mais importantes na expressão de sentimentos é a expressão facial, que também é avaliada em várias escalas de dor de neonatos. Vários sistemas foram revisados, e um levantamento das bases utilizadas foi realizado. De todas as bases revisadas, todos os autores foram contatados e os dados solicitados, e apenas a base COPE foi disponibilizada. Por se tratar de uma população vulnerável, a coleta de dados de recém nascidos é difícil e necessita de vários níveis de aprovação, que geralmente são aprovações para que apenas quem coletou os dados possa utilizá-los. Isso que implica na pouca existência de bases de dados nessa área, sendo que apenas uma dessas bases é púbica.

Dois métodos de detecção de face foram testados na base original, e ambos não tiveram um desempenho satisfatório. Esses resultados eram esperados pelos métodos serem amplamente utilizados para detecção de faces de adultos e não terem sido treinados com faces de bebês. Os resultados mostram que existe a necessidade de sistemas específicos para esse corte populacional.

Quatro métodos diferentes de extração de características foram testados na base processada, métodos de processamento de imagem e métodos de visão computacional. No desempenho geral, características provenientes de redes neurais

- Características provenientes de redes neurais:
    1. Tiveram melhor desempenho nas imagens sem ruído;
    2. Camadas mais altas se mostraram com melhor desempenho no geral;
    3. Camadas rasas e intermediárias se mostraram mais robustas a ruído.
- Características de processamento de imagem:
    1. Tiveram deltas menores nas comparações de ruídos.

A base COPE, apesar de bem consolidada, é uma base pequena, com apenas 204 imagens. Os testes realizados tinham como objetivo comparar métodos tradicionais de imagens de e métodos mais recentes de visão computacional, o que com feito com sucesso. Os métodos de visão computacional se mostraram superiores, o que era esperado.

Trabalhos futuros envolvem avanço para comparações mais robustas alterando a complexidade da parte de classificação, que para termos de comparação foi baseline para todas as características.

REFERÊNCIAS

Anand, K. J., Hickey, P. R. et al. (1987). Pain and its effects in the human neonate and fetus. *N Engl j Med*, 317(21):1321–1329.

Anand, K. J., Stevens, B. J. e McGrath, P. (2007). *Pain in Neonates and Infants*. Elsevier, 3th edition.

Arias, M. C. C. e Guinsburg, R. (2012). Differences between uni-and multidimensional scales for assessing pain in term newborn infants at the bedside. *Clinics*, 67(10):1165–1170.

Balda, R. D. C. X., Almeida, M. F. B. d., Peres, C. d. A. e Guinsburg, R. (2009). Fatores que interferem no reconhecimento por adultos da expressão facial de dor no recém-nascido. *Revista Paulista de Pediatria*.

Balda, R. d. C. X., Guinsburg, R., de Almeida, M. F. B., de Araújo Peres, C., Miyoshi, M. H. e Kopelman, B. I. (2000). The recognition of facial expression of pain in full-term newborns by parents and health professionals. *Archives of pediatrics & adolescent medicine*, 154(10):1009–1016.

Boyle, E. M., Bradshaw, J. e Blake, K. I. (2018). Persistent pain in neonates: challenges in assessment without the aid of a clinical tool. *Acta Paediatrica*, 107(1):63–67.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Brahnam, S., Chuang, C. e F. Shin, M. S. (2005). Machine recognition and representation of neonatal facial displays of acute pain. *Artificial Intelligence in Medicine*.

Breaus, L. M., McGrath, P. J. ., Craig, K. D., Santor, D., Cassidy, K. e Reid, G. J. (2001). Facial expression of children receiving immunizations: A principal components analysis of the child facial coding system. *The Clinical Journal of Pain*, 17:178–186.

Bussotti, E. A., Guinsburg, R. e Pedreira, M. d. L. G. (2015). Cultural adaptation to brazilian portuguese of the face, legs, activity, cry, consolability revised (flaccr) scale of pain assessment. *Revista latino-americana de enfermagem*, 23(4):651–659.

Chandrasekaran, B. e Keuneke, A. (1987). Classification problem solving: a tutorial from an ai perspective. Em *Pattern Recognition Theory and Applications*, páginas 393–409. Springer.

Chermont, A. G., Guinsburg, R., Balda, R. d. C. X. e Kopelman, B. I. (2003). O que os pediatras conhecem sobre avaliação e tratamento da dor no recém-nascido? *Jornal de Pediatria*.

Craig, K. D., Whitfield, M. F., Grunau, R. V., Linton, J. e Hadjistavropoulos, H. D. (1993). Pain in the preterm neonate: behavioural and physiological indices. *Pain*, 52(3):287–299.

Cremillieux, C., Makhlouf, A., Pichot, V., Trombert, B. e Patural, H. (2018). Objective assessment of induced acute pain in neonatology with the newborn infant parasympathetic evaluation index. *European Journal of Pain*, 22(6):1071–1079.

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

Dequeker, S., Van Lancker, A. e Van Hecke, A. (2018). Hospitalized patients' vs. nurses' assessments of pain intensity and barriers to pain management. *Journal of advanced nursing*, 74(1):160–171.

Devlin, J. W., Skrobik, Y., Gélinas, C., Needham, D. M., Slooter, A. J., Pandharipande, P. P., Watson, P. L., Weinhouse, G. L., Nunnally, M. E., Rochwerg, B. et al. (2018). Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the icu. *Critical care medicine*, 46(9):e825–e873.

Elias, L. S., Guinsburg, R., Peres, C. A., Balda, R. C., Santos, A. et al. (2008). Disagreement between parents and health professionals regarding pain intensity in critically ill neonates. *Jornal de pediatria*, 84(1):35–40.

Fitzgerald, M. e McIntosh, N. (1989). Pain and analgesia in the newborn. *Archives of Disease in Childhood*, 64(4 Spec No):441.

Gholami, B., Haddad, W. M. e Tannenbaum, A. R. (2010). Relevance vector machine learning for neonate pain intensity assessment using digital imaging. *IEEE Transactions on biomedical engineering*, 57(6):1457–1466.

Grunau, R. E. e Craig, K. D. (1987). Pain expression in neonates: facial action and cry. *Pain*, 28:395–410.

Grunau, R. E., Oberlander, T., Holsti, L. e Whitfield, M. F. (1998). Bedside application of the neonatal facial coding sytem in pain assessment of premature neonates. *Pain*, 76:277–286.

Guinsburg, R. (1999). Avaliação e tratamento da dor no recém-nascido. *J Pediatr (Rio J)*, 75(3):149–60.

Guinsburg, R., Balda, R. d. C. X., Berenguel, R. C., Almeida, M. F. B., Tonelloto, J., Santos, A., Kopelman, B. I. et al. (1997). Aplicação das escalas comportamentais para a avaliação da dor em recém-nascidos. *J Pediatr*, 73(6):411–8.

Guinsburg, R. e Cuenca, M. C. (2010). A linguagem da dor no recém-nascido. *São Paulo: Sociedade Brasileira de Pediatria.[Internet]*.

Hazelhoff, L., Han, J., Bambang-Oetomo, S. e de With, P. H. N. (2009). Behavioral state detection of newborns bases on facial expression analysis. *Advanced Concepts for Intelligent Vision systems*.

He, K., Zhang, X., Ren, S. e Sun, J. (2016). Deep residual learning for image recognition. Em *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 770–778.

Hjelmås, E. e Low, B. K. (2001). Face detection: A survey. *Computer vision and image understanding*, 83(3):236–274.

Holsti, L., Grunau, R. E., Oberlander, T. F., Whitfield, M. F. e Weinberg, J. (2005). Body movements: an important additional factor in discriminating pain from stress in preterm infants. *The Clinical journal of pain*, 21(6):491.

Hossin, M. e Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. e Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hudson-Barr, D., Capper-Michel, B., Lambert, S., Palermo, T. M., Morbeto, K. e Lombardo, S. (2002). Validation of the pain assessment in neonates (pain) scale with the neonatal infant pain scale (nips). *Neonatal Network*, 21(6):15–22.

Hummel, P. e van Dijk, M. (2006). Pain assessment: current status and challenges. Em *Seminars in Fetal and Neonatal Medicine*, volume 11, páginas 237–245. Elsevier.

James, G., Witten, D., Hastie, T. e Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jesus, J. A. L. d. (2011). Condutância da pele como indicador de dor aguda no recém-nascido: estudo comparativo com frequência cardíaca, saturação de oxigênio e escalas comportamentais de dor.

Kamel, M. e Campilho, A. (2009). *Image Analysis and Recognition: 6th International Conference, ICIAR 2009, Halifax, Canada, July 6-8, 2009, Proceedings*, volume 5627. Springer Science & Business Media.

Kazemi, V. e Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, páginas 1867–1874.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

Lawrence, J., Alcock, D., McGrath, P., Kay, J., MacMurray, S. B. e Dulberg, C. (1993). The development of a tool to assess neonatal pain. *Neonatal network: NN*, 12(6):59–66.

Liu, Z.-Q., Cai, J.-H. e Buse, R. (2012). *Handwriting recognition: soft computing and probabilistic approaches*, volume 133. Springer.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Lu, G., Li, X. e Li, H. (2008a). Facial expression recognition for neonatal pain assessment. Em *2008 International Conference on Neural Networks and Signal Processing*, páginas 456–460.

Lu, G., Yang, C., Chen, M. e Li, X. (2016). Sparse representation based facial expression classification for pain assessment in neonates. Em *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, páginas 1615–1619.

Lu, G., Yuan, L., Li, X. e Li, H. (2008b). Facial expression recognition of pain in neonates. Em *2008 International Conference on Computer Science and Software Engineering*, volume 1, páginas 756–759.

Mansor, M. N., Jamil, S. H.-F. S. A., Junoh, A. K., Rejab, M. N., Jamil, A. H.-F. S. A. e Ahmad, J. (2012a). Fast infant pain detection method. Em *2012 International Conference on Computer and Communication Engineering (ICCCE)*, páginas 918–921.

Mansor, M. N., Jamil, S. H. F. S. M., Rejab, M. N. e Jamil, A. H. F. S. M. (2012b). K-nn algorithm for fast infant pain detection. Em *2012 International Symposium on Instrumentation Measurement, Sensor Network and Automation (IMSNA)*, volume 2, páginas 358–360.

Mansor, M. N., Junoh, A. K., Ahmed, A. e Osman, M. K. (2014). Single scale self quotient image and pnn for infant pain detection. Em *2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014)*, páginas 553–555.

Mansor, M. N. e Rejab, M. N. (2013a). Infant pain recognition system with glcm features and gann under unstructed lighting condition. Em *2013 IEEE International Conference on Control System, Computing and Engineering*, páginas 243–248.

Mansor, M. N. e Rejab, M. N. (2013b). Phase congruency image and sparse classifier for newborn classifying pain state. Em *2013 IEEE International Conference on Control System, Computing and Engineering*, páginas 450–454.

Mansor, M. N., Syam, S. H. F., Rejab, M. N. e b, A. H. F. S. (2012c). Ar model for infant pain anxiety recognition using fuzzy k-nn. Em *2012 International Symposium on Instrumentation Measurement, Sensor Network and Automation (IMSNA)*, volume 2, páginas 374–376.

Manworren, R. C. B. e Hynan, L. S. (2003). Clinical validation of flacc: Preverbal patient pain scale. *Pediatric Nursing*, 29(2):140–146.

Marchi, A., Vellucci, R., Mameli, S., Piredda, A. R. e Finco, G. (2009). Pain biomarkers. *Clinical drug investigation*, 29(1):41–46.

Maxwell, L. G., Malavolta, C. P. e Fraga, M. V. (2013). Assessment of pain in the neonate. *Clinics in perinatology*, 40(3):457–469.

Michie, D., Spiegelhalter, D. J. e Taylor, C. C. (1994). Machine learning, neural and statistical classification.

Pereira, A. L. d. S. T., Guinsburg, R., Almeida, M. F. B. d., Monteiro, A. C., Santos, A. M. N. d. e Kopelman, B. I. (1999). Validity of behavioral and physiologic parameters for acute pain assessment of term newborn infants. *São Paulo medical journal*, 117(2):72–80.

Peters, J. W. B., Koot, H. M., Grunau, R. E., Boer, J., van Druenen, M. J., Tibboel, D. e Duivenvoorden, H. (2002). Neonatal facial coding system for assessing postoperative pain in infants: Item reduction is valid and feasible. *The Clinical Journal of Pain*.

Presbytero, R., da Costa, M. L. V. e Santos, R. C. S. (2010). Os enfermeiros da unidade neonatal frente ao recém-nascido com dor. *Revista da rede de enfermagem do Nordeste*, 11(1):125–132.

Prkachin, K. M. (2009). Assessing pain by facial expression: Facial expression as nexus. *Pain Researh & Management: The Journal of the Canadian Pain Society*, 14(1):53–58.

Prkachin, K. M., Solomon, P., Hwang, T. e Mercer, S. R. (2001). Does experience influence judgements of pain behaviour? evidence from relatives of pain patients and therapists. *Pain Research and Management*, 6(2):105–112.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Santos, L. M., Pereira, M. P., dos Santos, L. F. N. e de Santana, R. C. B. (2012). Avaliação da dor no recém-nascido prematuro em unidade de terapia intensiva. *Revista Brasileira de Enfermagem*, 65(1):27–33.

Schechter, N. L., Berde, C. B. e Yaster, M. (2002). *Pain in Infants, Children and Adolescents*. Lippincott Williams & Wilkins, 2nd edition.

Schiavenato, M., Byers, J. F., Scovanner, P., McMahon, J. M., Xia, Y., Lu, N. e He, H. (2008). Neonatal pain expression: Evaluating the primal face of pain. *Pain*, 138:240–271.

Slater, L., Asmerom, Y., Boskovic, D. S., Bahjri, K., Plank, M. S., Angeles, K. R., Phillips, R., Deming, D., Ashwal, S., Hougland, K. et al. (2012). Procedural pain and oxidative stress in premature neonates. *The journal of pain*, 13(6):590–597.

Sun, X., Wu, P. e Hoi, S. C. (2018). Face detection using deep learning: An improved faster rcnn approach. *Neurocomputing*, 299:42–50.

Sweet, S. D. e McGrath, C. P. J. (1998). Relative importance of mothers versus medical staffs behavior in the prediction of infant immunization pain behavior. *Journal of Pediatric Psychology*, 23(4):249–256.

Turk, D. C. e Melzack, R. (2011). *Pain in Infants, Children and Adolescents*. Guilford Publications, 3nd edition.

Viola, P. e Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Voepel-Lewis, T., Shayevitz, J. R. e Malviya, S. (1997). for scoring postoperative pain in young children. *Pediatric nursing*, 23(3).

Walters, A. R. (2018). Current evidence for pediatric triage pain protocols in the emergency department.

Wong, D. L. e Hockenberry, M. J. (2000). *Wong's Clinical Manual of Pediatric Nursing*. Mosby.

Yuan, L., Bao, F. S. e Lu, G. (2008). Recognition of neonatal facial expressions of acute pain using boosted gabor features. Em *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, volume 2, páginas 473–476.

Zamzmi, G., Goldgof, D., Kasturi, R. e Sun, Y. (2018). Neonatal pain expression recognition using transfer learning. *arXiv preprint arXiv:1807.01631*.

Zamzmi, G., Kasturi, R., Goldgof, D., Zhi, R., Ashmeade, T. e Sun, Y. (2017). A review of automated pain assessment in infants: Features, classification tasks, and databases. *IEEE Reviews in Biomedical Engineering*.

Zhong, Y. e Liu, L. (2011). Remote neonatal pain assessment system based on internet of things. Em *2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing*, páginas 629–633.

Zhu, X. e Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. Em *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, páginas 2879–2886. IEEE.