UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

DOUTORADO EM ENGENHARIA DE TELEINFORMÁTICA

MANUEL GONÇALVES DA SILVA NETO

MODELOS DE CLASSIFICAÇÃO E PROGNÓSTICO PARA AVALIAÇÃO DO BEM-ESTAR FETAL

FORTALEZA

2021

MANUEL GONÇALVES DA SILVA NETO

MODELOS DE CLASSIFICAÇÃO E PROGNÓSTICO PARA AVALIAÇÃO DO BEM-ESTAR FETAL

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Orientador: Prof. Dr. Danielo Gonçalves Gomes

Coorientador: Prof. Dr. João Paulo do Vale Madeiro

FORTALEZA

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S581m Silva Neto, Manuel Gonçalves da.
Modelos de classificação e prognóstico para avaliação do bem-estar fetal / Manuel Gonçalves da Silva Neto. – 2021.
152 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2021.
Orientação: Prof. Dr. Danielo Gonçalves Gomes.
Coorientação: Prof. Dr. João Paulo do Vale Madeiro.

1. Modelo de prognóstico. 2. Classificação. 3. Frequência cardíaca fetal. 4. Avaliação do estado fetal. I. Título.

CDD 621.38

MANUEL GONÇALVES DA SILVA NETO

MODELOS DE CLASSIFICAÇÃO E PROGNÓSTICO PARA AVALIAÇÃO DO BEM-ESTAR FETAL

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Aprovada em: 01 de Junho de 2021

BANCA EXAMINADORA

Prof. Dr. Danielo Gonçalves Gomes (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. João Paulo do Vale Madeiro (Coorientador)
Universidade Federal do Ceará (UFC)

Prof. Dra. Debora Christina Muchaluat Saade
Universidade Federal Fluminense (UFF)

Prof. Dr. João Alexandre Lôbo Marques
University of Saint Joseph - Macau (USJ)

Prof. Dr. Francisco Herlânio Costa Carvalho
Universidade Federal do Ceará (UFC)

Prof. Dr. Paulo César Cortez
Universidade Federal do Ceará (UFC)

À minha família.

AGRADECIMENTOS

Ao Prof. Dr. Danielo Gonçalves Gomes pelas orientações ao longo deste doutorado.

Aos Professores Doutores João Paulo do Vale Madeiro e Alexandre Lobo Marques pelas orientações e disponibilização dos biossinais referentes a base de dados rotulada *DB-Trium*.

Ao Prof. Dr. Flávio Rubens de Carvalho Sousa, pela disponibilidade dos biossinais referentes a base de dados não rotulada *DB-HeraBeat*.

Ao aluno de Doutorado em Engenharia de Teleinformática e amigo de laboratório, Bruno Riccelli dos Santos Silva, pelas discussões e parceria ao longo das madrugadas de pesquisa.

A minha querida esposa, a minha filha, aos meus familiares e amigos, que nos momentos em que estive ausente devido aos estudos, compreenderam que o futuro é feito a partir da constante dedicação no presente!

“Todas as vitórias ocultam uma abdicação.”

(Simone de Beauvoir)

RESUMO

Biossinais provenientes do monitoramento fetal podem ser adotados como indicadores para avaliação do seu bem-estar. A análise assistida por computador dos padrões em biossinais como a frequência cardíaca fetal (FCF) de forma isolada ou combinada com dados clínicos dos pacientes tem sido utilizadas como ferramentas de suporte a decisão ou em modelos de prognóstico nos ambientes clínicos. Atualmente os sistemas baseados em biossinais possuem funcionalidades que variam desde a simples exibição de indicadores de bem estar fetal até a capacidade de classificar e predizer o estado do feto baseado em padrões. No entanto, a construção destes modelos de prognóstico enfrenta a ausência de padrões ou consensos entre os parâmetros e blocos de construção adotados em seu desenvolvimento, dificultando a sua comparação de forma objetiva assim como o desenvolvimento de novas soluções. Nesta tese, um processo de avaliação dos blocos de construção dos sistemas de suporte a diagnóstico fetal foi desenvolvido. Foi proposto ainda um modelo de prognóstico baseado em técnicas de processamento avançado de sinais e aprendizado de máquina como ferramenta de suporte na avaliação do bem estar do feto. Primeiramente, foram adotados guias bem estabelecidos para realizar um mapeamento sistemático da literatura a fim de estabelecer uma visão geral do estado da arte relacionados aos principais parâmetros e blocos de construção utilizados no desenvolvimento de sistemas de apoio ao diagnóstico fetal baseados em biossinais. Em seguida, um processo de avaliação destes blocos de construção foi desenvolvido, onde algoritmos de aprendizado de máquina supervisionados foram avaliados em cenários baseados em séries temporais e engenharia de *features* dentro de três esquemas de segmentação dos sinais. O processo de avaliação utilizou ainda de aprendizagem de máquina semi-supervisionada para análise do estado fetal, utilizando bases de dados de biossinais anotadas em conjunto com bases de dados não anotadas. Por fim, um modelo de prognóstico foi desenvolvido tomando por base a combinação dos parâmetros e blocos de construção mais bem avaliados. O processo de avaliação resultou em uma combinação de blocos de construção que melhoraram a performance do modelo de prognóstico final, atingindo resultados competitivos em ambientes com forte limitação de dados em comparação a soluções de avaliação do bem estar fetal do estado da arte.

Palavras-chave: Modelo de prognóstico. Classificação. Frequência cardíaca fetal. Avaliação do estado fetal

ABSTRACT

Fetal monitoring signals such as fetal heart rate (FHR) are critical indicators of fetal well-being. Computer-assisted analysis of FHR patterns alone or combined with patient' clinical data has been used as a decision supporting-tool and prognostic models in the clinical environment. The biosignal-based systems comprise functionalities that varying from fetal wellbeing indicators exhibition to predictive and pattern classification capabilities; However, the prognostic model design suffers from the lack of gold standards for the building blocks decision-making. Thus impairing the direct comparison of proposals and the development of new solutions. In this thesis we propose an evaluation process for the building blocks of the decision supporting-tools and a prognostic model based on advanced signal processing techniques and machine learning algorithms for the fetal state assessment. First, we employed well-established guidelines to systematically map the literature in order to establish an overview of the state of the art related to the main parameters and building blocks used in the development of support systems for biosignal-based fetal diagnosis systems. Then, a process for evaluating these building blocks was developed, in which supervised machine learning algorithms were evaluated separately for time series and feature engineering within three signal segmentation schemes. The evaluation process also utilizes a semi-supervised machine learning algorithm to analyze the fetal state using annotated biosignal databases in conjunction with non-annotated databases. Finally, a prognostic model was developed based on the combination of parameters and the best-evaluated building blocks. The evaluation process provided a combination of techniques and methods that increased the final prognostic model performance, achieving relevant results with a smaller amount of data when compared to the state-of-the-art fetal status assessment solutions.

Keywords: Prognostic model. Classification. Fetal heart rate. Fetal state assessment

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Esta tese de doutorado propõe um processo de avaliação e modelo de prognóstico baseado em técnicas de processamento avançado de sinais e aprendizado de máquina como ferramentas de apoio a avaliação do bem-estar fetal. Este Capítulo de Introdução traz a contextualização, descrição do problema em foco, hipóteses, questões de pesquisa, objetivos e estrutura deste manuscrito.

1.1 Contextualização

Com mais de 50 anos desde sua primeira adoção em ambiente clínico, o monitoramento da frequência cardíaca fetal (FCF) e contrações uterinas (UC) conquistou um espaço permanente nos cuidados obstétricos e no processo de decisão dos médicos (AYRES-DE-CAMPOS, 2018). Cardiotocografia (CTG) é um dos métodos principais para monitoramento biofísico do desenvolvimento fetal intra-uterino por registro simultâneo de sinais FCF e função UC (CZABANSKI *et al.*, 2016).

A identificação a tempo dos fetos oxigenados inadequadamente permitem a tomada de ação apropriada antes da ocorrência de danos. O sofrimento fetal é caracterizado pela diminuição da concentração de oxigênio no sangue arterial, conhecida como hipoxemia, esta falta de oxigenação de forma severa atinge os tecidos e é conhecida por hipóxia (AYRES-DE-CAMPOS *et al.*, 2015a). As observações no CTG são amplamente utilizadas para avaliar o bem-estar do feto, apoiando a detecção precoce de uma condição patológica e auxiliando na previsão de complicações futuras ou intervenção pré-invasiva no feto (IRAJI, 2019). Apesar da existência de guias de avaliação bem estabelecidos no apoio a inspeção visual dos registros de CTG pelos clínicos, como os guias da *International Federation of Gynecology and Obstetrics* (FIGO) (AYRES-DE-CAMPOS *et al.*, 2015b), essas análises visuais podem levar à variabilidade intra e interobservador, contribuindo para partos cirúrgicos desnecessários (ABRY *et al.*, 2018).

Para minimizar a variabilidade da avaliação baseada em especialistas, pesquisadores de campos multidisciplinares como engenharia, ciência da computação e medicina combinam esforços para projetar sistemas de análise auxiliados por computador que visam o suporte aos médicos no diagnóstico do estado fetal. Assim, o desenvolvimento de sistemas de monitoramento fetal inteligentes que podem interpretar padrões de sinais biomédicos com precisão tem sido o tema de pesquisas em andamento (RICCIARDI *et al.*, 2020).

Parâmetros e indicadores pós-nascimento provenientes de monitoramentos retrospectivos, podem ser usados como referência para verificar a eficiência dos sistemas automatizados de classificação do estado fetal. Embora esta abordagem não seja totalmente livre de erros porque um resultado anormal pode ser o resultado de complicações durante o trabalho de parto, continua sendo o método de referência mais objetivo para determinar o grau de certeza da interpretação do estado fetal nos sistemas automatizados (CZABANSKI *et al.*, 2016).

Os modelos de prognóstico podem incluir habilidades preditivas e reconhecimento de padrões em que dados de avaliação fetal retrospectiva são usados como base para a interpretação da informação fetal atual (BARQUERO-PEREZ *et al.*, 2017; MARQUES *et al.*, 2019; ALSAGGAF *et al.*, 2020; IRAJI, 2019; ZHAO *et al.*, 2019; GEORGOULAS *et al.*, 2017; COMERT *et al.*, 2018). Nestes modelos, a capacidade de reconhecimento de casos patológicos e saudáveis deve ser a mais alta possível para um modelo de prognóstico eficiente e as taxas de verdadeiro-positivo precisam exceder 60% para benefícios clínicos satisfatórios (PETROZZIELLO *et al.*, 2018).

1.2 Motivação

O processo de desenvolvimento de ferramentas de suporte a decisão clínica para avaliação do estado fetal, em especial os sistemas dotados de habilidades como classificação e predição, enfrentam desafios que impactam no seu desempenho final (LI *et al.*, 2019). Nesta seção, dois problemas relacionados ao desenvolvimento de modelos de prognóstico são destacados: (i) ausência de padrões para caracterização do estado fetal nos sistemas de avaliação automática do feto (Seção 1.2.1) e (ii) a escolha de uma combinação ótima de blocos de construção para compor o modelo de prognóstico (Seção 1.2.2).

1.2.1 Problema 1: Ausência de padrões para caracterização do estado fetal nos sistemas de avaliação automática do feto

Em relação a caracterização do estado normal e anormal baseado em dados de monitoramento fetais retrospectivos, ainda não existe um padrão ouro em relação aos critérios de separação de classes para dividir os biossinais (ALSAGGAF *et al.*, 2020), assim como os critérios padronizados de segmentação de séries temporais provenientes dos biossinais para análise (BARQUERO-PEREZ *et al.*, 2017; ABRY *et al.*, 2018). Desse modo, torna-se difícil a

comparação das soluções existentes para avaliação do estado fetal de forma objetiva.

1.2.2 Problema 2: Necessidade de uma combinação eficiente de blocos de construção para compor o modelo de prognóstico

Nas fases de projeto e desenvolvimento destes sistemas, existem complexas tomadas de decisão em relação às abordagens e tecnologias empregadas que afetam a eficiência final do modelo. Pode-se aplicar engenharia de *features*, em que algoritmos baseados em conhecimento de especialistas são usados para extrair *features* de biossinais que podem representar informações relevantes para problemas específicos (SUPRATAK *et al.*, 2016). Por outro lado, o modelo pode usar abordagens de autoaprendizagem para extrair informações representativas diretamente das séries temporais nos biossinais (ZHAO *et al.*, 2019).

A eficiência de um modelo de prognóstico também pode ser afetada por problemas relacionados ao balanceamento de classes nos dados de monitoramento retrospectivo empregadas para o treinamento dos classificadores (RICCIARDI *et al.*, 2020) e pela escolha entre diferentes algoritmos de classificação. Também existem lacunas na avaliação da capacidade de generalização dos modelos entre diferentes conjuntos de dados retrospectivos para comprovar sua eficiência (ABRY *et al.*, 2018). Desta forma, a escolha dos blocos de construção para proposição de novas soluções torna-se um ponto crítico no projeto destes modelos. Um processo de avaliação destes blocos de construção de forma sistemática e extensível é necessária para o desenvolvimento de um modelo de prognóstico eficiente. O processo de avaliação mencionado determina sistematicamente os blocos de construção ótimos para o modelo final de prognóstico empregado para a avaliação do estado fetal.

1.3 Questões de Pesquisa

A partir dos problemas apresentados na seção 1.2, formulamos as seguintes questões de pesquisa **(QP)** associadas a avaliação do bem-estar fetal.

QP#1: Como estimar/identificar desfechos adversos para a vitalidade fetal, tais como sofrimento fetal, hipoxia, a partir de sinais biológicos fetais isolados e/ou combinados?

QP#2: Quais blocos de construção podem ser utilizados no desenvolvimento das ferramentas de auxílio no diagnóstico médico para reconhecimento e predição do estado fetal baseado em sinais biológicos?

QP#3: Utilizar um processo de avaliação dos blocos de construção de forma sistemática auxilia na construção de modelos de prognóstico eficientes e generalizáveis?

1.3.1 Hipóteses

Para cada questão de pesquisa, é possível associar uma hipótese abordada ao longo desta tese, conforme segue.

Hipótese #1 associada à **QP#1**: a primeira hipótese é que modelos supervisionados e semi-supervisionados de aprendizado de máquina podem ser usados para prever desfechos adversos para a vitalidade fetal, cumprindo os requisitos mínimos de 60% para taxas de verdadeiros positivos.

Hipótese #2 associada à **QP#2**: a segunda hipótese desta tese é a de que seja possível extrair informações relevantes e atualizadas da literatura no que diz respeito aos blocos de construção e parâmetros utilizados para o desenvolvimento de soluções para o suporte a decisão médica na avaliação do bem estar fetal baseados em biossinais.

Hipótese #3 associada à **QP#3**: a terceira hipótese é que o uso de um processo de avaliação bem definido e sistemático pode prover uma combinação eficiente de blocos de construção para construção de modelos de prognóstico com alto grau de generalização.

1.4 Objetivos

O objetivo principal desta tese é a proposição de metodologias para determinar sistematicamente os blocos de construção ótimos para o modelo de prognóstico, visando a avaliação dos níveis de bem-estar do feto. As referidas metodologias englobam desde a aquisição e agrupamento de conhecimento sobre a área de pesquisa até a avaliação dos blocos de construção e desenvolvimento dos modelos finais de prognóstico para avaliação do estado fetal.

Para atingir esse objetivo geral, foram estabelecidos objetivos específicos:

1. prover uma visão panorâmica da literatura atual sobre os blocos de construção empregados no projeto de sistemas para a interpretação assistida por computador de padrões de biossinais fetais;
2. integrar e avaliar diferentes fontes de dados no processo de desenvolvimento dos modelos de prognóstico;
3. extrair um conjunto de *features* representativas dos biossinais para avaliação do bem-estar

fetal;

4. fragmentar os biossinais em janelas de tempo referentes a hora do parto e avaliar sua capacidade de representação do bem-estar fetal;
5. avaliar algoritmos de aprendizado de máquina supervisionados e semi-supervisionados na discriminação de dados de monitoramento fetal retrospectivos entre dois estados fetais: normal e patológico;
6. investigar e determinar sistematicamente os blocos de construção ótimos para compor um modelo de prognóstico visando a avaliação dos níveis de bem-estar do feto;
7. obter um modelo de prognóstico para caracterização dos níveis de bem-estar do feto.

1.5 Metodologia da Pesquisa

A metodologia aplicada ao longo desta tese pode ser sumarizada conforme segue:

1. mapeamento sistemático da literatura para o levantamento do estado da arte sobre os blocos de construção e parâmetros utilizados para separação dos estados normal e patológico dos fetos nos modelos de prognóstico;
2. algoritmo de pré-processamento e integração de diferentes bases de biossinais em um único fluxo de processamento;
3. implementação e validação do processo de avaliação baseada em aprendizado de máquina supervisionado com algoritmos baseados em engenharia de *features* e em séries temporais;
4. análise do impacto da alteração na duração do biossinal em relação a hora do parto sobre o desempenho final do modelo;
5. análise do impacto do uso de técnicas de balanceamento de dados durante o processo de treinamento dos algoritmos no que se refere ao desempenho final do modelo;
6. implementação e avaliação de algoritmos de aprendizado de máquina semi-supervisionados na avaliação do bem-estar fetal;
7. análise do impacto da escolha da base de dados anotada no desempenho final do modelo semi-supervisionado;
8. análise do desempenho do modelo de prognóstico final em relação a modelos da literatura.

1.6 Contribuições

A principal contribuição desta tese é um modelo de prognóstico generalizável utilizando biossinais fetais, que visa fornecer suporte à decisão médica com base em métodos avançados de processamento de sinais e algoritmos de aprendizado de máquina (AM) supervisionados (Capítulos 5 e 6) e semi-supervisionados (Capítulo 7). Além desta contribuição central, podemos destacar algumas contribuições secundárias: (i) um panorama atualizado sobre os blocos de construção e parâmetros empregados no desenvolvimento de sistemas de suporte a decisão médica para avaliação do bem-estar fetal; (ii) um processo de avaliação para obtenção da combinação ótima dos blocos de construção empregados na composição do modelo de prognóstico; (iii) bases de dados rotuladas a partir de *features* extraídas durante os experimentos, incluindo a rotulagem automática com auxílio dos modelos semi-supervisionados. A Figura 1 apresenta uma visão geral dos modelos desenvolvidos no decorrer esta tese.

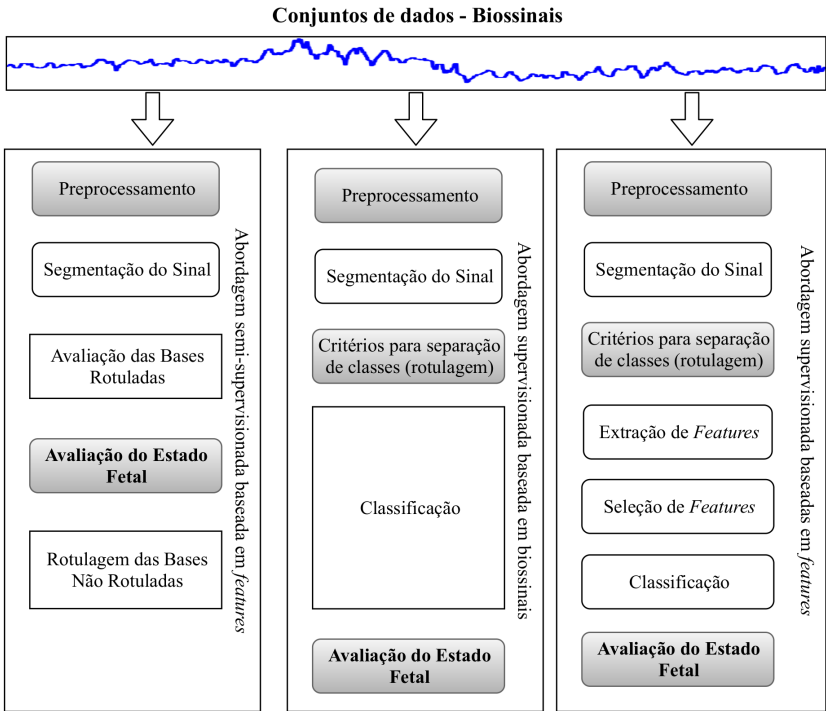


Figura 1 – Visão geral dos modelos desenvolvidos. Fonte: Elaboração própria.

1.7 Organização da Tese

A Figura 2 mostra a organização da tese em relação ao fluxo dos principais procedimentos aqui executados.

O texto subsequente está estruturado da seguinte forma. O Capítulo 2 apresenta os

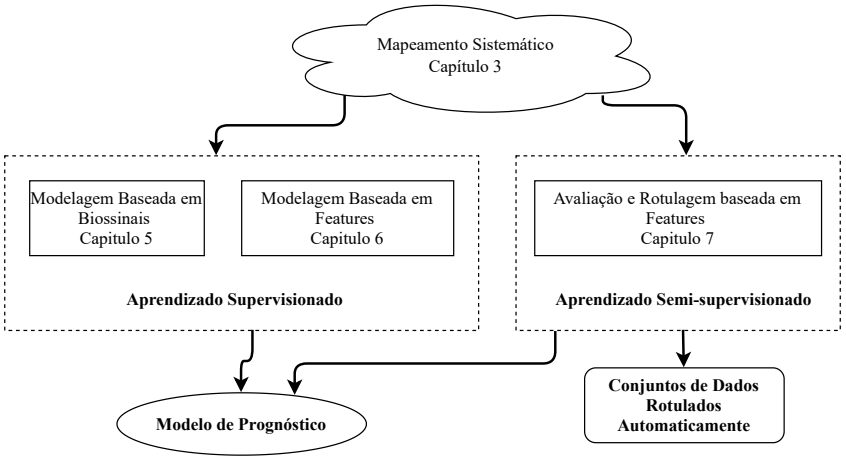


Figura 2 – Fluxo dos procedimentos executados. Fonte: Elaboração própria.

conceitos fundamentais que compõem a base teórica desta tese. O Capítulo 3 apresenta um mapeamento sistemático, o qual é utilizado como linha de base para o desenho dos experimentos e fundamentação das hipóteses, apresentamos ainda neste Capítulo uma discussão sobre os trabalhos relacionados. No Capítulo 4 são apresentados materiais e métodos de forma global e comuns aos cenários experimentais ao longo desta tese. O Capítulo 5 relaciona a avaliação baseada em biossinais (séries temporais) em conjunto com seus materiais, métodos e resultados. O Capítulo 6 relaciona a avaliação baseada em *features* em conjunto com seus materiais, métodos e resultados. No Capítulo 7

é apresentada a modelagens semi-supervisionada juntamente com seus materiais, métodos e resultados. As conclusões e trabalhos futuros são apresentados no Capítulo 8. No apêndice A, a lista de estudos selecionados para o mapeamento sistemático é sumarizada.

2 FUNDAMENTAÇÃO TEÓRICA

Este Capítulo aborda as principais bases teóricas necessárias para compreensão desta tese. Na primeira seção abordamos a avaliação computadorizada do estado fetal. Na segunda seção são apresentados os conceitos fundamentais relacionados as técnicas de aprendizado de máquina e métodos estatísticos empregados no desenho dos experimentos.

2.1 Avaliação do Estado Fetal

Um dos principais objetivos do monitoramento fetal é o de evitar um resultado adverso relacionado a hipoxia ou acidemia, uma vez que a produção inadequada de energia celular causada pela hipoxia ou acidose tem potencial para comprometer as funções celulares, causando a morte destas (AYRES-DE-CAMPOS *et al.*, 2015a). Apesar da maioria dos fetos nascerem com algum sinal de acidose metabólica, eles conseguem se recuperar rapidamente sem a ocorrência de complicações a curto ou longo prazo. No entanto, em casos onde a hipoxia/acidose fetal possui intensidade e duração suficiente para causar o mal funcionamento de órgãos importantes, existe o risco de danos permanentes e a hipoxia/acidose é fortemente associada a desfechos fetais adversos como complicações neurológicas de curto prazo (encefalopatia hipóxico-isquêmica) e de longo prazo (paralisia cerebral), podendo acarretar na morte do feto (AYRES-DE-CAMPOS *et al.*, 2015a).

Uma das principais fontes de informação sobre o bem-estar fetal é a sua atividade cardíaca, uma das formas de obtenção desta informação em ambiente clínico é por meio da cardiotocografia (CTG), a qual permite que os obstetras detectem se o feto esta em estado de deterioração (por exemplo, hipóxia fetal em curso), o que pode ocorrer mesmo em gravidez de baixo risco (CHUDACEK *et al.*, 2014).

Existem três abordagens principais que visam otimizar o uso das informações provenientes do monitoramento fetal, que são (CHUDACEK *et al.*, 2014):

1. melhorando a baixa concordância inter e intra-observador na avaliação dos CTGs por meio de guias bem estabelecidos como o guia FIGO (AYRES-DE-CAMPOS *et al.*, 2015b), os quais se baseiam na avaliação de características morfológicas da frequência cardíaca fetal (FCF) e sua relação com medições tocográficas das contrações uterinas (UC);
2. buscando por soluções que acrescentem avaliações mais objetivas e quantitativas sobre o estado fetal, utilizando de novas métricas representativas para seu estado como amostras

de sangue fetal ou eletrocardiogramas fetais (fECG) por meio de *ST waveform analysis* (STAN); e

3. realizando a avaliação computadorizada dos traçados de monitoramento por meio da extração de informações para uso em sistemas de análise automatizada.

Quando a hipóxia/acidose fetal é antecipada ou suspeita, uma ação é necessária para evitar danos ao feto e embora haja uma forte associação entre certos padrões de FCF e hipóxia/acidose, sua capacidade de discriminar situações com ou sem acidose metabólica ainda é limitada (AYRES-DE-CAMPOS *et al.*, 2015b). Neste contexto, esta tese alinha-se aos objetivos do monitoramento fetal que é identificar situações adversas de forma a evitar lesões ao feto.

2.2 Aprendizado de Máquina

Aprendizado de máquina ou *machine learning* pode ser definido como um conjunto de métodos capazes de detectar de forma automática padrões em dados e então utilizá-los para predizer novos padrões em dados futuros, podendo ainda realizar tomada de decisão sob incerteza (MURPHY, 2012). Assim, um algoritmo de aprendizado de máquina é aquele capaz de aprender com os dados (GOODFELLOW *et al.*, 2016). Em termos gerais, o aprendizado de máquina divide-se em dois ramos principais baseados no tipo de experiência sobre os dados que lhes é concedido durante o processo de treinamento: Aprendizado supervisionado e aprendizado não-supervisionado. O termo aprendizagem supervisionada origina-se da visão do resultado sendo fornecido por um instrutor ou professor que mostra ao sistema de aprendizagem o que fazer. Na aprendizagem não supervisionada, não há instrutor ou professor, e o algoritmo deve aprender a dar sentido aos dados sem este guia (GOODFELLOW *et al.*, 2016).

No aprendizado supervisionado, o objetivo é criar um mapeamento a partir das entradas x para uma saída y, dado um conjunto rotulado de pares com entradas-saídas $D = \{(x_i, y_i)\}_{i=1}^{N}$ onde D é chamado de conjunto de treinos e N é o número de amostras de treino. Cada entrada de treinamento x_i pode ser um vetor D-dimensional numérico representado por *features*, uma representação de dados complexa ou uma sequência em forma de série temporal, entre outros. Em relação as saídas (variável de resposta), em geral assume-se que y_i é uma variável categórica pertencente a um conjunto finito de opções $y_i \in \{1, ..., C\}$ ou um valor numérico escalar. Quando y_i é categórico, o problema de aprendizado é denominado de classificação ou reconhecimento de padrões, e quando y_i é um valor escalar real, o problema de aprendizado é conhecido como de regressão (MURPHY, 2012).

A aprendizagem não supervisionada envolve a observação de vários exemplos de um vetor aleatório x, na tentativa de aprender implícita ou explicitamente a distribuição de probabilidade $p(x)$, ou algumas propriedades dessa distribuição (GOODFELLOW *et al.*, 2016). Aqui, recebemos apenas entradas desprovidas de qualquer rótulo, $D = \{(x_i)\}_{i=1}^{N}$, e o objetivo é encontrar padrões nos dados, o qual é conhecido como descoberta de conhecimento (MURPHY, 2012).

A aprendizagem semi-supervisionada encontra-se entre as abordagens supervisionadas e não supervisionadas. Além dos dados não rotulados, o algoritmo recebe algumas informações de supervisão (dados rotulados), mas não necessariamente para todas as amostras. Em geral, essas informações serão as saídas (ou alvos) associados a algumas das amostras. Nesse caso, o conjunto de dados $X = (x_i)_{i=1}^{N}$ pode ser dividido em duas partes: os pontos $X_l = (x_1, ..., x_l)$, para os quais os rótulos $Y_l = (y_1, ..., y_l)$ são fornecidos, e os pontos $X_u = (x_l + 1, ..., x_l + u)$, cujos rótulos não são conhecidos (CHAPELLE *et al.*, 2010).

2.2.1 Algoritmos de Aprendizagem de Máquina

A caracterização do estado fetal pode ser interpretada como uma tarefa de classificação, na qual é possível mapear entradas x em saídas y, onde $y \in \{1, ..., C\}$, sendo C o número de classes. Nesta tese, empregou-se $C = 2$, o que é conhecido como classificação binária (MURPHY, 2012).

Foram empregados cenários distintos de avaliação como fundamento para construção dos modelos de prognóstico, o que envolveu algoritmos de classificação (supervisionados e semi-supervisionados). O funcionamento dos principais algoritmos empregados são apresentados a seguir.

2.2.1.1 MinMax Scaler

Algoritmos de aprendizado de máquina são sensíveis a grandes variações em sua escala de valores de entrada, uma vez que escalas diferentes não contribuem de forma igualitária no aprendizado do modelo. Desta forma, uma prática comum é aplicar algoritmos de escala nas *features* para garantir que elas estejam entre um determinado valor de máximo e mínimo, em geral entre 0 (zero) e 1 (um) (MURPHY, 2012). Seja $x = \{v_1, v_2, ..., v_i\}$ um vetor representando uma *feature* de entrada e os valores $[a, b]$ indicando os valores de limite máximos e mínimos entre os quais os novos valores escalados devem permanecer, tal que $\min(x) \geq a$ e $\max(x) \leq b$.

A escala min-max ou normalização min-max realiza a seguinte operação (KUHN *et al.*, 2013):

$$x_{scaled} = a + \frac{(x - \min(x))(b-a)}{\max(x) - \min(x)} \tag{2.1}$$

2.2.1.2 Recursive Feature Eliminator

Para diminuir o espaço de *features* utilizado pelos algoritmos de classificação, é importante que se obtenha a melhor combinação das *features* disponíveis. O algoritmo *Recursive Feature Eliminator* (RFE) (GUYON *et al.*, 2002) implementa uma seleção de *features* com base no *ranking* da importância de cada *feature*. As menos importantes são eliminadas sequencialmente antes da modelagem com objetivo de encontrar um subconjunto que produza um modelo preciso usando diferentes combinações de *features* (FERGUS *et al.*, 2018).

O algoritmo 1 apresenta o funcionamento geral do RFE, onde dado um conjunto de *features* $F = \{f_1, ..., f_S\}$ cada *feature* f recebe um *ranking* usando sua importância para o modelo, onde S é número total de *features* de onde são derivados os valores candidatos para o número de *features* a reter $(S_1, S_2, ...)$. Esse processo é repetido e as *features* com a melhor *ranking* do S_i são mantidas.

Algoritmo 1: Recursive Feature Eliminator. Adaptado de (FERGUS *et al.*, 2017).

```
Treine o modelo no conjunto de treinamento utilizando todas as features ;
Compute o desempenho do modelo ;
Compute a importância de cada feature ;
for cada subconjunto de tamanho S_i, i = 1...S do
    Mantenha as S_i features mais importantes ;
    Treine o modelo no conjunto de treinamento utilizando Si;
    Calcule o desempenho do modelo;
    Recalcule a importância de cada feature ;
end
Calcule o desempenho geral sobre S_i ;
Determine o número apropriado de features ;
Utilize o modelo baseado no valor ótimo para S_i ;
```

2.2.1.3 Synthetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique (SMOTE) (BOWYER *et al.*, 2002) é uma técnica de reamostragem de dados empregada para compensar o desequilíbrio nos dados em relação a distribuição das classes. Ele opera na classe minoritária criando dados artificiais a partir da mesma. O SMOTE é baseado em dados reais pertencentes à classe minoritária e opera no

espaço de *features*, para cada instância da classe minoritária ele apresenta uma amostra sintética ao longo das linhas que unem essa instância particular com seus k vizinhos mais próximos. Normalmente, após o SMOTE, o conjunto de dados tem aproximadamente o mesmo número de amostras em cada classe (KUHN *et al.*, 2013). O algoritmo 2 apresenta o funcionamento geral do SMOTE.

Algoritmo 2: SMOTE. Adaptado de (FERGUS *et al.*, 2017).

Data: Dados minoritários $D^{(t)} = \{x_i \in R^d\}, i = 1,2,...,T$ número de instâncias minoritárias (T), porcentagem SMOTE (N), número de vizinhos próximos (k)
Result: Return dados sintéticos S

```
for i = 1, ...,T do
    Encontre os k vizinhos (classe minoritária) mais próximos de x_i;
    N̂ = N/100 ;
    while N̂ ≠ 0 do
        Selecione um dos k vizinhos mais próximos, x̄ ;
        Selecione um número aleatório α ∈ [0,1] ;
        x̂ = x_i + α(x̄ − x_i) ;
        Append x̂ a S ;
    end
end
```

2.2.1.4 Support Vector Machine

Support vector machine (SVM) consiste em um modelo de aprendizado supervisionado baseado em hiperplanos no espaço de *features*, cuja ideia principal é a divisão do espaço em dois subespaços, minimizando o erro empírico e maximizando a margem entre as instâncias mais próximas e os hiperplanos (CORTES; VAPNIK, 1995). Sejam N observações e um conjunto $D = \{(x_i, y_i)\}_{i=1}^{N}$, onde x_i é a entrada e $y_i = \{0,1\}$ representa as classes resultantes. O algoritmo busca por um hiperplano w, que maximize a distancia (margem) entre este hiperplano e as amostras próximas a ele (CORTES; VAPNIK, 1995). Estas amostras são chamadas de vetores de suporte (*support vector*) onde as *features* podem ser mapeadas em outros espaços, os quais são aumentados utilizando uma função de *kernel*.

2.2.1.5 K-Nearest Neighbors

K-Nearest Neighbors (kNN) é um algoritmo baseado em instancias, cuja técnica de classificação é baseada nos k pontos mais próximos ou conjunto de pontos no espaço de

features (ALTMAN, 1992). O kNN usa métricas de distância, geralmente distância euclidiana, para encontrar a vizinhança ótima de atributos em relação aos rótulos de classe dos dados de treinamento (MURPHY, 2012). Este algoritmo se caracteriza como baseado em instância por não possuir um estágio de treinamento ou processo de aprendizagem propriamente dito, em vez disso, no momento da validação/teste, para produzir uma saída y para uma nova entrada de x, o kNN encontra os k-vizinhos mais próximos da amostra x nos dados de treinamento X. Em seguida, ele retorna a média dos valores y correspondentes no conjunto de treinamento (GOODFELLOW *et al.*, 2016).

2.2.1.6 Bagging

Bootstrap aggregation (Bagging) (BREIMAN, 1996) é um método de *ensemble* que adota a técnica de amostragem *Bootstrap* na construção de modelos básicos, em nosso caso, *Decision Trees*. Ele gera novos conjuntos de dados por amostragem do conjunto de dados original com substituição e treina classificadores de base nos conjuntos de dados amostrados. Para obter um classificador de *ensemble*, ele combina todos os classificadores básicos por votação majoritária (AGGARWAL, 2014), ou seja, ele treina *M decision trees* diferentes em subconjuntos de dados distintos, escolhidos aleatoriamente com substituição e, em seguida, calcula o seguinte *ensemble*:

$$f(x) = \sum_{m=1}^{N} \frac{1}{M} f_m(x) \tag{2.2}$$

Onde f_m é a m-ésima *tree* (MURPHY, 2012).

2.2.1.7 Random Forests

Random Forest (BREIMAN, 2001) pode ser considerada uma variante da abordagem de Bagging. Ele segue as principais etapas do Bagging e usa algoritmos *Decision Trees* para construir os classificadores básicos. Além da amostragem Bootstrap e votação majoritária usada no Bagging, o classificador Random Forest incorpora ainda a seleção de espaço de *features* aleatório na construção do conjunto de treinamento para promover a diversidade dos classificadores de base, o algoritmo 3 descreve o procedimento geral do Random Forest.

Algoritmo 3: Random Forest. Adaptado de (FERGUS *et al.*, 2017).

Data: Seja um conjunto de treinamento $((x1, y1), ..., (x_N, y_N)$, onde $x_i \in R^d$ e $y_i \in \{norm, path\}$

1 *Define o número de árvores na floresta, B, e o número de random features a ser selecionada, m.*;
2 **for** *b = 1, ...,B* **do**
3 Utilizando o conjunto de treinamento e amostragem com reposição, criar um conjunto de amostras bootstrap de tamanho *n*; alguns padrões serão replicados, enquanto outros podem ser omitidos;
4 Construir um classificador *Decision tree*, $n_b(x)$ utilizando as amostras *bootstrap* como dados de treinamento, selecionar de forma randômica para cada nó da árvore *m features* que serão consideradas para divisão;
5 Classificar os padrões não bootstrap (dados out-of-bag) utilizando o classificador $n_b(x)$;
6 Assinalar x_i para a classe mais recorrente pelo classificador $n'_b(x)$, onde b' se refere às amostras bootstrap que não contém x_i.
7 **end**

2.2.1.8 Gradient Tree Boosting

Gradient Boosting Decision Tree (GTB) é um algoritmo *ensemble* que se utiliza de *Decision Trees* como classificadores base os quais são treinados em sequência e em cada iteração, o seu aprendizado se baseia no ajuste de gradientes negativos conhecidos como erros residuais (KE *et al.*, 2017). Especificamente, em cada iteração, uma subamostra dos dados de treinamento é formada aleatoriamente (sem substituição) do conjunto de dados de treinamento completo. Esta subamostra selecionada aleatoriamente é então usada no lugar da amostra completa para o classificador base e computa a atualização do modelo para a iteração atual (FRIEDMAN, 2002). Uma das implementações do GTB é a baseada em histogramas, a qual busca os melhores pontos de separação dos dados por meio dos valores ordenados das *features*. O algoritmo baseado em histograma agrupa valores contínuos de *features* em *bins* e os utiliza para construir histogramas de *features* durante o treinamento (KE *et al.*, 2017). Apresentamos no algoritmo 4 uma visão geral da implementação baseada em histogramas do *gradient tree boosting*.

2.2.1.9 Multi-layer Perceptron

Feed-forward neural networks ou *multi-layer perceptron* (MLP) (RUMELHART *et al.*, 1986) pode ser visto como um aproximador universal contendo múltiplas camadas cujo objetivo da *feed-forward neural network* é de aproximar uma função f^* para um classificador,

Algoritmo 4: GTB baseado em Histogramas. Adaptado de (KE *et al.*, 2017)

```
   Entrada: I: dados de treinamento, d: profundidade máxima
   Entrada: m: dimensão das features
 1 nodeSet = {0}, número de nós no nível corrente;
 2 rowSet = {{0,1,2,...}} índice dos dados nos nós da árvore ;
 3 for i = 1, ...,d do
 4     for node in nodeSet do
 5         usedRows = rowSet[node] ;
 6         for k = 1,...,m do
 7             H = new Histogram() ;
 8             Construa histogram ;
 9             for j in usedRows do
10                 bin = I.f[k][j].bin ;
11                 H[bin].y = H[bin].y + I.y[j] ;
12                 H[bin].n = H[bin].n + 1 ;
13             end
14             Encontre a melhor divisão no histograma H ;
15         end
16     end
17     Atualize rowSet e nodeSet de acordo com os melhores pontos de divisão ;
18 end
```

$y = f^*(x)$ que mapeia uma entrada x para uma saída y utilizando uma função mapeadora $y = f(x;\theta)$ que ajusta os valores do parâmetro θ que resultem na melhor aproximação (MURPHY, 2012).

Este tipo de modelo é chamado de redes (*networks*) pela composição de diferentes funções sendo associado a um grafo direcional acíclico que descreve como estas funções são compostas, sejam $f^{(1)}, f^{(2)}, ..., f^{(n)}$ conectadas em forma de corrente, $f(x) = f^{(n)}(f^{(\cdots)}(f^{(1)}(x))))$ onde $f^{(1)}$ é a primeira camada da rede, $f^{(2)}$ a segunda, e assim por diante.

Uma camada pode ser abstraída como um conjunto de muitas unidades (neurônios) que agem em paralelo, cada uma representando uma função vetor-para-escalar, a última camada da rede é chamada de camada de saída e as camadas intermediárias entre a primeira e a última são chamadas de camadas ocultas ou *hidden layers*, onde são aplicadas funções de ativação (*activation functions*) para computar os valores das camadas ocultas (GOODFELLOW *et al.*, 2016).

2.2.1.10 Convolutional Neural Networks (CNN)

Convolutional neural networks (CNN) (LECUN *et al.*, 1998) é um tipo de rede neural indicada para processamento de dados organizados em forma de *grid*, como séries temporais que podem ser vistas como um *grid* 1D (uni-dimensional) com amostras em intervalos regulares de tempo e dados de imagens que podem ser representados como *grid* de pixels 2D (bi-dimensional) (GOODFELLOW *et al.*, 2016). O modelo CNN normalmente compreende três camadas principais: camadas de convolução, *pooling* e camadas completamente conectadas (*fully-connected layer*).

As camadas convolucionais, que atuam como camadas ocultas, são os componentes centrais do modelo CNN (ZHAO *et al.*, 2019). Nas camadas de convolução, uma operação de mesmo nome é aplicada aos dados de entrada e as *features* extraídas são passadas para a próxima camada, composta da saída de múltiplos *feature maps*. O CNN aprende com *filters*, também conhecidos como *kernels* de tamanhos diferentes, os quais são usados na camada oculta para dividir as entradas, que são analisadas usando vários *feature maps*. A classificação nas camadas sucessivas é feita usando as *features* extraídas. As camadas de *pooling* reduzem a dimensão dos *feature maps* mantendo as informações mais representativas. Por fim, camadas *fully-connected* representam as conexões e neurônios de uma rede neural *feed-forward* a qual recebe como dados de entrada a os *feature maps* provenientes das camadas anteriores.

Existem ainda camadas *dropout*, aplicadas para evitar *overfitting* no processo de treinamento das redes neurais, esta camada é responsável por descartar parte dos valores atribuídos durante o treinamento (ZHAO *et al.*, 2019).

2.2.1.11 Multi-head Convolutional Neural Networks

Modelos de rede neural com várias cabeças (*multi-headed*) utilizam o princípio fundamental de que uma cabeça representa sequências de camadas de uma rede neural trabalhando de forma independente, a qual é usada para processamento das entradas em separado das outras cabeças, sendo que as saídas geradas por cada uma delas convergem para uma camada em comum, na qual são agrupadas para processamento em camadas subsequentes, produzindo assim a previsão final (KAUSHIK *et al.*, 2020). No caso da *multi-head* CNN, nosso trabalho empregou os elementos de redes convolucionais como as camadas de convolução e *pooling* para montar as cabeças, e utilizamos uma camada *concatenate* para agrupar os *feature maps* resultantes para o

processamento em camadas *fully-connected* para classificação do estado fetal.

2.2.1.12 Convolutional Long Short-term Memory Neural Networks (CNN-LSTM)

Redes neurais *Long short-term memory* (LSTM) (HOCHREITER; SCHMIDHUBER, 1997) são projetadas para lidar com dados sequenciais, sendo um tipo de *recurrent neural network* (RNN) conhecidos como *gated* RNNs os quais se fundamentam em camadas ocultas chamadas *memory blocks* ou *cells* (GOODFELLOW *et al.*, 2016). Os *memory cells* são compostos por gates, sendo que cada *memory cell* é composto por 3 (três) tipos de gates: *input gate*, *forget gate* e *output gate* com uma conexão auto-recorrente (*self-recurrent connection*) (TAN *et al.*, 2018).

- *Input gate*: responsável por qual informação deve ser armazenada no *memory block*.
- *Forget gate*: responsável por quanto de cada informação deve ser retida ou esquecida pelo *memory block*.
- *Output gate*: responsável por quando a informação armazenada deve ser utilizada.

O modelo *convolutional Long Short-term Memory* (CNN-LSTM) empregado nesta tese empilha camadas CNN em camadas LSTM, adicionamos ainda uma camada *fully-connected* para processar a avaliação final do estado fetal.

2.2.1.13 LabelSpreading

O *LabelSpreading* (ZHOU *et al.*, 2004) é um algoritmo de aprendizado semi-supervisionado baseado em grafos, cujo funcionamento se fundamenta na ideia de construir um grafo cujos nós são pontos de dados (rotulados e não rotulados) e as arestas representam semelhanças entre pontos. Os rótulos conhecidos são usados para propagar informações através do grafo a fim de rotular todos os nós. O funcionamento geral do algoritmo depende da geometria dos dados induzidos por amostras rotuladas e não rotuladas. Esta geometria pode ser representada naturalmente por um grafo empírico $g = (V,E)$ onde os nós $V = \{1,...,n\}$ representam os dados de treinamento e as arestas E representam semelhanças entre eles, na qual as similaridades são representadas por uma matriz de pesos $W : W_{ij}$ que é não-zero se x_i e x_j são vizinhos, ou seja, a aresta (i, j) de E é representada por sua matriz de pesos W_{ij} (CHAPELLE *et al.*, 2010).

Dados um conjunto $X = \{x1,...,x_l,x_l+1,...,x_n\}$, e um conjunto de rótulos $L = \{1,...,c\}$, os primeiros l pontos $x_i(i \leq l)$ são rotulados tal que $y_i \in L$ e os pontos remanescentes $x_u(l+1 \leq u \leq n)$ são não rotulados. Para predizer os rótulos dos pontos não rotulados, seja F um conjunto de matrizes *n x c* com entradas não negativas. Uma matriz $F = [F_1^T,...,F_n^T]^T \in F$

corresponde a classificação na base de dados X pela rotulagem de cada ponto x_i com um rótulo $y_i = argmax_j \leq_c F_{ij}$. Definindo uma matriz $Y_{n\times c} \in F$ com $Y_{ij} = 1$ se x_i for rotulado com $y_i = j$ e $Y_{ij} = 0$ caso contrário. O algoritmo 5 é apresentado a seguir, (ZHOU *et al.*, 2004).

Algoritmo 5: *Label spreading*. Adaptado de (CHAPELLE *et al.*, 2010).

1 Compute a matriz afinidade $W_{ij} = \exp(\frac{-\|x_i - x_j\|^2}{2\sigma^2})$ para $i \neq j$ e $W_{ii} = 0$;
2 Compute a matriz $S = D^{-1/2}WD^{-1/2}$ onde D é uma matriz diagonal em quê $D_{ii} = \sum_j W_{ij}$;
3 Iterar $F(t+1) = \alpha SF(t) + (1-\alpha)Y$ até a convergência, onde o parâmetro $\alpha \in [0,1)$;
4 Seja F^* uma representação do limite da sequência $\{F(t)\}$. Rotular cada ponto x_i com o rótulo $y_i = argmax_j \leq_c F^*_{ij}$

2.2.2 Métricas de avaliação

Foram empregadas métricas de avaliação derivadas de representações gráficas bem estabelecidas na literatura como a matriz de confusão e curva de características operacionais do receptor (ROC).

A matriz de confusão inclui quatro métricas principais: os valores de verdadeiros positivos (TP), falsos positivos (FP), falsos negativos (FN) e verdadeiros negativos (TN). TP e TN são o número de amostras que foram classificadas corretamente como positivas (patológicas) e negativas (normais), respectivamente. FN e FP são o número de amostras positivas e negativas que foram erroneamente classificadas como negativas e positivas, respectivamente.

Uma curva ROC é um gráfico dentro de um espaço quadrado unitário em que o eixo horizontal (x) denota a taxa de falso-positivo FPR ou (1 - Especificidade) e o eixo vertical (y) denota as taxas de verdadeiro-positivo (TPR) ou a sensibilidade de um classificador. A diagonal entre o canto esquerdo inferior (0,0) e o canto direito superior (1,1) representa a igualdade TPR = FPR para todas as amostras. O desempenho do classificador que está junto a essa diagonal atua como um classificador que atribui rótulos positivos e negativos às instâncias aleatoriamente. Segue-se que os classificadores situados acima desta diagonal têm um desempenho melhor do que o aleatório, no caso dos modelos de prognóstico, seria o desejado (JAPKOWICZ; SHAH, 2011).

Foram utilizadas métricas de desempenho bem estabelecidas para classificação binária no domínio médico (COMERT *et al.*, 2018; FERGUS *et al.*, 2018): Sensibilidade (SE), especificidade (SP), a média geométrica (GM) entre sensibilidade e especificidade e a área sob a

curva característica de operação do receptor (AUC). Para efeitos de completude, computamos ainda a métrica referente a acurácia (ACC). Essas métricas agregam informações unificadas derivadas das representações gráficas mencionadas.

- Sensibilidade (SE): $SE = \frac{TP}{TP+FN}$, a capacidade do modelo de discriminar as amostras patológicas.
- Especificidade (SP): $SP = \frac{TN}{TN+FP}$, a capacidade do modelo de discriminar as amostras saudáveis.
- Média geométrica (GM): $GM = \sqrt{SP * SE}$, uma forma unificada para representar o balanço entre SE e SP pela sua média geométrica.
- Área sob a curva ROC (AUC): Métrica que representa a capacidade do modelo em discriminar as classes normal e patológica de forma unificada. Performances em torno de 50% representam valores em torno da diagonal, ou seja, um classificador aleatório.
- Acurácia (ACC): $\frac{TP+TN}{TP+TN+FP+FN}$, representa a precisão do modelo, a proporção da classificação correta para todas as instâncias que são usadas.

2.2.3 Métodos estatísticos

A literatura fornece métodos para comparação entre pares e entre grupos de valores na avaliação de algoritmos de aprendizado de máquina em cenários de avaliação de bases de dados isoladas e combinadas (múltiplos domínios). Os testes não paramétricos de Friedman (FRIEDMAN, 1937) e Wilcoxon (WILCOXON, 1945) são adequados para comparar o desempenho de classificadores em múltiplos domínios, uma vez que esses testes podem ser aplicados a qualquer medida para avaliação de classificadores e não assumem distribuições normais ou homogeneidade de variância (DEMSAR, 2006). O teste Mann–Whitney U (MANN; WHITNEY, 1947) é apropriado para comparação de amostras pertencentes a duas populações independentes (SHESKIN, 2007).

2.2.3.1 Teste de Wilcoxon

O teste *signed rank* de Wilcoxon é uma alternativa não paramétrica ao teste t pareado e é adequado para avaliações de desempenho do classificador em esquemas que adotem bases de dados sozinhas ou agrupadas (JAPKOWICZ; SHAH, 2011). Este teste compara duas populações em torno de suas medianas classificando as diferenças de pares, ignorando os sinais e comparando as classificações para as diferenças positivas e negativas sob a hipótese nula de que as diferenças

são iguais a zero (WILCOXON, 1945).

Considerando duas listas contendo os valores de desempenho emparelhadas C_1 e C_2, seja d_i a i-th diferença entre o n valor do desempenho do i-th resultado disponível para $i \in \{(C_{2i} - C_{1i}), (C_{2(i+1)} - C_{1(i+1)}), ...n\}$. O teste de Wilcoxon(WILCOXON, 1945) ranqueia $|d_i|$ e assinala uma média para cada empate d_i. Os ranks W_{s1} e W_{s2} são definidos a seguir, as diferenças com valor zero entre W_{s1} e W_{s2} são divididas igualmente ignorando uma das diferenças zero se existir um número ímpar delas.

$$W_{s1}^{+} = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i), \tag{2.3}$$

$$W_{s2}^{-} = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i). \tag{2.4}$$

Seja r = *número de ranks assinalados*, a estatística T_{wilcox} é calculada como $T_{wilcox} = min(W_{s1}, W_{s2})$, e em caso de r pequenos $(r \leq 25)$, valores críticos exatos provenientes da tabela de estatística Wilcoxon com r degraus de liberdade podem ser usadas para verificar se a hipótese nula é rejeitada. Para valores altos de r, a distribuição T_{wilcox} pode ser aproximada pela normal com a seguinte expressão:

$$\mu_{wilcox} = \frac{r(r+1)}{4}; \tag{2.5}$$

$$\sigma_{wilcox} = \sqrt{\frac{r(r+1)(2r+1)}{24}}; \tag{2.6}$$

$$z_{wilcox} = \frac{T_{wilcox} - \mu_{wilcox}}{\sigma_{wilcox}}. \tag{2.7}$$

Em ambos os casos em que valores pequenos de r quanto para valores grandes de r, a hipótese nula pode ser rejeitada se T_{wilcox} for menor ou igual ao valor crítico listado para r nas tabelas de distribuições relativas ao nível de confiança pré-estabelecido (SHESKIN, 2007).

2.2.3.2 Teste Mann–Whitney U

O teste Mann–Whitney U(MANN; WHITNEY, 1947) é um equivalente não paramétrico ao teste t para amostras independentes. O teste Mann–Whitney U não é o mesmo teste Wilcoxon signed-rank, apesar de ambos serem não paramétricos e envolverem a soma de *ranks*. O teste Mann–Whitney U é aplicado a amostras independentes enquanto o teste de Wilcoxon signed-rank é aplicado a amostras pareadas ou dependentes (SHESKIN, 2007).

Mann–Whitney U teste avalia se duas populações diferem entre si baseado em suas medianas, desta forma, a sua hipótese nula H_0 é que duas populações possuem os valores de

mediana iguais. Temos então a hipótese alternativa não direcional H_1 de que as duas medianas diferem entre si. Inicialmente, o teste para comparar dois grupos envolve os três passos a seguir:

1. todos os valores são ordenadas pela magnitude (independente do seu grupo de origem);
2. são atribuídos *ranks* aos valores ordenados, sendo 1 o de menor magnitude, 2 o segundo maior, e assim por diante, até o total de valores N;
3. todos os valores de mesma magnitude tem seus respectivos *rank* substituídos pela média dos mesmos;

Desta forma, sejam duas sequencias de valores $L_1 = \{v1_1, v1_2, v1_3 ... n_1\}$ e $L_2 = \{v2_1, v2_2, v2_3 ... n_2\}$ pertencentes a dois grupos distintos, não necessariamente com a mesma quantidade de itens, contendo n_1 e n_2 elementos respectivamente num total de $n_1 + n_2 = N$ elementos conjuntos. Sejam R_1 os valores referentes aos *ranks* ordenados de L_1 e R_2 os valores referentes aos *ranks* ordenados de L_2. Os valores para U_1 e U_2 empregados na definição da estatística U são computados conforme segue (SHESKIN, 2007):

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1; \tag{2.8}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2; \tag{2.9}$$

$$U = min\{U_1, U_2\}. \tag{2.10}$$

O menor valor entre U_1 e U_2 representa a estatística U cujos intervalos são tabulados a partir da estatística Mann–Whitney U de valores críticos. Para que as diferenças entre os grupos seja estatísticamente significante, o valor obtido de U deve ser menor ou igual aos valores críticos tabulados para o nível de significância desejado.

2.2.3.3 Teste de Friedman

O teste de Friedman é um equivalente não paramétrico da ANOVA unilateral com medidas repetidas e é adequado para comparação de vários classificadores em cenários com múltiplas bases de dados (DEMSAR, 2006). Este teste ranqueia o desempenho dos classificadores para cada conjunto de dados separadamente. No caso de empate, o teste atribui a média dos ranques. Com k classificadores e n bases de dados, o teste de Friedman envolve o ranqueamento de uma matriz de dimensão $\{n \times k\}$ onde cada $\{r_i^j\}$ item representa o ranque j-th de k classificadores do i-th de n bases de dados. o ranque médio R_j e a estatística de Friedman

χ_F^2 são apresentados conforme segue (SHESKIN, 2007):

$$R_j = \sum_{i=1}^{n} r_i^j; \tag{2.11}$$

$$\chi_F^2 = \frac{12}{nk(k+1)} \left[\sum_{j=1}^{k} (\sum R_j)^2 \right] - 3n(k+1). \tag{2.12}$$

A hipótese nula é os desempenhos dos classificadores são equivalentes e, portanto, seus ranques médios são iguais. A estatística de Friedman segue uma estimativa aproximada pela distribuição χ^2 com $k-1$ degraus de liberdade para valores $(n > 10)$ e $(k > 5)$. Para valores pequenos de n e k, valores críticos exatos especificamente tabulados para o teste de Friedman podem ser usados (DEMSAR, 2006). Para rejeitar a hipótese nula, o valor calculado χ_F^2 deve ser igual ou maior que o valor crítico apresentado no nível de significância pré-especificado (SHESKIN, 2007).

2.2.3.4 Teste de Nemenyi

O teste de Nemenyi (NEMENYI, 1962) pode ser aplicada como análise *post-hoc* quando o teste de Friedman rejeita a hipótese nula (DEMSAR, 2006). O teste de Nemenyi identifica a diferença mínima exigida entre a soma das classificações de dois classificadores no nível de significância pré-especificado. O desempenho dos dois classificadores C_1 e C_2 é significativamente diferente se as médias dos ranques correspondentes diferem pelo menos pela diferença crítica do teste, ou, equivalentemente:

$$\left[\frac{R_{jC_1}}{\sqrt{\frac{k(k+1)}{6n}}} - \frac{R_{jC_2}}{\sqrt{\frac{k(k+1)}{6n}}} \right] > q_a \sqrt{\frac{k(k+1)}{6n}}. \tag{2.13}$$

Os valores q_a são tabulados a partir da estatística usada no teste de Tukey divididos por $\sqrt{2}$, com $(n-1)(k-1)$ degraus de liberdade (DEMSAR, 2006). A diferença entre dois valores é estatisticamente significativa se a inequação acima for mantida (SHESKIN, 2007).

2.3 Síntese do Capítulo

Este Capítulo realizou uma breve apresentação dos fundamentos teóricos relacionados a esta pesquisa. Os temas abordados foram avaliação do bem-estar fetal, aprendizado de máquina e métodos estatísticos empregados no desenho e validação dos experimentos.

3 REVISÃO DA LITERATURA

Neste Capítulo, é apresentado um panorama do estado da arte relacionado aos parâmetros e tecnologias empregadas no projeto de sistemas de avaliação do estado fetal. Foi realizado um mapeamento sistemático para identificar e sintetizar pesquisas recentes a respeito dos principais blocos de construção necessários para o desenvolvimento de soluções computadorizadas para avaliação do bem-estar do feto. Os resultados deste mapeamento foram utilizados como fundamento para o desenho dos experimentos e hipóteses desta tese. É apresentado ainda um comparativo dos modelos de prognósticos produzidos nesta tese frente a trabalhos da literatura.

3.1 Mapeamento Sistemático da Literatura

Um estudo empírico que investiga uma questão de pesquisa específica é chamado de estudo primário. Já os estudos secundários são aqueles que revisam os estudos primários relacionados a uma questão em busca de evidências sobre um determinado problema (KITCHENHAM *et al.*, 2010). Mapeamento sistemático é um tipo de estudo secundário, também conhecido como estudo de escopo, que visa identificar e classificar estudos sobre uma área ou tema de interesse. Os estudos de mapeamento têm como objetivo fornecer uma visão geral de uma área de tópico e identificar grupos de evidências para orientar pesquisas futuras (KITCHENHAM, 2007; KITCHENHAM *et al.*, 2011).

Mapeamentos sistemáticos têm sido adotados com sucesso para extração e classificação de conhecimento em diversos domínios de aplicação, como nos trabalhos de Idri *et al.* (IDRI *et al.*, 2018), Bischoff *et al.* (BISCHOFF *et al.*, 2019), Mehta *et al.* (MEHTA *et al.*, 2019) e Silva Neto *et al.* (SILVA NETO *et al.*, 2019). Até onde sabemos, nenhum outro mapeamento sistemático sobre os componentes básicos dos sistemas de avaliação do estado fetal com foco nos estágios de desenvolvimento e nos parâmetros para separação do estado fetal adotados no projeto destes sistemas foi realizado até o momento. A rápida evolução deste assunto complexo e relevante necessita de uma visão multidisciplinar do estado da arte destes estudos.

O mapeamento foi executado com base nas diretrizes bem estabelecidas e nas melhores práticas recomendadas por Petersen *et al.* (PETERSEN *et al.*, 2015), Kitchenham (KITCHENHAM, 2007), e Wohlin (WOHLIN, 2014).

3.2 Protocolo de Pesquisa

3.2.1 Objetivo e questões de pesquisa do mapeamento

Este mapeamento objetivou um panorama atualizado dos blocos de construção e parâmetros utilizados para separação dos estados normais e patológicos nos sistemas de apoio à avaliação do estado fetal. Para isso, foram definidas três questões principais de pesquisa seguidas de três subquestões, conforme mostrado na Tabela 1. Essas questões reúnem evidências sobre os blocos de construção que compõem os sistemas de avaliação do estado fetal baseados em biossinais sob a perspectiva da tomada de decisão durante o desenvolvimento destes sistemas em relação às técnicas, métodos, parâmetros de configuração críticos e o uso de dados retrospectivos para avaliação fetal.

Tabela 1 – Resumo das questões de pesquisa.

ID	Questão de pesquisa	Descrição
RQ1	Quais são os tipos de contribuições nos estudos sobre o sistemas de avaliação de estado fetal baseado em biossinais?	Agrupar as publicações por tipo de contribuição na área de projeto e desenvolvimento de sistemas para avaliação do estado fetal.
RQ2	Quais dados são usados nos estudos?	Examinar os dados dos estudos em relação ao número de amostras, duração do sinal, critérios de separação de classes, segmentação do biossinal e disponibilidade de dados de forma pública.
RQ3	Em quais etapas se divide o processo de desenvolvimento dos sistemas de avaliação do estado fetal?	Explorar as principais etapas do projeto de sistemas de avaliação do estado fetal e suas tecnologias, técnicas e métodos habilitadores.
RQ3-1	Quais abordagens de preparação de dados (DP) são empregadas?	Identificar as abordagens de preparação de dados empregadas nos sistemas de avaliação do estado fetal.
RQ3-2	Quais abordagens de transformação de dados (DT) são empregadas?	Identificar as abordagens para representação dos dados empregadas nos sistemas de avaliação do estado fetal.
RQ3-3	Quais métodos de construção de modelo (MC) são empregados?	Identificar os métodos que fornecem as habilidades de predição ou classificação aos sistemas de avaliação do estado fetal.

Fonte: Elaboração própria

3.2.2 Estratégia de pesquisa

Esta tese se utilizou de bases de dados eletrônicas ou bibliotecas digitais para obtenção dos estudos primários. Considerando as questões de pesquisa e as dimensões P e I da PICO (População, Intervenção, Comparação e resultado) sugeridas por Kitchenham (KITCHENHAM, 2007), foram identificadas palavras-chave e formulada a *string* de pesquisa. Nos estudos de mapeamento, as outras dimensões (comparação e resultado) podem restringir a busca e remover

artigos relevantes da área temática (PETERSEN *et al.*, 2015).

- *População:* a população representa estudos que abordam o desenvolvimento de sistemas de avaliação do estado fetal baseados em biossinais.
- *Intervenção:* no contexto deste estudo, intervenção refere-se aos blocos de construção do sistema de avaliação do estado fetal, como métodos, abordagens, parâmetros de configuração, uso de dados retrospectivos e tecnologias facilitadoras.

Empregou-se o *booleano* AND para vincular os termos principais e o *booleano* OR para agrupar os correlatos. Assim, foi definida a consulta de pesquisa ou *search query* (SQ), aplicada às bibliotecas digitais conforme mostra a Tabela 2.

Tabela 2 – Consulta de pesquisa aplicada em bibliotecas digitais. Fonte: Elaboração própria.

Consulta de pesquisa
(((("electronic fetal monitoring") OR ("electronic foetal monitoring") OR (cardiotocography) OR (cardiotocographic) OR (cardiotocogram) OR ("fetal heart rate") OR ("foetal heart rate") OR ("FECG") OR ("Foetal surveillance ") OR ("foetal monitoring signals") OR ("FHR signal")) AND ((("computer-aided") OR ("computer-assisted") OR ("computer based") OR ("computer-based") OR ("computerised analysis") OR ("computer analysis") OR ("computerized analysis") OR ("computational analysis") OR (algorithm) OR (software) OR (automatic) OR (online) OR ("computerized diagnostic") OR (classification) OR (classifier) OR (prediction) OR (recognition) OR ("decision-support") OR ("decision support") OR ("predictive models") OR ("pattern recognition") OR ("machine learning") OR ("data mining") OR ("deep learning") OR ("artificial inteligence") OR ("intelligent system") OR ("prognostic model") OR ("intelligent assessment")) AND ((("risk assessment") OR ("fetal distress") OR ("foetal distress") OR ("hypoxia") OR ("suffering") OR ("pathological") OR ("fetal state") OR ("fetal outcome") OR ("foetal state") OR ("fetal acidemia") OR ("fetal asphyxia") OR ("fetal risk") OR ("acidaemia")))

A busca foi realizada em cinco bibliotecas digitais que englobam domínios multidisciplinares, incluindo as engenharias, ciências da computação e informática médica. As bibliotecas escolhidas foram **Engineering Village** (Compendex), **Scopus**, **Web of Science**, **IEEE Xplore** digital library, e **Pubmed**. Em seguida, para complementar o conjunto de resultados, foi aplicada a técnica *forward snowballing* nos artigos aprovados. *Forward snowballing* se refere à identificação de novos artigos com base nos artigos que citam os estudos que estão sendo examinados (WOHLIN, 2014). A técnica de *forward snowballing* complementa o conjunto de resultados obtidos inicialmente nas bibliotecas digitais, cobrindo artigos relevantes que citaram os estudos iniciais aprovados. Desta forma, reduz a necessidade de empregar um grande número de fontes eletrônicas de pesquisa.

As bibliotecas digitais empregadas nesta tese indexam artigos relevantes de diferentes domínios de aplicação e foram aplicadas com sucesso (PETERSEN *et al.*, 2015) em estudos

baseados em evidência. Por fim, é importante ressaltar que foram utilizados os mecanismos das bases *Scopus* e *Web of Science* para analisar as citações dos artigos aprovados durante os procedimentos de *snowballing*.

3.2.3 Seleção dos estudos

Critérios de inclusão e exclusão devem ser definidos para garantir que estudos não relacionados a questões de pesquisa sejam eliminados do processo de busca, bem como assegurar a seleção daqueles relacionados para análise (KITCHENHAM, 2007). A seleção de estudos primários relevantes foi dividida em etapas de acordo com as recomendações de Petersen *et al.* (PETERSEN *et al.*, 2015). Na primeira etapa, foram aplicados critérios de seleção por meio da leitura por pares do título e resumo dos estudos primários extraídos de bibliotecas digitais. Em caso de divergência entre os avaliadores, o estudo segue para análise detalhada na próxima etapa. Na segunda etapa, revisamos por pares o texto na íntegra dos artigos pré-selecionados. A Tabela 3 apresenta resumidamente os critérios de seleção empregados para filtragem dos estudos primários.

Tabela 3 – Critérios de inclusão e exclusão. Fonte: Elaboração própria.

Critério de inclusão
1: Artigos de periódicos ou artigos de conferências/workshops que tenham um tema central ou tenham seções explicitamente dedicadas ao projeto de soluções de avaliação do estado fetal baseadas em biossinais. Esses estudos devem indicar claramente os benefícios do uso de cada bloco de construção nos sistemas de avaliação do estado fetal.
Critérios de exclusão
EC1: Publicações duplicadas (apenas uma é considerada).
EC2: *Short Papers* (até cinco páginas), onde é difícil extrair respostas para questões de pesquisa com clareza.
EC3: Estudos primários sem relação com as questões de pesquisa.
EC4: Estudos secundários e terciários: pesquisas, revisões de literatura, revisões sistemáticas e mapeamentos sistemáticos.
EC5: Material promocional, conselhos editoriais, resumos e pôsteres.
EC6: Estudos primários publicados em um idioma diferente do inglês.
EC7: Estudos primários publicados antes de janeiro de 2016.
EC8: Estudos em que o texto completo não está disponível.

Buscou-se apenas por artigos publicados em inglês porque este é o idioma dominante em eventos relevantes, mesmo em regiões geográficas onde outros idiomas são usados. Definimos 2016 como o limite inicial para obter apenas artigos recentes com aproximadamente cinco anos de publicação. A lista completa dos estudos primários analisados e os arquivos exportados das bibliotecas digitais estão disponíveis em nosso repositório de dados (SILVA NETO; GOMES, 2021). A Figura 3 apresenta uma visão geral do processo de seleção empregado para filtrar a

literatura relevante.

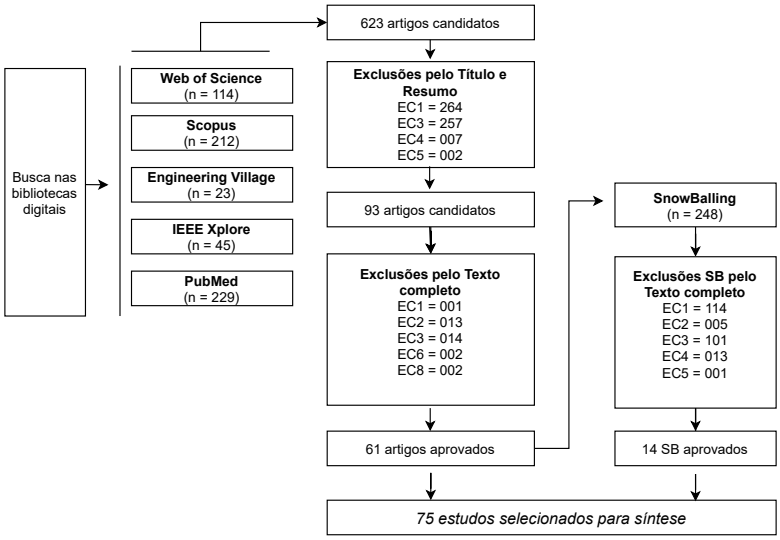


Figura 3 – Processo de seleção de estudos. Fonte: Elaboração própria.

3.2.4 Extração e classificação de dados

Nesta etapa, foram extraídos dados dos artigos selecionados para reunir informações relevantes a fim de fornecer respostas às questões de pesquisa.

RQ1: canal de publicação, ano e tipo de pesquisa. Conforme apresentado por Idri *et al.* (IDRI *et al.*, 2018), foram adotados tipos de contribuição nas pesquisas.

- **Técnica (T)**: uma técnica nova ou aprimorada relativa à avaliação do estado fetal assistida por computador.
- **Comparação (C)**: uma comparação entre técnicas.
- **Validação (V)**: uma avaliação de desempenho de técnicas existentes.
- **Método (M)**: aplicação de um conjunto de técnicas existentes para resolver um determinado problema.
- **Outros (OT, OG e OM)**: ferramentas ou *other-tools*, guias ou *other-guides* e modelos ou *other-models*.

RQ2: esta questão está centrada nos dados empregados para desenvolver e validar as propostas. Cobrindo o intervalo do biossinal em relação ao parto, a duração, a estratégia de segmentação do sinal, a disponibilidade dos dados e os critérios de separação de classes adotados na análise de monitoramentos retrospectivos.

RQ3: os estágios de desenvolvimento do sistema foram categorizados nas tarefas de preparação de dados (seleção de segmento ou pré-processamento), transformação de dados

(caracterização de sinal, extração de informação e redução de dimensão) e construção de modelo (habilidades preditivas e de classificação) (COMERT; KOCAMAZ, 2017; ZHAO *et al.*, 2019; COMERT; KOCAMAZ, 2018). A seguinte categorização foi empregada.

- **Preparação dos dados ou *data preparation* (DP)**: seleção de segmentos e procedimentos de pré-processamento relevantes.
- **Transformação dos dados ou *data transformation* (DT)**: representação e extração de informação.
- **Construção do modelo ou *model construction* (MC)**: tecnologias, métodos e abordagens para previsão e classificação.

RQ3-1: nesta subquestão, as atividades de pré-processamento adotadas por Comert *et al.* (COMERT *et al.*, 2019) e Chudacek *et al.* (CHUDACEK *et al.*, 2009) foram combinadas para extrair as abordagens de preparação de dados da seguinte forma:

- seleção de segmento: critérios para remover artefatos e manter a qualidade do sinal para análise posterior;
- redução de artefatos: detecção de *outliers* e ajuste das mudanças abruptas no sinal;
- ajustes extras: *detrending*, *smoothing* e técnicas de filtragem para melhoria na qualidade do sinal;

RQ3-2: esta subquestão centra-se nas abordagens de representação de dados para desenvolver e validar as propostas. Ela cobre o domínio da representação com um foco particular na proposição de novos índices de representação de dados.

RQ3-3: nesta subquestão, foram analisados os métodos que fornecem habilidades preditivas aos sistemas de avaliação do estado fetal. A questão cobre os critérios de rotulagem de classe, métodos de classificação, intervalo de dados, número de registros utilizados na avaliação dos modelos e o desempenho alcançado.

3.3 Resultados do Mapeamento Sistemático

Esta seção apresenta os resultados desse mapeamento sistemático e discute as principais descobertas. Para identificar grupos de evidências, as publicações foram classificadas com base em seus fatores comuns em relação a cada questão de pesquisa ou *research question* (RQ) deste mapeamento.

3.3.1 Resultados da seleção dos estudos

Identificamos 623 artigos candidatos na fase de busca em bibliotecas digitais. Depois de aplicar um critério de filtragem de duas etapas, os revisores mantiveram 61 estudos. Então, 248 artigos candidatos foram identificados por meio da técnica de *SnowBalling* (WOHLIN, 2014) no qual 14 artigos relevantes foram retidos. O processo de seleção completo resultou em 75 trabalhos selecionados conforme apresentado na Figura 3.

Utilizamos esses 75 estudos selecionados para responder às questões de pesquisa. A lista completa dos artigos selecionados com os resultados da classificação geral é apresentada na Tabela 45 do apêndice A e em nosso repositório de dados (SILVA NETO; GOMES, 2021).

3.3.2 RQ1: Quais são os tipos de contribuições nos estudos sobre o sistemas de avaliação de estado fetal baseado em biossinal?

A Figura 4 retrata o tipo de publicação entre os 75 estudos selecionados. O tipo de publicação mais representativo foi **método**, predominando ao longo do período. Publicações do tipo **comparação** estiveram presentes ao longo de todo o período, havendo carência de **validações** de 2019 a 2021. O ano de 2019 foi o que apresentou o maior número de publicações, onde o maior número foram **métodos** e **técnicas**. Note que mais da metade dos estudos selecionados refere-se ao método como um tipo de publicação, e o tipo de publicação outro - ferramentas está presente ao longo dos anos; no entanto, existem lacunas referentes a validação destes estudos.

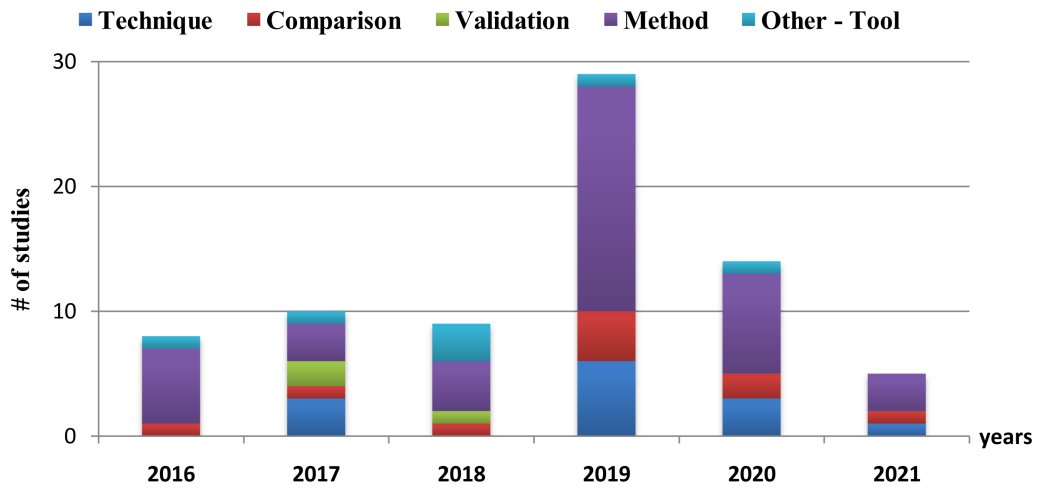


Figura 4 – Tipo de publicação por ano. Fonte: Elaboração própria.

Dentre os estudos avaliados, os periódicos mais recorrentes são Computers in Biology

and Medicine, Acta Obstetricia et Gynecologica Scandinavica, e IEEE Access. As cinco principais fontes de publicação estão listadas na Tabela 4.

Tabela 4 – Top 5 fontes de publicação. Fonte: Elaboração própria.

Fonte da publicação	Tipo	# estudos
Computers in Biology and Medicine	J	4
Acta Obst. et Gynec. Scandinavica	J	3
IEEE Access	J	3
Applied Artificial Intelligence	J	2
Biomedical Signal Processing and Control	J	2

J = Jornal

3.3.3 RQ2: Quais dados são usados nos estudos?

Esta questão investiga os conjuntos de dados empregados para desenvolver ou validar os sistemas de avaliação do estado fetal. A Tabela 5 mostra o tipo de dado, as abordagens de rotulagem frequentes, o número de estudos, o número de registros e a fonte da publicação onde é possível obter os detalhes de cada base de dados empregada.

Embora alguns estudos empreguem vários conjuntos de dados simultaneamente (JEZEWSKI *et al.*, 2016; ANISHA *et al.*, 2021; PETROZZIELLO *et al.*, 2019; ZARMEHRI *et al.*, 2019), consideramos cada uma das bases de dados individualmente. Os conjuntos de dados empregados dividem-se principalmente em derivados dos biossinais representados por *features* e nos que utilizaram os biossinais diretamente.

Sobre os dados contendo sinais, o CTU-UHB (CHUDACEK *et al.*, 2014), UTSB (BOUDET *et al.*, 2019b) e a base de dados SpaM (GEORGIEVA *et al.*, 2017) fornecem sinais de frequência cardíaca fetal (FCF) e contração uterina (UC). A base de dados HUFA fornece apenas o sinal de FCF. Em relação as bases de dados NIFECGDB (GOLDBERGER *et al.*, 2000), ADFECGDB (GOLDBERGER *et al.*, 2000), e PCCDB (GOLDBERGER *et al.*, 2000), estas contêm registros multicanal de eletrocardiograma fetal (FECG).

As bases DNL-IUGR (SIGNORINI *et al.*, 2020a) e UCI-CTG (DUA; GRAFF, 2017) fornecem um conjunto de *features* pré-extraídas e prontas para uso. O UCI-CTG consiste em medições de frequência cardíaca fetal (FCF) e contração uterina (UC) derivados da cardiotocografia e computadas pelo programa ©SisPorto 2.0 e rotulados por especialistas. O programa ©SisPorto[1] encontra-se atualmente em sua versão 4.0. A base DNL-IUGR contém um conjunto

[1] http://www.omniview.eu/ing

Tabela 5 – Conjuntos de dados usados nos estudos selecionados. Fonte: Elaboração própria

Base de dados	Tipo	Rotulagem	# artigos	# registros	Fonte/Ref.
CTU-UHB	sinal	pH, BDecf, BE, BW, Apg1, Apg5, TD, Experts	28	552	(CHUDACEK *et al.*, 2014)
DNL-IUGR	feature	IUGR parameters	1	120	(SIGNORINI *et al.*, 2020a)
HUFA	sinal	pH and Apg5	1	32	(BARQUERO-PEREZ *et al.*, 2017)
NIFECGDB	sinal	FIGO	1	55	(GOLDBERGER *et al.*, 2000)
ADFECGDB	sinal	FIGO	1	5	(GOLDBERGER *et al.*, 2000)
PCCDB	sinal	FIGO	1	175	(GOLDBERGER *et al.*, 2000)
SpaM	sinal	pH	2	300	(GEORGIEVA *et al.*, 2017)
UCI-CTG	feature	Experts	28	2126	(DUA; GRAFF, 2017)
UTSB	sinal	Experts	2	155	(BOUDET *et al.*, 2019b)
private	-	parametros IUGR, pH, BW, Apg5, and Experts	16	Min: 34, Max: 35.429	(LU *et al.*, 2019; SIGNORINI; MAGENES, 2016; MARQUES *et al.*, 2019; GIULIANO *et al.*, 2017; LU *et al.*, 2018; GEORGIEVA *et al.*, 2017; WOLF *et al.*, 2019; WARMERDAM *et al.*, 2018; STROUX *et al.*, 2017; GAO; LU, 2019; CZABANSKI *et al.*, 2016; ITO *et al.*, 2021; ZARMEHRI *et al.*, 2019; WU *et al.*, 2019; ROMANO *et al.*, 2016; GYLLENCREUTZ *et al.*, 2018)

TD = Tipo de parto, BDecf = base deficit in the extracellular fluid, BE = Base excess, BW = birth weight, Apg{1,5} = Apgar scores 1 min e 5 min, IUGR = Intrauterine growth restriction, FIGO = International federation of gynecology and obstetrics guidelines.

de índices lineares e não lineares em conjunto com o resultado do diagnóstico de IUGR realizado após o nascimento.

Os conjuntos de dados fornecem informações de rotulagem de classe variando de marcadores bioquímicos (pH, BDecf, BE), indicadores subjetivos(Apg1, Apg5), limiares de índice (FIGO, IUGR parameters), e consenso entre experts. Os dados mais utilizados foram CTU-UHB e UCI-CTG. Além disso, é fácil perceber pela Tabela 5 que existe uma variação considerável no número de registros entre os conjuntos de dados empregados na literatura, e a

quantidade mais significativa (35.429 registros) para avaliação retrospectiva de dados pertencem a dados privados.

3.3.4 RQ3: Em quais etapas se divide o processo de desenvolvimento dos sistemas de avaliação do estado fetal?

A questão RQ3 investiga os estágios do projeto do sistema e suas estratégias para produzir soluções de avaliação do estado fetal. Primeiramente, foram identificadas quais etapas foram adotadas nos estudos. Em segundo lugar, cada abordagem foi detalhada em relação ao seus estágios com foco em suas particularidades e sua contribuição para a avaliação do estado fetal. Esse esquema permitiu a análise de diferentes abordagens da literatura de forma objetiva.

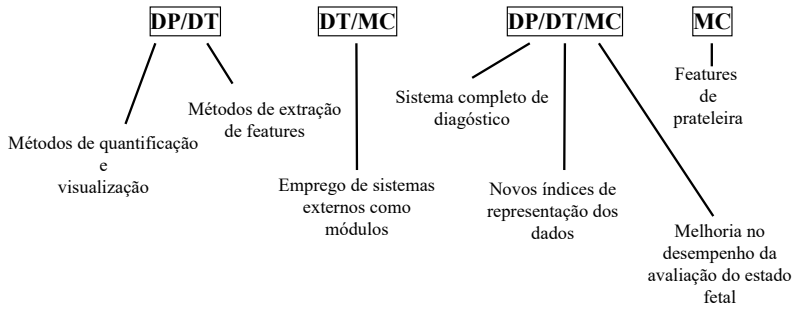


Figura 5 – Adoção dos estágios. Fonte: Elaboração própria

A Figura 5 apresenta os estudos agrupados por abordagem em relação à adoção dos estágios no desenvolvimento dos sistema em uma taxonomia de dois níveis: Preparação de dados (DP), transformação de dados (DT) e construção de modelo (MC). Para as publicações que empregaram a lista de estágios completa DP/DT/MC, identificamos três abordagens:

- **Sistemas completos de diagnóstico auxiliado por computador:** Apresentação de um sistema completo de diagnóstico auxiliado por computador (CAD) e inclui funcionalidades para realizar a preparação de dados, transformação de *features* e processos de classificação (ZHAO *et al.*, 2018; COMERT; KOCAMAZ, 2017; ANISHA *et al.*, 2021; COMERT; KOCAMAZ, 2018).
- **Novos índices de representação de dados:** Propostas de novos índices de representação de dados (DRI) e apresentação de sua validação na avaliação do estado fetal (DAS *et al.*, 2020c; FUENTEALBA *et al.*, 2019a; FUENTEALBA *et al.*, 2019b; ZENG *et al.*, 2021; ZHAO *et al.*, 2019; ZHAO *et al.*, 2019; STROUX *et al.*, 2017; BARQUERO-PEREZ *et al.*, 2017; KIM *et al.*, 2017; ZARMEHRI *et al.*, 2019; ALSAGGAF *et al.*, 2020; COMERT *et*

al., 2018; WU *et al.*, 2019).

- **Melhorar a avaliação do estado fetal:** Empregam uma combinação de métodos para melhorar o desempenho da avaliação do estado fetal (IFSD) (WARMERDAM *et al.*, 2018; INTAN *et al.*, 2019; GAO; LU, 2019; COMERT; KOCAMAZ, 2019; CZABANSKI *et al.*, 2016; MA'SUM *et al.*, 2019; FERGUS *et al.*, 2020; PETROZZIELLO *et al.*, 2019; DAS *et al.*, 2020b; RAMANUJAM *et al.*, 2020; COMERT *et al.*, 2019; FENG *et al.*, 2018).

Os artigos que não executaram a etapa de preparação de dados foram agrupados separadamente. Nestes artigos, foram empregados sistemas externos (EDP) para transformação dos dados como um fluxo de processos para executar as etapas de DT e MC diretamente (SIGNORINI; MAGENES, 2016; JEZEWSKI *et al.*, 2016; CZABANSKI *et al.*, 2020; JEZEWSKI *et al.*, 2019; ABBAS *et al.*, 2018; SIGNORINI *et al.*, 2020b). Também foram identificados estudos que empregaram dados de prateleira (ODF) com um conjunto de *features* pré-estabelecidas para avaliar o desempenho da construção de seus modelos apenas com o estágio MC (HUDDAR; SONTAKKE, 2019; KANNAN *et al.*, 2021; MOLLA *et al.*, 2021; AGRAWAL; MOHAN, 2019; BATRA *et al.*, 2017; SONTAKKE *et al.*, 2019; SUBASI *et al.*, 2020; POTHARAJU *et al.*, 2019; SHAH *et al.*, 2015; KADHIM; ABED, 2020; DAS *et al.*, 2020a; NAGENDRA *et al.*, 2017; PIRI; MOHAPATRA, 2019; ALSAYYARI, 2019; PIRI *et al.*, 2020; AFRIDI *et al.*, 2019; KEDDACHI; THELJANI, 2016; ZHANG; ZHAO, 2017; YILMAZ, 2016; CHAMIDAH; WASITO, 2015; CHEN *et al.*, 2019; HUANG *et al.*, 2020; KAUR *et al.*, 2019; IRAJI, 2019; DAS *et al.*, 2019; XUE, 2019; HOODBHOY *et al.*, 2019). Por fim, foram agrupados os estudos que não fornecem recursos preditivos e se utilizaram apenas dos estágios de preparação e transformação de dados isoladamente em duas abordagens, como segue:

- **Métodos de quantificação e visualização:** Fornecem métodos para quantificar e visualizar (QVM) parâmetros e indicadores do bem-estar fetal provenientes do biossinal (BOUDET *et al.*, 2020; LU *et al.*, 2019; FUENTEALBA *et al.*, 2017; MARQUES *et al.*, 2019; GIULIANO *et al.*, 2017; LU *et al.*, 2018; GEORGIEVA *et al.*, 2017; WOLF *et al.*, 2019; BOUDET *et al.*, 2019a; ITO *et al.*, 2021; ROMANO *et al.*, 2016; GYLLENCREUTZ *et al.*, 2018).
- **Métodos para extração de *features*:** Apresentaram novos métodos de extração de *features* (FEM) para a avaliação do estado fetal (WANG *et al.*, 2020).

3.3.5 RQ3-1: Quais abordagens de preparação de dados (DP) são empregadas?

Foi observados que os estudos empregaram as abordagens de preparação de dados em forma de cadeia de componentes ou de forma individual. A partir dos estudos selecionados, foram agrupadas as técnicas e abordagens empregadas para rejeição de registros de baixa qualidade na seleção de segmentos, redução de anomalias (artefatos) e ajustes extras no sinal em seu estágio de preparação de dados referente ao desenvolvimento de sistemas de avaliação do estado fetal, conforme apresentado na Tabela 6 e Tabela 7.

Na rejeição de artefatos durante o processo de seleção de sinal, os estudos empregaram critérios para descartar segmentos de dados não confiáveis, conforme apresentado na Tabela 6. Esses critérios compreendem a atribuição de limites para componentes ou valores anômalos, duração da perda de sinal e a porcentagem de leituras ausentes em relação a duração do sinal.

Tabela 6 – Abordagens para rejeição de biossinais. Fonte: Elaboração própria.

Rejeição de artefatos no biossinal	Pub. ID
Componentes FCM	S030, S063
Componentes FCM e intervalo (> 30s) que diferem (> 25 bpm)	S002
Valores ($\leq$ 20 bpm)	S075
Valores (> 200 bpm ou < 50 bpm)	S013
Lacunas no inicio/final	S020, S060
Perda (> 5%) ou (> 5) pontos consecutivos	S041, S055
Perda (> 20%)	S031
Perda (> 50%)	S023,S025,S026
Lacunas (> 3 s)	S071
Lacunas ($\geq$ 10 s)	S001
Lacunas (> 15 s)	S005, S022, S029, S035, S044, S061, S064, S067, S068, S072
Lacunas (> 15 s) e valores (> 200 bpm ou < 50 bpm)	S054
Valores ausentes	S010
Lacunas (> 2 min) ou segmentos (> 20 s) com perda de 40%	S004

FCM = frequência cardíaca materna.

Estudos empregaram a rejeição direta da frequência cardíaca materna (FCM) como componentes anômalos (BOUDET *et al.*, 2020; ANISHA *et al.*, 2021; DAS *et al.*, 2020b) e a rejeição de quaisquer valores ausentes (MARQUES *et al.*, 2019) do sinal FCF. Para valores anômalos, os limites empregados para rejeição foram pontos (< 20 bpm) (GYLLENCREUTZ *et al.*, 2018) e o par (> 200 bpm ou < 50 bpm) (FUENTEALBA *et al.*, 2019a; FUENTEALBA *et al.*, 2019b). A rejeição por tempo de perda de sinal variando de cinco pontos consecutivos (BARQUERO-PEREZ *et al.*, 2017), duração de três segundos (ROMANO *et al.*, 2016), dez segundos (ZHAO *et al.*, 2018), quinze segundos (COMERT; KOCAMAZ, 2019; FUENTEALBA

et al., 2019b; ALSAGGAF *et al.*, 2020; COMERT; KOCAMAZ, 2018; COMERT *et al.*, 2019; COMERT *et al.*, 2018; COMERT; KOCAMAZ, 2017; ZHAO *et al.*, 2019; ZHAO *et al.*, 2019; INTAN *et al.*, 2019; FENG *et al.*, 2018) e dois minutos (WANG *et al.*, 2020) para duração de lacunas. Identificamos a rejeição de segmentos por intervalo de sinal (> 30 s) que contém pontos que diferem (> 25 bpm) (BOUDET *et al.*, 2020). Para a porcentagem de perda de sinal, os limites variam de cinco por cento (BARQUERO-PEREZ *et al.*, 2017; SIGNORINI *et al.*, 2020b), vinte por cento (WARMERDAM *et al.*, 2018), quarenta por cento (WANG *et al.*, 2020) e cinquenta por cento (GIULIANO *et al.*, 2017; GEORGIEVA *et al.*, 2017; WOLF *et al.*, 2019). Identificamos também a rejeição direta de intervalos específicos, como os últimos cinco minutos (ZARMEHRI *et al.*, 2019).

A abordagem de remoção de artefato mais frequente foram as lacunas de sinal (> 15 s).

De acordo com a Tabela 7, as abordagens de redução de artefato e ajuste extra evitam a rejeição de artefato ao melhorar a qualidade geral do sinal. Tanto a interpolação linear quanto a de Hermite foram empregadas para ajustar as mudanças abruptas no sinal, preencher os pontos ausentes e as lacunas do sinal e para correção de *outliers*.

Para ajustar as mudanças abruptas no sinal, a diferença (> 25 bpm) entre os pontos adjacentes foi linearmente e Hermit interpolado. Estudos fizeram uso de abordagens alternativas para ajustes de sinal, como a média dos cinco pontos mais próximos (SIGNORINI *et al.*, 2020b), a substituição de valores ausentes por zero (MA'SUM *et al.*, 2019), substituição dos valores ausentes por regressão baseada na média (STROUX *et al.*, 2017), e a média entre o predecessor e o sucessor quando os pontos adjacentes diferem (> 20 bpm) (MARQUES *et al.*, 2019).

Os valores de limiar (<60 bpm), (> 200 bpm ou <50 bpm), (> 220 bpm ou <50 bpm) foram linearmente e Hermit spline interpolados como *outliers*.

Os estudos também empregaram *detrending* polinomiais de segunda ordem, *detrending* polinomiais de terceira ordem, filtros Gaussianos, *smoothing* de média móvel, unscented Kalman filter baseado em mínimos quadrados (UKF) e o filtro adaptativo de mínimos quadrados (LMS) aprimorado para alavancar a qualidade do sinal.

3.3.6 RQ3-2: Quais abordagens de transformação de dados (DT) são empregadas?

A principal atividade de transformação de dados é a extração de *features* dos sinais biológicos como forma de representação de dados. Os estudos apresentaram representação de

Tabela 7 – Abordagens para ajustes nos biossinais. Fonte: Elaboração própria.

Redução de artefatos e ajustes extras	Pub. ID
Interpolação de Hermite dos valores ausentes	S058
Interpolação linear de lacunas	S075
Média móvel	S020
Interpolação linear de lacunas ($\leq$ 15s); Interpolação de Hermite (> 15s); Interpolação linear de até 5 pontos que diferem entre si (> 25 bpm).	S016
Interpolação linear de intervalos R–R que diferem (> 25 bpm) ou (> 200 ms)	S031
Substituir lacunas (< 5) pontos pela média 5 valores seguintes.	S055
Interpolação linear de intervalos que diferem (> 25 bpm) entre si; Interpolação de Hermite de lacunas ($\leq$ 15 s); Detrending baseada em Baseline	S072
Interpolação linear de valores (> 220 bpm ou < 50 bpm); Interpolação de Hermite dos valores adjacentes que diferem entre si (> 25 bpm).	S001
Interpolação linear dos valores ausentes	S002
Interpolação linear de lacunas ($\leq$ 15 s) ou valores adjacentes que diferem (> 25 bpm); Interpolação de Hermite de valores (> 200 bpm ou < 50 bpm)	S005,S022,S029, S035
Interpolação de Hermite de valores (> 200 bpm ou < 50 bpm)	S008
Média entre o valor antecessor e sucessor quando valores adjacentes diferem (> 20 bpm)	S010
Interpolação de Hermite em lacunas ($\leq$ 15 s).	S013
Filtro Gaussiano	S040
Interpolação linear de valores (< 60 bpm) ou adjacentes que diferem (> 25 bpm)	S041
Interpolação linear de valores (> 220 bpm ou < 50 bpm) e segmentos (> 30 s) com valores adjacentes que diferem (> 25 bpm)	S042
Interpolação de Hermite em lacunas ($\leq$ 15 s); Filtro de mediana e Detrending polinomial de segunda ordem	S044
Interpolação linear de segmentos (> 2 s) com valores adjacentes que diferem (> 25 bpm) ou valores (> 220 bpm ou < 50 bpm)	S060
Interpolação de Hermite de lacunas ($\leq$ 15 s); Interpolação linear de valores (> 200 bpm ou < 50 bpm); Detrending polinomial de segunda ordem	S061,S067, S068
Interpolação de Hermite de lacunas ($\leq$ 15 s) para FHR e ($\leq$ 25 s) para UC; Filtro de médias móveis na UC	S054
Interpolação linear de valors adjacentes que diferem (> 25 bpm); Interpolação de Hermite em lacunas ($\leq$ 15 s); Detrending polinomial de segunda ordem	S064
Componente de sistema Black-box	S003,S007,S021, S024,S050,S051, S065
Interpolação linear de valores ausentes; Filtro Gaussiano	S070
Filtro adaptativo baseado em Least Mean Square (LMS)	S030
Interpolação linear de pontos adjacentes que diferem (> 25 bpm); Detrending polinomial de terceira ordem; Downsample de sinal para 1Hz	S004
Unscented Kalman filter (UKF) baseado em Least-square	S049
Substituição de valores ausentes por zero	S053
Interpolação linear de valores ausentes; Downsample para 0.25Hz	S059
Substituição de valores ausentes por regressão baseada em médias	S032
Ajustes no limite de picos, faixa e frequência da contração uterina	S012
Interpolação linear de lacunas ($\leq$ 3 s)	S071

MHR = *maternal heart rate* (frequência cardíaca materna).

dados morfológicos, temporais, de frequência, não lineares, clínicos e do domínio do tempo-frequência.

A partir dos estudos selecionados, foi identificado o uso de *features* bem estabelecidas na literatura, bem como a proposição de novos índices para representação das informações

extraídas dos biossinais. As *features* bem estabelecidas pertencem aos seguintes domínios:

- **Morfológico:** apresenta informações em torno dos valores da linha de base e contrações uterinas (ZHAO *et al.*, 2018; BOUDET *et al.*, 2020; LU *et al.*, 2019; COMERT; KOCAMAZ, 2017; SIGNORINI; MAGENES, 2016; FUENTEALBA *et al.*, 2017; MARQUES *et al.*, 2019; JEZEWSKI *et al.*, 2016; CZABANSKI *et al.*, 2020; GIULIANO *et al.*, 2017; LU *et al.*, 2018; WOLF *et al.*, 2019; GAO; LU, 2019; BOUDET *et al.*, 2019a; CZABANSKI *et al.*, 2016; JEZEWSKI *et al.*, 2019; ITO *et al.*, 2021; COMERT; KOCAMAZ, 2018; DAS *et al.*, 2020b; COMERT *et al.*, 2019; COMERT *et al.*, 2018; ROMANO *et al.*, 2016; FENG *et al.*, 2018; GYLLENCREUTZ *et al.*, 2018) .
- **Domínio do tempo:** representa as informações dependentes do tempo (ZHAO *et al.*, 2018; COMERT; KOCAMAZ, 2017; SIGNORINI; MAGENES, 2016; MOLLA *et al.*, 2021; DAS *et al.*, 2020c; GIULIANO *et al.*, 2017; LU *et al.*, 2018; WOLF *et al.*, 2019; ANISHA *et al.*, 2021; WARMERDAM *et al.*, 2018; STROUX *et al.*, 2017; INTAN *et al.*, 2019; CZABANSKI *et al.*, 2016; JEZEWSKI *et al.*, 2019; MA'SUM *et al.*, 2019; FERGUS *et al.*, 2020; PETROZZIELLO *et al.*, 2019; COMERT; KOCAMAZ, 2018; ALSAGGAF *et al.*, 2020; RAMANUJAM *et al.*, 2020; COMERT *et al.*, 2019; COMERT *et al.*, 2018; FENG *et al.*, 2018).
- **Frequência:** extrai padrões baseados na frequência dos sinais (ZHAO *et al.*, 2018; COMERT; KOCAMAZ, 2017; GIULIANO *et al.*, 2017; WARMERDAM *et al.*, 2018; SIGNORINI *et al.*, 2020b; ROMANO *et al.*, 2016; FENG *et al.*, 2018).
- **Não linear:** emprega técnicas não lineares para extrair representações de sinal úteis (ZHAO *et al.*, 2018; COMERT; KOCAMAZ, 2017; SIGNORINI; MAGENES, 2016; GIULIANO *et al.*, 2017; GEORGIEVA *et al.*, 2017; WARMERDAM *et al.*, 2018; SIGNORINI *et al.*, 2020b; COMERT; KOCAMAZ, 2018; RAMANUJAM *et al.*, 2020; COMERT *et al.*, 2019; ROMANO *et al.*, 2016; FENG *et al.*, 2018).
- **Frequência de tempo:** extrai informações que são dependentes de frequência que variam com o tempo dos sistemas biológicos (COMERT; KOCAMAZ, 2017; COMERT; KOCAMAZ, 2018; COMERT *et al.*, 2019).
- **Clínico:** complementa a representação dos dados derivados do sinal com informações clínicas sobre a gestação (DAS *et al.*, 2020c; GIULIANO *et al.*, 2017; ABBAS *et al.*, 2018).

Os domínios morfológico e temporal são as representações de dados mais frequentes

entre os artigos avaliados. O domínio do tempo compreende qualquer representação de dados dependente do tempo, como variabilidade da frequência cardíaca (VFC) adulto, estatísticas do biossinal na forma de série temporal e índices baseados em eletrocardiograma (ECG) fetal.

Em combinação com representações de dados bem estabelecidas, os estudos selecionados também propuseram novos índices de representação de dados nos domínios morfológico (DAS *et al.*, 2020c), não linear (WANG *et al.*, 2020; ZHAO *et al.*, 2019; BARQUERO-PEREZ *et al.*, 2017; KIM *et al.*, 2017; WU *et al.*, 2019), frequência de tempo (FUENTEALBA *et al.*, 2019a; ZENG *et al.*, 2021; POTHARAJU *et al.*, 2019; ZHAO *et al.*, 2019; COMERT; KOCAMAZ, 2019; FUENTEALBA *et al.*, 2019b; ZARMEHRI *et al.*, 2019; COMERT *et al.*, 2018), e domínio de tempo (STROUX *et al.*, 2017; ALSAGGAF *et al.*, 2020).

No que se refere ao domínio morfológico, Das *et al.* (DAS *et al.*, 2020c) introduziu parâmetros baseados em UC para melhorar o desempenho na identificação do estagio do parto. No domínio de tempo, Stroux *et al.* (STROUX *et al.*, 2017) avaliou a capacidade discriminativa de marcadores FCF para o diagnóstico da IUGR através da estimativa da *short term variability* (STV) e *long term variability* (LTV) correspondendo aos estados ativo e repouso dos fetos. Comert *et al.* (ALSAGGAF *et al.*, 2020) decompôs o sinal FCF em séries temporais multicanal para obter descritores e índices para o *common spatial patterns*.

As novas propostas do domínio de tempo e frequência, fizeram uso de representações derivadas da energia do espectro no sinal. Representações baseadas na energia espectral via *ensemble empirical mode decomposition with adaptive noise* (CEEMDAN) e modelagem auto-regressiva variável no tempo (TV-AR) foram propostas (FUENTEALBA *et al.*, 2019a; FUENTEALBA *et al.*, 2019b). Representações baseadas na análise do espectro via *Fast Fourier Transform* também foram empregadas (ZARMEHRI *et al.*, 2019). *Short-Time Fourier Transform* foi usado para produzir imagens de espectrogramas para uso direto (COMERT; KOCAMAZ, 2019) nas avaliações do estado fetal e para derivar *features* (POTHARAJU *et al.*, 2019; COMERT *et al.*, 2018). Representações do domínio de tempo e frequência baseado em descritores de imagens foram obtidos através de *continuous wavelet transform* (CWT), *wavelet coherence* (WTC), e *cross-wavelet transform* (XWT) (POTHARAJU *et al.*, 2019; COMERT; KOCAMAZ, 2019; COMERT *et al.*, 2018). Esses métodos extrairam novos índices de frequência de tempo a partir de espectogramas e incorporaram descritores úteis para a avaliação do estado fetal.

A análise da dinâmica não linear da frequência cardíaca fetal foi explorada por meio de gráficos de visibilidade (WANG *et al.*, 2020), *normalized compression distance* (BARQUERO-

PEREZ *et al.*, 2017), via *unscented Kalman filter* (UKF) (KIM *et al.*, 2017), pela dimensão fractal de imagens digitalizadas (WU *et al.*, 2019) e pela reconstrução do espaço de fase da série temporal em imagens de gráfico de recorrência bidimensional (ZHAO *et al.*, 2019).

Nos artigos avaliados, os domínios que agregaram novas proposta de índices mais frequentes foram os de frequência de tempo e o domínio não linear.

3.3.7 RQ3-3: Quais métodos de construção de modelo (MC) são empregados?

O estágio referente a construção de modelos (MC) agrega inteligência aos sistemas de avaliação do estado fetal assistida por computador, fornecendo-lhes habilidades de classificação ou preditivas. Os estudos selecionados utilizaram dados retrospectivos para avaliar seus modelos e utilizaram diferentes técnicas para cada etapa do projeto. Visando uma análise objetiva dos métodos e técnicas utilizadas, separamos os estudos que empregaram apenas a etapa de construção do modelo de forma isolada, dos estudos que empregaram a cadeia de multiestágios em seu processo de desenvolvimento.

Os estudos selecionados que empregaram apenas o estágio MC de forma isolada se utilizaram da base de dados UCI-CTG (DUA; GRAFF, 2017) com foco na combinação de métodos para melhorar a capacidade preditiva dos modelos. O UCI-CTG fornece 21 *features* provenientes dos domínios morfológicos e de tempo prontos para uso ao longo de 2126 registros rotulados por especialistas como normais (1655), suspeitos (295) e patológicos (176). Os métodos utilizados e o desempenho alcançado em cada estudo que empregou o estágio MC isoladamente através das *features* prontas para uso do UCI-CTG foram resumidos na Tabela 8.

Diversas técnicas foram empregadas com o objetivo de potencializar a capacidade discriminativa dos modelos, dentre elas, identificamos o uso de algoritmos de classificação sozinhos ou em uma combinação de algoritmos de classificação com técnicas de balanceamento de dados, seleção de *features*, redução dimensional e técnicas de clusterização em conjunto com os algoritmos de classificação.

Nos estudos que empregaram algoritmos de classificação isoladamente, *artificial neural network* (ANN) (HUDDAR; SONTAKKE, 2019; XUE, 2019), *random forest* (RF) (MOLLA *et al.*, 2021; SONTAKKE *et al.*, 2019; KAUR *et al.*, 2019), *decision trees* (DT) (AGRAWAL; MOHAN, 2019), *support vector machines* (SVM) (BATRA *et al.*, 2017), *legendre neural network* (LNN) (ALSAYYARI, 2019), *multi-support vector domain description* (SVDD) (KEDDACHI; THELJANI, 2016), *probabilistic neural network* (PNN) (YILMAZ, 2016), *weighted random*

forest (WRF) (CHEN *et al.*, 2019), *adaptive neuro-fuzzy inference systems* (ANFIS) (HUANG *et al.*, 2020), *long short-term memory* (LSTM) (DAS *et al.*, 2019), e *deep stacked sparse auto-encoders* (DSSAEs) (IRAJI, 2019) foram utilizados.

Para lidar com problemas de balanceamento de classes, o *Synthetic Minority Over-sampling Technique* (SMOTE) foi aplicado em conjunto com *Extreme gradient boosting* (XGBoost) (HOODBHOY *et al.*, 2019) e *k-nearest neighbors* (KNN) (POTHARAJU *et al.*, 2019). A técnica de redução de dimensionalidade *principal component analysis* (PCA) foi aplicada em conjunto com o classificador *adaptive boosting* (Adaboost) (ZHANG; ZHAO, 2017).

Sobre a adoção de técnicas de seleção de *features*, o algoritmo *firefly* (KADHIM; ABED, 2020), *minimum redundancy maximum relevance* (MRMR) (DAS *et al.*, 2020a), *correlation based feature selection* (CFS) (AFRIDI *et al.*, 2019) e técnicas de seleção de *features* baseadas em ranking (FR) techniques (NAGENDRA *et al.*, 2017; PIRI; MOHAPATRA, 2019) foram empregadas. Houve ainda o uso de técnicas baseadas em modelos evolutivos como *Particle Swarm Optimization* (PSO) (KANNAN *et al.*, 2021) e *multi-objective genetic algorithm* (MOGA) (PIRI *et al.*, 2020), os quais foram empregados em conjunto com classificadores RF and XGboost. Técnicas de *Bagging ensemble* também foram adotadas em conjunto com o classificador *random forest* (SUBASI *et al.*, 2020; SHAH *et al.*, 2015).

Os modelos com melhor desempenho atingiram acima de 99% de acurácia utilizando *Particle Swarm Optimization* (PSO) e RF (KANNAN *et al.*, 2021), *Bagging* e RF (SUBASI *et al.*, 2020), SMOTE e KNN (POTHARAJU *et al.*, 2019), MRMR e RF (DAS *et al.*, 2020a), FR e RF (NAGENDRA *et al.*, 2017), RF (KAUR *et al.*, 2019), WRF (CHEN *et al.*, 2019), e DSSAEs (IRAJI, 2019). É fácil perceber que o classificador *random forest* de forma isolada ou combinados com outras técnicas foi os método de classificação com o melhor desempenho no conjunto de *features* da base de dados UCI-CGT.

Para os estudos que empregaram vários estágios no processo de desenvolvimento, resumimos na Tabela 10 e Tabela 9 os blocos de construção empregados na construção dos modelos considerando parâmetros chave como bases de dados, critérios para separação das classes em patológicas e normais, intervalo do biossinal utilizado, abordagens para anotação das classes, tipo de representação dos dados usada pelos algoritmos de classificação e as métricas empregadas para medir o desempenho dos modelos para avaliação do bem-estar fetal.

Com uma rápida análise da Tabela 9 e Tabela 10, é possível perceber que o CTU-UHB foi o conjunto de dados mais recorrente, o número de registros em cada conjunto de dados

Tabela 8 – Construção de modelos no UCI-CTG. Fonte: Elaboração própria.

Referência	Método	*Score* (%)
(HUDDAR; SONTAKKE, 2019)	ANN	Acc=74.6
(KANNAN *et al.*, 2021)	PSO+RF	Acc=99.57; Pr=99.6; Rc=99.6
(MOLLA *et al.*, 2021)	RF	Acc=94.8; Pr=94.8; Rc=94.8
(AGRAWAL; MOHAN, 2019)	DT	Acc=91.54
(BATRA *et al.*, 2017)	SVM	Acc=92.39
(SONTAKKE *et al.*, 2019)	RF	Acc=93.4
(SUBASI *et al.*, 2020)	Bagging+RF	Acc=99.02; AUC=99.9
(POTHARAJU *et al.*, 2019)	SMOTE+KNN	Acc=99.05
(SHAH *et al.*, 2015)	Bagging+RF	Acc=94.73
(KADHIM; ABED, 2020)	FFly+NB	Acc=86.54
(DAS *et al.*, 2020a)	MRMR+RF	Acc=99.91
(NAGENDRA *et al.*, 2017)	FR+RF	Acc=99.28
(PIRI; MOHAPATRA, 2019)	FR+CBA	Acc=84.02
(ALSAYYARI, 2019)	LNN	MSE=0.001
(PIRI *et al.*, 2020)	MOGA+XGboost	Acc=94
(AFRIDI *et al.*, 2019)	CFS+NB	Acc=83.06; Pr=92.2;Rc=83.10
(KEDDACHI; THELJANI, 2016)	Multi-SVDD	Acc=82.20
(ZHANG; ZHAO, 2017)	PCA+Adaboost	AUC=95.6
(YILMAZ, 2016)	PNN	Acc=92.05
(CHAMIDAH; WASITO, 2015)	K-means+SVM	Acc=90.64
(CHEN *et al.*, 2019)	WRF	Acc=99.71; F1=97.85
(HUANG *et al.*, 2020)	ANFIS	Acc=97.54
(KAUR *et al.*, 2019)	RF	Acc=97; Pr=97; Rc=99
(IRAJI, 2019)	DSSAEs	Acc=99.5; Se=99.7; Sp=97.5
(DAS *et al.*, 2019)	LSTM	Acc=98
(XUE, 2019)	ANN	Acc=91.85
(HOODBHOY *et al.*, 2019)	SMOTE+XGBoost	Pr=92; Rc=92; F1=92

PSO = Particle Swarm Optimization, SVM = support vector machine, ANN = artificial neural network, RF = random forest, KNN = k-nearest neighbors, FFly = Fire Fly, MRMR = Minimum Redundancy Maximum Relevance, FR = feature rank, DT = decision tree, NB = Naive Bayes, CBA = classification based on association, LNN = legendre neural network, CFS = correlation based feature selecion, SVDD = support vector domain description , MOGA = multi-objective genetic algorithm, XGboost = Extreme gradient boosting, Adaboost = Adaptive Boosting, PCA = Principal component analysis, PNN = probabilistic neural network, WRF = weighted random forest, ANFIS = adaptive neuro-fuzzy inference system, DSSAEs = deep stacked sparse auto-encoders, LSTM = long short-term memory.

variou de unidades a milhares, o intervalo de sinal analisado diferiu de 15 min até o uso de sinal completo. A rotulagem de classe mais frequente foi o valor do pH e o tipo de classificação mais recorrente foi o baseado em *features*. Percebe-se ainda que o método de classificação mais frequente foi SVM, e as métricas mais utilizadas foram sensibilidade e especificidade.

Para rotulagem de classe, os valores de pH diferem de 7,05, 7,1, 7,15 até 7,20. Os valores de BDecf variaram entre 8 e 12, os escores de Apgar variaram entre 4, 5 e 7 e o BW variou de 5 a 10. A anotação de classes com base em especialistas, o tipo de parto e diagnóstico patológico específico como IUGR também foram empregados nos estudos. Estudos demonstram ainda o emprego de mais de um critério para separação de classes onde os valores de pH foram usados em conjunto com o BE (ZENG *et al.*, 2021), com o BDecf (FUENTEALBA *et al.*, 2019a;

Tabela 9 – Construção de modelos em estudos multiestágio baseados em *features*. Fonte: Elaboração própria.

ID	Dados	Divisão(N/P)	Parâmetros			Score (%)
			Intervalo	Rotulagem	Método	
S001	CTU-UHB	(447/105)	Full	pH<7.15	Adaboosting	Se=92;Sp=90
S005	CTU-UHB	(272/44)	30 min	Experts	SVM	Se=89;Sp=81
S012	CTU-UHB	(275)	30 min	Experts	Fuzzy+ANN	AUC=90.9
S013	CTU-UHB	(354/18)	60 min	pH<7.05;Bf≥12 pH>7.20;Bf<12	SVM	Se=79.5; Sp=86.45
S016	CTU-UHB	(442/27)	30 min	pH≤7.05;BE≤ −10	ECSVM	Se=85.2;Sp=66.1
S018	CTU-UHB	-	-	pH≤7.15;Bf≥8	RF	Se=87.5;Sp=100
S020	CTU-UHB UCI-CTG	(323/228) (II)(1655/176)	-	pH<7.2;A1<7;BW<10 (II)Experts	FCM	Se=59.61;Sp=78.46 II(Acc=87.81)
S021	CTU-UHB	(507/44)	-	pH≤7.05	FCM+	Se=71.29;Sp=93.73
S030	NIFECGDB ADFECGDB PCCDB	(300)	-	Experts	SVM	Se=98;Sp=96
S031	*private*	(80/20)	45 min	pH<7.05	SVM	Se=81;Sp=77
S032	*private*	(1163/1163)	60 min	IUGR diag.	LR	AUC=76
S035	CTU-UHB	(447/105)	-	pH<7.15	NB+Bagging	Pr=34; Rc=65
S041	HUFA	(17/15)	Full	pH<7.05;A5≤7	NCD+KNN	Se=92;Sp=85
S050	*private*	(685/92)	Full	pH<7.1;BW≤5;A5<5	LSVM	Acc=84.9;QI=67.25
S051	CTU-UHB	(443/108)	Full	pH<7.1;A5≤4;BW≤5	CPP	Se=69.14;Sp=88.33
S054	CTU-UHB	(354/18)	20-60 min	pH>7.05;Bf≥12 pH>7.20;Bf<12	SVM	QI=83.2
S055	DLN-IUGR	(60/60)	-	IUGR diag.	RF	Se=90.2;Sp=91.9
S060	(I)*private* (II)SpaM (III)CTU-UHB	(I)(142/6) (II)(240/60) (III)(552)	50 min	pH<7.05	FFT	(I)Se=100;Sp=85.1 (II)Se=67;Sp=80 (III)Se=63.6;Sp=80.1
S061	CTU-UHB	(375/177)	15 min	pH<7.2	SVM	Se=76.83; Sp=78.27
S063	CTU-UHB	(552)	Full	Experts	Fuzzy+RF	Acc=93
S064	CTU-UHB	(447/105)	30 min	pH<7.15	SVM	Se=74.29;Sp=99.55
S065	CTU-UHB	(552)	-	Expert	RFE+DT	Acc=97.81
S067	CTU-UHB	(375/177)	15 min	pH<7.2	SVM	Se=77.4;Sp=93.86
S068	CTU-UHB	(439/113)	30 min	pH<7.15	GA+LS-SVM	Se=63.45;Sp=65.88
S070	*private*	(1859/161)	20 min	Expert	SVM	Rc=1
S072	CTU-UHB	(358/62)	30 min	pH<7.1;pH>7.2	Deep GP	Se=91;Sp=82

N = normal, P = patológico, Bf = BDecf = déficit de base no fluído extracelular, BE = Excesso de base, BW = Peso ao nascer, A{1,5} = Apgar scores no 1st e 5th minutos, IUGR = *Intrauterine growth restriction*, SVM = *support vector machine*, ANN = *artificial neural network*, ECSVM = *Ensemble Cost-sensitive* SVM, RF = *random forest*, FCM = *fyzzy c-means*, CPP = *clustering with pair of prototypes*, LR = *logistic regression*, LSVM = *Lagrangian* SVM, FFT = f*ast fourier transform*, NCD= *Normalized compresion distance*, DT = *decision tree*, NB = Naive Bayes, LS-SVM = Least Square SVM.

ABBAS *et al.*, 2018; FUENTEALBA *et al.*, 2019b), com BW e Apgar (JEZEWSKI *et al.*, 2016; CZABANSKI *et al.*, 2016; JEZEWSKI *et al.*, 2019).

Analisando a divisão de dados, tipo de representação, métodos e desempenho dos modelos, é importante ressaltar que as abordagens de autoaprendizagem empregaram mais dados e entre seus modelos de maior desempenho houve o uso de representações dos dados derivadas de imagens em conjunto com classificadores 2-D-CNN. Por outro lado, as abordagens baseadas

Tabela 10 – Construção de modelos em estudos multi-estágio baseados em extração automática. Fonte: Elaboração própria.

ID	Dados	Divisão(N/P)	Parâmetros			Score (%)
			Intervalo	Rotulagem	Método	
S022	CTU-UHB	(21,000/21000)	20 min	pH<7.15	2-D-CNN	Se=99.29;Sp=98.10
S029	CTU-UHB	(2,632/630)	20 min	pH<7.15	2-D-CNN	Se=98.22;Sp=94.87
S040	*private*	(272/20)	-	Experts	LSTM	Se=100; Sp=82.2
S044	CTU-UHB	(508/44)	15 min	pH$\leq$7.05	2-D-CNN	Se=56.16;Sp=96.51
S049	CTU-UHB	(139/107)	20 min	pH<7.15	MDL	Acc=93
S053	CTU-UHB	(447/105)	Full	pH$\leq$7.15	2-D-CNN	F1=81
S058	CTU-UHB	(506/46)	Full	Deliv. type	1-D-CNN	Se=80;Sp=79
S059	(I)*private*	(35,429)	60 min	pH<7.05	MCNN	(I)AUC=73
	(II)SpaM	(160/40)				(II)AUC=91
	(III)CTU-UHB	(552)				(III)AUC=77

N = normal, P = patológico, 2-D = duas dimensões, CNN = *convolutional neural network*, MDL = *minimum description lenght*, MCNN = *multimodal* CNN.

em *features* alcançaram seus melhores resultados com menos dados, empregando algoritmos de classificação bem estabelecidos na literatura, como SVM em conjunto com métodos de *ensemble*.

Vale ressaltar que mesmo estudos diferentes que adoraram o mesmo conjunto de dados ainda apresentaram forte variação nos parâmetros de configuração e blocos de construção auxiliares, o que prejudica a comparação direta das soluções existentes. O uso de abordagens de autoaprendizagem alcançou os melhores desempenhos sob a perspectiva apenas das métricas de avaliação, mas o uso destes modelos exigiu grandes quantidades de dados e etapas extras de transformação de dados.

3.4 Comparativo com Resultados da Literatura

Nesta seção, são apresentados trabalhos relacionados com a proposta desta tese os quais complementam os artigos avaliados em nosso mapeamento sistemático. Aqui exibimos um sumário das suas contribuições e por fim, apresentamos as vantagens desta tese frente ao estado da arte.

Apesar de uma comparação exata com a literatura publicada recentemente ser difícil devido a variações nos parâmetros chave no que se refere ao desenho dos experimentos, a abordagem proposta é comparável ou mesmo supera trabalhos relevantes da mesma.

A Tabela 11 resume os resultados de trabalhos correlatos que se utilizaram de bases de dados retrospectivos de forma isolada e a Tabela 12 apresenta o resumo dos trabalhos que realizaram abordagens *cross-dataset*. São comparadas informações relevantes do seu desenho experimental, os quais devem ser tratados com cautela, uma vez que diferentes critérios e

Tabela 11 – Trabalhos com bases de dados isoladas abordando a avaliação fetal. Fonte: Elaboração própria.

Ref.	**Dados**	**# of (N/P)**	**Métodos**			**Escore (%)**
			SS	**SC**	**CL**	
(ZENG *et al.*, 2021)	CTU-UHB	(442/27)	30min	pH<7.05; BE≤-10	ECSVM	Sp: 66.01 Se: 85.20
(ALSAGGAF *et al.*, 2020)	CTU-UHB	(447/105)	30min	pH < 7.15	SVM	Sp: 99.55 Se: 73.33
(ZHAO *et al.*, 2019)	CTU-UHB	(21000/21000)	Full	pH < 7.15	2D-CNN	Se: 99.29 Sp: 98.10
(LI *et al.*, 2019)	Privado	(3012/1461)	20min	Experts	1D-CNN	Sp: 84.77 Se: 98.40
(COMERT; KOCAMAZ, 2019)	CTU-UHB	(552/44)	15min	pH < 7.05	2D-CNN	Sp: 96.51 Se: 56.15
(COMERT *et al.*, 2018)	CTU-UHB	(439/113)	30min	pH < 7.15	LS-SVM	Sp: 65.88 Se: 63.45
(BARQUERO-PEREZ *et al.*, 2017)	HUFA	(15/17)	60min	pH < 7.05	KNN	Sp: 92.0 Se: 85.0
(GEORGOULAS *et al.*, 2017)	CTU-UHB	(508/44)	30min	pH < 7.05	LS-SVM	Sp: 65.30 Se: 72.12
(SPILKA *et al.*, 2017)	Privado	(1251/37)	20min	pH < 7.05	S-SVM	Sp: 75.0 Se: 73.0
Este trabalho	DB-Trium (I)* CTU-UHB (II) HUFA (III) SpAM** DB-HeraB**	(112/47) (358/40) (14/13)	30min	Experts* (pH<7.05) Sem rótulos**	SVM	(I)Sp: 76.1 (I)Se: 63.7 (II)Sp: 85.6 (II)Se: 67.5 (III)Sp: 75.0 (III)Se: 73.3

N = normal, P = patológico, SC = critério de separação das classes, SS = esquema de segmentação, CL = melhores classificadores.

combinações de blocos de construção estão envolvidos.

Em nossos cenários avaliativos, empregamos a Especificidade (SP), a qual representa a capacidade de reconhecimento para classes saudáveis, enquanto a Sensibilidade (SE) representa esta capacidade para classes patológicas. A AUC representa ambas as capacidades onde valores acima de 50% indicam um modelo com desempenho melhor do que um classificador aleatório. Recomenda-se um mínimo de 60% para Sensibilidade (PETROZZIELLO *et al.*, 2018), com um cenário ideal sendo SE e SP os mais altos possíveis (PETROZZIELLO *et al.*, 2019). Os trabalhos relacionados foram comparados tendo por base estas recomendações em conjunto com os parâmetros e blocos de construção empregados nos experimentos.

Em relação aos trabalhos que superaram os modelos mais bem classificados desta tese, a comparação direta é difícil. Alguns desses trabalhos (BARQUERO-PEREZ *et al.*, 2017; ALSAGGAF *et al.*, 2020) propuseram novos índices diagnósticos (*features*) para detecção de patologias que foram avaliados em conjuntos de dados de forma isolada. Barquero-Pérez *et*

Tabela 12 – Trabalhos *cross-dataset* abordando avaliação fetal. Fonte: Elaboração própria.

Ref.	**Dados**	**# of (N/P)**	**Métodos**			**Escore (%)**
			SS	**SC**	**CL**	
(PETROZZIELLO *et al.*, 2019)	Privado (I) CTU-UHB (II) SpAM (III)	(30.000/5.000) (508/44) (160/40)	60min	pH < 7.05	MCNN	(I)AUC: 77.0 (I/II)AUC: 82.0 (I/III)AUC: 92.0
(ABRY *et al.*, 2018)	Privado (I) CTU-UHB (II)	(1.021/29) (330/14)	20min	pH < 7.05	S-SVM	(I)Sp: 74.0 (I)Se: 66.0 (I/II)Sp: 80 (I/II)Se: 64
(SPILKA *et al.*, 2016)	Privado (I) CTU-UHB (II)	(1.251/37) (400/20)	20min	pH < 7.05	S-SVM	(I)Sp: 76.0 (I)Se: 70.0 (II)Sp: 68.0 (II)Se: 70 (I/II)Sp: 83 (I/II)Se: 60
Este trabalho	DB-Trium* (I) CTU-UHB (II) HUFA (III) SpAM DBHeraBeat	(112/47) (358/40) (14/13) (NR) (NR)	30min	Experts* (pH<7.05) Automática**	SVM LabSp**	(I)Sp: 76.1 (I)Se: 63.7 (II)Sp: 85.6 (II)Se: 67.5 (III)Sp: 75.0 (III)Se: 73.3 (I/II)Sp: 71.6 (I/II)Se: 61.7 (II/I)Sp: 92.3 (II/I)Se: 42.7 (II/III)Sp: 81.5 (II/III)Se: 50.7 (III/I)Sp: 74.1 (III/I)Se: 63.4 (III/II)Sp: 64.6 (III/II)Se: 68.4 ***(I/II)Sp: 80.1* ***(I/II)Sp:62.5*

NR = Não rotulada, N = normal, P = patológico, SC = critério de separação das classes, SS = esquema de segmentação, CL = melhores classificadores, (a/b): treinamento em (a) e validação em (b). ** = Indica o uso da abordagem semi-supervisionada ao avaliar bases de dados rotuladas de forma cruzada.

al. (BARQUERO-PEREZ *et al.*, 2017) avaliou o uso da distância de compressão normalizada (NCD) + KNN alcançando SP = 92% e SE = 85% em um esquema de segmentação de sinal com divisões entre quatro horas a uma hora da hora do parto. Alsaggaf *et al.*(ALSAGGAF *et al.*, 2020) fez uso de eletroencefalografia multicanal (EEG) como um método de extração de *features* FCF para obter as padrões espaciais (CSP), onde o SP foi igual a 99,55%, e SE foi igual a 73,33%. Zeng *et al.* (ZENG *et al.*, 2021) empregou descritores de imagem de frequência de tempo como *feature* FCF e obteve SP = 66,1% e SE = 85,2%. Nesta tese, optou-se por não incluir as *features* mencionadas na avaliação dos modelos devido a necessidade de tarefas específicas de pré-processamento envolvidas em sua extração, acrescidas ao fato de ainda não serem representações bem estabelecidas no domínio da avaliação fetal, as quais foram testadas apenas em bases isoladas com configurações específicas. É importante ressaltar que a variação

Tabela 13 – *Features* empregadas nos trabalhos relacionados. Fonte: Elaboração própria.

Ref.	*Features* selecionadas
(ZENG *et al.*, 2021)	FHR(mean, median, std, meanAD, rms), CWTF, WTCF e XWTF(flux, flatness, energy concentration, NRen, ShannonEn)
(ALSAGGAF *et al.*, 2020)	Baseline, # of ACC, # of DCC, LTI, STV, II, Delta, FHR(mean, std, median, skewness, kurtosis), meanAD, medianAD, ApEn(2,0.15-20), SampEn(2,0.15-20), LZC, # of contractions, UC(mean, std), CSP1 e CSP2
(COMERT *et al.*, 2018)	FHR(meanAD, medianAD, baseline, rms, # of ACC, LTV, STV), IBTF(correlation-LF0, homogeneity-MF0, energy-VLF90, energy-LF0, correlation-MF0, energy-LF135, energy-MF45, energy-HF135)
(BARQUERO-PEREZ *et al.*, 2017)	FHR(mean, std, LTI, STV), moment(kth-order, central) e NCD
(GEORGOULAS *et al.*, 2017)	VLF, LF, Pointcaré SD2
(SPILKA *et al.*, 2017)	meadianAD, baseline ($\beta 0$), Hurst parameter
(ABRY *et al.*, 2018)	T_{stress}, DWT (c1, c2), medianAD, HF, LF, baseline ($\beta 1$)
(SPILKA *et al.*, 2016)	baseline ($\beta 0, \beta 1$), # of (ACC, DCC), medianAD, T_{stress}, DCC_{area}, LTV, STV, VLF, LF, HF, HF, LF/HF, spectral index (α), DWT (H, h_{min}, c1, c2, c3, c4)
Este trabalho	meanAD, FD_Higushi, TRI, NN20, pNN20 e CVNN

Abreviações: CWTF continuous wavelet transform, WTCF wavelet coherence , XWTF cross-wavelet transform, NRen normalized Renyi entropy, CSP common spatial patters, IBTF image based time-frequency, NCD normalized compression distance, ACC accelerations, DCC decelerations, LF Low Frequency, VLF Very Low Frequency, HF Hight frequency, DWT discrete wavelet transform multifractal parameters, T_{stress} averaged FHR duration, SD desvio padrão, AD desvio absoluto, FD fractal dimension, TRI triangular index, NNx número de pares com sucessivos intervalos NN que diferem mais que x, pNNx porcentagem da NNx, CV coeficiente de variação.

entre os limites de adotados para separação das classes prejudica uma comparação direta dos desempenhos. Nesta tese, as melhores pontuações baseadas em *features* foram SP = 85,6% para a base CTU-UHB e SE = 73,3% no conjunto de dados HUFA.

Imagens extraídas dos sinais FCF em conjunto com redes neurais convolucionais bidimensionais (2-DCNN) foram utilizadas para a avaliação do estado fetal. Zhao *et al.* (ZHAO *et al.*, 2019) fez uso de 42.000 imagens bidimensionais de gráfico recorrente (RP) provenientes da FCF para treinar classificadores 2-DCNN, alcançando SP = 98,10% e SE = 99,29%. Comert e Konamaz (COMERT; KOCAMAZ, 2019) fizeram uso da *Short-time Fourrier transform* (STFT) para extrair imagens de espectrogramas e aplicá-las em um classificador 2-DCNN, obtendo SP = 96,51% e SE = 56,15%. Neste último, percebe-se uma capacidade limitada para o reconhecimento das classes patológicas. Essas abordagens são promissoras, no entanto, a classificação de imagens

requer uma grande quantidade de dados para treinamento e procedimentos específicos de pré-processamento e transformação dos dados, o que esta além do escopo deste estudo.

Li *et al.* (LI *et al.*, 2019) superou os modelos de séries temporais (SP = 84,77 e SE = 98,40) ao empregar o CNN unidimensional (1-DCNN) nos sinais FCF para classificar a hipóxia fetal. No entanto, os resultados aqui obtidos indicam que 1-DCNN alcançou um bom desempenho no modelo baseado em séries temporais (SP = 71,3 e SE = 58,2), mesmo em um conjunto de dados dez vezes menor, reforçando a sugestão de avaliar os modelos baseados em séries temporais utilizando um conjunto de dados maior.

O desempenho dos melhores cenários de avaliação nesta tese parecem estar de acordo com os resultados de Georgoulas *et al.* (GEORGOULAS *et al.*, 2017), Spilka *et al.* (SPILKA *et al.*, 2017) e Comert *et al.* (COMERT *et al.*, 2018). Neste último, os cenários aqui melhor classificados alcançaram escores superiores do que em sua proposta de novo índice de prognóstico.

O desempenho deste trabalho parece estar de acordo com os resultados de Petrozzielo *et al.* (PETROZZIELLO *et al.*, 2018; PETROZZIELLO *et al.*, 2019), Abry *et al.* (ABRY *et al.*, 2018) e Spilka *et al.* (SPILKA *et al.*, 2016) em cenários de avaliações individuais, mesmo com nossos dados sendo 68 vezes e duas vezes menores, respectivamente, aos empregados nos referidos trabalhos. Para os cenários *cross-dataset*, nossa melhor SE = 68,4, foi compatível com a SE = 60,0 de Spilka *et al.* (SPILKA *et al.*, 2016) e com a SE = 64,0 de Abry *et al.* (ABRY *et al.*, 2018). A avaliação *cross-dataset* aqui realizada empregou ainda uma abordagem semi-supervisonada a qual apresentou SP = 80,1 e SE = 62,5. Petrozzielo *et al.* (PETROZZIELLO *et al.*, 2018; PETROZZIELLO *et al.*, 2019) alcançou AUC de até 92,0 na avaliação *cross-dataset*, superando as pontuações deste trabalho com a ressalva de ter aplicado um conjunto de dados dez vezes maior para treinamento e empregando uma arquitetura diferente para o modelo baseado em séries temporais.

Vale ressaltar que todas as avaliações de *cross-dataset* citadas nos trabalhos referenciados executaram o treinamento na base de dados com uma quantidade de amostras mais considerável para validar as bases menores (unidirecional). O processo de avaliação *cross-dataset* aqui realizado empregou todos os conjuntos de dados disponíveis para treinamento e testes, independentemente da quantidade de dados e aplicou a abordagem semi-supervisionada a fim de validar a qualidade dos dados para rotulagem automática, fazendo assim uma avaliação abrangente da generalização do desempenho do modelo e sua capacidade de reconhecimento em

cenários com restrição no quantitativo de registros.

3.5 Síntese do Capítulo

Este estudo de mapeamento sistemático objetivou criar uma visão geral dos estudos primários no que diz respeito aos blocos de construção que compõem os sistemas de avaliação do estado fetal. Um total de 75 artigos foram selecionados e os estudos foram classificados em diferentes categorias para permitir uma avaliação objetiva e reproduzível das publicações. Uma das novidades deste mapeamento é que, até onde sabemos, este é o primeiro panorama dos estudos sobre os componentes básicos dos sistemas de avaliação do estado fetal com foco nas etapas de projeto destes sistemas.

As principais conclusões do mapeamento são resumidas da seguinte forma:

- *(RQ1) Quais são os tipos de contribuições nos estudos sobre o sistemas de avaliação de estado fetal baseado em biossinal?* A maioria dos estudos selecionados apresentaram o tipo de publicação *método*. Os tipos *outros - ferramentas* e *técnica* também estão presentes ao longo dos anos; entretanto, a ausência de de estudos do tipo *validação* no mesmo período indica uma lacuna para estudos que validem estas propostas em cenários reais ou ambiente médico.
- *(RQ2) Quais dados são usados nos estudos?* O CTU-UHB foi o conjunto de dados baseado em biossinais mais frequente entre os estudos analisados e UCI-CTG foi o conjunto de dados com *features* prontas para uso mais frequente. Os estudos apontam uma variação considerável no número de registros entre os conjuntos de dados empregados para avaliação retrospectiva, sendo que a quantidade mais significativa (35.429) pertenceu a uma base de dados de caráter privada no que se refere ao acesso para outros pesquisadores.
- *(RQ3) Em quais etapas se divide o processo de desenvolvimento dos sistemas de avaliação do estado fetal?* Os estudos selecionados utilizaram os estágios de preparação de dados (DP), transformação de dados (DT) e construção de modelo (MC). Os estudos foram agrupados pelo uso de cada estágio em uma taxonomia de dois níveis (apresentado na Figura 5).
- *(RQ3-1) Quais abordagens de preparação de dados (DP) são empregadas?* A abordagem mais recorrente foi a rejeição de lacunas de sinal com duração maior que (15 s). A maioria dos estudos utilizou interpolação linear e de Hermite para atividades de redução de artefatos: (i) O ajuste da diferença (> 25 bpm) entre pontos adjacentes; (ii) Ajuste dos

valores *outlier* (<60 bpm), (> 200 bpm ou <50 bpm), (> 220 bpm ou <50 bpm); (iii) no preenchimento dos valores ausentes; (iv) no preenchimento de lacunas de sinal (<15 s).

– *(RQ3-2) Quais abordagens de transformação de dados (DT) são empregadas?* A atividade de transformação de dados mais recorrente foi a extração de *features*. Em combinação com representações de dados bem estabelecidas na literatura, os estudos selecionados também propuseram novos índices de representação. Os domínios morfológicos e temporais são as representações de dados mais frequentes entre as *features* bem estabelecidas. Por outro lado, a maioria das novas propostas de índices representativos pertence a domínios de frequência de tempo e domínio não linear. A maioria dos novos índices abordou a representação dos sinais baseados em espectrogramas.
– *(RQ3-3) Quais métodos de construção de modelo (MC) são empregados?* Para os estudos que empregaram o estágio MC diretamente com o conjunto de *features* prontas para uso provenientes da base de dados UCI-CTG, o classificador *Random Forest* sozinho ou em combinação com outras técnicas foi o modelo de classificação de melhor desempenho. A maioria dos estudos que empregaram múltiplos estágios de desenvolvimento fez uso do conjunto de dados CTU-UHB e a rotulagem de classe mais frequente foi o (pH leq 7,15). Nos modelos multiestágio o método de classificação mais recorrente foi o SVM dentro de esquemas baseados em *features*, e as métricas de desempenho mais usadas foram sensibilidade e especificidade.

Vale ressaltar que apesar do uso do mesmo conjunto de dados entre estudos diferentes, os parâmetros que compõem os blocos de construção dos sistemas nestes estudos apresentaram variações, dificultando sua comparação direta. Os resultados do mapeamento indicaram que embora o uso de abordagens baseadas diretamente nos biossinais com auto-extração de informações tenha alcançado os melhores desempenhos do estado da arte, exigiu grandes quantidades de dados e etapas extras de transformação de dados. Os modelos baseados em *features* alcançaram desempenhos competitivos e empregaram menores quantidades de dados, o que deve ser levado em consideração na tomada de decisão pelos blocos de construção dos sistemas de avaliação do estado fetal.

Em relação ao comparativo dos modelos de prognósticos desta tese com os trabalhos relacionados, em suma, nosso processo de avaliação e os modelos de prognóstico resultantes apresentam os seguintes destaques:

1. os modelos mais bem classificados desta tese alcançaram resultados comparáveis ou até

melhores na avaliação do estado fetal do que outros métodos do estado da arte, o que parece ser promissor como ferramenta de apoio à decisão médica;

2. o processo de avaliação de blocos de construção é sistemático e extensível a qualquer número de conjuntos de dados, classificadores e combinação de técnicas de segmentação de sinal;
3. apresentou-se em primeira mão (i) abordagem de três esquemas de segmentação de dados em três conjuntos de dados rotulados e dois não rotulados no tocante ao processo de construção dos modelo de prognóstico para avaliação do bem estar fetal; (ii) avaliação de seis classificadores em cenários baseados em engenharia de *features*, três classificadores em cenários baseados diretamente em biossinais (séries temporais) e um algoritmo semi-supervisionado para os esquemas de segmentação mencionados (iii) avaliação da capacidade de generalização por meio de *cross-dataset* (bidirecional) para três conjuntos de dados nas abordagens supervisionadas e semi-supervisionadas.

4 MATERIAIS E MÉTODOS

Este Capítulo apresenta os materiais e aspectos metodológicos utilizados nesta tese de forma global e comuns em diferentes cenários avaliativos.

4.1 Plataforma analítica

Nesta tese, foram empregados diferentes ambientes para execução dos experimentos baseados em engenharia de *features* e para execução dos baseados em biossinais (séries temporais) como segue. Os experimentos baseados em engenharia de *features* foram realizados em um sistema operacional Ubuntu Linux 20.04 LTS executando em um Intel©i7-7500U CPU 2.70GHz (4CPUs), 16 GB RAM, com *graphic processing unit* (GPU) NVIDIA ©GeForce 940MX com 4 GB de memória RAM dedicada.

A plataforma *Amazon Elastic Compute Cloud* (Amazon EC2) da Amazon Web Services (AWS) foi adotada nos experimentos baseados em biossinais (séries temporais) com alto custo computacional. Nela, um sistema operacional AWS Deep Learning Amazon Image (DLAMI) baseado no Ubuntu 18.04 rodando sobre uma instância AWS EC2 padrão *g4dn.2xlarge*. A instância possui processadores Intel R Xeon escaláveis (8 vCPU), 32 GB de memória RAM, e placa de vídeo NVIDIA T4 Tensor Core GPU com 16Gb de memória RAM dedicada.

O software foi desenvolvido na linguagem de programação Python e empregou bibliotecas para análise de dados de código aberto: Scikit-learn[1] 0.24, Scipy [2] 1.5.4, e Keras[3] 2.3.0 com suporte a processamento na GPU. Também empregamos pacotes baseados em Python para processamento dos biossinais: Hfda [4], Entropy [5], hrv-analysis [6] para tarefas de extração de *features* baseadas na variação cardiaca adulta, e a biblioteca imbalanced-learn [7] para métodos de balanceamento de classes.

[1] https://scikit-learn.org/
[2] https://www.scipy.org/
[3] https://keras.io/
[4] https://pypi.org/project/hfda/
[5] https://raphaelvallat.com/entropy/
[6] https://pypi.org/project/hrv-analysis/
[7] https://imbalanced-learn.org/stable/about.html

4.2 Descrição das bases de dados

Foram utilizadas cinco bases de dados contendo sinais biológicos nesta tese. Três destas bases com informações úteis para rotulagem foram empregadas nos experimentos com aprendizado de máquina supervisionado, CTU-UHB (CHUDACEK *et al.*, 2014), HUFA (BARQUERO-PEREZ *et al.*, 2017) e DB-TRIUM (MARQUES *et al.*, 2019). As duas bases de dados remanescentes não possuem qualquer informação sobre o desfecho fetal ou indicadores para sua rotulagem e foram empregadas nos experimentos da abordagem semi-supervisionada. São elas a base de dados SpAM (GEORGIEVA *et al.*, 2017) e a base de dados DB-HeraBeat. Esta última trata-se de uma base não anotada inédita que foi avaliada em primeira mão nos estudos desta tese.

As bases avaliadas contém diferentes quantidades de registros e foram adquiridas com diferentes tecnologias. Apesar da presença de sinais de FCF e UC nas bases de dados CTU-UHB, DB-TRIUM e SpAM, visando a interoperabilidade do modelo de prognóstico, apenas o sinal de FCF foi empregado em nossos experimentos.

4.2.1 Conjunto de dados CTU-UHB

A base de dados CTU-UHB consiste em um total de 552 registros intra-parto adquiridas na enfermaria obstétrica do Hospital Universitário em Brno, República Tcheca. Os sinais foram gravados com os monitores fetais STAN S21 e S31 e Avalon FM40 e FM50. Todas as gravações têm no máximo 90 minutos de duração e começam no máximo 90 minutos antes do parto. Cada registro fornece um sinal de frequência cardíaca fetal (FCF) e contrações uterinas (UC), ambas amostradas a 4 Hz. O conjunto de dados fornece informações objetivas úteis para a avaliação do status do resultado fetal e rotulagem das classes, como excesso de base (BE), valores de pH da artéria umbilical e déficit de base no líquido extracelular (BDecf). As pontuações do Apgar no 1º e no 5º minutos também são disponibilizadas. O CTU-UHB está disponível para acesso público.

4.2.2 Conjunto de dados HUFA

Para a base de dados HUFA, apenas o sinal FCF está disponível. Este biossinal foi adquirido com um cardiotocógrafo Philips no Hospital Universitario Fundación de Alcorón, Madrid, Espanha. Os sinais compreendem registros FCF amostrados a 4 Hz com duração

variável de até 4h antes do parto, totalizando 32 registros. O conjunto de dados também contém indicadores para avaliação do resultado fetal, como os valores de pH da artéria umbilical e o índice de Apgar 10 minutos após o parto. Estes dados também encontram-se disponíveis publicamente[8].

4.2.3 DB-TRIUM dataset

Para o conjunto de dados DB-TRIUM, sinais de FCF e UC com duração variável de até 3 horas, ambos amostrados a 4 Hz, foram adquiridos usando um sistema de cardiotocografia GE Corometrics Série 250CX na Klinikum rechts der Isar, Technische Universität München, Munich, Alemanha. Empregamos 159 registros contendo anotação prévia do estado fetal realizada por especialistas, que foi executada com apoio do Trium Analysis Online GmBH[9]. As amostras avaliadas do DB-TRIUM não fornecem informações sobre quais registros correspondem ao monitoramento anteparto e intraparto. O status de disponibilidade deste conjunto de dados é privado e os indicadores de resultado fetal como pH umbilical e scores de Apgar não estão disponíveis para esta base de dados.

4.2.4 Conjunto de dados SpAM

Esta base de dados provém do *CTG challenge* realizado no *Signal Processing and Monitoring (SPaM) in Labour workshop*[10]. Trata-se de um banco de dados de validação contendo registros intraparto com pelo menos uma hora de duração, gravidez unifetal (> 36 semanas), originadas de três hospitais diferentes do Reino Unido, França e República Tcheca. São 300 registro amostrados em 4 Hz com nível de qualidade do sinal alto, sendo que parte dos registros possui pH < 7,05 (sem indicação de quais registros). O conjunto de dados de desafio é fornecido como uma validação; Não há indicação de desfecho fetal ou rotulagem em classes.

4.2.5 Conjunto de dados DB-HeraBeat

A base de dados DB-HeraBeat contém registros de FCF anonimizados com duração variável de até 5 horas, amostrados a 1 Hz, os quais foram adquiridos no período de julho de 2019 a setembro de 2020 utilizando o dispositivo *HeraBEAT*[11] *personal ultrasound fetal*

[8] https://sites.google.com/site/hufahypoxia/
[9] https://www.trium.de/
[10] http://users.ox.ac.uk/~ndog0178/spam2017.htm
[11] https://herabeat.com/how-it-works/

doppler durante procedimentos de rotina para triagem clínica de atendimentos ambulatorial. O *HeraBEAT personal ultrasound fetal doppler* é um monitor de batimento cardíaco materno e fetal autoguiado, de baixo custo, sem fio, projetado para auto-administração a partir das 12 semanas de gestação e seus registros FCF mostraram-se equivalentes aos obtidos via cardiotocografia em ambiente clínico durante protocolos de avaliação para gestações de baixo risco (PORTER *et al.*, 2021). Foram selecionados sinais com duração mínima de 30 minutos, referentes a gravidez unifetal de baixo risco com idade gestacional (> 30 semanas), totalizando 1519 registros. Este conjunto de dados não possui indicações sobre rotulagem ou desfecho fetal para divisão em classes.

4.3 Pré-processamento de sinal

As medições em cenários reais estão sujeitas a erros e, para lidar com biossinais brutos, são necessárias tarefas de pré-processamento que garantam a qualidade desejada nos sinais (SUPRATAK *et al.*, 2016). Nesta tese, foi adotado um esquema de rejeição de artefato em três etapas, conforme sugerido por Georgoulas *et al.*(GEORGOULAS *et al.*, 2017) e Zhao *et al.*(ZHAO *et al.*, 2019), com adaptações necessárias para iteroperabilidade das nossas bases de dados conforme segue. Seja $x(i)$ uma amosta pertencente a um sinal biomédico X como o FCF em batimentos por minuto (bpm). Com X contendo N amostras, onde $i = 1,2,..N$ e N é o número total de amostras do sinal.

1. Foram removidas lacunas de sinal ou valores ausentes consecutivos com duração maior que cinco segundos.
2. Foram interpolados linearmente os valores de *outliers*, $x(i) \leq 50$ or $x(i) \geq 200$. Também foram interpolados linearmente as lacunas de sinal com duração menor que cinco segundos.
3. Quando a diferença entre $x(i)$ e quaisquer amostra adjacente $x(i-1)$ ou $x(i+1)$ excede 25 bpm, essa amostra foi substituída com a interpolação de Hermite.

A Figura 6 apresenta um segmento de sinal em sua forma original e seu correspondente pré-processado contendo os primeiros 30 minutos do registro ID = 1001 da base CTU-UHB.

Para garantir a compatibilidade entre bases de dados provenientes de diferentes métodos de aquisição, atividades de pré-processamento extras foram executados em bases específicas conforme segue:

1. na base de dados HUFA, devido a restrições impostas pela equipamento de monitora-

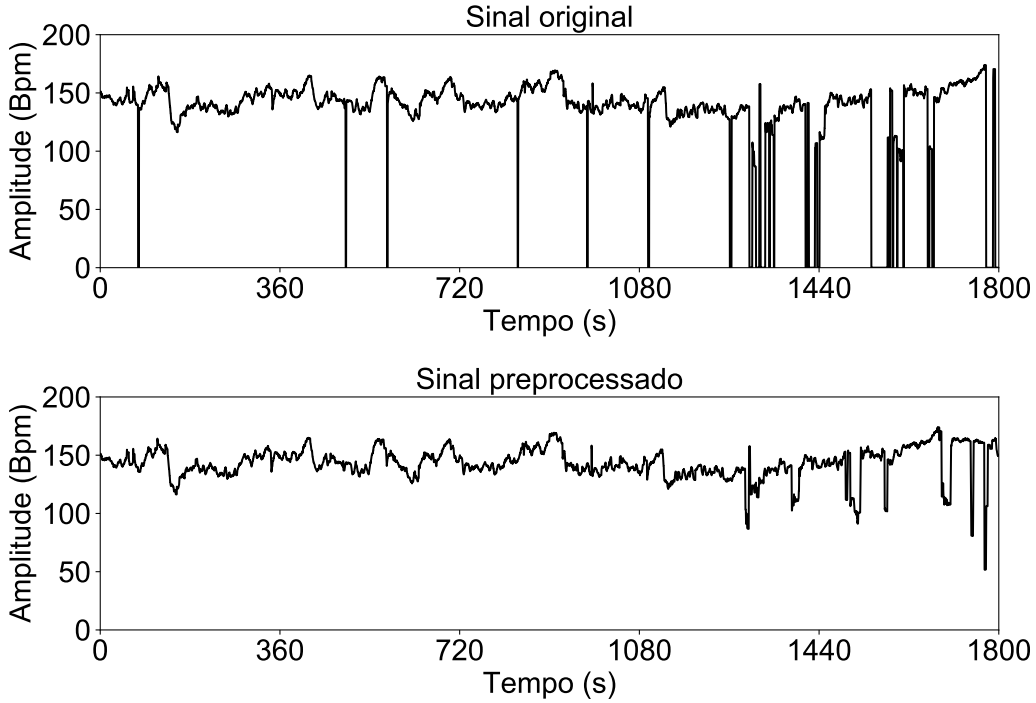


Figura 6 – Sinal FCF original e pré-processado. Fonte: Elaboração própria.

mento, a amplitude dos valores recuperados é de quatro vezes o valor estimado em bpm (BARQUERO-PEREZ *et al.*, 2017), desta forma os valores FCF foram multiplicados por 0,25 (ou divididos por 4) para obter o valores em bpm compatíveis com as bases de dados remanescentes.

2. para base DB-HeraBeat, os dados originais amostrados em 1 Hz foram reamostrados para 4 Hz aplicando a interpolação de Hermite.

As atividades de pré-processamento explanadas auxiliam a manutenção da qualidade de sinal e permitem uma avaliação objetiva do estado do feto ao utilizar uniformemente bases de dados retrospectivas de monitoramento fetal de origem e características distintas.

4.4 Segmentação do sinal

No ambiente clínico, a duração dos sinais FCF avaliados por especialistas varia entre 10 e 30 minutos para produzir uma interpretação consistente (ZHAO *et al.*, 2019). No entanto, a pesquisa biomédica apresenta diferentes critérios para o duração dos traçados de FCF adotados para análise computacional.

Estudos da literatura podem empregar critérios de seleção baseados na qualidade do sinal, os quais envolvem avaliação da duração mínima do segmento, taxa de perda e a quantidade

total de *outliers* presente em uma determinada janela de tempo (RICCIARDI *et al.*, 2020). Existem critérios baseados na análise de intervalos pré-estabelecidos de acordo com a duração dos sinais pertencentes a base de dados analisada.

Nestes cenários, as abordagens variam entre o uso do sinal em sua completude (*full-data*) com processamento contínuo em janelas fixas, por exemplo, com 20 minutos de duração (ZHAO *et al.*, 2018; LI *et al.*, 2019), cenários *full-data* com processamento de segmentos com 60 minutos de duração (BARQUERO-PEREZ *et al.*, 2017), o uso direto do esquema *full-data* em um processamento contínuo (FERGUS *et al.*, 2018; FERGUS *et al.*, 2017) e a separação do sinal pelo estágio do parto (estágio I e II) (ABRY *et al.*, 2018). Por fim, sob a perspectiva que as maiores variações no sinal de FCF ocorrem próximo a hora do parto, existem abordagens de segmentação que focam na proximidade do parto, como o uso dos últimos 15, 20 ou 30 minutos do sinal (COMERT *et al.*, 2018; ALSAGGAF *et al.*, 2020).

Nesta tese, foi empregada uma abordagem de segmentação baseada em intervalos de tempo para avaliar a capacidade do sinal em discriminar os registros saudáveis dos patológicos. No monitoramento de sinais intra-parto esta abordagem proporciona uma análise relativa a proximidade com a hora do nascimento e em monitoramentos de sinais anteparto a abordagem citada resulta na avaliação das amostras em relação a duração do sinal, sendo as mais recentes localizadas no final do mesmo. Optamos por uma segmentação com duração de 30 minutos do sinal da FCF, um total de 7200 amostras devido à frequência de amostragem em 4 Hz (ALSAGGAF *et al.*, 2020). Desta forma, para cada base de dados foram avaliadas três abordagens de segmentação de sinal.

1. análise dos primeiros 30 minutos de sinal;
2. análise dos últimos 30 minutos de sinal;
3. análise do sinal em sua totalidade (*full-data*).

A avaliação de diferentes esquemas de segmentação dos dados de monitoramento retrospectivos objetivou se aproximar de um cenário real em que o tempo restante até o parto é desconhecido (BARQUERO-PEREZ *et al.*, 2017). Para cada base de dados, derivamos três subconjuntos de dados que foram avaliados de forma independente.

Nos cenários baseados em engenharia de *features*, foram extraídas *features* para cada esquema de segmentação mencionado. Assim, para cada base de dados avaliado CTU, HUFA, DB-TRIUM, SpAM e DB-HeraBeat, foram criados três conjuntos de dados derivados extraindo as *features* do sinal nos primeiros 30 minutos, outro conjunto de dados extraindo os recursos do

sinal nos últimos 30 minutos e o último conjunto extraindo as *features* do sinal em sua totalidade (*full-data*). Este procedimento resultou em quinze conjuntos de dados baseados em *features*, três para cada dado original, os quais foram nomeados de CTU-first30, CTU-last30 e CTU-full para base CTU-UHB e foi adotado o mesmo procedimento para todas bases de dados.

Também foi realizada derivação nos esquemas *first30*, *full-data* e *last30* para os cenários baseados diretamente em biossinais (séries temporais). Sendo que o *first30* e *last30* se resumem na separação dos primeiros e últimos 30 minutos dos sinais para análise dos biossinais. No esquema de dados completos (*full-data*), o sinal completo foi segmentado em parcelas iguais com duração de 30 minutos cada e mantendo o rótulo original do sinal (normal ou patológico), desta forma, no esquema de séries temporais *full-data* cada registro de duração t em minutos produziu n novos registros individuais com 30 minutos de duração, onde $n = \frac{t}{30}$.

4.5 Separação do estado fetal em classes

Os registros das bases de dados de monitoramento fetal retrospectivos CTU-UBH, HUFA e DB-Trium foram rotulados em classes para uso nos algoritmos de classificação de aprendizado de máquina supervisionado. Estas bases de dados biomédicos provêem informações clínicas úteis para a rotulagem da classe e posterior reconhecimento dos novos dados. As bases de dados SpAM e DB-HeraBeat foram empregadas para aprendizagem semi-supervisionada e portanto, não possuem rotulagem ou indicadores pré-estabelecidos para este fim.

Os rótulo para classes dos dados de avaliação retrospectiva do estado fetal podem ser determinados com base na análise de medidas específicas coletadas após o nascimento, como o índice de Apgar, que representa o resultado da avaliação visual do recém-nascido no primeiro, quinto e décimo minutos e o peso ao nascer (BW), expresso em percentis para uma determinada população de recém-nascidos (CZABANSKI *et al.*, 2016). Existem também marcadores bioquímicos, como a medição do pH proveniente do sangue do cordão umbilical (pH), o excesso de base (BE) e os valores do déficit de base no líquido extracelular (BDecf) (ZHAO *et al.*, 2019). Uma abordagem alternativa é usar uma rotulagem de classe baseada na análise visual dos traçados FCF por especialistas (MARQUES *et al.*, 2019), na qual os especialistas costumam fazer uso de diretrizes bem estabelecidas, como os guias FIGO (AYRES-DE-CAMPOS *et al.*, 2015b).

O pH do sangue do cordão umbilical é um forte indicador de condições patológicas do feto, como hipoxemia e eventos neurológicos adversos (KUMAR *et al.*, 2016). Nos estudos

referentes a análise computadorizada da FCF, o valor do pH foi usado considerando diferentes valores, variando entre 7,05, 7,15 e 7,20. (ROTARIU *et al.*, 2014; GEORGOULAS *et al.*, 2017; COMERT *et al.*, 2018; ALSAGGAF *et al.*, 2020). Esses estudos empregam limites únicos (COMERT *et al.*, 2018) para abordagens de classificação binária (normal e patológica) ou adotam limites múltiplos (CZABANSKI *et al.*, 2016) para dividir as classes em classes normais, suspeitas e patológicas.

Tabela 14 – Distribuição das classes. Fonte: Elaboração própria.

Dados baseados em *features*		
Dados	Segmentação	Distribuição (N/P)
DB-Trium	First30/Last30/Full-data	(112/47)
CTU-UHB	First30/Last30/Full-data	(358/40)
HUFA	First30/Last30/Full-data	(14/13)
Dados baseados em biossinais		
DB-Trium	First30	(112/47)
	Last30	(112/47)
	Full-data	(584/506)
CTU-UHB	First30	(358/40)
	Last30	(358/40)
	Full-data	(530/53)
HUFA	First30	(14/13)
	Last30	(14/13)
	Full-data	(203/143)

N = normal, P = patológico

Nas abordagens para rotulagem dos registros em duas classes (binária) baseados nos valores do pH, os estudos podem adotar todo o conjunto de dados disponíveis utilizando um único valor como divisor entre as classes normal e patológica (COMERT; KOCAMAZ, 2018), incluindo os registros intermediários (suspeitos) nos rótulos normais ou patológicos dependendo do valor limite empregado (COMERT *et al.*, 2018). Por outro lado, abordagens de rotulagem binária baseadas no valor de pH podem se utilizar de múltiplos limites para identificar e excluir os registros intermediários (suspeitos), produzindo conjuntos de treinamento mais objetivos (FENG *et al.*, 2018; FUENTEALBA *et al.*, 2017; FUENTEALBA *et al.*, 2019a; FUENTEALBA *et al.*, 2019b; PETROZZIELLO *et al.*, 2019).

Nesta tese, foi utilizada uma abordagem de rotulagem binária em classes normal e patológica. Para as bases de dados CTU-UHB e HUFA, foi adotado o valor de pH como critério objetivo com 7,05 e 7,20 para os limiares inferior e superior, respectivamente (FUENTEALBA *et al.*, 2019a; FUENTEALBA *et al.*, 2019b). Os registros com (pH < 7,05) foram rotulados como

patológicos, enquanto (pH > 7,20) representam condições normais de saúde fetal. Desta forma, os registros pertencentes a valores intermediários (pH > 7,05 e pH < 7,20) ou suspeitos foram identificados e removidos das bases com rotulagem baseada em pH. A base de dados DB-Trium forneceu rótulos pré-estabelecidos baseados na avaliação de especialistas para a separação do estado fetal. A Tabela 14 apresenta a distribuição final dos rótulos por classe para cada esquema de segmentação de sinal. Para os conjuntos de dados de série temporal, os esquema *full-data* apresenta uma quantidade maior dos registro devido à divisão do sinal em segmentos de igual duração.

4.6 Algoritmos

Algoritmos de aprendizado de máquina tem se mostrado eficazes na solução de problemas do domínio biomédico (FERGUS *et al.*, 2018; TANG *et al.*, 2018; TAN *et al.*, 2018; RICCIARDI *et al.*, 2020; ALSAGGAF *et al.*, 2020; MARQUES *et al.*, 2020). Nas abordagens de engenharia de *features* e na de séries temporais, um conjunto de algoritmos de classificação binários supervisionados (duas classes) que operam de maneiras diferentes foram utilizados nesta tese para discriminar o status do feto em classes normais e patológicas. Também fez-se uso de um algoritmo semi-supervisionado para complementar o processo avaliativo.

Foram incluídos no processo de avaliação algoritmos que diferem em metodologia e abordagens para separação das classes.

(i) o SVM é não probabilístico e baseia-se em hiperplanos;
(ii) o KNN é caracterizado como *lazy learning* e baseado em distâncias;
(iii) o GTB, BG e RF são *ensemble methods* que adotam árvores de decisão, diferindo na forma em que as predições individuais são combinadas para produzir a classificação final;
(iv) o MLP baseia-se na aproximação de funções não lineares;
(v) os classificadores CNN, CNN-LSTM e multi-head CNN são baseados em convoluções para extração de informações representativas dos dados;
(vi) o LabelSpreading é um algoritmo baseado na difusão dos rótulos das classes ao longo de um grafo.

Na abordagem supervisionada baseada em *features*, foram empregados: *Support Vector Machine* (SVM) (CORTES; VAPNIK, 1995), *K-Nearest Neighbors* (KNN) (ALTMAN, 1992), *Randon Forest classifier* (RF) (BREIMAN, 2001), *Gradient Tree Boosting* (GTB) (FRIEDMAN, 2002; KE *et al.*, 2017), *Multi-layer perceptron* (MLP) (RUMELHART *et al.*, 1986)

e o classificador *Bagging* (BG) (BREIMAN, 1996). Fez-se ainda uso do algoritmo *Recursive feature elimination* (RFE) (GUYON *et al.*, 2002) para seleção de *features*.

Em relação ao cenário semi-supervisionado, foi empregado o algoritmo LabelSpreading (ZHOU *et al.*, 2004) para rotulagem das bases de dados DB-HeraBeat e SpAM.

Para resolver os problemas relacionados ao desequilíbrio na distribuição entre as classes normal e patológica, fez-se uso do algoritmo *Synthetic Minority Oversampling Technique* (SMOTE) (BOWYER *et al.*, 2002) em que a classe de quantidade minoritária foi aumentada para equilibrar a distribuição de classes saudáveis e patológicas. O SMOTE tem sido aplicado com sucesso na pesquisa médica para lidar com conjuntos de dados não balanceados (FERGUS *et al.*, 2017; FERGUS *et al.*, 2018; RICCIARDI *et al.*, 2020). Nesta tese, o conjunto de treinamento sofreu *oversampling* isoladamente para manter o conjunto de teste confiável e evitar viés durante a avaliação.

Nos cenários de avaliação baseados diretamente nos biossinais (séries temporais), foram adotados três algoritmos de classificação. Primeiro, redes neurais convolucionais de uma dimensão (CNN) (LECUN *et al.*, 1998), segundo, uma variação de CNN chamada *multi-head* CNN, e por fim, um modelo híbrido composto por CNN e camadas *Long short-term memory* (HOCHREITER; SCHMIDHUBER, 1997) (CNN-LSTM)(TAN *et al.*, 2018).

Os modelos baseados em redes, como MLP, CNN, multi-head CNN e CNN-LSTM não possuem uma estrutura padrão pronta para uso e, portanto, a arquitetura dos modelos foi definida em um processo de *grid-search*. As estruturas arquiteturais resultantes são descritas nas subsessões a seguir.

4.6.1 Estrutura da MLP

A estrutura do modelo MLP consiste em uma camada de entrada e duas camadas ocultas totalmente conectadas (fc), seguidas por uma camada de saída conforme apresentado na Figura 7. Foram empregados trinta e oito neurônios na camada de entrada quando usamos o conjunto completo de *features* e seis neurônios quando avaliou-se o esquema de seleção de *features*. Cada neurônio de entrada representa uma *feature*.

Utilizou-se um número fixo de 76 neurônios na primeira e na segunda camadas ocultas totalmente conectadas. Finalmente, um neurônio na camada de saída representa a saída como normal (N) ou patológica (P) para o rótulo de classe estimado. Fez-se uso da função de ativação Relu em camadas ocultas e uma função de ativação sigmoide na camada de saída. Por

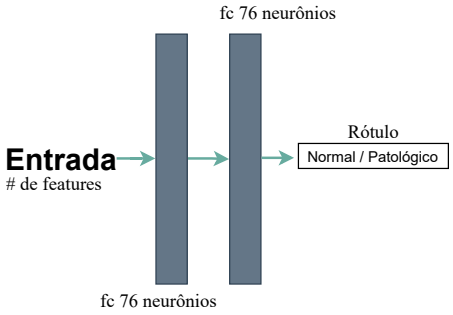


Figura 7 – *Arquitetura Multi-layer perceptron.* Fonte: Elaboração própria

tratar-se de uma abordagem de classificação binária, foi aplicado o *binary-cross-entropy* como função de avaliação.

4.6.2 Estrutura CNN

Nesta tese, foi projetada a estrutura do modelo CNN apresentada na Figura 8 para classificar dados unidimensionais (1-D) como o sinal FCF, o qual é processado em intervalos de 30 minutos de duração ou (1 x 7200 amostras de sinal 1-D) a 4 Hz. A arquitetura contém quatro

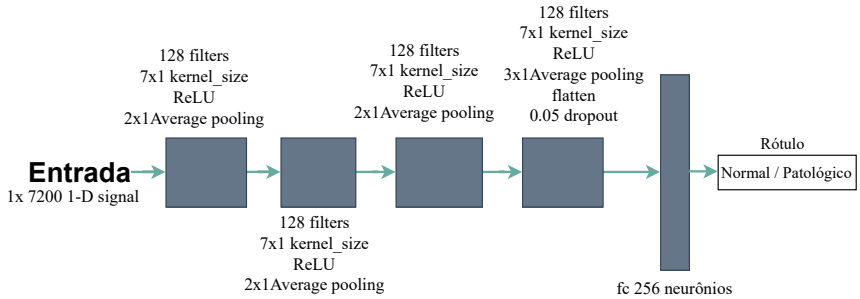


Figura 8 – Estrutura CNN. Fonte: Elaboração própria.

blocos de camadas convolucionais acrescidas de camadas de *pooling* e finalmente, uma camada totalmente conectada.

4.6.3 Estrutura Multi-head CNN

Em arquiteturas *multi-head*, cada série de entrada é tratada por um modelo separado (*head*), e a saída de cada um desses *heads* é combinada para que seja realizada a classificação (KAUSHIK *et al.*, 2020). O modelo *multi-headed* CNN processa a mesma série FCF utilizando-se de *kernel* de diferentes tamanhos para cada *head*. Foi projetada uma arquitetura *two-headed*, conforme apresentado na Figura 9.

Cada *head* é composta por um bloco de camada convolucional acrescido de uma

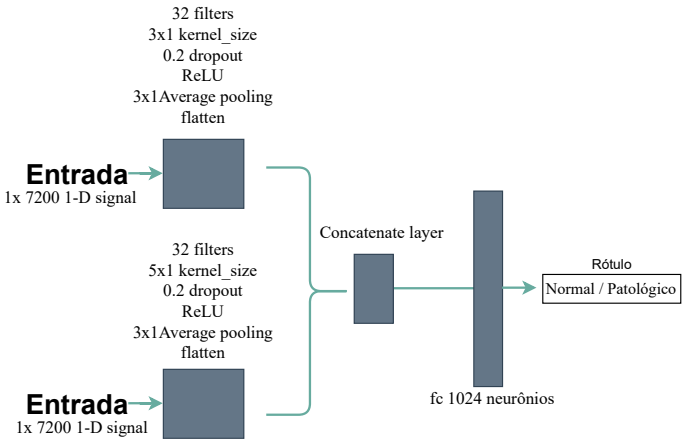


Figura 9 – Estrutura multi-head CNN. Fonte: Elaboração própria.

camada *average pooling*. As interpretações de cada *head* são concatenadas e interpretadas por uma camada totalmente conectada.

4.6.4 Estrutura CNN-LSTM

Os modelos CNN-LSTM usam o CNN para extrair informações proeminentes do sinal e aplicar camadas *long short-term memory* (LSTM) para um aprendizado sequencial e assim produzir a classificação (ULLAH *et al.*, 2020). A Figura 10 apresenta a estrutura utilizada.

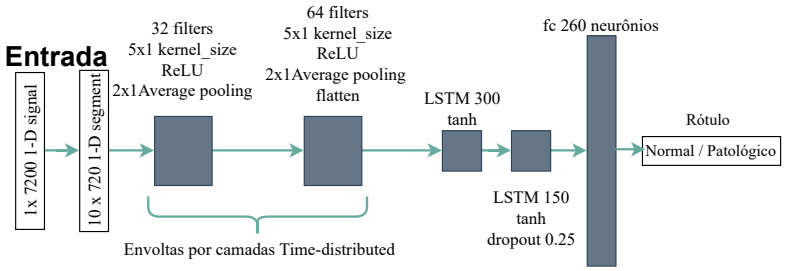


Figura 10 – Estrutura CNN-LSTM. Fonte: Elaboração própria.

Para o modelo CNN-LSTM, cada sinal de entrada 1 x 7200 1-D foi dividido em 10 x 720 segmentos 1-D. A arquitetura do modelo contém dois blocos de camadas convolucionais acrescidos de uma camada *pooling*. Esses blocos foram gerenciados por camadas *TimeDistributed* que permitem o processamento de cada sub-divisão de segmento em um bloco contínuo. Em seguida, os resultados foram processados por camadas LSTM para classificação de sequências. Finalmente, a saída das camadas LSTM foi processada em uma camada totalmente conectada para estimar o rótulo da classe.

4.7 Síntese do Capítulo

Este Capítulo apresentou uma visão geral das técnicas empregadas ao longo desta tese. Uma visão detalhada do desenho dos experimentos e cenários avaliativos pode ser encontrada nas seções referente aos métodos em cada Capítulo subsequente.

5 MODELAGEM SUPERVISIONADA BASEADA EM BIOSSINAIS

Neste Capítulo é apresentada a construção de modelos nos quais seus principais blocos de construção tem como base o uso direto dos biossinais em forma de séries temporais. Primeiramente, os materiais e métodos empregados na avaliação dos blocos de construção para estes modelos são apresentados, em seguida, a avaliação de um conjunto de classificadores em diferentes esquemas de segmentação de dados é exibida. É realizada ainda uma avaliação referente a capacidade de generalização das melhores combinações obtidas. São apresentados resultados referentes ao reconhecimento dos rótulos baseados em pH com valores pertencentes originalmente a categoria suspeita (*grey zone*) no esquema de classificação binário. Por fim, os resultados em relação a combinação mais eficiente dos blocos de construção para avaliação do bem-estar fetal nestes modelos são sumarizados.

A Figura 11 apresenta em destaque o esquema empregado neste Capítulo em relação aos modelos desenvolvidos nesta Tese.

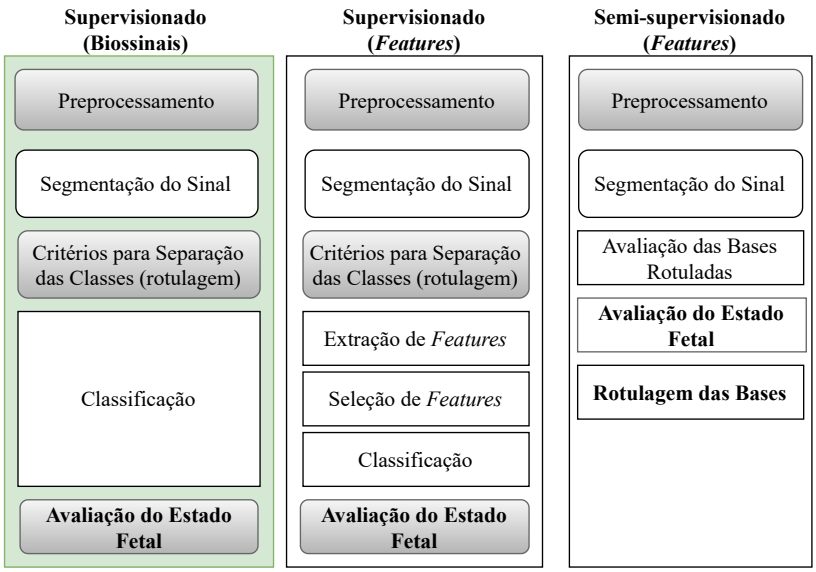
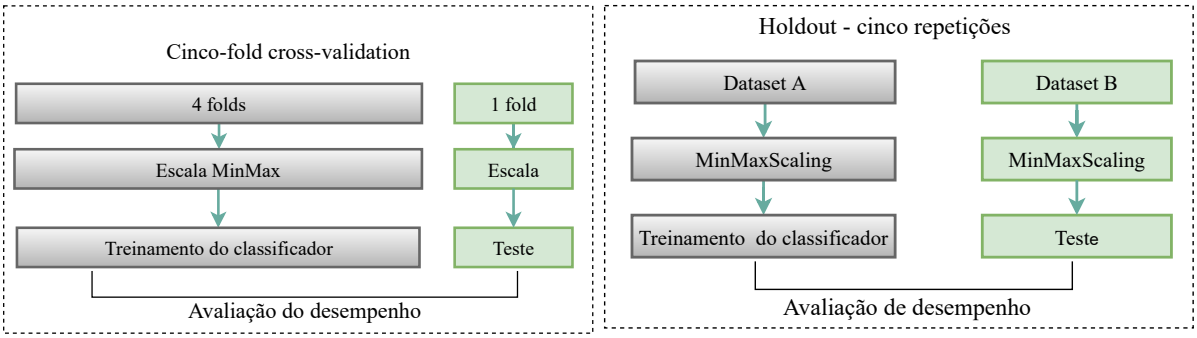


Figura 11 – Modelos supervisionados baseados em biossinais (em destaque). Fonte: Elaboração própria.

5.1 Métodos para Modelos Baseados em Biossinais

Foi utilizado o *k-fold cross-validation* (CV) estratificado com k = 5 para avaliação dos esquemas de segmentação em cada uma das bases de dados isoladamente para os modelos baseado em séries temporais. Para otimizar o aprendizado da rede, os valores das amostras

foram reduzidos dentro de um pequeno intervalo especificado [0,1], para isto foi usada a escala *Min-Max* apenas nos dados de treinamento para posterior aplicação na parcela de teste dos dados, evitando viés na avaliação. Os modelos baseados em redes que aplicam aprendizado diretamente dos biossinais (séries temporais) como os de *deep learning* necessitam de uma grande quantidade de dados para serem eficazes. Neste trabalho, este problema foi minimizado empregando valores de *batch-sizes* pequenos variando entre 8 e 16 nas etapas de treinamento dos modelos.



(a) Cross-validation (single dataset) (b) Holdout (cross-datasets)

Figura 12 – Modelos de séries temporais com bases de dados únicas e base de dados cruzadas. Fonte: Elaboração própria.

Para avaliar o grau de generalização dos modelos, foi adotada a validação cruzada entre bases de dados (*cross-dataset*), onde um classificador foi treinado em uma combinação de blocos de construção para um determinado conjunto de dados e seu desempenho foi validado em outra base de dados completa. Foi empregada a técnica *holdout* e o processo de treinamento e testes foi repetido por cinco vezes. A base de dados de treinamento tambem foi escalada aplicando o sistema de normalização *Min-Max* e a escala resultante foi aplicada na base de testes. A Figura 12 apresenta uma visão geral da avaliação para as séries temporais em dois conjuntos de dados empregando a validação *cross-dataset*.

O *cross-dataset* foi adaptado para avaliação dos intervalos de pH categorizados como suspeitos (*grey zone*) de forma que, apenas nos dados com rotulagem baseados no valor do pH (CTU-UHB e HUFA), o conjunto de validação foi representado pelos registros da mesma base pertencentes a área intermediária, os quais haviam sido removidos durante o processo inicial de rotulagem destas bases (ver Seção 4.5). Desta forma, na avaliação *cross-dataset* da *grey zone*, a base com rotulagem binária foi empregada para treinamento e os registros marcados como suspeitos da mesma base foram usados para validação a fim de verificar a qual o rótulo seria

atribuído (normal e patológico) em cada registro originalmente rotulado como suspeito.

5.2 Resultados de Modelos Baseados em Biossinais

Nos cenários baseados diretamente em biossinais (séries temporais), foram avaliados três classificadores: CNN, Multi-head CNN, e CNN-LSTM para cada esquema de segmentação de dados nas bases rotuladas DB-TRIUM, CTU-UHB e HUFA.

A Tabela 15 apresenta o desempenho dos classificadores no esquema de segmentação *first30*. Para a base CTU-UHB, todos os classificadores apresentaram valores de desempenho altos para acurácia (ACC) e especificidade (SP), com valores baixos para sensibilidade (SE). O área sob a curva ROC (AUC) teve um desempenho similar a classificação aleatória para o CNN e CNN-LSTM. O Multi-head CNN (M-CNN) obteve um desempenho para a AUC levemente melhor do que o classificador aleatório.

A avaliação dos esquemas de segmentação na base de dados HUFA apresentou baixo desempenho para todos os classificadores e, no DB-TRIUM, o classificador M-CNN alcançou a AUC um pouco melhor do que o classificador aleatório ($> 50\%$) com valores altos para SP e baixo SE. O desempenho geral dos classificadores indicam uma capacidade fraca em discriminar classes patológicas (valores do SE) ao utilizar os biossinais diretamente e de forma não balanceada.

Tabela 15 – Cenários *first30* - CTU-UHB, HUFA e DB-TRIUM. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
CNN	0.818 (0.043)	0.887 (0.056)	0.200 (0.100)	0.406 (0.086)	0.543 (0.031)
M-CNN	0.876 (0.021)	0.955 (0.027)	0.175 (0.127)	0.354 (0.199)	0.565 (0.058)
CNN-L	0.831 (0.037)	0.912 (0.046)	0.050 (0.061)	0.132 (0.162)	0.484 (0.019)
CNN	0.326 (0.116)	0.066 (0.13)	0.566 (0.133)	0.094 (0.188)	0.316 (0.110)
M-CNN	0.406 (0.137)	0.300 (0.266)	0.466 (0.266)	0.209 (0.259)	0.383 (0.145)
CNN-L	0.560 (0.135)	0.533 (0.244)	0.566 (0.326)	0.432 (0.239)	0.549 (0.113)
CNN	0.579 (0.080)	0.698 (0.112)	0.295 (0.112)	0.444 (0.084)	0.496 (0.075)
M-CNN	**0.685** (0.039)	**0.839** (0.046)	**0.313** (0.135)	**0.497** (0.116)	**0.576** (0.066)
CNN-L	0.629 (0.064)	0.776 (0.065)	0.275 (0.104)	0.452 (0.101)	0.525 (0.067)

Topo: CTU-UHB, Meio: HUFA, Base: DB-TRIUM. Valores em destaque correspondem aos de melhor desempenho.

Os resultados da avaliação dos classificadores aplicados ao esquema de dados completos (*full-data*) são apresentados na Tabela 16. A base CTU-UHB manteve alto ACC e SP com baixos valores para SE e GM. Em relação a AUC, obteve-se o desempenho próximo ao aleatório. Para HUFA e DB-TRIUM, os classificadores alcançaram resultados médios-baixos para SE, SP

e GM. Lá, os classificadores CNN e CNN-L obtiveram AUC > 50%.

Tabela 16 – Cenários *full-data* - CTU-UHB, HUFA e DB-TRIUM. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
CNN	0.855 (0.023)	0.920 (0.020)	0.209 (0.123)	0.389 (0.203)	0.564 (0.063)
M-CNN	0.885 (0.020)	0.958 (0.194)	0.149 (0.121)	0.325 (0.193)	0.553 (0.063)
CNN-L	**0.891** (0.013)	**0.956** (0.017)	**0.250** (0.141)	**0.470** (0.131)	**0.603** (0.066)
CNN	0.529 (0.051)	0.625 (0.059)	0.392 (0.104)	0.489 (0.070)	0.508 (0.055)
M-CNN	0.526 (0.052)	0.654 (0.079)	0.343 (0.059)	0.471 (0.052)	0.499 (0.049)
CNN-L	0.537 (0.056)	0.634 (0.131)	0.397 (0.076)	0.492 (0.041)	0.516 (0.042)
CNN	0.541 (0.019)	0.652 (0.174)	0.412 (0.214)	0.430 (0.216)	0.532 (0.027)
M-CNN	0.485 (0.018)	0.482 (0.017)	0.484 (0.035)	0.483 (0.017)	0.485 (0.017)
CNN-L	0.538 (0.016)	0.748 (0.209)	0.296 (0.242)	0.320 (0.261)	0.522 (0.022)

Topo: CTU-UHB, Meio: HUFA, Base: DB-TRIUM. Valores em destaque representam os melhores desempenhos.

Na Tabela 17 são apresentados os resultados do esquema *last30* para o CTU-UHB, HUFA e DB-TRIUM. Este esquema obteve os melhores desempenhos gerais para as três bases de dados avaliadas em comparação aos esquemas de segmentação remanescentes. O CTU-UHB apresentou escores de SE baixos, com AUC atuando próximo ao padrão aleatório. A base de dados HUFA apresentou equilíbrio nos valores obtidos para SE, SP e GM com AUC melhor do que o aleatório (> 50%). No DB-TRIUM, todos os classificadores obtiveram melhor equilíbrio entre SP e SE do que nas outras bases de dados. O classificador CNN foi o de melhor desempenho para os três conjuntos de dados ao se empregar as métricas GM e AUC como critérios de avaliação.

Tabela 17 – Esquema *last30* - CTU-UHB, HUFA e DB-TRIUM

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
CNN	0.836 (0.031)	0.916 (0.030)	0.125 (0.079)	0.299 (0.0159)	0.520 (0.046)
M-CNN	0.894 (0.009)	0.977 (0.006)	0.150 (0.093)	0.337 (0.180)	0.563 (0.046)
CNN-L	0.861 (0.016)	0.930 (0.022)	0.250 (0.111)	0.468 (0.108)	0.590 (0.050)
CNN	0.586 (0.097)	0.700 (0.266)	0.533 (0.339)	0.479 (0.242)	0.616 (0.066)
M-CNN	0.513 (0.165)	0.433 (0.294)	0.533 (0.339)	0.343 (0.286)	0.483 (0.169)
CNN-L	0.659 (0.215)	0.666 (0.182)	0.599 (0.388)	0.551 (0.327)	0.633 (0.227)
CNN	**0.672** (0.114)	**0.713** (0.120)	**0.582** (0.180)	**0.636** (0.131)	**0.648** (0.122)
M-CNN	0.711 (0.085)	0.856 (0.083)	0.364 (0.096)	0.556 (0.091)	0.610 (0.082)
CNN-L	0.698 (0.066)	0.821 (0.079)	0.402 (0.095)	0.570 (0.083)	0.611 (0.069)

Topo: CTU-UHB, Meio: HUFA, Base: DB-TRIUM. Valores em destaque correspondem aos melhores desempenhos.

5.2.0.1 Comparativo inter-esquemas de segmentação

As rodadas de cada *k-fold* para cada classificador emparelhado com sua contraparte entre dois esquemas de segmentação distintos, foram usadas para comparar seus valores de AUC. As avaliações de série temporal empregaram *5-fold cross-validation*, resultando em cinco estimativas de desempenho para cada classificador. A Tabela estatística de Wilcoxon fornece um nível de confiança $\alpha = 0.10$ para n = 5 (correspondendo as rodadas *k-fold* emparelhadas) com o valor crítico exato para hipótese *two-tailed* $T_{wc} = 0$. As comparações entre esquemas para dois valores de AUC são estatisticamente significativas se $T_w \leq T_{wc}$.

Tabela 18 – Comparações inter-esquema - Wilcox ($n = 5$, 90% CI). Fonte: Elaboração própria.

Algoritmo	CTU-UHB					
	AUC - média					
	A	B	C	T_wAB	T_wAC	T_wBC
CNN	0.543	0.564	0.520	5.00	6.00	**0.00**
M-CNN	0.565	0.553	0.563	5.00	5.00	7.00
CNN-L	0.484	0.603	0.590	**0.00**	**0.00**	6.00
Algoritmo	HUFA					
	AUC - média					
	A	B	C	T_wAB	T_wAC	T_wBC
CNN	0.316	0.508	0.616	**0.00**	**0.00**	1.00
M-CNN	0.383	0.499	0.483	1.00	1.50	5.00
CNN-L	0.549	0.516	0.633	6.00	3.50	5.00
Algoritmo	DB-TRIUM					
	AUC - média					
	A	B	C	T_wAB	T_wAC	T_wBC
CNN	0.496	0.532	0.648	6.00	1.00	1.00
M-CNN	0.576	0.485	0.610	1.00	6.00	**0.00**
CNN-L	0.525	0.522	0.611	7.00	3.00	1.00

Esquemas: A (*first30*), B (*full-data*), C (*last30*). Valores em destaque correspondem a diferenças estatísticamente significantes ($T_w \leq T_wc$)

A Tabela 18 apresenta a comparação entre esquemas por AUC para cada classificador e sua contraparte nos esquemas *first30*, *full-data* e *last30*: A, B e C, respectivamente. Não há diferença estatística entre esquemas na base CTU-UHB, exceto para CNN-L em (T_wAB) e (T_wAC) e para CNN em (T_wBC). Na base HUFA, não há diferença estatística entre os esquemas *full-data* e *last30* (T_wBC) para todos os classificadores. No entanto, este conjunto de dados apresenta diferença estatística apenas para CNN ao comparar os *first30* e *full-data* (T_wAB) e ao comparar os esquemas *first30* e *last30* (T_wAC). No conjunto de dados DB-TRIUM, não há diferença significativa entre os esquemas *firts30* e *full-data* (T_wAB) e entre o *first30* e *last30* (T_wAC) para todos os classificadores. Por outro lado, o M-CNN apresentou diferença significativa entre os esquemas *full-data* e *last30* (T_wBC).

5.2.0.2 Comparativo de desempenho dos classificadores

Adotou-se o teste estatístico de Friedman (FRIEDMAN, 1937) para comparar os valores de AUC dos classificadores em todos os conjuntos de dados em diferentes esquemas de segmentação. Foram incluídos nesta comparação os esquemas de segmentação *last30* e de dados completos (*full-data*) por terem apresentado os melhores desempenhos.

Utilizando um nível de confiança de 95%, com o valor crítico exato da Tabela estatística de Friedman $F_c = 7,00$ para hipótese *two-tailed* com k = 3 classificadores e n = 6 (correspondendo aos esquemas *full-data* e *last30* para todos os três conjuntos de dados), o teste resulta em $chi_F^2 = 8,52$. Como $chi_F^2 > F_c$, as avaliações indicam que pelo menos um par de AUC entre os classificadores obteve diferenças estatisticamente significativas. Desta forma, foi realizada uma análise *post-hoc* para identificar estes pares aplicando o teste de Nemenyi.

Na análise *post-hoc* de Nemenyi (NEMENYI, 1962) com um nível de confiança de 95 %, k = 3 classificadores e n = 6, obteve-se uma diferença crítica de hipótese *two-tailed* CD = 2,73. Assim, a comparação de um par de desempenhos dos classificadores é estatisticamente significativo se a diferença absoluta de seus ranques médios correspondentes for maior que 2,73, conforme mostrado na Tabela 19.

Tabela 19 – Análise post-hoc para o teste de Friedman. Fonte: Elaboração própria.

Dados	AUC - média		
	CNN	MULTI-CNN	CNN-LSTM
D1-B	0.564	0.553	0.603
D2-B	0.508	0.499	0.516
D3-B	0.532	0.485	0.522
D1-C	0.520	0.563	0.590
D2-C	0.616	0.483	0.633
D3-C	0.648	0.610	0.611
média	0.565	0.532	0.579
Algoritmo	Comparações de desempenho para todo o conjunto de dados		
	CNN	MULTI-CNN	CNN-LSTM
CNN	-	-	-
MULTI-CNN	**10.39**	-	-
CNN-LSTM	**5.19**	**15.58**	-

Bases de dados: D1 (CTU-UHB), D2 (HUFA) e D3 (DB-TRIUM). Esquemas: B (*full-data*), C (*last30*). Valores em destaque correspondem a diferenças estatisticamente significantes (diferenças absoluta entre a soma dos ranques > 2.73)

As comparações mostram que os desempenhos dos classificadores CNN, MULTI-CNN e CNN-LSTM diferem estatisticamente.

5.2.0.3 Avaliação em bases cruzadas para cenários baseados em biossinais

Foi avaliada a capacidade de generalização dos modelos baseados no uso direto dos biossinais treinando os classificadores nos conjuntos de dados com maior número de registros, o CTU-UHB e DB-TRIUM, e validando nas bases remanescentes. Devido ao seu melhor desempenho nos experimentos com as bases de dados isoladas, optou-se por utilizar o esquema *last30* na avaliação da generalização. Devido à pouca quantidade de registros, a base de dados HUFA foi usada apenas para validação nestes cenários avaliativos. A Tabela 20 apresenta os resultados da avaliação *cross-dataset* para os classificadores baseados diretamente em biossinais (séries temporais).

Tabela 20 – Avaliação *cross-dataset* de séries temporais - Fonte: Elaboração própria.

Trein./Aval.	Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensitividade mean (std)	Gmean mean (std)	AUC mean (std)
I/III	CNN	0.518 (0.033)	0.938 (0.030)	0.128 (0.053)	0.339 (0.073)	0.533 (0.032)
	Multi-CNN	0.555 (0.000)	1.000 (0.000)	0.142 (0.000)	0.377 (0.000)	0.571 (0.000)
	CNN-LSTM	0.533 (0.018)	0.984 (0.030)	0.114 (0.057)	0.324 (0.073)	0.549 (0.016)
I/II	CNN	0.700 (0.020)	0.939 (0.030)	0.131 (0.045)	0.345 (0.061)	0.535 (0.022)
	Multi-CNN	0.725 (0.007)	0.962 (0.011)	0.161 (0.010)	0.394 (0.011)	0.562 (0.005)
	CNN-LSTM	0.701 (0.006)	0.944 (0.025)	0.123 (0.051)	0.333 (0.064)	0.534 (0.014)
II/III	CNN	**0.607** (0.055)	**0.800** (0.115)	**0.428** (0.110)	**0.574** (0.077)	**0.614** (0.055)
	Multi-CNN	0.592 (0.000)	0.923 (0.110)	0.285 (0.000)	0.513 (0.000)	0.604 (0.000)
	CNN-LSTM	0.518 (0.052)	0.861 (0.057)	0.200 (0.083)	0404 (0.096)	0.530 (0.051)
II/I	CNN	0.635 (0.020)	0.657 (0.023)	0.440 (0.025)	0.537 (0.017)	0.548 (0.016)
	Multi-CNN	**0.782** (0.015)	**0.827** (0.017)	**0.380** (0.059)	**0.558** (0.042)	**0.603** (0.029)
	CNN-LSTM	0.728 (0.024)	0.761 (0.025)	0.434 (0.056)	0.574 (0.040)	0.598 (0.033)

I = CTU-UHB, II = DB-Trium, III = HUFA. Valores em destaque correspondem aos de melhor desempenho.

A adoção do conjunto de dados CTU-UHB como um conjunto de treinamento resultou em valores altos para SP e baixo para SE, com valores de AUC melhores do que o aleatório ao validarmos no DB-TRIUM e HUFA. Lá, o Multi-CNN foi o classificador de melhor desempenho.

Ao empregar a base de dados DB-TRIUM como um conjunto de treinamento em bases cruzadas, alguns pontos devem ser considerados. A base de dados DB-TRIUM fez uso de rotulagem de classe baseada em especialistas para designar os registros normais e patológicos. Então, ao validar os conjuntos de dados HUFA e CTU-UHB baseados em marcadores bioquímicos de pH, os resultados mostraram uma alta capacidade de reconhecer o registro de normais (SP) e capacidade média-baixa para discriminar os patológicos (SE) com o AUC tendo um desempenho melhor do que o aleatório. Os classificadores CNN e Multi-CNN foram os de melhor desempenho nesses cenários. Assim, esses resultados indicaram um potencial

discriminativo generalizável na rotulagem com base em especialistas, mesmo quando aplicada na avaliação de dados retrospectivos rotulados com base em marcadores bioquímicos.

5.2.0.4 Avaliação dos registros pertencentes a área cinza

Nesta tese, objetivando a generalização dos modelos, o processo de rotulagem baseado nos valores de pH empregou limites inferiores (< 7,05) e superiores (> 7,20) para caracterizar os registros em classes normais e patológicas, desta forma, os registros com valores intermediários ($\geq 7,05$ e $\leq 7,20$) que se referem a categoria suspeita ou *grey zone*, foram removidos dos cenários de avaliação binária (ver seção 4.5).

Verificou-se quais rótulos foram atribuídos a registros pertencentes a *grey zone* quando avaliados por modelos treinados nas categorias normal e patológico. A Tabela 21 apresenta detalhes da distribuição dos rótulos estimados no conjunto de registros referentes a *grey zone*. O classificador CNN e o cenário de segmentação de dados *last30* foram empregados na estimativa dos rótulos por terem obtido os melhores desempenhos em avaliações individuais.

Tabela 21 – Classificação da *grey zone*. Fonte: Elaboração própria.

CTU-UHB (**154 registros**)	
Algoritmo	Distribuição das classes estimadas
CNN	(N = 135, P = 19) N = 55 (pH < 7.15), N = 80 (pH $\geq$ 7.15) P = 10 (pH < 7.15), P = 9 (pH $\geq$ 7.15)
HUFA (**5 registros**)	
Algoritmo	Distribuição das classes estimadas
CNN	(N = 3, P = 2) N = 1 (pH < 7.15), N = 2 (pH $\geq$ 7.15) P = 2 (pH < 7.15), P = 0 (pH $\geq$ 7.15)

N = normal, P = patológico.

A literatura emprega o valor de pH = 7,15 para separação das classes normais e patológicas com limiar único, onde os valores de pH < 7,15 são marcados como patológicos e valores $\geq 7,15$ são rotulados como normais (COMERT; KOCAMAZ, 2018; COMERT *et al.*, 2018). A base CTU-UHB apresentou 154 (cento e cinquenta e quatro) registros marcados originalmente como suspeitos (*grey zone*), os quais foram estimados pelo modelo binário em 135 registros normais e 19 registros patológicos (N = 135, P = 19). Avaliando-se o limiar de pH = 7,15 nos grupos estimados, observou-se que entre os registros rotulados como normais, 55 apresentaram valores de pH < 7,15 e 80 apresentaram pH $\geq$ 7,15. Entre os registros estimados como patológicos, 10 apresentaram valores de pH < 7,15 e 9 apresentaram pH $\geq$ 7,15.

A base de dados HUFA apresentou 5 (cinco) registros marcados como suspeitos, os quais foram estimados como 3 normais e 2 patológicos (N = 3, P = 2). Na análise pelo limiar de pH = 7,15 observou-se que entre os registros rotulados como normais, 1 apresentou valores de pH < 7,15 e 2 (dois) apresentaram pH $\geq$ 7,15. Os registros rotulados como patológicos se dividiram em 2 (dois) com pH < 7,15 e 0 (zero) com valor de pH $\geq$ 7,15.

De forma geral, a avaliação dos rótulos suspeitos na base CTU-UHB apresentou 64 (55 + 9) estimativas incorretas e 90 (80 + 10) estimativas corretas. Na base de dados HUFA, foram encontrados 1 estimativa incorreta e 4 (quatro) estimativas corretas. Este valores mostraram-se promissores na análise dos rótulos estimados para *grey zone* quando empregado o limiar de pH = 7,15 para o grau de certeza.

5.3 Síntese do Capítulo

Os esquemas de segmentação de dados *first30*, *full-data* e *last30* foram usados para modelar e avaliar o desempenho dos classificadores baseados diretamente em biossinais (séries temporais) CNN, CNN-LSTM e MULTI-CNN. O desempenho da generalização dos modelos foi avaliado por treinamento e validação *cross-dataset*.

A distribuição do número de registros em cada classe do conjunto de dados é desbalanceada em favor das classes saudáveis em diferentes níveis. Assim, a avaliação dos classificadores nos esquemas de segmentação de sinal contendo um número superior de registros normais em relação aos patológicos resultou de forma geral em desempenho reduzido para Sensibilidade (capacidade do reconhecimento dos casos patológicos) e alta Especificidade (capacidade do reconhecimento dos casos normais).

Apesar de apresentar equivalência estatística entre os esquemas de segmentação, os classificadores de melhor desempenho geral foram o CNN e o CNN-LSTM na segmentação *last30* para todos os conjuntos de dados. Na base de dados CTU-UHB, o classificador CNN-LSTM atingiu 0,930 para especificidade (SP), 0,250 para sensibilidade (SE), 0,468 para média geométrica (GM) e 0,590 para AUC. Na avaliação da base HUFA, o classificador CNN-LSTM atingiu 0,666 para a especificidade, 0,599 para a sensibilidade, 0,551 para a média geométrica e 0,633 para a AUC. Na base DB-TRIUM, o classificador CNN atingiu 0,713 para especificidade, 0,582 para sensibilidade, 0,636 para média geométrica e 0,648 para AUC, sendo estes últimos os de melhores desempenhos gerais nos cenários baseados em biossinais.

Na avaliação da capacidade de generalização do aprendizado via *cross-dataset*, os

melhores resultados para treinamento/validação foram DB-TRIUM/HUFA, no classificador CNN com 0,800 para Especificidade, 0,428 para Sensibilidade, 0,574 para Média Geométrica e 0,614 para AUC. Para DB-TRIUM/CTU-UHB o classificador com melhor desempenho foi o Multi-CNN com 0,827 para Especificidade, 0,380 para Sensibilidade, 0,558 para Média geométrica, 0,603 para AUC.

Para algoritmos baseados em séries temporais existe a necessidade de uma grande quantidade de registros para treinamento e avaliação. desta forma, nossos cenários avaliativos apresentaram um quantitativo relativamente baixo de registros para treinar/validar os modelos. Outro sim, o uso de bases de dados com distribuição de classes não balanceadas enfraqueceu o desempenho dos modelos para avaliação baseada em séries temporais, em específico na sua capacidade de reconhecimento de casos patológicos.

De forma geral, o desempenho dos classificadores na base de dados CTU-UHB alcançaram os resultados mais deficitários. Os classificares avaliados nas bases de dados HUFA e o DB-TRIUM alcançaram melhores desempenhos mesmo nos cenários de dados não balanceados e com forte restrição no número de registros. Desta forma, os resultados indicaram a necessidade de uma análise mais aprofundada com provável adoção de um quantitativo maior de registros no processo de treinamento e testes dos modelos. Com o cenário de avaliação atual, os classificadores CNN e CNN-LSTM aplicados no esquema de segmentação *last30* (últimos 30 minutos) do segmento FCF indicaram a combinação de blocos de construção de melhor desempenho para uso direto nos biossinais. No entanto, a baixa capacidade de reconhecimento dos casos patológicos exige uma investigação mais aprofundada em um maior número de registros a fim de habilitar o uso destes modelos em ambiente clínico.

6 MODELAGEM SUPERVISIONADA BASEADA EM *FEATURES*

Este Capítulo apresenta a construção de modelos nos quais seus principais blocos de construção têm como base o uso de engenharia de *features* com a extração e uso de informações representativas a respeito do estado normal e patológico derivadas dos biossinais. Primeiramente, os materiais e métodos empregados na avaliação dos blocos de construção para estes modelos são apresentados. Em seguida, os resultados da avaliação de blocos de construção fundamentais como os esquemas de segmentação, bases de dados, uso de balanceamento nos dados e os algoritmos de classificação são apresentados. É feita uma avaliação referente à capacidade de generalização das melhores combinações obtidas. São apresentadas ainda discussões referentes ao reconhecimento dos rótulos baseados em pH com valores pertencentes a categoria suspeita (*grey zone*) no esquema de classificação binário. Por fim, sumarizamos os resultados em relação à combinação ótima dos blocos de construção para avaliação do bem-estar fetal destes modelos nos conjuntos de dados avaliados.

A Figura 13 apresenta em destaque o esquema empregado neste Capítulo em relação aos modelos desenvolvidos nesta Tese.

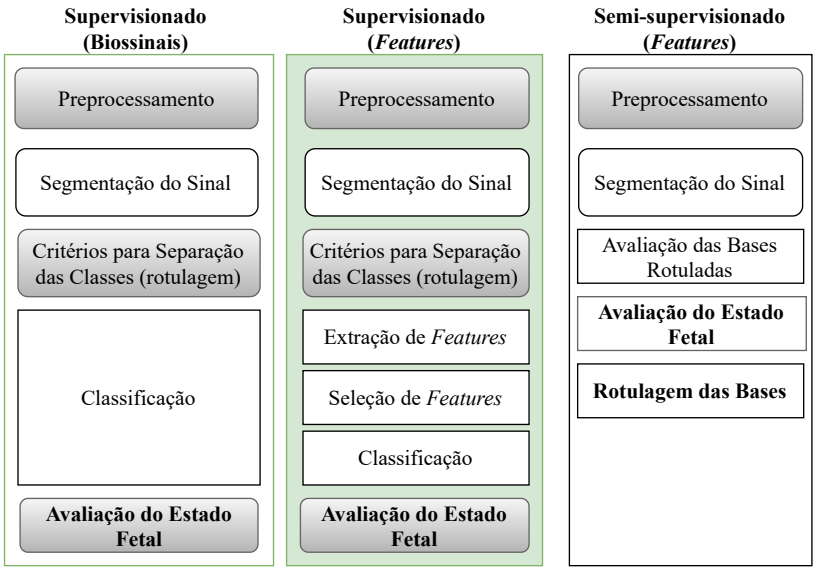


Figura 13 – Modelos supervisionados baseados em features (em destaque). Fonte: Elaboração própria.

6.1 Materiais e Métodos para Modelos Baseados em *Features*

Esta seção apresenta os materiais e métodos empregados nos modelos baseados em engenharia de *features*. O desenho experimental é sumarizado incluindo a extração das *features* e seleção do conjunto mais representativo destas para uso na avaliação dos blocos de construção.

6.1.1 Extração de features

Para a abordagem baseada em engenharia de *features*, um conjunto representativo foi extraído proveniente dos domínios morfológico, linear, frequência e não-linear, incluindo representações baseadas na variabilidade da frequência cardíaca de adultos (VFC). As *features* aplicadas nesta tese foram empregadas com sucesso pela literatura (AYRES-DE-CAMPOS *et al.*, 2015b; GONÇALVES *et al.*, 2006; SPILKA *et al.*, 2012; SIGNORINI *et al.*, 2003; ZHAO *et al.*, 2018) como indicadores do bem-estar fetal. A Tabela 22 apresenta as *features* extraídas.

6.1.1.1 Domínio morfológico

As *features* do domínio morfológico apresentadas nos guias FIGO (AYRES-DE-CAMPOS *et al.*, 2015b) descrevem padrões no sinal de FCF que dependem em sua maioria da linha de base. Para computação da linha de base (*baseline*), foi aplicado o algoritmo apresentado por Fergus *et al.* (FERGUS *et al.*, 2017). Seja um sinal de FCF X de comprimento N, onde $X = \{x_n, n = 1,2...,N\}$, em que a média $\overline{x}$ da linha de base virtual (VBL) é definida como:

$$\overline{x} = \frac{\sum_{n=1}^{N} x_n}{N}; \tag{6.1}$$

O valor de $\overline{x}$ é utilizado para remover as variações superiores e inferiores H, L = 10, e assim computar a baseline real (RBL), o número de acelerações Acc_{total} e de desacelerações Dec_{total} (GEORGOULAS *et al.*, 2017):

$$RBL = \frac{\int_L^H X}{N}; \tag{6.2}$$

$$Acc_{total} = \exists x_i \in X, x_i \geq RBL + 15 \& D \geq 15; \tag{6.3}$$

$$Dec_{total} = \exists x_i \in X, x_i \leq RBL - 15 \& D \geq 15. \tag{6.4}$$

em que x_i é o i-ésimo elemento do sinal X, RBL é a *baseline* real, e D é o tempo de duração em que x_i se mantém acima da RBL + 15 ou abaixo da RBL - 15 (AYRES-DE-CAMPOS *et al.*, 2015b).

Tabela 22 – *Features*. Fonte: Elaboração própria.

Domínio	*Features*
Morfológico (AYRES-DE-CAMPOS *et al.*, 2015b; COMERT *et al.*, 2018)	Baseline ACC_{total} DEC_{total}
Linear (COMERT; KOCAMAZ, 2016; COMERT *et al.*, 2018)	$FHR(mean, std, rms)$ $FHR(meanAD)$ $FHR(medianAD)$ STV, LTV, LTI, and II
Frequência (GONÇALVES *et al.*, 2006)	VLF, LF, MF, HF LF/(MF+HF), TP
Não linear (SPILKA *et al.*, 2012)	FD_Higushi ApEn, SampEn LZC
VFC adulta (SHAFFER; GINSBERG, 2017; ZHAO *et al.*, 2018)	pointcaré(SD1, SD2) p_ratio (SD2/SD1) TRI, $RR(mean, median)$ RMSSD, SDNN, SDSD NN50, pNN50, NN20, pNN20 CVNN, CVSD, $NN(range)$

Abreviações: SD desvio padrão, RMS root mean squares, AD desvio absoluto, STV short-term variation, LTV long-term variation, LTI long-term irregularity, II interval index, TP total power, VLF very-low,LF low, MF mid, and HF high frequencies. ApEn approximate entropy, SampEn sample entropy, LZC lempel-ziv complexity, FD fractal dimension, pointcaré(SD1, SD2) desvio padrão, TRI triangular index. RMSSD rms de diferenças sucessivas, SDNN desvio padrão do NN, SDSD desvio padrão das diferenças entre sucessivos NN, NNx número de pares com sucessivos intervalos NN que diferem mais que x, pNNx porcentagem da NNx, CV coeficiente de variação.

6.1.1.2 Domínio linear

Para o domínio de tempo (linear), foram adotadas as formulações matemáticas presente em (GONÇALVES *et al.*, 2006; COMERT; KOCAMAZ, 2016; FERGUS *et al.*, 2017; COMERT *et al.*, 2018) para extração das *features* provenientes do sinal de frequência cardíaca fetal (*fetal heart rate*). Computamos a média (FCF_{mean}), desvio padrão (FCF_{std}), *root mean square* (FCF_{rms}), *mean absolute deviation* (FCF_{meanAD}) e *median absolute deviation*

($FCF_{medianAD}$).

Seja $x(n)$ um sinal FCF com amostras distribuídas em $n = 1,2,...,N$. Assim, foram computadas as seguintes expressões:

$$FCF_{mean} = \bar{x} = \frac{\sum_{n=1}^{N} x(i)}{N}; \tag{6.5}$$

$$FCF_{std} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x(i)-\bar{x})^2}; \tag{6.6}$$

$$FCF_{rms} = \sqrt{\frac{1}{N}(x_1^2 + x_2^2 + ... + x_N^2)}; \tag{6.7}$$

$$FCF_{meanAD} = \sum_{i=1}^{N}\frac{|x_i - \bar{x}|}{N}; \tag{6.8}$$

$$FCF_{medianAD} = median(x_i - median(X)). \tag{6.9}$$

Seja M o número de minutos em um sinal, e $sm(i)$ o i-ésimo bloco com duração de 2,5 segundos do sinal de FCF. Seja ainda o IQR a representação do *inter-quartile range* e *std* o desvio padrão do mesmo sinal. Implementamos a *short-term variability* (STV), *long-term variability* (LTV), *long-term irregularity* (LTI) e o *interval-index* (II) conforme segue:

$$STV = \frac{1}{24M}\sum_{i=1}^{24M}|sm(i+1) - sm(i)|; \tag{6.10}$$

$$LTV = \frac{1}{M}\sum_{i=1}^{M}[\underset{i\in M}{max}(x(i)) - \underset{i\in M}{min}(x(i))]; \tag{6.11}$$

$$LTI = IQR(\sqrt{x^2(i) + x^2(i+1)}); \tag{6.12}$$

$$II = \frac{STV}{std[sm(i)]}. \tag{6.13}$$

6.1.1.3 Domínio de frequência

Para a obtenção das *features* do domínio de frequência, a análise do espectro foi usada. Assim, estimou-se a área abaixo do espectro utilizando a implementação do método de *Welch* do *scipy*, com janela de *Hanning* aplicada nos sinais FCF sobre sequências de comprimento 256, sem sobreposições. Este trabalho considerou quatro bandas de frequência: VLF (0–0.03 Hz), LF (0.03–0.15 Hz), MF (0.15–0.50 Hz) e HF (0.50–1.00 Hz), correspondente às frequências muito baixas, baixas, médias e altas. Também foi calculada as potência total (TP) e sua razão LF/(MF + HF), conforme apresentado em (SIGNORINI *et al.*, 2003; GONÇALVES *et al.*, 2006).

6.1.1.4 Domínio não-linear

Do domínio não-linear, foram extraídas *features* capazes de quantificar a complexidade do sinal ou sua irregularidade (SPILKA *et al.*, 2012), tais como: Dimensão fractal de Higushi com (k = 5), *approximate entropy* (ApEn) e *sample entropy* (SampEn) com valores m = 2; r = 0,2 * desvio padrão. Obtivemos ainda a complexidade Lempel-Ziv normalizada, conforme as definições apresentadas em (SPILKA *et al.*, 2012; COMERT; KOCAMAZ, 2016). A *ApEn* e *SampEn* são extraídas de forma similar, com a diferença principal que *SampEn* não contém *self-matches*. O parâmetro m representa a dimensão embutida e r age como um limite. Assim, dado um sinal $X = x_i, x_2, ..., x_N$ com N amostras, dadas as subsequências de X de tamanho m representadas por $u_m(i)$, e, em termos Euclidianos, os vetores $u_m((i)$ e $u_m(j)$ representados por $n_i^m(r)$, em que a probabilidade destes vetores serem próximos é expressa por $C_i^m(r) = n_i^n/(N-m+1)$, resultam-se a seguintes definições:

$$ApEn(m,r,N) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} ln(C_r^m(i)) - \frac{1}{N-m}\sum_{i=1}^{N-m} ln(C_r^{m+1}(i)); \tag{6.14}$$

$$SampEn(m,r) = \lim_{N\to\infty} -\ln\frac{C^{m+1(r)}}{C^m(r)}; \tag{6.15}$$

A complexidade Lempel-Ziv (LZC) estima os padrões repetidos em uma serie temporal (COMERT; KOCAMAZ, 2016). Seja n o tamanho da série e $c(n)$ o número de itens únicos na sequência. A complexidade Lempel-Ziv pode ser definida como:

$$C(n) = \frac{c(n)log_2(n)}{n}. \tag{6.16}$$

A dimensão fractal pelo método de Higushi utiliza o tamanho estimado do sinal. No caso, seja um sinal $X = x_1, x_2, ..., x_N$ com N amostras. Um novo sinal $X_k^m = \{x_m, x_{m+k}, ..., x_{m+[(N-m)/k])}\}$ com $m = 1, 2, ..., k)$ é construído onde [] representa a notação de Gauss, m define o tempo inicial, e k o intervalo de tempo. Para cada m, o tamanho $L_m(k)$ de X_k^m é computado. O tamanho da curva para o intervalo k é $\langle L(k)\rangle$, que é definido como a média sobre k conjuntos de $L_m(k)$. A curva de tamanho $\langle L(k)\rangle$ e relacionada à dimensão fractal D é representada pela fórmula exponencial $\langle L(k)\rangle \alpha k^{-D}$, em que a dimensão fractal de Higushi é estimada pela inclinação de uma curva de regressão ajustada de $\langle L(k)\rangle$ versus k (SPILKA *et al.*, 2012):

$$\langle L(k)\rangle = \sum_{m=1}^{k} \frac{L_m(k)}{k}. \tag{6.17}$$

6.1.1.5 Baseadas na variabilidade da frequência cardíaca adulta

Nas *features* derivadas da frequêcia cardíaca adulta e sua variabilidade (VFC), foi necessário converter os valores da FCF amostrada para intervalos NN ou RR em medições época a época. Desta forma, antes de calcular *features* baseadas em variabilidade da frequência cardíaca, a versão de época a época do sinal FCF foi extraída conforme apresentado em (ZHAO *et al.*, 2018):

$$RR_i = \frac{60.000}{FCF_i}\, milissegundos \tag{6.18}$$

Desta forma, o sinal representa os intervalos RR ou NN (*normal-to-normal*), ou seja, os intervalos de tempo entre dois picos R (RR *intervals*) no complexo QRS ao longo do tempo em milissegundos, os quais podem ser utilizados para calcular índices temporais equivalentes ao domínio de análise cardíaca adulta conforme apresentado em (SHAFFER; GINSBERG, 2017; ZHAO *et al.*, 2018). Para fins de nomenclatura em algumas formulações matemáticas, assume-se que as denominações NN e RR representam o mesmo sinal, ou seja um sinal $RR = \{RR_i, RR_{i+1}, ..., RR_{i+n}\}$ ou $NN = \{NN_i, NN_{i+1}, ..., NN_{i+n}\}$ em milissegundos de tamanho n. Inicialmente, computamos a média (μ), *root mean square*, a amplitude entre os valores máximos e mínimos, desvio padrão dos intervalos NN (SDNN), valor RMS (*root mean square*) das sucessivas diferenças entre intervalos (RMSSD), desvio padrão da diferença entre intervalos sucessivos (SDSD):

$$RR_{mean} = \overline{NN} = \frac{\sum_{i=1}^{n} RR_{(i)}}{n}; \tag{6.19}$$

$$SDNN = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(NN(i) - \overline{NN})^2}; \tag{6.20}$$

$$RR_{rms} = \sqrt{\frac{1}{n}(RR_1^2 + RR_2^2 + ... + RR_n^2)}; \tag{6.21}$$

$$RR_{amp} = \max(RR) - \min(RR); \tag{6.22}$$

$$RMSSD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(NN_i - NN_{i+1})^2}; \tag{6.23}$$

$$SDSD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[(NN_i - NN_{i+1}) - (\frac{\sum_{i=1}^{n} NN_i - NN_{i+1}}{n})]^2}. \tag{6.24}$$

Foram extraídos o coeficiente de variação (CVNN), o coeficiente de variação das diferenças sucessivas (CVSD), e o *triangular index* (TRI), no qual o valor de $D(NN)$ representa a densidade da distribuição máxima no histograma para o intervalo NN_i mais frequente. Foram

computados ainda o desvio padrão da linha perpendicular à identidade do plot de Poincaré (Poincaré SD1), desvio-padrão ao longo da linha de identidade do plot Poincaré (Poincaré SD2), e a razão Poincaré (SD1/SD2):

$$CVNN = \frac{SDNN}{\overline{NN}}; \tag{6.25}$$

$$CVSD = \frac{SDSD}{\frac{\sum_{i=1}^{n}(NN_i - NN_{i+1})}{n}}; \tag{6.26}$$

$$TRI = \frac{n}{D(NN)}; \tag{6.27}$$

$$SD1 = \sqrt{\frac{1}{2}SDSD^2}; \tag{6.28}$$

$$SD2 = \sqrt{2SDNN^2 - \frac{1}{2}SDSD^2}; \tag{6.29}$$

$$SD_{ratio} = \frac{SD2}{SD1}. \tag{6.30}$$

Por fim, foi computado o número de intervalos sucessivos que diferem entre si mais que x (50 e 20) milissegundos (NNx) e a porcentagem de NNx (pNNx) em relação ao número de amostras n:

$$NN20 = \exists NN_i \in NN, NN_i > 20; \tag{6.31}$$

$$NN50 = \exists NN_i \in NN, NN_i > 50; \tag{6.32}$$

$$pNN20 = \frac{NN20}{n}; \tag{6.33}$$

$$pNN50 = \frac{NN50}{n}. \tag{6.34}$$

Apesar das representações de Poincaré pertencerem tipicamente ao domínio não-linear, aqui, por empregar o sinal época a época, tais representações foram abordadas como *features* provenientes da variabilidade da frequência cardíaca adulta.

6.1.2 Seleção de features

Com as *features* extraídas, foi realizada a seleção da combinação mais eficiente para classificação dos estados normal e patólogico dos fetos. Para isto, foi empregado o algoritmo *recursive feature elimination* (RFE) (GUYON *et al.*, 2002) com o classificador Random Forest atuando como seletor, e usando a Área Sob a Curva (AUC) como métrica ou função de custo.

O algoritmo RFE seleciona um número fixo de *features* com melhor desempenho nas tarefas de classificação ao calcular a mudança em uma função de custo causada pela remoção de uma determinada *feature*. O algoritmo de seleção foi usado nos três esquemas de segmentação para cada base de dados e as rodadas que obtiveram melhor desempenho da AUC foram registradas. Para encontrar um subconjunto de *features* úteis e produzir um modelo preciso, o RFE remove as menos essenciais sequencialmente (FERGUS *et al.*, 2018).

Nesta tese, a fim de obter um conjunto de *features* generalizável para as várias bases de dados utilizadas, primeiramente as 20 *features* mais representativas foram selecionadas para cada conjunto de dados rotulados CTU-UHB, HUFA e DB-TRIUM. Em seguida, foram ranqueadas entre estas as *features* comuns aos três conjuntos de dados. A Tabela 23 destaca os 20 principais *features* para cada conjunto de dados e as seis que compuseram o conjunto final de *features* a serem empregadas nos experimentos.

As seis *features* selecionadas (extraídas das 20 selecionadas em cada base de dados) são meanAD, FD_Higushi, TRI (*triangular index*), NN20, pNN20 e CVNN. Estas *features* foram usadas em nosso processo de avaliação.

6.1.3 Avaliação dos modelos baseados em features

A Figura 14 apresenta uma visão geral da avaliação dos blocos de contrução nos cenários baseados em engenharia de *features*. Para avaliar os blocos de construção incluindo os esquemas de segmentação das bases de dados rotuladas e algoritmos de classificação, foram empregados processos de avaliação nas bases de dados de forma individual e cruzadas (*cross-dataset*). Na avaliação das bases de forma isolada, a abordagem *cross-validation* aninhada foi adotada, que consiste em um *loop* interno e outro externo. Esta abordagem contribui para tornar o processo de avaliação imparcial durante a seleção de parâmetros necessários para otimizar os algoritmos de classificação (JAPKOWICZ; SHAH, 2011).

Duas repetições do *5-fold cross-validation* (2x5-*fold* cv) foram usadas em sua versão estratificada no *loop* externo para avaliação do desempenho global. Desta forma, a avaliação de desempenho foi executada dez vezes (duas repetições de cinco vezes) e, em seguida, foi calculada a média e desvio-padrão das métricas no conjunto de rodadas. Em cada execução do *loop* externo, o processo dividiu os dados em cinco parcelas iguais, quatro delas foram usadas para treinamento (80%) e uma (20%) para teste ou validação. Para garantir que os valores *features* se mantenham dentro de um pequeno intervalo especificado, como [0,1], as parcelas de

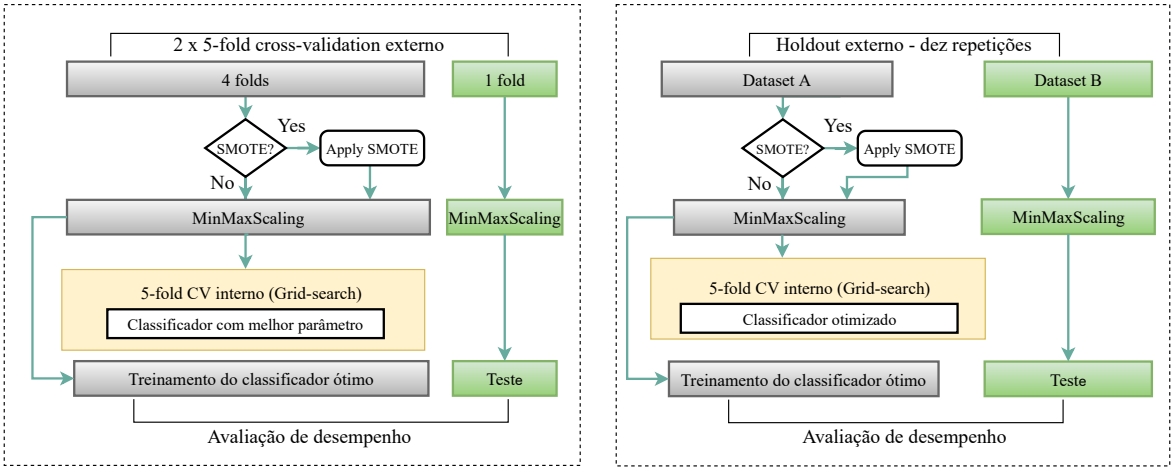
Tabela 23 – Top 20 *features*. Fonte: Elaboração própria.

Feature	Selecionadas por base de dados		
	CTU-UHB	HUFA	DB-TRIUM
Baseline	True	False	False
# of accelerations	True	False	False
$FHR(std)$	True	False	True
FHR_(MeanAD)	**True**	**True**	**True**
$FHR(medianAD)$	True	False	True
STV	False	True	True
LTV	True	False	True
LTI	True	False	True
II	True	True	False
LZC	False	False	False
ApEn	True	True	False
SampEn	True	False	False
FD_Higushi	**True**	**True**	**True**
TP	False	True	False
VLF	False	True	True
LF	False	True	True
MF	True	True	False
HF	True	True	False
LF/(MF+HF)	True	True	False
pointcare_SD1	False	False	True
pointcare_SD2	False	True	True
p_ratio(SD2/SD1)	True	True	False
TRI	**True**	**True**	**True**
RMSSD	False	True	True
$RR(mean)$	True	False	False
SDNN	False	True	True
SDSD	False	True	True
NN20	**True**	**True**	**True**
PNN20	**True**	**True**	**True**
CVNN	**True**	**True**	**True**
$NN(range)$	False	True	True

Em destaque as *features* mais representativas selecionadas simultaneamente nas três bases de dados.

treinamento foram dimensionadas aplicando a normalização *Min-Max* e a escala resultante foi reutilizada nas parcelas de teste. Antes de treinar o classificador, buscou-se pelos seus melhores hiper-parâmetros empregando *grid-search* no *cross-validation* (CV) estratificado mais interno, o qual foi executado apenas nas parcelas de treinamento. Por fim, o classificador foi treinado com os parâmetros otimizados nas parcelas de treinamento e as métricas de avaliação foram computadas para as parcelas de teste.

Para a avaliação *cross-dataset*, foram empregadas dez repetições do método nas



(a) Cross-validation aninhado (Uma base de dados). (b) Abordagem *Holdout* (*cross-datasets*).
Figura 14 – Cenários de *features* em bases únicas e em *cross-dataset*. Fonte: Elaboração própria.

Tabela 24 – Intervalos de hiper-parâmetros para *grid-search*

Classificador	Hyper-parâmetros
GTB	**max_iter**: [250, 500]
KNN	**n_neighbors**: [5-15]
SVM	**Kernel**: ['poly', 'rbf'] **C**: [0.001, 0.01, 0.1, 1.0, 10] **gamma**: [0.001, 0.01, 0.1]
RF	**n_estimators**: [100, 500] **criterion**:['gini', 'entropy'] **max_features**:['sqrt', 'log2']
MLP	**epochs**:[50, 100] **batch_size**: [16, 32]
BG	**n_estimators**: [10,100,1000]

Gradient Tree Boosting (GBT), Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Multi-layer perceptrom (MLP), Bagging (BG).

quais cada algoritmo de classificação foi treinado em um conjunto de dados específico e o desempenho do modelo foi testado nas bases de dados remanescentes. Os dados de treinamento também foram dimensionados aplicando a escala *MinMax* e a escala resultante foi reutilizada no conjunto de dados de teste. Fez-se uso do *5-fold cross-validation* estratificado internamente nos dados de treinamento para buscar os hiperparâmetros dos classificadores via grid-search. Em seguida, o classificador de parâmetro ideal foi treinado no conjunto de dados de treinamento e as métricas de avaliação foram computadas no conjunto de dados de teste. Finalmente, a média dos resultados das avaliações foram computados.

A Tabela 24 apresenta o intervalo de parâmetros otimizados em cada iteração do *k-fold* com os nomes equivalentes a implementação dos algoritmos nas bibliotecas *scikit-learn* e

Keras. Foi empregado o conhecimento de especialistas para especificar os intervalos de valores utilizados para ajustar os hyper-parâmetros de cada algoritmo (OLSON *et al.*, 2018).

Para o estudo dos intervalos de pH categorizados como suspeitos *grey zone*, o esquema de avaliação *cross-dataset* foi adaptado de forma que, apenas nos dados com rotulagem baseados no valor do pH (CTU-UHB e HUFA), o conjunto de validação foi representado pelos registros da mesma base pertencentes a área intermediária, os quais haviam sido removidos durante o processo inicial de rotulagem destas bases (ver Seção 4.5). Desta forma, foi realizada uma avaliação *cross-dataset* onde a base com rotulagem binária foi empregada para treinamento e os registros marcados como suspeitos da mesma base foram usados para validação a fim de verificar qual o rótulo seria atribuído (normal e patológico) para classes suspeitas após o treinamento do modelo.

6.2 Resultados de modelos baseados em *features*

Para cada intervalo de dados, foram avaliados os cenários de segmentação relacionados ao desempenho de classificação dos dados em sua forma original (desbalanceados) e dos dados balanceados com a técnica SMOTE.

- Cenário 1: *Features* selecionadas via RFE nos dados não balanceados;
- Cenário 2: *Features* selecionadas via RFE com dados balanceados (SMOTE).

Esses cenários de avaliação objetivaram identificar o esquema de segmentação de dados, o algoritmo de classificação e a combinação de blocos de construção mais adequadas para o modelo de prognóstico baseado em *features*.

6.2.0.1 Cenários de avaliação para base CTU-UHB

A Tabela 25 apresenta a avaliação dos classificadores no esquema de segmentação *first30* para CTU-UHB. Na parte superior, os resultados dos dados originais não balanceados demonstram que todos os classificadores alcançaram alta especificidade (capacidade de reconhecer classes saudáveis) e baixa Sensibilidade (capacidade de reconhecer casos patológicos). A média geométrica (equilíbrio relativo entre SE e SP) foi baixa, e os resultados da AUC foram próximos ao aleatório (50%) apesar do alto valor de acurácia (ACC). Na parte inferior, os dados balanceados com SMOTE (BOWYER *et al.*, 2002) mostra uma ligeira melhora nos valores SE e uma diminuição nos valores de ACC e SP. Houve um aumento na GM, e a AUC teve um

desempenho ligeiramente melhor do que o acaso. Considerando todas as métricas, o SVM obteve os melhores resultados para o esquema de segmentação *first30* quando executado no conjunto de dados balanceado.

Tabela 25 – CTU-UHB - esquema first30. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
KNN	0.895 (0.009)	0.980 (0.016)	0.075 (0.082)	0.189 (0.192)	0.531 (0.034)
SVM	0.899 (0.000)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.500 (0.000)
RF	0.894 (0.012)	0.986 (0.010)	0.075 (0.061)	0.210 (0.172)	0.530 (0.032)
MLP	0.896 (0.005)	0.995 (0.006)	0.012 (0.037)	0.035 (0.105)	0.504 (0.017)
GTB	0.870 (0.011)	0.963 (0.127)	0.037 (0.057)	0.103 (0.158)	0.500 (0.027)
BG	0.886 (0.020)	0.970 (0.019)	0.137 (0.067)	0.342 (0.128)	0.554 (0.037)
KNN	0.582 (0.051)	0.600 (0.060)	0.425 (0.061)	0.502 (0.031)	0.512 (0.030)
SVM	**0.732** (0.055)	**0.770** (0.067)	**0.387** (0.180)	**0.526** (0.128)	**0.579** (0.081)
RF	0.783 (0.037)	0.845 (0.040)	0.237 (0.162)	0.409 (0.179)	0.541 (0.080)
MLP	0.733 (0.056)	0.779 (0.063)	0.325 (0.114)	0.493 (0.091)	0.552 (0.061)
GTB	0.776 (0.038)	0.839 (0.047)	0.212 (0.137)	0.384 (0.166)	0.526 (0.061)
BG	0.760 (0.049)	0.818 (0.056)	0.237 (0.141)	0.387 (0.205)	0.528 (0.069)

Topo: dados originais não balanceados, Base: dados balanceados com SMOTE. Valores em destaque mostram os melhores desempenhos.

Na Tabela 26 é apresentada a avaliação dos blocos de construção no esquema de dados completos (*full-data*). Os dados balanceados apresentaram aumento nos resultados SE, GM e AUC relacionados aos escores desbalanceados. Os classificadores SVM e KNN apresentaram as melhores pontuações.

Tabela 26 – CTU-UHB - esquema full-data. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
KNN	0.898 (0.008)	0.997 (0.008)	0.012 (0.037)	0.035 (0.106)	0.504 (0.019)
SVM	0.899 (0.000)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.500 (0.000)
RF	0.912 (0.018)	0.988 (0.015)	0.225 (0.093)	0.444 (0.159)	0.606 (0.049)
MLP	0.913 (0.021)	0.986 (0.017)	0.262 (0.130)	0.475 (0.182)	0.624 (0.066)
GTB	0.899 (0.015)	0.976 (0.008)	0.212 (0.158)	0.394 (0.227)	0.594 (0.078)
BG	0.914 (0.015)	0.984 (0.017)	0.287 (0.112)	0.520 (0.106)	0.636 (0.053)
KNN	0.742 (0.031)	0.755 (0.042)	0.625 (0.176)	0.676 (0.098)	0.690 (0.078)
SVM	**0.786** (0.031)	**0.801** (0.034)	**0.650** (0.165)	**0.713** (0.102)	**0.725** (0.080)
RF	0.861 (0.035)	0.900 (0.035)	0.512 (0.226)	0.662 (0.147)	0.706 (0.111)
MLP	0.820 (0.043)	0.850 (0.047)	0.550 (0.139)	0.677 (0.084)	0.700 (0.070)
GTB	0.849 (0.028)	0.896 (0.029)	0.425 (0.187)	0.600 (0.140)	0.660 (0.091)
BG	0.847 (0.032)	0.889 (0.037)	0.475 (0.183)	0.634 (0.131)	0.682 (0.086)

Topo: dados originais não balanceados, Base: dados balanceados com SMOTE. Valores em destaque mostram os melhores desempenhos.

Os resultados da avaliação no esquema de segmentação *last30* são apresentados na Tabela 27. Os dados desbalanceados originais mostram alto desempenho para ACC e SP, assim como um baixo desempenho na SE, refletindo nos resultados de GM e AUC. Apesar da ligeira

diminuição nos valores de ACC e SP, os dados balanceados apresentaram resultados promissores para todos os classificadores em que o SVM e o RF foram os de melhor desempenho.

Tabela 27 – CTU-UHB - esquema *last30*. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
KNN	0.912 (0.013)	0.993 (0.012)	0.187 (0.139)	0.374 (0.211)	0.590 (0.067)
SVM	0.900 (0.003)	1.000 (0.000)	0.012 (0.037)	0.035 (0.106)	0.506 (0.018)
RF	0.912 (0.019)	0.979 (0.015)	0.312 (0.128)	0.540 (0.119)	0.645 (0.064)
MLP	0.914 (0.017)	0.984 (0.011)	0.287 (0.125)	0.517 (0.123)	0.636 (0.063)
GTB	0.910 (0.020)	0.972 (0.015)	0.362 (0.130)	0.584 (0.104)	0.667 (0.066)
BG	0.922 (0.020)	0.976 (0.018)	0.437 (0.150)	0.640 (0.126)	0.706 (0.073)
KNN	0.751 (0.036)	0.765 (0.043)	0.625 (0.111)	0.688 (0.056)	0.695 (0.052)
SVM	**0.837** (0.021)	**0.856** (0.023)	**0.675** (0.139)	**0.755** (0.081)	**0.765** (0.066)
RF	0.875 (0.034)	0.904 (0.039)	0.612 (0.117)	0.740 (0.071)	0.758 (0.056)
MLP	0.836 (0.047)	0.867 (0.049)	0.562 (0.128)	0.693 (0.085)	0.714 (0.070)
GTB	0.861 (0.036)	0.892 (0.040)	0.587 (0.137)	0.717 (0.079)	0.739 (0.059)
BG	0.856 (0.037)	0.890 (0.046)	0.550 (0.127)	0.693 (0.077)	0.720 (0.056)

Topo: dados originais não balanceados, Base: dados balanceados com SMOTE. Valores em destaque mostram os melhores desempenhos.

6.2.0.2 Cenários de avaliação na base HUFA

Comparando nossos três conjuntos de dados com rótulos, o HUFA apresenta a menor quantidade de dados e uma distribuição de classes quase equilibrada. A Tabela 28 mostra a avaliação dos classificadores no esquema de segmentação *first30* para a base de dados HUFA.

Tabela 28 – HUFA - esquema *first30*. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
KNN	0.410 (0.166)	0.283 (0.197)	0.566 (0.300)	0.300 (0.255)	0.425 (0.172)
SVM	0.500 (0.118)	0.000 (0.000)	0.950 (0.150)	0.000 (0.000)	0.500 (0.118)
RF	0.556 (0.138)	0.433 (0.290)	0.700 (0.233)	0.457 (0.250)	0.566 (0.138)
MLP	0.489 (0.173)	0.333 (0.288)	0.683 (0.283)	0.376 (0.252)	0.508 (0.172)
GTB	0.440 (0.048)	0.400 (0.489)	0.600 (0.489)	0.000 (0.000)	0.500 (0.000)
BG	0.516 (0.201)	0.616 (0.350)	0.449 (0.258)	0.454 (0.246)	0.533 (0.204)
KNN	0.426 (0.148)	0.350 (0.203)	0.516 (0.262)	0.334 (0.234)	0.433 (0.143)
SVM	0.486 (0.106)	0.066 (0.133)	0.866 (0.266)	0.066 (0.133)	0.466 (0.066)
RF	**0.590** (0.149)	**0.416** (0.291)	**0.750** (0.226)	**0.480** (0.263)	**0.583** (0.166)
MLP	0.436 (0.241)	0.400 (0.280)	0.499 (0.341)	0.418 (0.228)	0.450 (0.250)
GTB	0.480 (0.074)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.500 (0.000)
BG	0.440 (0.154)	0.400 (0.317)	0.499 (0.268)	0.317 (0.267)	0.450 (0.171)

Topo: dados originais não balanceados, Base: dados balanceados com SMOTE. Valores em destaque mostram os melhores desempenhos.

Nos esquema *first30* para os dados semi-balanceados, os classificadores mostram uma melhor capacidade de classificar os casos patológicos (SE) do que os saudáveis (SP), e a AUC obteve resultados próximos ao aleatório. Nos dados balanceados com a técnica SMOTE,

os classificadores apresentaram desempenho baixo com um pequeno aumento nos escores do GM em relação aos dados originais. O classificador GTB reconheceu incorretamente todos os registros patológicos nos dados balanceados e o RF foi o classificador de melhor desempenho.

Tabela 29 – HUFA - esquema *full-data*. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
KNN	0.443 (0.074)	0.716 (0.258)	0.233 (0.200)	0.266 (0.225)	0.475 (0.065)
SVM	0.480 (0.074)	0.400 (0.489)	0.666 (0.421)	0.115 (0.230)	0.533 (0.066)
RF	0.643 (0.135)	**0.733** (0.249)	**0.616** (0.279)	**0.630** (0.138)	**0.675** (0.131)
MLP	0.516 (0.092)	0.599 (0.326)	0499 (0.258)	0.454 (0.164)	0.550 (0.100)
GTB	0.440 (0.048)	0.400 (0.489)	0.600 (0.489)	0.000 (0.000)	0.500 (0.000)
BG	0.553 (0.129)	0.583 (0.291)	0.583 (0.291)	0.497 (0.203)	0.583 (0.129)
KNN	0.550 (0.159)	0.816 (0.240)	0.333 (0.324)	0.355 (0.309)	0.575 (0.126)
SVM	0.556 (0.116)	0.833 (0.223)	0.333 (0.197)	0.443 (0.231)	0.583 (0.091)
RF	**0.573** (0.119)	**0.633** (0.348)	**0.566** (0.290)	**0.467** (0.255)	**0.600** (0.116)
MLP	0.539 (0.099)	0.666 (0.298)	0.466 (0.145)	0.523 (0.094)	0.566 (0.116)
GTB	0.480 (0.074)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.500 (0.000)
BG	0.593 (0.137)	0.566 (0.249)	0.650 (0.216)	0.579 (0.142)	0.608 (0.144)

Topo: dados originais não balanceados, Base: dados balanceados com SMOTE. Valores em destaque mostram os melhores desempenhos.

Na Tabela 29, é apresentada a avaliação de cada classificador no esquema de dados completos (*full-data*) da base HUFA . O classificador *random forest* (RF) obteve melhores desempenhos tanto nos dados originais como nos balanceados e apresentou pouca variação nos valores da SE, SP, GMEAN e AUC.

Tabela 30 – HUFA - esquema *last30*. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
KNN	0.726 (0.138)	0.750 (0.271)	0.733 (0.249)	0.704 (0.160)	0.741 (0.141)
SVM	0.536 (0.124)	0.433 (0.388)	0.700 (0.348)	0.299 (0.305)	0.566 (0.089)
RF	0.633 (0.204)	0.566 (0.366)	0.733 (0.290)	0.550 (0.311)	0.650 (0.216)
MLP	0.596 (0.153)	0.533 (0.296)	0.683 (0.283)	0.491 (0.283)	0.608 (0.144)
GTB	0.440 (0.048)	0.400 (0.489)	0.600 (0.489)	0.000 (0.000)	0.500 (0.000)
BG	0.653 (0.167)	0.566 (0.366)	0.766 (0.260)	0.550 (0.311)	0.666 (0.182)
KNN	**0.726** (0.138)	**0.750** (0.271)	**0.733** (0.249)	**0.704** (0.160)	**0.741** (0.141)
SVM	0.560 (0.154)	0.700 (0.339)	0.416 (0.359)	0.347 (0.300)	0.558 (0.144)
RF	0.690 (0.199)	0.683 (0.353)	0.733 (0.290)	0.624 (0.286)	0.708 (0.194)
MLP	0.670 (0.151)	0.666 (0.288)	0.666 (0.197)	0.635 (0.155)	0.666 (0.149)
GTB	0.480 (0.074)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.500 (0.000)
BG	0.693 (0.214)	0.633 (0.406)	0.766 (0.260)	0.587 (0.349)	0.700 (0.221)

Topo: dados originais não balanceados, Base: dados balanceados com SMOTE. Valores em destaque mostram os melhores desempenhos.

A Tabela 30 mostra a avaliação de cada classificador no esquema de segmentação *last30* para a base HUFA. O classificador KNN obteve desempenho equivalente nos dados semi-balanceados e balanceados. Em relação ao desempenho geral para todas as métricas, os

classificadores RF e KNN alcançaram os melhores resultados nos dados balanceados.

6.2.0.3 Cenários de avaliação na base DB-TRIUM

Na Tabela 31 é apresentado o desempenho dos algoritmos na combinação de blocos de construção baseada no esquema de segmentação *first30* na base DB-TRIUM. O cenário de dados não balanceado apresenta pontuações SE e GM baixas, com o desempenho da AUC próximo ao aleatório. Neste esquema de segmentação, o cenário de dados balanceados manteve o baixo desempenho geral com os valores AUC em torno do aleatório. O classificador GTB atingiu a pontuação mais equilibrada no conjunto de dados balanceados.

Tabela 31 – DB-TRIUM - esquema *first30*. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
KNN	0.625 (0.057)	0.848 (0.082)	0.097 (0.102)	0.213 (0.188)	0.473 (0.579)
SVM	0.704 (0.014)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.500 (0.000)
RF	0.660 (0.045)	0.861 (0.056)	0.182 (0.068)	0.371 (0.130)	0.521 (0.036)
MLP	0.651 (0.046)	0.887 (0.079)	0.085 (0.079)	0.205 (0.171)	0.486 (0.032)
GTB	0.635 (0.062)	0.799 (0.058)	0.241 (0.118)	0.426 (0.120)	0.520 (0.079)
BG	0.616 (0.049)	0.795 (0.088)	0.192 (0.115)	0.338 (0.179)	0.493 (0.046)
KNN	0.478 (0.062)	0.473 (0.075)	0.487 (0.134)	0.473 (0.075)	0.480 (0.072)
SVM	0.384 (0.088)	0.241 (0.181)	0.722 (0.210)	0.345 (0.159)	0.481 (0.068)
RF	0.597 (0.051)	0.696 (0.101)	0.358 (0.132)	0.485 (0.073)	0.527 (0.050)
MLP	0.550 (0.068)	0.566 (0.097)	0.510 (0.136)	0.528 (0.071)	0.538 (0.069)
GTB	**0.600** (0.070)	**0.687** (0.085)	**0.394** (0.124)	**0.513** (0.090)	**0.541** (0.076)
BG	0.572 (0.036)	0.655 (0.076)	0.370 (0.110)	0.482 (0.060)	0.512 (0.037)

Topo: dados originais não balanceados, Base: dados balanceados com SMOTE. Valores em destaque mostram os melhores desempenhos.

A Tabela 32 mostra os resultados da avaliação de algoritmos no esquema de dados completos (*full-data*). Os dados balanceados apresentaram ligeira diminuição da ACC e SP com aumento dos valores da SE em relação à versão não balanceada. Com os dados balanceados, todos os classificadores alcançaram uma pontuação equilibrada entre as métricas e com a GM e AUC alcançando bons resultados. Os classificadores SVM e BG compreendem os melhores desempenhos para o esquema de segmentação *full-data*.

Na Tabela 33 é exibida a avaliação da combinação de blocos de construção para segmentação *last30* na base DB-TRIUM. Os classificadores apresentaram valores elevados de ACC e SP, com valores de SE moderados. o GM e AUC alcançaram escores aceitáveis mesmo no cenário de dados não balanceados. Nos dados balanceados, a avaliação dos classificadores resulta em uma diminuição na ACC e SP seguida por um aumento na SE, GM e AUC, obtendo um bom desempenho geral. Todos os classificadores tiveram um bom desempenho e os melhores

Tabela 32 – DB-TRIUM - esquema *full-data*. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
KNN	0.688 (0.031)	0.888 (0.062)	0.208 (0.134)	0.388 (0.168)	0.548 (0.050)
SVM	0.710 (0.037)	0.982 (0.029)	0.064 (0.139)	0.107 (0.221)	0.532 (0.061)
RF	0.736 (0.073)	0.856 (0.081)	0.443 (0.170)	0.601 (0.128)	0.649 (0.090)
MLP	0.735 (0.049)	0.883 (0.079)	0.378 (0.181)	0.558 (0.114)	0.630 (0.072)
GTB	0.714 (0.046)	0.856 (0.081)	0.443 (0.170)	0.601 (0.128)	0.649 (0.090)
BG	0.735 (0.053)	0.842 (0.068)	0.472 (0.162)	0.619 (0.096)	0.657 (0.074)
KNN	0.663 (0.097)	0.668 (0.089)	0.651 (0.134)	0.658 (0.106)	0.659 (0.106)
SVM	**0.726** (0.103)	**0.761** (0.128)	**0.637** (0.176)	**0.686** (0.117)	**0.699** (0.106)
RF	0.685 (0.075)	0.748 (0.106)	0.526 (0.182)	0.613 (0.104)	0.637 (0.087)
MLP	0.707 (0.073)	0.743 (0.099)	0.615 (0.194)	0.663 (0.104)	0.679 (0.091)
GTB	0.657 (0.066)	0.701 (0.070)	0.547 (0.177)	0.605 (0.124)	0.624 (0.091)
BG	0.707 (0.063)	0.735 (0.079)	0.634 (0.153)	0.676 (0.081)	0.685 (0.077)

Topo: dados originais não balanceados, Base: dados balanceados com SMOTE. Valores em destaque mostram os melhores desempenhos.

classificados foram o MLP e o KNN nos dados balanceados.

Tabela 33 – DB-TRIUM - esquema *last30*. Fonte: Elaboração própria.

Algoritmo	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
KNN	0.764 (0.059)	0.945 (0.063)	0.332 (0.135)	0.546 (0.117)	0.639 (0.070)
SVM	0.745 (0.066)	0.986 (0.020)	0.175 (0.233)	0.256 (0.323)	0.581 (0.110)
RF	0.732 (0.072)	0.857 (0.063)	0.438 (0.143)	0.605 (0.111)	0.648 (0.087)
MLP	0.773 (0.056)	0.887 (0.058)	0.499 (0.133)	0.659 (0.090)	0.693 (0.071)
GTB	0.682 (0.050)	0.780 (0.051)	0.447 (0.105)	0.587 (0.071)	0.614 (0.059)
BG	0.742 (0.056)	0.861 (0.045)	0.459 (0.126)	0.623 (0.090)	0.660 (0.071)
KNN	0.676 (0.049)	0.665 (0.050)	0.701 (0.128)	0.679 (0.067)	0.683 (0.067)
SVM	0.739 (0.055)	0.884 (0.084)	0.396 (0.230)	0.536 (0.223)	0.640 (0.092)
RF	0.688 (0.049)	0.754 (0.053)	0.534 (0.139)	0.628 (0.081)	0.644 (0.068)
MLP	**0.695** (0.084)	**0.727** (0.093)	**0.616** (0.173)	**0.661** (0.103)	**0.672** (0.098)
GTB	0.676 (0.057)	0.753 (0.093)	0.489 (0.143)	0.597 (0.074)	0.621 (0.062)
BG	0.695 (0.058)	0.754 (0.060)	0.553 (0.137)	0.640 (0.081)	0.653 (0.074)

Topo: dados originais não balanceados, Base: dados balanceados com SMOTE. Valores em destaque mostram os melhores desempenhos.

6.2.0.4 Comparações de desempenho entre esquemas

O desempenho entre esquemas de classificadores individuais foi comparado estatisticamente. Entre nossas métricas de avaliação adotadas, tanto a AUC quanto a média geométrica (GM) representam o equilíbrio entre a capacidade discriminativa das classes saudáveis (Especificidade) e patológica (Sensibilidade). A análise empregou a AUC como critério de comparação. Foi aplicado o teste de ranqueamento de Wilcoxon (WILCOXON, 1945) para comparar as rodadas dentro de uma execução *k-fold* para cada desempenho de classificador emparelhado com sua contraparte entre dois esquemas. Foram adotados os desempenhos provenientes dos cenários

com dados balanceados para comparar as métricas. Nas avaliações baseadas em engenharia de *features*, para *2x5-fold cross-validation*, cada classificador resulta em dez estimativas de desempenho. Assim, para o nível de confiança $alpha = 0,05$, $n = 10$ (correspondendo às execuções k-fold emparelhadas), o valor crítico exato de uma hipótese *two-tailed* é $T_{wc} = 8$ (SHESKIN, 2007). A comparação entre dois esquemas de classificadores é estatisticamente significativa se nosso $T_w \leq T_{wc}$.

Tabela 34 – Comparações inter-esquema via Wilcox - ($n = 10$, 95% CI). Fonte: Elaboração própria.

Algritmo	CTU-UHB					
	AUC por esquema					
	A	B	C	T_wAB	T_wAC	T_wBC
KNN	0.512	0.690	0.695	**0.00**	**0.00**	20.0
SVM	0.579	0.725	0.765	**2.00**	**3.00**	12.0
RF	0.541	0.706	0.758	**0.00**	**0.00**	15.0
MLP	0.552	0.700	0.714	**0.00**	**0.00**	19.0
GTB	0.526	0.660	0.739	**4.00**	**0.00**	1.00
BG	0.528	0.682	0.720	**2.00**	**0.00**	8.00
Algoritmo	HUFA					
	AUC por esquema					
	A	B	C	T_wAB	T_wAC	T_wBC
KNN	0.433	0.575	0.741	**3.00**	**0.00**	**0.00**
SVM	0.466	0.583	0.558	**0.00**	**3.50**	**7.00**
RF	0.583	0.600	0.708	12.0	**3.00**	10.5
MLP	0.450	0.566	0.666	12.5	**3.00**	**3.00**
GTB	0.500	0.500	0.500	27.5	27.5	27.5
BG	0.450	0.608	0.700	9.50	**0.00**	8.50
Algoritmo	DB-TRIUM					
	AUC por esquema					
	A	B	C	T_wAB	T_wAC	T_wBC
KNN	0.480	0.659	0.683	**3.00**	**0.00**	20.0
SVM	0.481	0.699	0.640	**1.00**	**5.00**	17.0
RF	0.527	0.637	0.644	**0.00**	**0.00**	25.0
MLP	0.538	0.679	0.672	**5.00**	**8.00**	27.0
GTB	0.541	0.624	0.621	**7.00**	**7.00**	25.0
BG	0.512	0.685	0.653	**0.00**	**1.00**	15.0

Esquemas: A (*First30*), B (*Full-data*), C (*Last30*). Valores em destaque representam diferença estatisticamente significante ($T_w \leq 8$).

A Tabela 34 apresenta a comparação entre esquemas por AUC para cada classificador e sua contraparte dos esquemas *firt30* (A), *full-data* (B) e *last30* (C). Nos conjuntos de dados CTU-UHB e DB-TRIUM, não há diferença significativa entre os esquemas *full-data* e *last30* (T_wBC). Todos os classificadores alcançaram significância estatística ao comparar os esquemas *first30* e *full-data* (T_wAB) assim como os esquemas *first30* e *last30* (T_wAC) nos conjuntos de dados CTU-UHB e DB-TRIUM. Para a base de dados HUFA, a diferença significativa pertence aos classificadores SVM e KNN em (T_wAB), e para SVM, KNN e MLP em (T_wBC).

A comparação entre os esquemas *first30* e *last30* (T_wAC) foi estatisticamente significativa para todos os classificadores, exceto para GTB, que teve um desempenho igualmente insatisfatório em todos os esquemas de segmentação para o conjunto de dados HUFA.

Para os conjuntos de dados CTU-UHB e DB-TRIUM, os escores dos classificadores apresentaram uma equivalência geral entre os esquemas *full-data* e *last30*. A avaliação do esquema de segmentação *first30* obteve um desempenho insatisfatório e diferiu do esquema *last30* (melhores desempenhos) na maioria das situações. O *last30* e a segmentação de intervalo *full-data* alcançaram os melhores desempenhos gerais. Assim, esses resultados apontaram que os esquemas *last30* e *full-data* possuem capacidade maior de discriminação dos estados fetais normal e patológico, sendo bons candidatos para modelagem prognóstica.

6.2.0.5 Comparação geral do desempenho dos classificadores

Para estimar a significância estatística geral das diferenças calculadas entre os desempenhos de cada par de classificador em todos os conjuntos de dados, foi empregado o teste de Friedman (FRIEDMAN, 1937) utilizando os valores de AUC. Não incluímos os esquemas *first30* devido às suas baixas pontuações gerais comprovadas nas comparações entre esquemas, evitando assim influenciar os resultados.

O desempenho de cada algoritmo para os esquemas de dados completos (*full-data*) e para os últimos 30 minutos de sinal (*last30*) foram comparados nas três bases de dados dos cenários balanceados. Usando um nível de confiança de 95%, com o valor crítico exato $F_c = 10,57$ da estatística de Friedman com hipótese *two-tailed* para (k = 6) e (n = 6), o teste resulta em $chi_F^2 = 20,11$. Como $chi_F^2 > F_c$, as escores *full-data* e *last30* em todas as bases de dados indicam que pelo menos um par de desempenhos entre classificadores obteve diferenças estatisticamente significativas. Foi realizada uma análise post-hoc para os pares a fim de apontar quais pares diferiram.

Na análise post-hoc, fez-se uso do teste de Nemenyi com um nível de confiança de 95% com valores de k = 6 classificadores e n = 6 (correspondendo a *full-data* e o *last30* esquemas para todos os três conjuntos de dados), calculou-se a diferença crítica para um teste de hipótese *two-tailed* = 3,075. Desta forma, a comparação de um par de desempenhos entre classificadores foi estatisticamente significativa se a diferença absoluta de suas escores médias correspondentes fosse maior que 3,075, conforme mostrado na Tabela 35.

De acordo com a Tabela 35, as comparações de pares de AUC entre classificadores

Tabela 35 – Post-hoc para o teste de Friedman. Fonte: Elaboração própria.

Dados	AUC					
	KNN	SVM	RF	MLP	GTB	BG
D1-B	0.690	0.725	0.706	0.700	0.660	0.682
D2-B	0.575	0.583	0.600	0.566	0.500	0.608
D3-B	0.659	0.699	0.637	0.679	0.624	0.685
D1-C	0.695	0.765	0.758	0.714	0.739	0.720
D2-C	0.741	0.558	0.708	0.666	0.500	0.700
D3-C	0.683	0.640	0.644	0.672	0.621	0.653
mean	0.674	0.662	0.675	0.666	0.607	0.675
Algoritmo	Comparativo entre AUCs					
	KNN	SVM	RF	MLP	GTB	BG
KNN	-	-	-	-	-	-
SVM	**3.70**	-	-	-	-	-
RF	2.77	0.92	-	-	-	-
MLP	1.85	**5.55**	**4.62**	-	-	-
GTB	**12.03**	**15.73**	**14.81**	**10.18**	-	-
BG	1.85	1.85	0.92	**3.70**	**13.88**	-

Bases de dados: D1 (CTU-UHB), D2 (HUFA) e D3 (DB-TRIUM). Esquemas: B (*Full-data*), C (*Last30*). Valores em destaque correspondem a diferenças estatisticamente significantes (soma das diferenças dos ranques absolutos que são > 3.075)

demonstram que SVM, RF, KNN e BG alcançaram a AUC de melhor desempenho geral, enquanto GTB alcançou a mais baixa. A diferença entre RF, SVM, KNN e GB não é estatisticamente significativa. A comparação pareada entre os resultados de MLP, KNN e BG também mostrou que eles não diferem estatisticamente. As comparações indicam que BG, SVM, KNN, MLP e RF tiveram desempenho semelhante, mas diferem do desempenho do GTB. Vale ressaltar que as escores gerais do GTB foram enfraquecidas em virtude do baixo desempenho na base de dados HUFA (D2-B, D2-C) que possui o menor número de registros.

6.2.0.6 Avaliação de bases cruzadas

O desempenho de generalização dos modelos foi avaliado treinando os classificadores em um conjunto de dados para posterior avaliação em cada base de dados remanescentes. As *features* selecionadas pelo RFE e os cenários de dados balanceados do esquema de segmentação *last30* foram adotados para este cenário. A Tabela 36 apresenta os resultados de cada classificador na avaliação *cross-dataset*.

É importante ressaltar que os três conjuntos de dados diferem em vários aspectos, como distribuição de classes (desequilíbrio), métodos de aquisição e número de amostras. Os conjuntos de dados CTU-UHB e HUFA empregaram os valores de pH para a rotulagem de classes. Por outro lado, o DB-TRIUM fez uso de rótulos baseados na opinião de especialistas. A avaliação dos classificadores *cross-dataset* no treinamento/avaliação das bases CTU-UHB/HUFA

Tabela 36 – Avaliaçao cross-dataset - cenários baseados em *features*. Fonte: Elaboração própria.

Trein./Aval.	Model	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
I/III	KNN	**0.655** (0.066)	**0.815** (0.078)	**0.507** (0.074)	**0.641** (0.069)	**0.661** (0.066)
	SVM	0.581 (0.016)	0.823 (0.035)	0.357 (0.000)	0.542 (0.011)	0.590 (0.017)
	RF	0.562 (0.022)	0.869 (0.035)	0.278 (0.038)	0.490 (0.032)	0.573 (0.022)
	MLP	0.599 (0.049)	0.838 (0.093)	0.378 (0.090)	0.556 (0.063)	0.608 (0.049)
	GTB	0.607 (0.024)	0.907 (0.046)	0.328 (0.057)	0.543 (0.420)	0.618 (0.024)
	BG	0.540 (0.033)	0.846 (0.034)	0.257 (0.057)	0.463 (0.050)	0.551 (0.033)
I/II	KNN	0.679 (0.032)	0.816 (0.042)	0.355 (0.026)	0.537 (0.026)	0.585 (0.026)
	SVM	**0.776** (0.010)	**0.923** (0.011)	**0.427** (0.017)	**0.628** (0.014)	**0.675** (0.011)
	RF	0.764 (0.009)	0.933 (0.016)	0.361 (0.040)	0.579 (0.030)	0.647 (0.016)
	MLP	0.738 (0.023)	0.915 (0.030)	0.317 (0.084)	0.532 (0.069)	0.616 (0.037)
	GTB	0.715 (0.012)	0.902 (0.017)	0.268 (0.037)	0.490 (0.032)	0.585 (0.017)
	BG	0.764 (0.014)	0.916 (0.016)	0.402 (0.027)	0.606 (0.022)	0.659 (0.016)
II/III	KNN	0.540 (0.041)	0.676 (0.046)	0.414 (0.069)	0.526 (0.052)	0.545 (0.040)
	SVM	**0.555** (0.074)	**0.884** (0.115)	**0.250** (0.250)	**0.310** (0.310)	**0.560** (0.067)
	RF	0.511 (0.036)	0.815 (0.051)	0.228 (0.053)	0.428 (0.053)	0.521 (0.036)
	MLP	0.477 (0.030)	0.692 (0.048)	0.278 (0.05)	0.436 (0.038)	0.485 (0.030)
	GTB	0.485 (0.030)	0.861 (0.046)	0.135 (0.038)	0.338 (0.051)	0.498 (0.030)
	BG	0.525 (0.032)	0.823 (0.060)	0.250 (0.047)	0.450 (0.044)	0.536 (0.032)
II/I	KNN	0.647 (0.024)	0.641 (0.027)	0.695 (0.021)	0.667 (0.016)	0.668 (0.015)
	SVM	0.804 (0.099)	0.847 (0.143)	0.414 (0.304)	0.510 (0.217)	0.631 (0.081)
	RF	0.713 (0.015)	0.727 (0.015)	0.590 (0.042)	0.654 (0.025)	0.658 (0.023)
	MLP	**0.706** (0.016)	**0.716** (0.021)	**0.617** (0.068)	**0.663** (0.032)	**0.667** (0.028)
	GTB	0.679 (0.009)	0.684 (0.011)	0.640 (0.027)	0.661 (0.013)	0.662 (0.012)
	BG	0.712 (0.011)	0.730 (0.014)	0.557 (0.031)	0.637 (0.015)	0.643 (0.012)
III/I	KNN	0.692 (0.095)	0.707 (0.115)	0.557 (0.109)	0.618 (0.032)	0.632 (0.031)
	SVM	**0.650** (0.123)	**0.646** (0.148)	**0.684** (0.105)	**0.652** (0.032)	**0.665** (0.022)
	RF	0.580 (0.041)	0.566 (0.049)	0.702 (0.034)	0.629 (0.015)	0.634 (0.012)
	MLP	0.573 (0.040)	0.565 (0.045)	0.645 (0.024)	0.603 (0.023)	0.605 (0.021)
	GTB	0.899 (0.000)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.500 (0.000)
	BG	0.529 (0.053)	0.505 (0.061)	0.745 (0.031)	0.611 (0.026)	0.625 (0.021)
III/II	KNN	0.637 (0.039)	0.717 (0.085)	0.444 (0.070)	0.559 (0.027)	0.581 (0.008)
	SVM	**0.710** (0.053)	**0.741** (0.091)	**0.634** (0.047)	**0.682** (0.026)	**0.688** (0.028)
	RF	0.672 (0.030)	0.696 (0.059)	0.614 (0.061)	0.651 (0.022)	0.655 (0.020)
	MLP	0.586 (0.019)	0.664 (0.026)	0.402 (0.075)	0.513 (0.046)	0.533 (0.031)
	GTB	0.704 (0.000)	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.500 (0.000)
	BG	0.638 (0.028)	0.669 (0.071)	0.565 (0.081)	0.610 (0.020)	0.617 (0.013)

I = CTU-UHB, II = DB-Trium, III = HUFA. Valores em destaque representam os melhores desempenhos.

mostrou alto desempenho para a SP, médio SE, e as escores AUC com desempenho melhor do que o acaso. O KNN obteve a maior AUC (0,661).

Na avaliação CTU-UHB/DB-TRIUM, os melhores desempenhos foram obtidos para o classificador SVM com AUC = 0,675. O treinamento e o teste para a avaliação *cross-dataset* DB-TRIUM/HUFA alcançaram sensibilidades muito baixas e pontuações AUC em torno do aleatório. A avaliação DB-TRIUM/CTU-UHB alcançou AUC = 0,667 para o classificador MLP. Na avaliação HUFA/CTU-UHB, o classificador SVM obteve o melhor resultado com AUC = 0,665. Na avaliação do HUFA/DB-TRIUM, o SVM também foi o melhor classificado, atingindo 0,688 para AUC.

Os resultados gerais dos cenários de avaliação *cross-dataset* mostraram a capacidade de classificar as classes de saudáveis (SP) e patológicas (SE) semelhantes às avaliações executadas nas bases de dados de forma isolada, exceto para avaliação DB-TRIUM/HUFA. Em geral, a AUC teve um desempenho melhor do que o aleatório e atingiu valores de AUC de até 0,688. Observamos um comportamento semelhante, mesmo ao realizar a avaliação de bases cruzadas do DB-TRIUM (rotulagem com base em especialistas) com o conjunto de dados rotulado com base em marcadores bioquímicos como o pH (CTU-UHB), atingindo assim 0,716, 0,617 e 0,667 para SP, SE e AUC, respectivamente. Com tal grau de diferença entre os dados disponíveis, os melhores desempenhos gerais na avaliação *cross-dataset* indicam uma capacidade de generalização aceitável, atingindo pontuações comparáveis àquelas alcançadas para os conjuntos de dados de forma isolada.

6.2.0.7 Hiper-parâmetros selecionados via grid-search

Nesta tese, os hiper-parâmetros para cada algoritmo de classificação foram selecionados dinamicamente de uma faixa de valores pré-definidos com uso de *grid-search*.

Tabela 37 – Hiper-parâmetros selecionados com frequência - esquema *last30*. Fonte: Elaboração própria.

Data	Classificador	Hiper-parâmetros
CTU-UHB	KNN	**n_neighbors**: [5]
	SVM	**Kernel**: ['rbf'], **C**: [10], **gamma**: [0.1]
	RF	**n_estimators**: [500], **criterion**:['gini'], **max_features**:['log2']
	MLP	**epochs**:[100], **batch_size**: [16]
	GTB	**max_iter**: [500]
	BG	**n_estimators**: [1000]
HUFA	KNN	**n_neighbors**: [5]
	SVM	**Kernel**: ['rbf'], **C**: [10], **gamma**: [0.1]
	RF	**n_estimators**: [100], **criterion**:['gini'], **max_features**:['log2']
	MLP	**epochs**:[100], **batch_size**: [32]
	GTB	**max_iter**: [250]
	BG	**n_estimators**: [100]
DB-Trium	KNN	**n_neighbors**: [15]
	SVM	**Kernel**: ['rbf'], **C**: [10], **gamma**: [0.1]
	RF	**n_estimators**: [100], **criterion**:['gini'], **max_features**:['sqrt']
	MLP	**epochs**:[100], **batch_size**: [16]
	GTB	**max_iter**: [250]
	BG	**n_estimators**: [1000]

A Tabela 37 apresenta os hiper-parâmetros selecionados com mais frequência nas rodadas do *k-fold* para cada classificador nas bases CTU-UHB, HUFA e DB-Trium para o esquema de segmentação *last30*, o qual obteve as melhores performances entre os cenários de avaliação. É importante ressaltar que objetivando a generalização em diferentes bases de dados,

o *grid-search* foi incluído em tempo de execução como parte do processo de treinamento dos modelos.

É possível perceber que apesar da semelhança entre os valores mais recorrentes apresentados, a base de dados HUFA apresenta valores distintos para *n_estimators* (RF e BG), assim como um valor de *batch_size* bem próximo ao seu numero total de registros, demonstrando o uso da estratégia *batch* no processo de treinamento (uso de todos os exemplo de treino a cada época), em contrapartida as bases CTU-UHB e DB-Trium se utilizaram de *bacth_sizes* pequenos, caracterizando a estratégia *mini-batch* (pequenas porções do treinamento a cada época). A variação dos valores selecionados nos hiper-parâmetros entre diferentes bases de dados indica relevância da sua seleção de forma dinâmica no processo de treinamento dos modelos.

6.2.0.8 Avaliação dos registros pertencentes a área cinza

Esta tese adotou bases de dados com rotulagens baseadas em valores de pH (CTU-UHB e HUFA) em conjunto com bases rotuladas por especialistas (DB-Trium). Objetivando a generalização dos modelos, o processo de rotulagem baseado nos valores de pH empregou limites inferiores (< 7,05) e superiores (> 7,20) para caracterizar os registros em classes normais e patológicas, desta forma, os registros com valores intermediários ($\geq 7,05$ e $\leq 7,20$) que se referem a categoria suspeita ou *grey zone*, foram removidos dos cenários de avaliação.

No entanto, é importante verificar quais rótulos são atribuídos a registros pertencentes a *grey zone* quando avaliados por modelos treinados nas categorias normal e patológico. A Tabela 38 apresenta uma visão geral dos registros pertencentes a *grey zone* de cada base de dados nos cenários de segmentação *last30*.

A Tabela 39 apresenta detalhes da distribuição dos rótulos estimados no conjunto de registros referentes a *grey zone*. O classificador SVM e o cenário de segmentação de dados *last30* foram empregados no processo de avaliação.

O valor de pH = 7,15 é amplamente empregado na literatura para separação das classes normais e patológicas com limiar único, desta forma os valores de pH < 7,15 são marcados como patológicos e valores $\geq 7,15$ são rotulados como normais (COMERT; KOCAMAZ, 2018; COMERT *et al.*, 2018). A base CTU-UHB apresentou 154 (cento e cinquenta e quatro) registros marcados como suspeitos (*grey zone*), os quais foram estimados pelo modelo binário em 98 registros normais e 56 registros patológicos (N = 98, P = 56). Adotando o limiar de pH = 7,15 para avaliar os grupos estimados separadamente, observou-se que entre os registros rotulados

Tabela 38 – Visão geral dos registros pertencentes a *grey zone*. Fonte: Elaboração própria.

CTU-UHB (**154 registros**)				
Cenário	*Feature*	Média (desvio padrão)	Max.	Min.
last30	meanAD	16.31 (4.86)	32.71	5.80
	FD_Higushi	1.17 (0.07)	1.39	1.06
	TRI	10.74 (4.70)	27.06	3.11
	NN20	272.46 (159.77)	862.00	17.00
	pNN20	3.78 (2.21)	11.97	0.23
	CVNN	0.20 (0.05)	0.35	0.06
	pH	7.14 (0.03)	7.20	7.05
HUFA (**5 registros**)				
Cenário	*Feature*	Média (desvio padrão)	Max.	Min.
last30	meanAD	18.20 (5.15)	25.46	14.13
	FD_Higushi	1.11 (0.03)	1.16	1.08
	TRI	13.45 (4.03)	18.84	9.51
	NN20	250.80 (105.30)	391.00	101.00
	pNN20	3.48 (1.46)	5.43	1.40
	CVNN	0.21 (0.06)	0.30	0.13
	pH	7.09 (0.06)	7.17	7.05

Tabela 39 – Classificação da *grey zone*. Fonte: Elaboração própria.

CTU-UHB (**154 registros**)	
Algoritmo	Distribuição das classes estimadas
SVM	(N = 98, P = 56) N = 38 (pH < 7.15), N = 60 (pH $\geq$ 7.15) P = 27 (pH < 7.15), P = 29 (pH $\geq$ 7.15)
HUFA (**5 registros**)	
Algoritmo	Distribuição das classes estimadas
SVM	(N = 3, P = 2) N = 2 (pH < 7.15), N = 1 (pH $\geq$ 7.15) P = 1 (pH < 7.15), P = 1 (pH $\geq$ 7.15)

N = normal, P = patológico.

como normais, 38 apresentaram valores de pH < 7,15 e 60 apresentaram pH $\geq$ 7,15, indicando adequação do modelo para de rótulos saudáveis na *grey zone*. Por outro lado, entre os registros estimados como patológicos, 27 apresentaram valores de pH < 7,15 e 29 apresentaram pH $\geq$ 7,15, indicando capacidade mediana ao estimar rótulos patológicos entre os registros suspeitos.

A base de dados HUFA apresentou 5 (cinco) registros marcados como suspeitos, os quais foram estimados como 3 normais e 2 patológicos (N = 3, P = 2). Na análise pelo limiar de pH = 7,15 observou-se que entre os registros rotulados como normais, 2 apresentaram valores de pH < 7,15 e 1 (um) apresentou pH $\geq$ 7,15. Os registros rotulados como patológicos se dividiram em 1 (um) com pH < 7,15 e 1 (um) valor de pH $\geq$ 7,15.

De forma geral, dado a grande diferença no número de registros, a avaliação dos rótulos suspeitos na base CTU-UHB apresentou melhor desempenho que na base HUFA. Os resultados na base CTU-UHB apresentaram 67 (38 + 29) estimativas incorretas e 87 (60 + 27) estimativas corretas, indicando valores promissores na estimativa da *grey zone*.

6.3 Síntese do Capítulo

Um conjunto de *features* foi extraído dos sinais FCF nos domínios morfológico, linear, de frequência, não linear e baseados em variabilidade de frequência cardíaca (VFC) adulta. Seis *features* foram selecionadas usando o algoritmo RFE. O conjunto resultante foi usado para modelar e avaliar o desempenho dos classificadores KNN, SVM, RF, MLP, BG e GTB nos esquemas de segmentação de dados *first30*, *full-data* e *last30*. O desempenho de generalização do modelo foi avaliado por treinamento e validação *cross-dataset*.

A distribuição de classes dos conjuntos de dados mostrou-se não balanceada em favor das classes saudáveis em diferentes níveis. Desta forma, a avaliação em dados não balanceados resultou em baixa sensibilidade (SE) e alta especificidade (SP). O algoritmo SMOTE foi utilizado para equilibrar a distribuição das classes saudáveis e patológicas nas parcelas de treinamento, resultando em uma pequena diminuição na SP com um aumento na SE. O esquema de segmentação de dados *first30* apresentou desempenho geral baixo, mesmo na versão de dados balanceados, indicando capacidade limitada de separação de classes em relação à proximidade da hora do parto. Os escores dos classificadores apresentaram equivalência geral entre os esquemas de segmentação de dados completos (*full-data*) e os últimos 30 minutos (*last30*), sendo estes os de melhores desempenhos em nossas avaliações. Os resultados apresentados nesta tese referentes a avaliação do intervalo de sinal mais representativo parece estar em conformidade com os resultados de Petrozziello *et al.* (PETROZZIELLO *et al.*, 2019) que obtiveram modelos de melhor desempenho nos últimos 60 minutos de CTG, independentemente da estágio do processo de parto.

Consideramos o equilíbrio entre sensibilidade e especificidade nas comparações de desempenho e avaliação das melhores combinações dos blocos de construção. Exceto para os resultados *full-data* na base de dados DB-TRIUM, todas as melhores pontuações foram obtidas nos esquemas referentes aos últimos 30 minutos (*last30*). No conjunto de dados CTU-UHB, o classificador de melhor desempenho foi o SVM com 0,856 para Especificidade, 0,675 para Sensibilidade, 0,755 para Média Geométrica e 0,765 para AUC.

Para o HUFA, o melhor resultado foi obtido utilizando KNN com 0,750 para Especificidade, 0,733 para Sensibilidade, 0,704 para Média Geométrica e 0,741 para AUC. No DB-TRIUM, o classificador de melhor desempeno foi o SVM no esquema *full-data* com 0,761 para Especificidade, 0,637 para Sensibilidade, 0,686 para Média Geométrica e 0,699 para AUC.

É importante ressaltar que comparando o desempenho do classificador individual

em todos os conjuntos de dados nos esquemas *full-data* e *last30*, o BG, SVM, KNN, MLP e RF atingiram desempenhos semelhantes estatisticamente, e todos diferiram do classificador GTB.

Na avaliação da capacidade de generalização dos modelos via *cross-dataset*, os melhores resultados para treinamento/validação no CTU-UHB/HUFA foram KNN com 0,815 para Especificidade, 0,507 para Sensibilidade, 0,641 para Média Geométrica e 0,661 para AUC. Para o CTU-UHB/DB-TRIUM foi o cassificador SVM com 0,923 para Especificidade, 0,427 para Sensibilidade, 0,628 para Média Geométrica e 0,675 para AUC. No DB-TRIUM/CTU-UHB, o de melhor desempenho foi o MLP com 0,716 para Especificidade, 0,617 para Sensibilidade, 0,663 para Média Geométrica e 0,667 para AUC. Na avaliação do classificador HUFA/CTU-UHB, o melhor resultado foi o SVM com 0,646 para Especificidade, 0,684 para Sensibilidade, 0,652 para Média Geométrica e 0,665 para AUC. Por fim, para HUFA/DB-TRIUM, o classificador de melhor desempenho foi o SVM com 0,741 para Especificidade, 0,634 para Sensibilidade, 0,682 para Média Geométrica e 0,688 para AUC.

7 MODELAGEM SEMI-SUPERVISIONADA

Neste Capítulo, uma abordagem semi-supervisionada para avaliação do estado fetal é explorada. Os materiais e métodos são apresentados e, em seguida, os resultados da avaliação. É apresentada em primeira mão uma análise da base de dados DB-HeraBeat para o reconhecimento do bem-estar fetal baseado em *features* provenientes dos biossinais, onde são sumarizados os resultados do modelo semi-supervisionado como parte integrante dos blocos de construção para avaliação do bem-estar fetal.

A Figura 15 apresenta em destaque o esquema empregado neste Capítulo em relação aos modelos desenvolvidos nesta Tese.

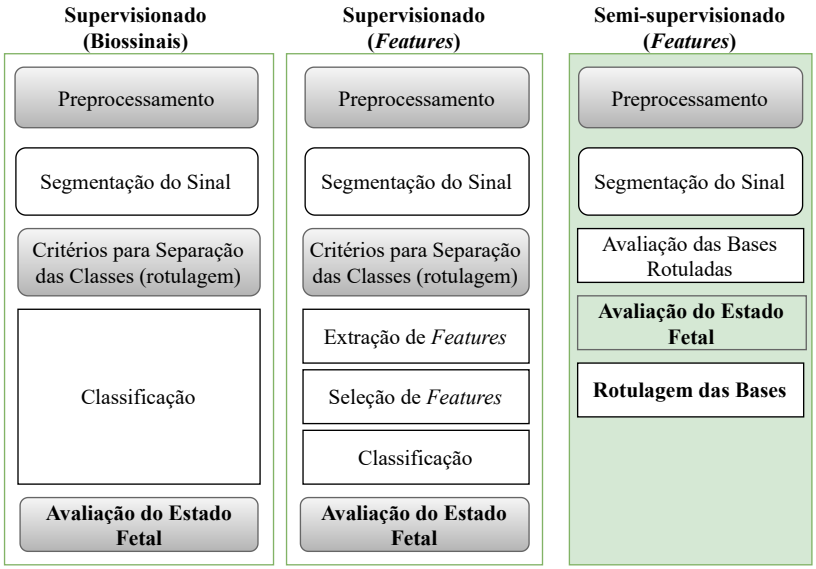


Figura 15 – Modelo semi-supervisionado (em destaque). Fonte: Elaboração própria.

7.1 Materiais e Métodos para Modelos Semi-supervisionados

No aprendizado semi-supervisionado, parte das amostras não possuem rótulos. Desta forma, foram usadas as bases de dados rotuladas com maior número de registros CTU-UHB e DB-Trium para rotular as bases não anotadas DB-HeraBeat e SpAM *dataset*.

Foi empregado o algoritmo semi-supervisionado *LabelSpreading* (ZHOU *et al.*, 2004) no processo de rotulagem. Para tal, foi elaborado um conjunto de cenários a fim de avaliar primeiramente o desempenho do método nas bases rotuladas para posterior aplicação na rotulagem propriamente dita, conforme apresentado nas subseções seguintes.

7.1.1 Avaliação das bases de dados rotuladas de forma isolada

Neste cenário, cada base rotulada CTU-UHB, DB-Trium e HUFA foi usada de forma isolada em uma auto avaliação semi-supervisionada. Seis *features* selecionadas com o RFE nos cenários de validação prévios foram usadas em conjunto com os esquemas de segmentação *first30*, de dados completos (*full-data*) e o *last30*. A Figura 16 apresenta uma visão geral da auto-avaliação semi-supervisionada nas bases de dados rotuladas.

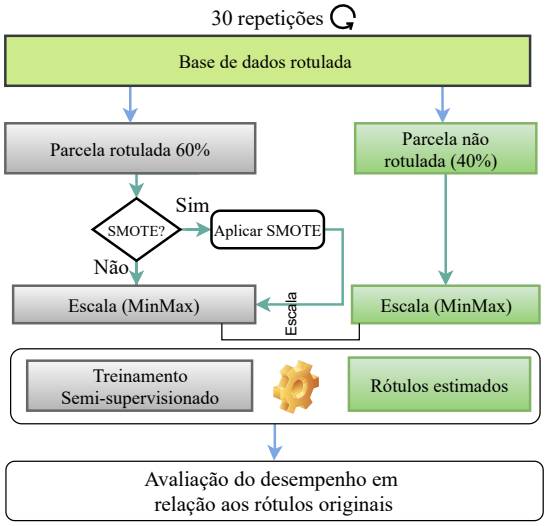


Figura 16 – Cenário de avaliação de bases rotuladas isoladamente. Fonte: Elaboração própria.

As bases de dados rotuladas foma divididas de forma estratificada em uma parcela para treinamento e outra para avaliação do modelo semi-supervisionado nas proporções 60/40. Em seguida, os rótulos da parcela de validação foram removidos a fim de permitir um ambiente de aprendizagem semi-supervisionado. Empregou-se a técnica SMOTE para balanceamento dos dados apenas na parcela referente aos dados rotulados (isolando a parcela de avaliação e evitando viés no processo), e em seguida, foi aplicada a normalização *Min-Max* na parcela rotulada para redução da amplitude dos valores em [0,1]. A escala foi reaplicada na parcela de avaliação para compatibilidade entre treinamento e avaliação.

O algoritmo *LabelSpreading* foi usado para rotular os dados de avaliação baseando-se na parcela rotulada da mesma base de dados. Por fim, foram computadas as métricas de acurácia, especificidade, sensibilidade, AUC e média geométrica para avaliar o desempenho entre os rótulos providos pelo algoritmo semi-supervisionado e os rótulos reais da parcela de validação. Todo o processo foi repetido em um total de 30 (trinta) vezes devido a variações provenientes dos parâmetros de inicialização em diferentes algoritmos de classificação empregados, obtendo

assim uma representatividade estatística. Por fim, foram computadas as médias e desvios-padrão para cada métrica de avaliação.

7.1.2 Avaliação das bases de dados rotuladas de forma cruzada

Na avaliação da abordagem semi-supervisionada de forma cruzada, foram empregadas as bases de dados rotuladas CTU-UHB, HUFA e DB-Trium para estimar o grau de certeza da avaliação do estado fetal entre bases de dados distintas. Neste cenário, optou-se por não utilizar o esquema de segmentação *first30* devido aos seus baixos desempenhos. Assim, fez-se uso apenas dos esquemas de dados completos (*full-data*) e *last30*. Foram utilizados ainda as 6 (seis) *features* provenientes da seleção via RFE. A Figura 17 apresenta uma visão geral do cenário para as bases de dados cruzadas.

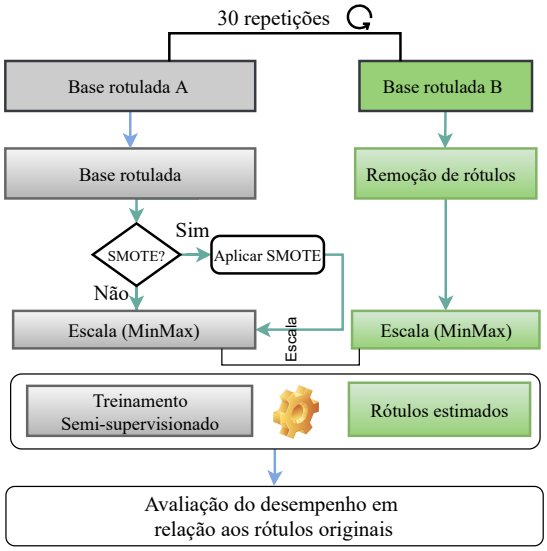


Figura 17 – Cenário de avaliação de bases rotuladas cruzadas. Fonte: Elaboração própria.

Dado um par de bases de dados rotuladas distintas, selecionou-se uma delas para validação, a qual teve seus rótulos removidos visando um ambiente de aprendizagem semi-supervisionado. Fez-se uso da técnica SMOTE para balanceamento dos dados apenas na parcela referente aos dados rotulados (isolando a parcela de validação e evitando viés), e em seguida foi aplicada a normalização *Min-Max* na parcela rotulada para redução da amplitude para valores entre [0,1]. A escala foi reaplicada na parcela de validação para compatibilidade entre as bases.

O algoritmo *LabelSpreading* foi usado para rotular os dados de validação tendo por parâmetro os rótulos da base que não sofreu alteração. Por fim, as métricas acurácia, especificidade, sensibilidade, AUC e média geométrica foram computadas para avaliar o desempenho

entre os rótulos providos pelo algoritmo semi-supervisionado e os rótulos originais da base dados de validação. Todo o processo foi repetido em um total de 30 (trinta) vezes a fim de obter uma representatividade estatística. Por fim, computamos as médias e desvios-padrão de cada métrica.

O cenário mencionado foi executado de forma que cada uma das bases de dados rotuladas, CTU-UHB, HUFA e DB-Trium, fosse aplicada pelo menos uma vez para rotulagem e validação das bases remanescentes nos esquemas de segmentação *full-data* e *last30*.

7.1.3 Cenário semi-supervisionado em bases não rotuladas

Neste cenário, fez-se uso das bases de dados rotuladas com maior número de registros DB-Trium e CTU-UHB individualmente para estimar os rótulos das bases de dados não rotuladas SpAM e DB-HeraBeat. Foram empregadas as seis *features* selecionadas via RFE e o esquema de segmentação *last30*, o qual demonstrou melhor capacidade para separação de classes nos cenários de avaliação anteriores. A Figura 18 apresenta uma visão geral deste cenário de avaliação.

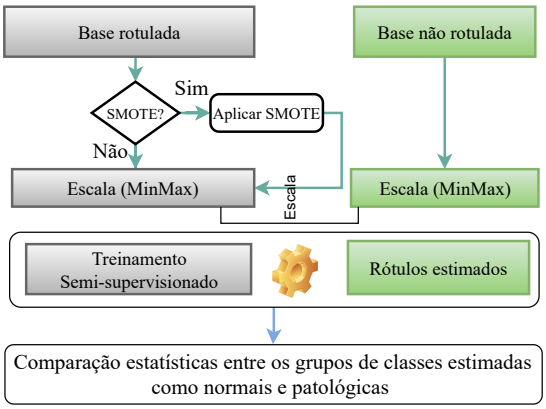


Figura 18 – Cenário em bases não rotuladas. Fonte: Elaboração própria.

O algoritmo *LabelSpreading* foi usado para estimar os rótulos de cada base não rotulada, DB-HeraBeat e SpAM, tendo por parâmetro os dados e rótulos das bases DB-Trium e CTU-UHB. Por tratar-se de um cenário semi-supervisionado real, ou seja, as bases de avaliação não possuem rótulos originais para comparação do grau de certeza por meio de métricas de desempenho, o teste estatístico Mann-Whitney U (MANN; WHITNEY, 1947) foi adotado para comparar as distribuições resultantes entre os rótulos estimados para cada *feature* usada. Este teste avalia o grau de entrelaçamento dos dados dos dois grupos após a ordenação onde a maior separação dos dados no conjunto indica que as amostras são distintas, rejeitando-se a hipótese de igualdade das medianas.

7.2 Resultados da Modelagem Semi-supervisionada

Esta seção apresenta os resultados obtidos em cada cenário semi-supervisionado. Primeiramente, são apresentados os resultados da avaliação de dados rotulados de forma isolada, em seguida são exibidos os resultados da abordagem entre bases cruzadas e por fim, os resultados dos cenários semi-supervisionados aplicados às bases não rotuladas são sumarizados.

7.2.1 Bases de dados rotuladas de forma isolada

São apresentadas na Tabela 40 as médias e desvios-padrão para acurácia, especificidade, sensibilidade, média geométrica e AUC provenientes das 30 (trinta) repetições do processo de auto-avaliação semi-supervisionada de cada base de dados rotulada CTU-UHB, DB-Trium e HUFA para os três esquemas de segmentação de dados: *first30*, *full-data* e *last30*.

Tabela 40 – Avaliação individual de bases rotuladas. Fonte: Elaboração própria.

Esquema	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
			CTU-UHB		
first30	0.707 (0.071)	0.740 (0.088)	0.408 (0.130)	0.539 (0.067)	0.574 (0.048)
full-data	0.774 (0.046)	0.786 (0.060)	0.662 (0.117)	0.716 (0.051)	0.724 (0.044)
last30	0.843 (0.028)	**0.871** (0.034)	**0.591** (0.110)	**0.714** (0.064)	**0.731** (0.052)
			DB-Trium		
first30	0.415 (0.054)	0.314 (0.184)	0.652 (0.228)	0.404 (0.078)	0.483 (0.054)
full-data	0.693 (0.065)	0.762 (0.115)	0.528 (0.138)	0.623 (0.070)	0.645 (0.057)
last30	**0.747** (0.038)	**0.823** (0.075)	**0.566** (0.142)	**0.674** (0.067)	**0.695** (0.052)
			HUFA		
first30	0.533 (0.161)	0.513 (0.309)	0.55 (0.248)	0.463 (0.213)	0.531 (0.164)
full-data	0.551 (0.141)	0.793 (0.161)	0.350 (0.228)	0.487 (0.197)	0.571 (0.136)
last30	0.663 (0.104)	**0.740** (0.211)	**0.600** (0.203)	**0.637** (0.129)	**0.670** (0.103)

Topo: CTU-UHB , Meio: DB-Trium, Base: HUFA. Valores em destaque representam os melhores desempenhos para as métricas ao se adotar os rótulos originais como linha de base.

Uma visão análise de alto nível destes resultados indica um comportamento similar aos resultados obtidos na avaliação supervisionada, onde os esquemas *first30* obtiveram os piores resultados e os esquemas *full-data* e *last30* alcançaram melhores desempenhos, obtendo valores similares entre si.

Os valores em destaque na Tabela 40 indicam que, para as três bases de dados, os melhores resultados foram obtidos no esquema de segmentação *last30* onde obteve-se AUCs com desempenho acima da classificação aleatória. O melhor valor de AUC foi 0.731, sendo obtido na auto-avaliação da base de dados CUT-UHB em seu esquema *last30*.

7.2.2 Bases de dados rotuladas de forma cruzada

Os resultados da avaliação das bases rotuladas em cenários semi-supervisionados de forma cruzada são apresentados na Tabela 41, onde são exibidas as médias e desvios padrão da acurácia, especificidade, sensibilidade, média geométrica e AUC provenientes das 30 (trinta) repetições no processo de avaliação. Neste cenário, foram avaliadas as bases de dados rotuladas CTU-UHB, DB-Trium e HUFA de forma cruzada apenas nos esquemas de segmentação *full-data* e *last30*. Assumiu-se valores zero para os desvios padrão (std = 0.00) onde ($std < 1.00 \times 10^{-14}$).

Tabela 41 – Avaliação cruzada (Base rotulada/base de validação). Fonte: Elaboração própria.

Data	Acurácia mean (std)	Especificidade mean (std)	Sensibilidade mean (std)	Gmean mean (std)	AUC mean (std)
			CTU/HUFA		
full-data	0.555 (0.00)	0.923 (0.00)	0.214 (0.00)	0.444 (0.00)	0.568 (0.00)
last30	**0.629** (0.00)	**0.846** (0.00)	**0.428** (0.00)	**0.602** (0.00)	**0.637** (0.00)
			CTU/DB-Trium		
full-data	0.679 (0.00)	0.937 (0.00)	0.063 (0.00)	0.244 (0.00)	0.500 (0.00)
last30	**0.805** (0.00)	**0.946** (0.00)	**0.468** (0.00)	**0.665** (0.00)	**0.707** (0.00)
			DB-Trium/CTU		
full-data	0.746 (0.00)	0.770 (0.00)	0.525 (0.00)	0.636 (0.00)	0.647 (0.00)
last30	**0.783** (0.00)	**0.801** (0.00)	**0.625** (0.00)	**0.707** (0.00)	**0.713** (0.00)
			DB-Trium/HUFA		
full-data	0.592 (0.00)	0.538 (0.00)	0.642 (0.00)	0.588 (0.00)	0.590 (0.00)
last30	**0.592** (0.00)	**0.769** (0.00)	**0.428** (0.00)	**0.574** (0.00)	**0.598** (0.00)
			HUFA/CTU		
full-data	0.306 (0.00)	0.234 (0.00)	0.949 (0.00)	0.472 (0.00)	0.592 (1×10^{-16})
last30	**0.582** (0.00)	**0.575** (0.00)	**0.649** (0.00)	**0.611** (0.00)	**0.612** (0.00)
			HUFA/DB-Trium		
full-data	0.517 (0.00)	0.589 (0.00)	0.340 (0.00)	0.447 (0.00)	0.464 (0.00)
last30	**0.641** (0.00)	**0.723** (0.00)	**0.446** (0.00)	**0.568** (0.00)	**0.585** (0.00)

Valores em destaque representam os melhores desempenhos para as métricas ao se adotar os rótulos originais como linha de base.

Pela Tabela 41, é possível perceber que as avaliações de desempenho nos esquemas de segmentação *last30* alcançaram os melhores resultados para todos os cenários de bases de dados rotuladas de forma cruzada. Os piores resultados são provenientes, de uma forma geral, da rotulagem pela base HUFA, a qual possui o menor número de registros e mostrou-se ineficiente na replicação dos rótulos de forma cruzada.

Os resultados apresentam valores elevados para Especificidade, indicando alta capacidade na estimativa dos rótulos normais. E valores de Sensibilidades médios, os quais indicam a capacidade mediana em reconhecer os rótulos patológicos. Os resultados que apresentaram melhor equilíbrio entre o reconhecimento de rótulos normais e patológicos pertenceram à utilização das bases DB-Trium e CTU-UHB, em ambos cenários de rotulagem e avaliação.

A Tabela 41 nos mostra ainda que todos os cenários semi-supervisionados de forma cruzada apresentaram valores de AUC superiores ao aleatório, onde o melhor desempenho geral foi proveniente da aplicação dos rótulos da base DB-Trium para estimar os rótulos da base CTU-UHB, alcançando AUC = 0.713 no esquema de segmentação *last30*.

7.2.3 Bases não rotuladas

Neste cenário, a abordagem semi-supervisionada empregou as bases de dados rotuladas com maior quantidade de registros, CTU-UHB e DB-Trium, na estimativa dos rótulos das bases não rotuladas DB-HeraBeat e SpAM. Para tal, fez-se uso do esquema de segmentação *last30* e as *features* obtidas via RFE.

São apresentadas na Tabela 42 as médias e desvios-padrão computadas para as *feature* após a estimativas de rótulos em normal e patológico da base de dados SpAM. É exibido ainda o valor de *p* referente à aplicação do teste estatístico Mann-Whitney U (MANN; WHITNEY, 1947) para hipótese alternativa H_1 *two-tailed* de que os valores em cada grupo diferem estatisticamente, indicando assim um certo nível de separabilidade entre os grupos de rótulos estimados.

Tabela 42 – Valores de Média (desvio padrão) por grupo normal e patológico para base SpAM. Fonte: Elaboração própria.

CTU-UHB/SpAM (N = 209, P = 091)				
Cenário	Feature	Normal Média (desvio padrão)	Patológico Média (desvio padrão)	*p*
last30	meanAD	12.72 (5.72)	19.90 (6.61)	$\mathbf{5.69 \times 10^{-17}}$
	FD_Higushi	1.23 (0.10)	1.21 (0.07)	0.197
	TRI	8.93 (4.54)	11.24 (5.19)	$\mathbf{0.3 \times 10^{-4}}$
	NN20	211.01 (171.00)	238.40 (164.45)	0.136
	pNN20	2.93 (2.37)	3.31 (2.28)	0.136
	CVNN	0.155 (0.070)	0.243 (0.074)	$\mathbf{1.61 \times 10^{-17}}$
DB-Trium/SpAM (N = 164, P = 136)				
Cenário	Feature	Normal Média (desvio padrão)	Patológico Média (desvio padrão)	*p*
last30	meanAD	10.93 (4.30)	19.68 (6.26)	$\mathbf{4.96 \times 10^{-31}}$
	FD_Higushi	1.25 (0.11)	1.20 (0.05)	$\mathbf{0.3 \times 10^{-4}}$
	TRI	7.71 (3.73)	11.95 (5.06)	$\mathbf{2.39 \times 10^{-15}}$
	NN20	143.73 (114.86)	310.47 (179.44)	$\mathbf{1.82 \times 10^{-19}}$
	pNN20	1.99 (1.59)	4.31 (2.49)	$\mathbf{1.82 \times 10^{-19}}$
	CVNN	0.134 (0.05)	0.240 (0.230)	$\mathbf{3.66 \times 10^{-30}}$

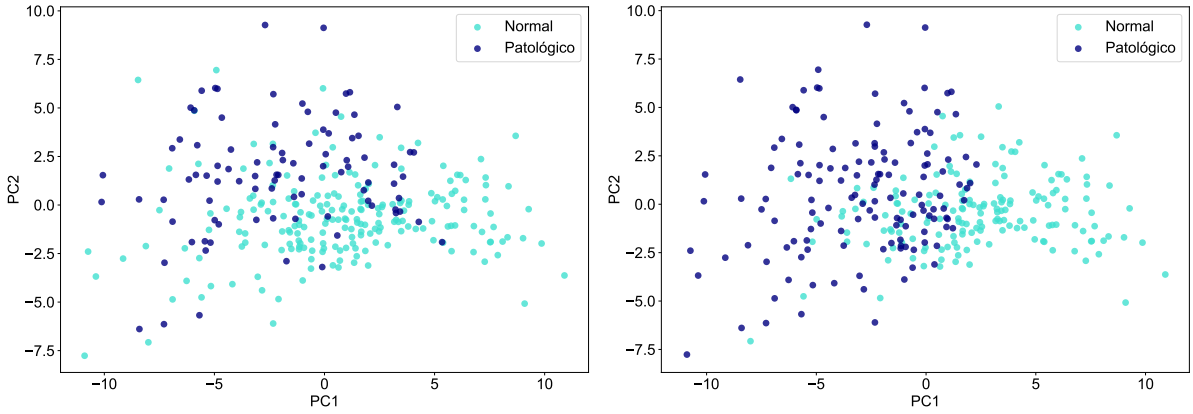
TRI = Triangular *index*, N = normal, P = patológicos. Os itens em destaque representam $p < 0.05$.

Conforme apresentado na Tabela 42, os rótulos da base de dados SpAM que foram estimados via base de dados CTU-UHB se distribuíram em 209 (duzentos e nove) registros normais e 91 (noventa e um) registros patológicos. A comparação entre os grupos normais

e patológicos com os rótulos estimados apresentaram diferença estatisticamente significante ($p < 0.05$) para as *features* meanAD, triangular index (TRI) e CVNN. Por outro lado, as *features* FD_Higushi, NN20 e pNN20 não apresentaram diferença estatística entre os grupos, indicando baixo nível de separabilidade ao empregar o CTU-UHB como rotulador.

Os rótulos da base de dados SpAM se distribuíram em 164 (cento e sessenta e quatro) registros normais e 136 (cento e trinta se seis) registros patológicos ao se adotar a base de dados DB-Trium como rotulador. Neste cenário semi-supervisionado, todas as *features* apresentaram diferença estatisticamente significante ($p < 0.05$) entre os grupos normal e patológico.

A Figura 19 apresenta uma visualização em alto nível da distribuição entre os grupos normais e patológicos dos rótulos estimados para a base de dados SpAM utilizando CTU-UHB e DB-Trium como rotuladores. Para tal, os dados foram reduzidos em duas dimensões (componentes) adotando a técnica *Principal Component Analysis* (PCA) para fins de visualização da distribuição final dos rótulos estimados.



(a) Estimados com CTU-UHB (b) Estimados com DB-Trium

Figura 19 – Base de dados SpAM reduzida a dois componentes. Fonte: Elaboração própria.

Para a base de dados DB-HeraBeat, são apresentados na Tabela 43 as médias e desvios-padrão para cada *feature* após as estimativas de rótulos em normal e patológico utilizando-se as bases CTU-UHB e DB-Trium como rotuladores no esquema de segmentação *last30*. Os valores de p referentes à aplicação do teste estatístico Mann-Whitney U para hipótese alternativa H_1 *two-tailed* de que os valores em cada grupo diferem estatisticamente também são apresentados.

Os rótulos estimados para a base de dados DB-HeraBeat via base de dados CTU-UHB se distribuíram em 1511 (um mil quinhentos e onze) registros normais e 8 (oito) registros patológicos. Na comparação entre os grupos normais e patológicos com os rótulos estimados,

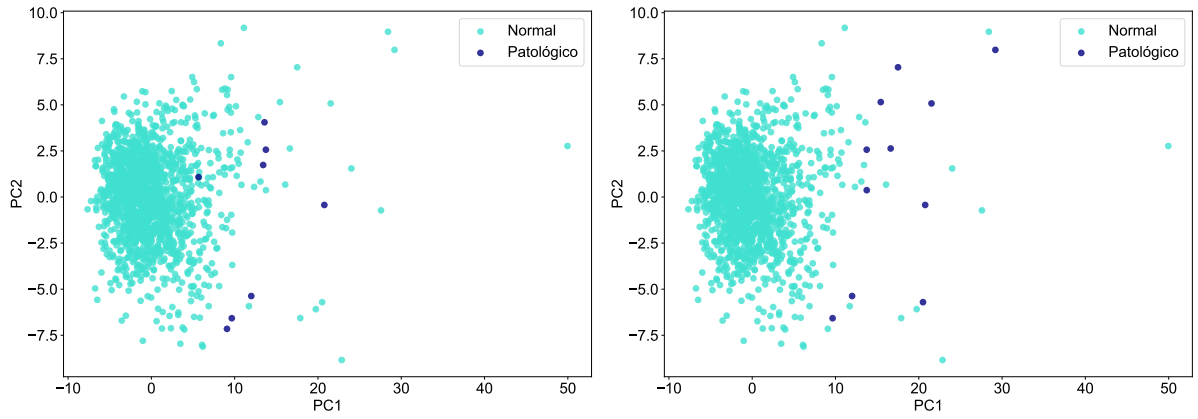
Tabela 43 – Valores de média (desvio padrão) por grupo normal e patológico para base DB-HeraBeat. Fonte: Elaboração própria.

CTU-UHB/DB-HeraBeat (N = 1511, P = 008)				
Cenário	Feature	Normal Média (desvio padrão)	Patológico Média (desvio padrão)	*p*
last30	meanAD	5.51 (2.25)	11.01 (5.05)	$\mathbf{0.2 \times 10^{-4}}$
	FD_Higushi	1.06 (0.014)	1.05 (0.012)	0.254
	TRI	5.03 (1.56)	7.45 (2.91)	**0.012**
	NN20	5.50 (13.98)	26.87 (23.58)	$\mathbf{0.4 \times 10^{-4}}$
	pNN20	0.076 (0.194)	0.373 (0.327)	$\mathbf{0.4 \times 10^{-4}}$
	CVNN	0.051 (0.019)	0.143 (0.019)	$\mathbf{1.33 \times 10^{-6}}$
DB-Trium/DB-HeraBeat (N = 1508, P = 011)				
Cenário	Feature	Normal Média (desvio padrão)	Patológico Média (desvio padrão)	*p*
last30	meanAD	5.48 (2.17)	14.19 (3.22)	$\mathbf{2.2 \times 10^{-8}}$
	FD_Higushi	1.06 (0.014)	1.04 (0.014)	$\mathbf{1.62 \times 10^{-5}}$
	TRI	5.02 (1.56)	7.37 (1.67)	$\mathbf{4.56 \times 10^{-5}}$
	NN20	5.13 (12.45)	71.09 (46.13)	$\mathbf{4.6 \times 10^{-7}}$
	pNN20	0.07 (0.173)	0.98 (0.64)	$\mathbf{4.6 \times 10^{-7}}$
	CVNN	0.051 (0.019)	0.132 (0.015)	$\mathbf{1.78 \times 10^{-8}}$

TRI = Triangular *index*, N = normal, P = patológicos. Os itens em destaque representam $p < 0.05$.

apenas FD_Higushi não apresentou diferença estatisticamente significante ($p < 0.05$).

Os rótulos da base de dados DB-HeraBeat se distribuíram em 1508 (um mil quinhentos e oito) registros normais e 11 (onze) registros patológicos ao se adotar a base de dados DB-Trium como rotulador. Neste cenário semi-supervisionado, todas as *features* apresentaram diferença estatisticamente significante ($p < 0.05$) entre os grupos normal e patológico.



(a) Estimados com CTU-UHB (b) Estimados com DB-Trium

Figura 20 – Base de dados DB-HeraBeat reduzida a dois componentes. Fonte: Elaboração própria.

A Figura 20 apresenta a distribuição entre os grupos normal e patológico dos rótulos estimados para a base de dados DB-HeraBeat ao se utilizarem as bases de dados CTU-UHB e DB-Trium como rotuladores. Para fins de visualização, os dados foram reduzidos a duas

dimensões (componentes) adotando a técnica *Principal Component Analysis* (PCA).

É fácil perceber pelas tabelas 42 e 43 que a rotulagem das bases SpAM e DB-Herabeat tomando por base os rótulos da base de dados DB-Trium apresentaram diferença estatisticamente significante entre as classes normal e patológicas, sendo estes resultados consistentes com os cenários de avaliação supervisionados e semi-supervisionados de forma cruzada. A Tabela 44 apresenta uma listagem dos identificadores referentes aos registros rotulados como patológicos ao se utilizar a base de dados DB-Trium como rotulador no cenário semi-supervisionado para as bases SpAM e DB-Herabeat.

Tabela 44 – Identificadores dos registros patológicos nas bases DB-HeraBeat e SpAM usando a rotulagem do DB-Trium. Fonte: Elaboração própria.

Base de dados	IDs dos registros patológicos
DB-Herabeat	'4973_16-01-2020', '3908_02-01-2020', '5206_20-01-2020', '4506_10-01-2020', '5311_21-01-2020', '18049_29-06-2020', '11720_17-04-2020', '11399_12-04-2020', '13708_11-05-2020', '25847_24-08-2020', '33284_24-09-2020'
SpAM	1, 2, 4, 5, 14, 17, 18, 21, 22, 24, 28, 32, 34, 36, 37, 40, 41, 45, 48, 49, 51, 52, 56, 57, 62, 64, 68, 69, 70, 71, 75, 77, 78, 79, 87, 90, 91, 93, 98, 100, 103, 104, 105, 109, 110, 115, 116, 117, 118, 119, 120, 122, 123, 124, 125, 126, 127, 131, 135, 139, 140, 141, 142, 143, 145, 150, 153, 158, 161, 162, 163, 164, 165, 166, 167, 169, 170, 171, 172, 173, 174, 175, 177, 178, 179, 180, 184, 185, 187, 189, 190, 193, 195, 197, 199, 205, 211, 212, 213, 218, 220, 221, 222, 224, 225, 230, 231, 232, 233, 234, 235, 236, 241, 244, 248, 249, 251, 253, 254, 258, 259, 264, 266, 268, 272, 274, 282, 283, 284, 288, 289, 290, 291, 295, 298, 299

7.3 Síntese do Capítulo

Foi avaliada uma abordagem semi-supervisionada com uso do algoritmo *LabelSpreading* na estimativa do estado fetal. Primeiramente, empregou-se as bases de dados rotuladas DB-Trium, HUFA e CTU-UHB em três esquemas de segmentação, *first30*, *full-data* e *last30*, de forma isolada, onde foram removidos 40% dos rótulos de cada base e comparados os rótulos estimados em relação aos rótulos reais por meio das métricas de acurácia, especificidade, sensibilidade, média geométrica e AUC. Os esquemas de segmentação *full-data* e *last30* obtiveram resultados similares entre si e superiores ao esquema *first30*. De forma geral, o esquema *last30* concentrou os melhores resultados para as três bases avaliadas atingindo AUC = 0.670 para a base HUFA, AUC = 0.695 para a base de dados DB-Trium e AUC = 0.731 para base CTU-UHB.

Em seguida, foi realizada a avaliação da capacidade de estimar os rótulos de forma cruzada entre as bases de dados. Foram empregados os esquemas de segmentação *full-data* e *last30* de forma que cada base rotulada DB-Trium, CTU-UHB e HUFA foi usada na função de

rotulação e validação entre si. Foi computada a acurácia, especificidade, sensibilidade, média geométrica e AUC utilizando os rótulos reais de cada base de validação para comparação com os rótulos estimados. Em todas as combinações cruzadas de rotulagem/validação, os esquemas de segmentação *last30* obtiveram os melhores resultados de classificação. As duas bases de dados com maior número de registros apresentaram os melhores resultados em ambas as combinações de rotuladora e validação, onde foram obtidos AUC = 0.707 ao utilizar a rotulagem do CTU-UHB na avaliação do DB-Trium e AUC = 0.713 ao se utilizar a rotulagem do DB-Trium na avaliação da base CTU-UHB.

Por fim, fez-se uso das bases de dados rotuladas com maior número de registros, CTU-UHB e DB-Trium, para estimar os rótulos das bases SpAM e DB-HeraBeat no esquema de segmentação *last30*. Foram comparadas as distribuições normal e patológica dos rótulos estimados por meio do teste estatístico Mann-Whitey U. Em ambas as bases de dados SpAM e DB-HeraBeat, foi obtida significância estatística ao se comparar os grupos normais e patológicos para todas as *features* empregadas ao se utilizar a base DB-Trium como rotuladora. Estes resultados são compatíveis com as avaliações entre bases rotuladas de forma cruzada, onde o uso da base DB-Trium para rotular a base CTU-UHB obteve o melhor desempenho. Utilizando-se da base DB-Trium para rotulagem das bases SpAM e DB-HeraBeat, a distribuição final da base de dados SpAM foi 164 (cento e sessenta e quatro) registros normais e 136 (cento e trinta e seis) registros patológicos. A base de dados DB-HeraBeat se distribuiu em 1508 (um mil quinhentos e oito) registros normais e 11 (onze) registros patológicos.

É importante reforçar que as bases de dados DB-HeraBeat e SpAM não possuem rótulos pré-estabelecidos para confirmação das classes estimadas. Além disto, a base DB-HeraBeat compreende os sinais coletados em monitoramentos de rotina envolvendo casos de baixo risco gestacional, o que justifica o baixo número de casos patológicos encontrados em relação ao seu número total de registros.

8 CONCLUSÕES

Neste Capítulo de conclusão, as questões de pesquisa e hipóteses levantadas na Introdução (Capítulo 1) desta tese são respondidas e confirmadas. São apresentadas ainda perspectivas para trabalhos futuros em conjunto com as publicações realizadas no trajeto deste curso de doutorado, assim como as considerações finais.

8.1 Respostas às Questões de Pesquisa (QP) e Hipóteses

8.1.1 Avaliação do estado fetal - QP #1

Foram apresentadas três abordagens para avaliação do estado fetal: (i) avaliação baseada em *features* empregando algoritmos de aprendizado de máquina supervisionados; (ii) avaliação com o uso direto dos biossinais empregando algoritmos de aprendizagem de máquina supervisionados; (iii) avaliação baseada em *features* empregando algoritmos de aprendizado de máquina semi-supervisionados. As abordagens propostas podem ser adaptadas e integradas a sistemas de monitoramento do estado fetal para auxílio na tomada de decisão médica (**Contribuição principal**). Conforme apresentado na Introdução (Capítulo 1) e Revisão de Literatura (Capitulo 3), a literatura recomenda valores de sensibilidade acima de 60% (reconhecimento de casos patológicos) para adoção dos modelos em uso clínico, com um cenário ideal apresentando o par sensibilidade e especificidade com os valores mais altos possíveis.

O cenário de avaliação para modelos baseado diretamente em biossinais obteve o melhor desempenho para o classificador CNN no esquema de segmentação *last30*, atingindo 67,2% para acurácia, 71,3% para especificidade, 58,2% para sensibilidade, 63,6% para média geométrica e 64,8% para AUC na base de dados DB-Trium. Desta forma, o modelo de prognóstico baseado em biossinais atingiu resultados de Se e Sp bem próximos ao desempenho necessário para adoção em uso clínico, outrossim, é importante ressaltar que o cenário de avaliação desta tese contou com uma quantidade de registros 68x (sessenta e oito vezes) menor que a base de dados privada baseada em biossinais de melhor desempenho empregada na literatura. Os resultados se mostraram compatíveis com os do mapeamento sistemático, os quais indicam a necessidade de grandes quantidades de dados neste tipo de modelo para que sejam obtidos altos níveis de desempenho.

No cenário de avaliação dos modelos baseados em *features*, no conjunto de dados

CTU-UHB, o classificador de melhor desempenho foi o SVM com 83,7% para acurácia, 85,6% para especificidade, 67,5% para sensibilidade, 75,5% para média geométrica e 76,5% para AUC. Na base HUFA, o melhor resultado foi obtido utilizando o classificador KNN com 72,6% para acurácia, 75,0% para especificidade, 73,3% para sensibilidade, 70,4% para média geométrica e 74,1% para AUC. No DB-TRIUM, o classificador de melhor desempeno foi o SVM com 69,5% para acurácia, 76,1% para especificidade, 63,7% para sensibilidade, 68,6% para média geométrica e 69,9% para AUC. Estes resultados se mostraram promissores para adoção em uso clínico.

A abordagem semi-supervisionada foi avaliada com o uso do algoritmo *LabelSpreading* para estimativa do estado fetal. Primeiramente, foram utilizadas as bases de dados rotuladas DB-Trium, HUFA e CTU-UHB de forma isolada onde foram removidos 40% dos rótulos de cada base e comparados os rótulos estimados em relação aos rótulos reais por meio das métricas de acurácia, especificidade, sensibilidade, média geométrica e AUC. Foram obtidos AUC = 67,0% para a base HUFA, AUC = 69,5% para a base de dados DB-Trium e AUC = 73,1% para base CTU-UHB.

Em seguida, foi executada a avaliação da capacidade de estimar os rótulos de forma cruzada entre as bases de dados, onde as duas bases de dados com maior número de registros apresentaram os melhores resultados em ambas as combinações de rotuladora e validação. Foram alcançadas acurácia de 80,5% e AUC = 70,7% ao utilizar a rotulagem do CTU-UHB na validação do DB-Trium e acurácia de 78,3% e AUC = 71,3% ao se utilizar a rotulagem do DB-Trium na validação da base CTU-UHB.

Por fim, as bases de dados rotuladas com maior número de registros, CTU-UHB e DB-Trium, foram empregadas para estimar os rótulos das bases SpAM e DB-HeraBeat. Em ambas as bases de dados SpAM e DB-HeraBeat, foi obtida significância estatística ao se compararem os grupos normais e patológicos para todas as *features* empregadas utilizando-se a base DB-Trium como rotuladora. Neste cenário, a distribuição final da base de dados SpAM foi 164 (cento e sessenta e quatro) registros normais e 136 (cento e trinta e seis) registros patológicos, enquanto a base de dados DB-HeraBeat se distribuiu em 1508 (um mil quinhentos e oito) registros normais e 11 (onze) registros patológicos.

Desta forma, com o uso de abordagens de aprendizado de máquina supervisionadas baseadas em *features* e em biossinais, e com a adoção de aprendizagem semi-supervisionada baseada em *features*, foi possível estimar desfechos adversos, atingindo-se desempenhos em

níveis aceitáveis para uso dos modelos de prognóstico em ambiente clínico para avaliação do bem-estar fetal, confirmando assim a **Hipótese#1** associada à questão **QP#1**.

8.1.2 Principais blocos de construção - QP #2

Nesta tese, foi realizado um mapeamento sistemático (Capítulo 3) para o levantamento do estado da arte sobre os blocos de construção empregados na avaliação do estado fetal, resultando em um panorama atualizado das principais técnicas e métodos empregados (**Contribuição secundária #1**). A literatura apresentou divergências sobre os parâmetros e técnicas que compõem os blocos de construção adotados em modelos de prognóstico. Até onde sabemos, este é o primeiro panorama dos estudos sobre os componentes básicos dos sistemas de avaliação do estado fetal com ênfase nas etapas de projeto destes sistemas, cujas principais conclusões podem ser resumidas conforme segue:

- a maioria dos estudos analisados propuseram novos métodos para avaliação do estado fetal, no entanto, os resultados indicam a ausência da validação destes estudos em ambiente clínico;
- a base de dados sobre biossinais CTU-UHB foi a mais frequente entre os estudos analisados, e UCI-CTG foi o conjunto de dados com *features* prontas para uso mais frequente. Os estudos apontam uma variação considerável no número de registros entre os conjuntos de dados, sendo que a quantidade mais significativa (35.429 registros) pertenceu a uma base de dados de caráter privativo;
- foi identificado entre os estudos selecionados a utilização dos estágios (i) preparação de dados (DP); (ii) transformação de dados (DT); (iii) e construção de modelo (MC), os quais foram agrupados pelo uso de cada estágio em uma taxonomia de dois níveis (vide Figura 5);
- a tarefa de preparação dos dados mais recorrente foi a remoção de lacunas no sinal com duração maior que (15 s). A maioria dos estudos utilizou interpolação linear e de Hermite para atividades de redução de artefatos: (i) o ajuste da diferença (> 25 bpm) entre pontos adjacentes; (ii) ajuste dos valores *outlier* (> 200 bpm ou <50 bpm); (iii) no preenchimento dos valores ausentes (< 15 s);
- a atividade de transformação de dados mais recorrente é a extração de *features*. Em combinação com representações de dados bem estabelecidas na literatura, os estudos selecionados propuseram novos índices de representação. Os domínios morfológicos

e temporais foram as representações de dados mais frequentes. A maioria das novas propostas de índices representativos pertence a domínios de frequência, de tempo e não-lineares. A maioria dos novos índices concentrou-se em representações de sinais baseadas em espectrogramas;

– para os estudos que empregaram o estágio de construção de modelo (MC) em conjunto com as *features* prontas para uso provenientes da base de dados UCI-CTG, o classificador *random forest* sozinho ou em combinação com outras técnicas foi o modelo de classificação de melhor desempenho. A maioria dos estudos que empregaram múltiplos estágios de desenvolvimento fez uso do conjunto de dados CTU-UHB e a rotulagem de classe mais frequente foi o pH. Nestes cenários, o método de classificação mais recorrente foi o SVM em cenários baseados em *features*, e as métricas de desempenho mais usadas foram sensibilidade e especificidade;
– O uso de abordagens de autoaprendizagem via *deep learning* apresentou altos desempenhos mas exigiu grandes quantidades de dados e etapas extras na transformação. Os modelos baseados em *features* apresentaram desempenhos competitivos e exigiram uma menor quantidade de dados no seu desenvolvimento.

Desta forma, com uso de uma metodologia sistemática e reproduzível, foi possível extrair informações relevantes e atualizadas da literatura sobre blocos de construção e parâmetros utilizados nas soluções computadorizadas para avaliação do bem-estar fetal, confirmando assim a **Hipótese#2** associada à **QP#2**.

8.1.3 Processo para construção dos modelos de prognóstico - QP #3

Nesta tese, foi aplicado um processo de avaliação para escolha dos blocos de construção em modelos de prognóstico a fim de identificar a combinação com melhor capacidade de discriminação entre os estados normal e patológico dos fetos (**Contribuição secundária #2**). A literatura apresentou variação na combinação de métodos e técnicas disponíveis para avaliação do bem-estar fetal. O método aplicado nesta tese compôs um processo de avaliação e escolha dos blocos de construção de forma sistemática e extensível onde foram incluídos no processo de construção dos modelos a escolha dos algoritmos de classificação, uma abordagem de segmentação baseada em intervalos de tempo para avaliar a capacidade do sinal em discriminar os registros saudáveis e patológicos e a avaliação da capacidade de generalização dos modelos para cada base de dados disponível.

Os resultados indicaram que o esquema de segmentação de dados *first30* apresentou desempenho geral baixo, mesmo após a aplicação de técnicas de balanceamento do número de registros pertencentes a cada classe, indicando capacidade limitada de separação de classes em relação à proximidade da hora do parto. Os escores dos classificadores apresentaram equivalência geral entre os esquemas de segmentação de dados completos (*full-data*) e os últimos 30 minutos (*last30*), sendo estes os de melhor desempenho em nossas avaliações. De maneira geral, os melhores resultados se concentraram no esquema de segmentação *last30*.

Para avaliação da generalização dos modelos, os cenários baseados em *features* obtiveram desempenhos superiores aos baseados em biossinais. O melhor desempenho foi alcançado para DB-TRIUM/CTU-UHB com o classificador MLP, atingindo 70,6% para acurácia, 71,6% para especificidade, 61,7% para sensibilidade, 66,3% para média geométrica e 66,7% para AUC.

Este resultado foi compatível com a avaliação semi-supervisionada baseada em *features*, na qual o uso da base DB-Trium obteve os melhores desempenhos quando aplicada na rotulagem de bases cruzadas e quando utilizada para rotulagem das bases SpAM e DB-HeraBeat.

Desta forma, ao se utilizar de um processo de avaliação dos blocos de construção de forma sistemática e reproduzível, foi possível construir modelos de prognóstico com desempenho de forma isolada e generalizada compatíveis com o uso em ambiente clínico na avaliação do bem-estar fetal, confirmando assim a **Hipótese#3** associada a questão **QP#3**.

8.2 Produção Científica

Esta seção informa manuscritos e artigos científicos direta ou indiretamente relacionados a esta tese.

8.2.1 Artigos diretamente associados à tese

(i) Artigos completos submetidos para periódicos (em processo de revisão):

1) **SILVA NETO, M. G. DA**; MADEIRO, J. P. DO VALE; MARQUES, J. A. L.; GOMES, D. G. Towards an Efficient Prognostic Model for Fetal State Assessment. Submetido ao *Measurement*[1];
2) **SILVA NETO, M. G. DA**; MADEIRO, J. P. DO VALE; GOMES, D. G. On designing a biosignal-based fetal state assessment system: A systematic mapping study.

[1] https://www.journals.elsevier.com/measurement

Submetido ao *Computer Methods and Programs in Biomedicine*[2].

8.2.2 Artigos tangenciais à tese

(i) Artigos completos publicados em periódicos

1) SILVA, BRUNO ; SILVEIRA, RICARDO ; **SILVA NETO, MANUEL** ; CORTEZ, PAULO ; GOMES, DANIELO . A comparative analysis of undersampling techniques for network intrusion detection systems design. Journal of Communication and Information Systems (JCIS),, v. 36, p. 31-43, 2021.
2) **NETO, MANUEL**; GOMES, DANIELO ; SOARES, JOSÉ . Credibility on Crowdsensing Data Acquisition: A Systematic Mapping Study. Journal of Communication and Information Systems (JCIS),, v. 34, p. 248-269, 2019.

(ii) Artigos completos publicados em anais de conferências

1) SILVA, BRUNO ; **NETO, MANUEL SILVA** ; CORTEZ, PAULO ; GOMES, DANIELO . Design of Network Intrusion Detection Systems with under-sampled datasets. In: 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 2019, Valparaiso. 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 2019. p. 1.
2) **SILVA NETO, MANUEL GONÇALVES DA**; G. GOMES, DANIELO . Network Intrusion Detection Systems Design: A Machine Learning Approach. In: XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, 2019, Gramado, RS. Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2019), 2019. p. 932.
3) Lima Filho, Joari S. ; **Silva Neto, Manuel Gonçalves da** ; Rego, Paulo A. L. ; Gomes, Danielo G. . Um Mecanismo de Offloading de Dados com Tomada de Decisão. In: 16º WPerformance (Workshop em Desempenho de Sistemas Computacionais e de Comunicação), 2017, São Paulo. Anais do 16º WPerformance (Workshop em Desempenho de Sistemas Computacionais e de Comunicação). Porto Alegre: SBC, 2017. p. 1740-1753.

[2] https://www.journals.elsevier.com/computer-methods-and-programs-in-biomedicine

8.3 Trabalhos futuros

São identificados trabalhos futuros visando aperfeiçoar ou expandir os resultados obtidos nesta tese, destacando-se:

(i) empregar bases com maior número de registros para construção dos modelos baseados em biossinais;

(ii) avaliar o uso de outros conjuntos de *features* propostas pela literatura, em especial as baseadas em espectogramas;

(iii) validar os rótulos estimados via aprendizagem semi-supervisionada com uso de avaliação de especialistas, a fim de obter uma informação de referência para uso destas bases no projeto de modelos de prognósticos;

(iv) indicar a patologia nos casos patológicos;

(v) avaliar outras faixas de duração para segmentação do sinal;

(vi) avaliar o uso conjunto dos sinais de FCF e UC na construção dos modelos de prognóstico;

8.4 Considerações finais

Nesta tese, buscou-se sistematicamente as melhores combinações de técnicas e métodos para construção de um modelo de prognóstico para apoio à tomada de decisão em ambiente clínico no tocante à avaliação do bem-estar fetal. Foram realizadas avaliações de cenários de classificação supervisionados e semi-supervisionados baseados em *features* e de cenários supervisionados baseados em biossinais na construção de modelos de prognóstico eficientes e generalizáveis.

Tendo por base o processo de avaliação para escolha dos blocos de construção, os resultados do mapeamento sistemático e os modelos de prognóstico resultantes desta tese, recomenda-se:

- preferencialmente empregar o esquema de segmentação *last30*, tendo por opção secundária o uso do esquema *full-data*;
- empregar os classificadores SVM ou MLP em conjunto com a técnica SMOTE em cenários de dados não balanceados baseados em *features*;
- fazer uso do modelo de classificação CNN em cenários baseados em biossinais, sendo que estes cenários exigem cautela na adoção de bases de dados não balanceadas ou que contenham um número reduzido de registros para treinamento e avaliação dos modelos.

É importante ressaltar que o uso da abordagem semi-supervisionada apresentou resultados promissores mas mostrou-se fortemente dependente dos dados empregados como rotuladores, exigindo extensa avaliação prévia destas bases.

Os modelos de melhor desempenho desta tese alcançaram resultados comparáveis ou até superiores, na avaliação do estado fetal, aos métodos do estado da arte conforme apresentado nos trabalhos relacionados (Seção 3.4), o que torna os modelos propostos promissores como ferramenta de apoio à decisão médica. Por fim, as ferramentas computacionais empregadas para a construção dos modelos são de código aberto *open source* e bem estabelecidas na área de ciência de dados, o que aumenta as possibilidades de adoção pelos pesquisadores e profissionais de domínios multidisciplinares, viabilizando a sua adaptação para o uso prático em ambiente clínico.

REFERÊNCIAS

ABBAS, R.; HUSSAIN, A. J.; AL-JUMEILY, D.; BAKER, T.; KHATTAK, A. Classification of foetal distress and hypoxia using machine learning approaches. In: **14th International Conference on Intelligent Computing (ICIC)**. Wuhan, China: Springer, 2018. p. 767–776.

ABRY, P.; SPILKA, J.; LEONARDUZZI, R.; CHUDACEK, V.; PUSTELNIK, N.; DORET, M. Sparse learning for intrapartum fetal heart rate analysis. **Biomedical Physics & Engineering Express**, 4, n. 3, p. 17, MAY 2018. ISSN 2057-1976.

AFRIDI, R.; IQBAL, Z.; KHAN, M.; AHMAD, A.; NASEEM, R. Fetal heart rate classification and comparative analysis using cardiotocography data and known classifiers. **International Journal of Grid and Distributed Computing**, v. 12, n. 1, p. 31–42, 2019. ISSN 2005-4262.

AGGARWAL, C. C. **Data Classification: Algorithms and Applications**. 1st. ed. New York, United States of America: Chapman & Hall/CRC, 2014. 704 p. ISBN 1466586745.

AGRAWAL, K.; MOHAN, H. Cardiotocography analysis for fetal state classification using machine learning algorithms. In: **2019 International Conference on Computer Communication and Informatics (ICCCI)**. [*S. l.*: *s. n.*], 2019. p. 1–6.

ALSAGGAF, W.; COMERT, Z.; NOUR, M.; POLAT, K.; BRDESEE, H.; TOğAçAR, M. Predicting fetal hypoxia using common spatial pattern and machine learning from cardiotocography signals. **Applied Acoustics**, v. 167, p. 107429, 2020. ISSN 0003-682X.

ALSAYYARI, A. Fetal cardiotocography monitoring using legendre neural networks. **Biomedical Engineering / Biomedizinische Technik**, v. 64, n. 6, p. 669–675, 2019. Disponível em: https://doi.org/10.1515/bmt-2018-0074.

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **The American Statistician**, Taylor & Francis, v. 46, n. 3, p. 175–185, 1992.

ANISHA, M.; KUMAR, S. S.; NITHILA, E. E.; BENISHA, M. Detection of fetal cardiac anomaly from composite abdominal electrocardiogram. **Biomedical Signal Processing and Control**, v. 65, p. 102308, 2021. ISSN 1746-8094.

AYRES-DE-CAMPOS, D. Electronic fetal monitoring or cardiotocography, 50 years later: whatś in a name? **American Journal of Obstetrics & Gynecology**, Elsevier, v. 218, n. 6, p. 545–546, Jun 2018. ISSN 0002-9378. Disponível em: https://doi.org/10.1016/j.ajog.2018.03.011.

AYRES-DE-CAMPOS, D.; ARULKUMARAN, S.; PANEL, F. I. F. M. E. C. Figo consensus guidelines on intrapartum fetal monitoring: Physiology of fetal oxygenation and the main goals of intrapartum fetal monitoring. **International Journal of Gynecology & Obstetrics**, v. 131, n. 1, p. 5–8, 2015. Disponível em: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1016/j.ijgo.2015.06.018.

AYRES-DE-CAMPOS, D.; SPONG, C. Y.; CHANDRAHARAN, E.; PANEL, F. I. F. M. E. C. Figo consensus guidelines on intrapartum fetal monitoring: Cardiotocography. **International Journal of Gynecology and Obstetrics**, v. 131, n. 1, p. 13–24, 2015. Disponível em: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1016/j.ijgo.2015.06.020.

BARQUERO-PEREZ, O.; SANTIAGO-MOZOS, R.; LILLO-CASTELLANO, J. M.; GARCIA-VIRUETE, B.; GOYA-ESTEBAN, R.; CAAMANO, A. J.; ROJO-ALVAREZ, J. L.; MARTIN-CABALLERO, C. Fetal Heart Rate Analysis for Automatic Detection of Perinatal Hypoxia Using Normalized Compression Distance and Machine Learning. **Frontiers in Physiology**, 8, FEB 28 2017. ISSN 1664-042X.

BARQUERO-PEREZ, O.; SANTIAGO-MOZOS, R.; LILLO-CASTELLANO, J. M.; GARCIA-VIRUETE, B.; GOYA-ESTEBAN, R.; CAAMANO, A. J.; ROJO-ALVAREZ, J. L.; MARTIN-CABALLERO, C. **HUFA hypoxia database**. 2017. Disponível em: https://sites.google.com/site/hufahypoxia/.

BATRA, A.; CHANDRA, A.; MATORIA, V. Cardiotocography analysis using conjunction of machine learning algorithms. In: **2017 International Conference on Machine Vision and Information Technology (CMVIT)**. [*S. l.*: *s. n.*], 2017. p. 1–6.

BISCHOFF, V.; FARIAS, K.; GONçALES, L. J.; BARBOSA, J. L. V. Integration of feature models: A systematic mapping study. **Information and Software Technology**, v. 105, p. 209 – 225, 2019. ISSN 0950-5849. Disponível em: http://www.sciencedirect.com/science/article/pii/S0950584916302178.

BOUDET, S.; Houzé l'Aulnoit, A.; DEMAILLY, R.; DELGRANCHE, A.; PEYRODIE, L.; BEUSCART, R.; Houzé de l'Aulnoit, D. A fetal heart rate morphological analysis toolbox for matlab. **SoftwareX**, v. 11, p. 12, 2020. ISSN 2352-7110. Disponível em: https://www.sciencedirect.com/science/article/pii/S2352711018302498.

BOUDET, S.; L'AULNOIT, A. H. de; DEMAILLY, R.; PEYRODIE, L.; BEUSCART, R.; L'AULNOIT, D. H. de. Fetal heart rate baseline computation with a weighted median filter. **Computers in Biology and Medicine**, v. 114, p. 103468, 2019. ISSN 0010-4825. Disponível em: https://www.sciencedirect.com/science/article/pii/S0010482519303403.

BOUDET, S.; L'AULNOIT, A. Houzé de; DEMAILLY, R.; DELGRANCHE, A.; PEYRODIE, L.; BEUSCART, R.; L'AULNOIT, D. Houzé de. Fetal heart rate signal dataset for training morphological analysis methods and evaluating them against an expert consensus. **Preprints**, v. 1, p. 39, 2019.

BOWYER, K. W.; CHAWLA, N. V.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002. ISSN 1076-9757. Disponível em: https://doi.org/10.1613/jair.953.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, Aug 1996. ISSN 1573-0565. Disponível em: https://doi.org/10.1007/BF00058655.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Disponível em: https://doi.org/10.1023/A:1010933404324.

CHAMIDAH, N.; WASITO, I. Fetal state classification from cardiotocography based on feature extraction using hybrid k-means and support vector machine. In: **2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)**. [*S. l.*: *s. n.*], 2015. p. 37–41.

CHAPELLE, O.; SCHLKOPF, B.; ZIEN, A. **Semi-Supervised Learning**. 1st. ed. London, United Kingdom: The MIT Press, 2010. 524 p. ISBN 0262514125.

CHEN, J.-y.; LIU, X.-c.; WEI, H.; CHEN, Q.-q.; HONG, J.-m.; LI, Q.-n.; HAO, Z.-f. Imbalanced cardiotocography multi-classification for antenatal fetal monitoring using weighted random forest. In: CHEN, H.; ZENG, D.; YAN, X.; XING, C. (Ed.). **Smart Health**. Cham: Springer International Publishing, 2019. p. 75–85. ISBN 978-3-030-34482-5.

CHUDACEK, V.; HUPTYCH, M.; KOUCKY, M.; SPILKA, J.; BAUER, L.; LHOTSKA, L. Fetal heart rate data pre-processing and annotation. In: **2009 9th International Conference on Information Technology and Applications in Biomedicine**. [*S. l.*: *s. n.*], 2009. p. 1–4.

CHUDACEK, V.; SPILKA, J.; BURSA, M.; JANKU, P.; HRUBAN, L.; HUPTYCH, M.; LHOTSKA, L. Open access intrapartum ctg database. **BMC Pregnancy and Childbirth**, v. 14, n. 1, p. 16, Jan 2014. ISSN 1471-2393. Disponível em: https://doi.org/10.1186/1471-2393-14-16.

COMERT, Z.; KOCAMAZ, A. F. Evaluation of fetal distress diagnosis during delivery stages based on linear and nonlinear features of fetal heart rate for neural network community. **International Journal of Computer Applications**, Foundation of Computer Science (FCS), NY, USA, New York, USA, v. 156, n. 4, p. 26–31, Dec 2016. ISSN 0975-8887. Disponível em: http://www.ijcaonline.org/archives/volume156/number4/26698-2016912417.

COMERT, Z.; KOCAMAZ, A. F. A novel software for comprehensive analysis of cardiotocography signals "ctg-oas". In: **2017 International Artificial Intelligence and Data Processing Symposium (IDAP)**. [*S. l.*: *s. n.*], 2017. p. 1–6.

COMERT, Z.; KOCAMAZ, A. F. Open-access software for analysis of fetal heart rate signals. **Biomedical Signal Processing and Control**, 45, p. 98–108, AUG 2018. ISSN 1746-8094.

COMERT, Z.; KOCAMAZ, A. F. Fetal hypoxia detection based on deep convolutional neural network with transfer learning approach. In: SILHAVY, R. (Ed.). **Software Engineering and Algorithms in Intelligent Systems**. Cham: Springer International Publishing, 2019. (CSOC2018 Computer Science on-line conference), p. 239–248. ISBN 978-3-319-91186-1.

COMERT, Z.; KOCAMAZ, A. F.; SUBHA, V. Prognostic model based on image-based time-frequency features and genetic algorithm for fetal hypoxia assessment. **Computers in Biology and Medicine**, 99, p. 85–97, AUG 1 2018. ISSN 0010-4825.

COMERT, Z.; SENGUR, A.; BUDAK, U.; KOCAMAZ, A. F. Prediction of intrapartum fetal hypoxia considering feature selection algorithms and machine learning models. **Health Information Science and Systems**, v. 7, n. 1, p. 17, Aug 2019. ISSN 2047-2501. Disponível em: https://doi.org/10.1007/s13755-019-0079-z.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, Sep 1995. ISSN 1573-0565. Disponível em: https://doi.org/10.1007/BF00994018.

CZABANSKI, R.; JEZEWSKI, M.; HOROBA, K.; JEZEWSKI, J.; LESKI, J. Fuzzy analysis of delivery outcome attributes for improving the automated fetal state assessment. **Applied Artificial Intelligence**, 30, n. 6, SI, p. 556–571, 2016. ISSN 0883-9514.

CZABANSKI, R.; JEZEWSKI, M.; LESKI, J. M.; KUPKA, T.; MARTINEK, R. Clustering with epsilon-hyperballs based simplification of fuzzy rules to support the assessment of fetal state. In: **2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)**. [*S. l.*: *s. n.*], 2020. p. 358–364.

DAS, S.; MUKHERJEE, H.; OBAIDULLAH, S. M.; SANTOSH, K. C.; ROY, K.; SAHA, C. K. Recurrent neural network based classification of fetal heart rate using cardiotocograph. In: SANTOSH, K. C.; HEGADI, R. S. (Ed.). **Recent Trends in Image Processing and Pattern Recognition**. Singapore: Springer Singapore, 2019. p. 226–234. ISBN 978-981-13-9184-2.

DAS, S.; MUKHERJEE, H.; OBAIDULLAH, S. M.; ROY, K.; SAHA, C. K. Ensemble based technique for the assessment of fetal health using cardiotocograph – a case study with standard feature reduction techniques. **Multimedia Tools and Applications**, v. 79, n. 47, p. 35147–35168, Dec 2020. ISSN 1573-7721. Disponível em: https://doi.org/10.1007/s11042-020-08853-2.

DAS, S.; MUKHERJEE, H.; SANTOSH, K. C.; SAHA, C. K.; ROY, K. Periodic change detection in fetal heart rate using cardiotocograph. In: **2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)**. [*S. l.*: *s. n.*], 2020. p. 104–109.

DAS, S.; OBAIDULLAH, S. M.; SANTOSH, K. C.; ROY, K.; SAHA, C. K. Cardiotocograph-based labor stage classification from uterine contraction pressure during ante-partum and intra-partum period: a fuzzy theoretic approach. **Health Information Science and Systems**, v. 8, n. 1, p. 16, Mar 2020. ISSN 2047-2501. Disponível em: https://doi.org/10.1007/s13755-020-00107-7.

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine learning research**, v. 7, n. Jan, p. 1–30, 2006.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository - Cardiotocography Data Set**. 2017. Disponível em: https://archive.ics.uci.edu/ml/datasets/cardiotocography.

FENG, G.; QUIRK, J. G.; DJURIć, P. M. Supervised and unsupervised learning of fetal heart rate tracings with deep gaussian processes. In: **2018 14th Symposium on Neural Networks and Applications (NEUREL)**. [*S. l.*: *s. n.*], 2018. p. 1–6.

FERGUS, P.; CHALMERS, C.; MONTANEZ, C. C.; REILLY, D.; LISBOA, P.; PINELES, B. Modelling segmented cardiotocography time-series signals using one-dimensional convolutional neural networks for the early detection of abnormal birth outcomes. **IEEE Transactions on Emerging Topics in Computational Intelligence**, p. 1–11, 2020.

FERGUS, P.; HUSSAIN, A.; AL-JUMEILY, D.; HUANG, D.-S.; BOUGUILA, N. Classification of caesarean section and normal vaginal deliveries using foetal heart rate signals and advanced machine learning algorithms. **Biomedical Engineering Online**, 16, p. 26, JUL 6 2017. ISSN 1475-925X.

FERGUS, P.; SELVARAJ, M.; CHALMERS, C. Machine learning ensemble modelling to classify caesarean section and vaginal delivery types using cardiotocography traces. **Computers in Biology and Medicine**, v. 93, p. 7 – 16, 2018. ISSN 0010-4825. Disponível em: http://www.sciencedirect.com/science/article/pii/S0010482517303918.

FRIEDMAN, J. H. Stochastic gradient boosting. **Computational Statistics & Data Analysis**, v. 38, n. 4, p. 367 – 378, 2002. ISSN 0167-9473. Nonlinear Methods and Data Mining. Disponível em: http://www.sciencedirect.com/science/article/pii/S0167947301000652.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. **Journal of the American Statistical Association**, Taylor & Francis, v. 32, n. 200, p. 675–701, 1937. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522.

FUENTEALBA, P.; ILLANES, A.; ORTMEIER, F. Analysis of the foetal heart rate in cardiotocographic recordings through a progressive characterization of decelerations. **Current Directions in Biomedical Engineering**, v. 3, n. 2, p. 423–427, 2017. Disponível em: https://doi.org/10.1515/cdbme-2017-0089.

FUENTEALBA, P.; ILLANES, A.; ORTMEIER, F. Cardiotocographic signal feature extraction through ceemdan and time-varying autoregressive spectral-based analysis for fetal welfare assessment. **IEEE Access**, v. 7, p. 159754–159772, 2019.

FUENTEALBA, P.; ILLANES, A.; ORTMEIER, F. Independent analysis of decelerations and resting periods through ceemdan and spectral-based feature extraction improves cardiotocographic assessment. **Applied Sciences**, v. 9, n. 24, 2019. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/9/24/5421.

GAO, W.; LU, Y. Fetal heart baseline extraction and classification based on deep learning. In: **2019 International Conference on Information Technology and Computer Application (ITCA)**. [*S. l.*: *s. n.*], 2019. p. 211–216.

GEORGIEVA, A.; ABRY, P.; CHUDáčEK, V.; DJURIć, P. M.; FRASCH, M. G.; KOK, R.; LEAR, C. A.; LEMMENS, S. N.; NUNES, I.; PAPAGEORGHIOU, A. T.; QUIRK, G. J.; REDMAN, C. W. G.; SCHIFRIN, B.; SPILKA, J.; UGWUMADU, A.; VULLINGS, R. **The Signal Processing and Monitoring (SPAM) in Labor Workshop 2017 Challenge Database**. 2017. Disponível em: http://users.ox.ac.uk/~ndog0178/CTGchallenge2017.htm.

GEORGIEVA, A.; REDMAN, C. W. G.; PAPAGEORGHIOU, A. T. Computerized data-driven interpretation of the intrapartum cardiotocogram: a cohort study. **Acta Obstetricia et Gynecologica Scandinavica**, 96, n. 7, p. 883–891, JUL 2017. ISSN 0001-6349.

GEORGOULAS, G.; KARVELIS, P.; SPILKA, J.; CHUDACEK, V.; STYLIOS, C. D.; LHOTSKA, L. Investigating ph based evaluation of fetal heart rate (fhr) recordings. **Health and Technology**, 7, n. 2-3, p. 241–254, NOV 2017. ISSN 2190-7188.

GIULIANO, N.; ANNUNZIATA, M. L.; ESPOSITO, F. G.; TAGLIAFERRI, S.; LIETO, A. D.; MAGENES, G.; SIGNORINI, M. G.; CAMPANILE, M.; ARDUINI, D. Computerised analysis of antepartum foetal heart parameters: New reference ranges. **Journal of Obstetrics and Gynaecology**, Taylor & Francis, v. 37, n. 3, p. 296–304, 2017.

GOLDBERGER, A. L.; AMARAL, L. A. N.; GLASS, L.; HAUSDORFF, J. M.; IVANOV, P. C.; MARK, R. G.; MIETUS, J. E.; MOODY, G. B.; PENG, C.-K.; STANLEY, H. E. Physiobank, physiotoolkit, and physionet. **Circulation**, v. 101, n. 23, p. e215–e220, 2000.

GONÇALVES, H.; ROCHA, A. P.; AYRES-DE-CAMPOS, D.; BERNARDES, J. Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery. **Medical and Biological Engineering and Computing**, v. 44, n. 10, p. 12, Sep 2006. ISSN 1741-0444. Disponível em: https://doi.org/10.1007/s11517-006-0105-6.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. London, United Kingdom: The MIT Press, 2016. 800 p. ISBN 0262035618.

GUYON, I.; WESTON, J.; BARNHILL, S.; VAPNIK, V. Gene selection for cancer classification using support vector machines. **Machine Learning**, v. 46, n. 1, p. 389–422, Jan 2002. ISSN 1573-0565. Disponível em: https://doi.org/10.1023/A:1012487302797.

GYLLENCREUTZ, E.; LU, K.; LINDECRANTZ, K.; LINDQVIST, P. G.; NORDSTROM, L.; HOLZMANN, M.; ABTAHI, F. Validation of a computerized algorithm to quantify fetal heart rate deceleration area. **Acta Obstetricia et Gynecologica Scandinavica**, v. 97, n. 9, p. 1137–1147, 2018.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Comput.**, MIT Press, Cambridge, MA, USA, v. 9, n. 8, p. 1735–1780, nov. 1997. ISSN 0899-7667. Disponível em: https://doi.org/10.1162/neco.1997.9.8.1735.

HOODBHOY, Z.; NOMAN, M.; SHAFIQUE, A.; NASIM, A.; CHOWDHURY, D.; HASAN, B. Use of machine learning algorithms for prediction of fetal risk using cardiotocographic data. **International Journal of Applied and Basic Medical Research**, v. 9, n. 4, p. 226–230, 2019.

HUANG, X.-q.; LI, L.; CHEN, Q.-q.; WEI, H.; HAO, Z.-f. Intelligent antenatal fetal monitoring model based on adaptive neuro-fuzzy inference system through cardiotocography. In: CAO, B.-y. (Ed.). **Fuzzy Information and Engineering-2019**. Singapore: Springer Singapore, 2020. p. 25–36. ISBN 978-981-15-2459-2.

HUDDAR, P. P.; SONTAKKE, S. A. Acquiring domain knowledge for cardiotocography: A deep learning approach. In: **2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)**. [*S. l.*: *s. n.*], 2019. p. 1–6.

IDRI, A.; BENHAR, H.; FERNáNDEZ-ALEMáN, J.; KADI, I. A systematic map of medical data preprocessing in knowledge discovery. **Computer Methods and Programs in Biomedicine**, v. 162, p. 69 – 85, 2018. ISSN 0169-2607. Disponível em: http://www.sciencedirect.com/science/article/pii/S0169260717313706.

INTAN, P. R. D.; MA'SUM, M. A.; ALFIANY, N.; JATMIKO, W.; KEKALIH, A.; BUSTAMAM, A. Ensemble learning versus deep learning for hypoxia detection in ctg signal. In: **2019 International Workshop on Big Data and Information Security (IWBIS)**. [*S. l.*: *s. n.*], 2019. p. 57–62.

IRAJI, M. S. Prediction of fetal state from the cardiotocogram recordings using neural network models. **Artificial Intelligence in Medicine**, v. 96, p. 33 – 44, 2019. ISSN 0933-3657. Disponível em: http://www.sciencedirect.com/science/article/pii/S0933365718304822.

ITO, A.; HAYATA, E.; NAKATA, M.; OJI, A.; FURUKAWA, T.; NAKAKUMA, M.; MORITA, M. iPREFACE score: Integrated score index to predict fetal acidemia by intrapartum fetal heart rate monitoring. **Journal of Obstetrics and Gynaecology Research**, 2021. Disponível em: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/jog.14652.

JAPKOWICZ, N.; SHAH, M. **Evaluating Learning Algorithms: A Classification Perspective**. New York, United States of America: Cambridge University Press, 2011. 424 p.

JEZEWSKI, M.; CZABANSKI, R.; HOROBA, K.; LESKI, J. Clustering with Pairs of Prototypes to Support Automated Assessment of the Fetal State. **Applied Artificial Intelligence**, 30, n. 6, SI, p. 572–589, 2016. ISSN 0883-9514.

JEZEWSKI, M.; CZABANSKI, R.; LESKI, J. M.; JEZEWSKI, J. Fuzzy classifier based on clustering with pairs of epsilon-hyperballs and its application to support fetal state assessment. **Expert Systems with Applications**, v. 118, p. 109–126, 2019. ISSN 0957-4174. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417418306080.

KADHIM, N. J. A.; ABED, J. K. Enhancing the prediction accuracy for cardiotocography (CTG) using firefly algorithm and naive bayesian classifier. **IOP Conference Series: Materials Science and Engineering**, IOP Publishing, v. 745, p. 012101, mar 2020. Disponível em: https://doi.org/10.1088/1757-899x/745/1/012101.

KANNAN, E.; RAVIKUMAR, S.; ANITHA, A.; KUMAR, S. A. P.; VIJAYASARATHY, M. Analyzing uncertainty in cardiotocogram data for the prediction of fetal risks based on machine learning techniques using rough set. **Journal of Ambient Intelligence and Humanized Computing**, Jan 2021. ISSN 1868-5145. Disponível em: https://doi.org/10.1007/s12652-020-02803-4.

KAUR, H.; KHULLAR, V.; SINGH, H.; BALA, M. Perinatal hypoxia diagnostic system by using scalable machine learning algorithms. **International Journal of Innovative Technology and Exploring Engineering**, v. 8, n. 12, p. 1954–1959, 2019.

KAUSHIK, S.; CHOUDHURY, A.; DASGUPTA, N.; NATARAJAN, S.; PICKETT, L. A.; DUTT, V. Ensemble of multi-headed machine learning architectures for time-series forecasting of healthcare expenditures. In: ____. **Applications of Machine Learning**. Singapore: Springer Singapore, 2020. p. 199–216. ISBN 978-981-15-3357-0. Disponível em: https://doi.org/10.1007/978-981-15-3357-0_14.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30, p. 3146–3154. Disponível em: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

KEDDACHI, K.; THELJANI, F. Fetal risk classification based on cardiotocography data: A kernel-based approach. In: ABRAHAM, A.; WEGRZYN-WOLSKA, K.; HASSANIEN, A. E.; SNASEL, V.; ALIMI, A. M. (Ed.). **Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015**. Cham: Springer International Publishing, 2016. p. 327–337. ISBN 978-3-319-29504-6.

KIM, S.-H.; YANG, H.-J.; LEE, S.-W. Fitmine: automatic mining for time-evolving signals of cardiotocography monitoring. **Data Mining and Knowledge Discovery**, v. 31, n. 4, p. 909–933, Jul 2017. ISSN 1573-756X. Disponível em: https://doi.org/10.1007/s10618-017-0493-2.

KITCHENHAM, B. **Guidelines for performing systematic literature reviews in software engineering**. Technical Report RT - EBSE-2007-01: Keele University and Durham University Joint Report, 2007. 65 p.

KITCHENHAM, B. A.; BRERETON, P.; TURNER, M.; NIAZI, M. K.; LINKMAN, S.; PRETORIUS, R.; BUDGEN, D. Refining the systematic literature review process—two participant-observer case studies. **Empirical Software Engineering**, v. 15, n. 6, p. 618–653, Dec 2010. ISSN 1573-7616. Disponível em: https://doi.org/10.1007/s10664-010-9134-8.

KITCHENHAM, B. A.; BUDGEN, D.; BRERETON, O. P. Using mapping studies as the basis for further research – a participant-observer case study. **Information and Software Technology**, v. 53, n. 6, p. 638 – 651, 2011. ISSN 0950-5849. Special Section: Best papers from the APSEC. Disponível em: http://www.sciencedirect.com/science/article/pii/S0950584910002272.

KUHN, M.; JOHNSON, K. *et al.* **Applied predictive modeling**. New York, United States of America: Springer-Verlag New York, 2013. v. 26. 615 p.

KUMAR, N.; SUMAN, A.; SAWANT, K. Relationship between immediate postpartum umbilical cord blood ph and fetal distress. **International Journal of Contemporary Pediatrics**, v. 3, n. 1, p. 113–119, 2016. ISSN 2349-3291. Disponível em: https://www.ijpediatrics.com/index.php/ijcp/article/view/344.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LI, J.; CHEN, Z.; HUANG, L.; FANG, M.; LI, B.; FU, X.; WANG, H.; ZHAO, Q. Automatic classification of fetal heart rate based on convolutional neural network. **IEEE Internet of Things Journal**, v. 6, n. 2, p. 1394–1401, 2019.

LU, Y.; GAO, Y.; XIE, Y.; HE, S. Computerised interpretation systems for cardiotocography for both home and hospital uses. In: **2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)**. [*S. l.*: *s. n.*], 2018. p. 422–427.

LU, Y.; QI, Y.; FU, X. A framework for intelligent analysis of digital cardiotocographic signals from iomt-based foetal monitoring. **Future Generation Computer Systems**, v. 101, p. 1130–1141, 2019. ISSN 0167-739X. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167739X19311434.

MANN, H. B.; WHITNEY, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 18, n. 1, p. 50 – 60, 1947. Disponível em: https://doi.org/10.1214/aoms/1177730491.

MARQUES, J. A. L.; CORTEZ, P. C.; MADEIRO, J. P. D. V.; FONG, S. J.; SCHLINDWEIN, F. S.; ALBUQUERQUE, V. H. C. D. Automatic cardiotocography diagnostic system based on hilbert transform and adaptive threshold technique. **IEEE Access**, v. 7, p. 73085–73094, 2019.

MARQUES, J. A. L.; HAN, T.; WU, W.; MADEIRO, J. P. d. V.; NETO, A. V. L.; GRAVINA, R.; FORTINO, G.; DE ALBUQUERQUE, V. H. C. Iot-based smart health system for ambulatory maternal and fetal monitoring. **IEEE Internet of Things Journal**, p. 1–1, NOV 2020. ISSN 2327-4662.

MA'SUM, M. A.; INTAN, P. R. D.; JATMIKO, W.; KRISNADHI, A. A.; SETIAWAN, N. A.; SUARJAYA, I. M. A. D. Improving deep learning classifier for fetus hypoxia detection in cardiotocography signal. In: **2019 International Workshop on Big Data and Information Security (IWBIS)**. [*S. l.*: *s. n.*], 2019. p. 51–56.

MEHTA, N.; PANDIT, A.; SHUKLA, S. Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. **Journal of Biomedical Informatics**, p. 103311, 2019. ISSN 1532-0464.

MOLLA, M. M. I.; JUI, J. J.; BARI, B. S.; RASHID, M.; HASAN, M. J. Cardiotocogram data classification using random forest based machine learning algorithm. In: ZAIN, Z. M.; AHMAD, H.; PEBRIANTI, D.; MUSTAFA, M.; ABDULLAH, N. R. H.; SAMAD, R.; NOH, M. M. (Ed.). **Proceedings of the 11th National Technical Seminar on Unmanned System Technology 2019**. Singapore: Springer Singapore, 2021. p. 357–369. ISBN 978-981-15-5281-6.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. London, United Kingdom: The MIT Press, 2012. 1098 p. ISBN 0262018020.

NAGENDRA, V.; GUDE, H.; SAMPATH, D.; CORNS, S.; LONG, S. Evaluation of support vector machines and random forest classifiers in a real-time fetal monitoring system based on cardiotocography data. In: **2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)**. [*S. l.*: *s. n.*], 2017. p. 1–6.

NEMENYI, P. Distribution-free multiple comparisons. In: INTERNATIONAL BIOMETRIC SOC. **Biometrics**. [*S. l.*], 1962. v. 18, n. 2, p. 263.

OLSON, R. S.; CAVA, W. L.; MUSTAHSAN, Z.; VARIK, A.; MOORE, J. H. Data-driven advice for applying machine learning to bioinformatics problems. **Pacific Symposium on Biocomputing**, v. 23, p. 192–203, 2018. ISSN 2335-6936. Disponível em: https://pubmed.ncbi.nlm.nih.gov/29218881.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and Software Technology**, v. 64, p. 1 – 18, 2015. ISSN 0950-5849. Disponível em: http://www.sciencedirect.com/science/article/pii/S0950584915000646.

PETROZZIELLO, A.; JORDANOV, I.; PAPAGEORGHIOU, T. A.; REDMAN, W. G. C.; GEORGIEVA, A. Deep learning for continuous electronic fetal monitoring in labor. In: **2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**. [*S. l.*: *s. n.*], 2018. p. 5866–5869. ISSN 1558-4615.

PETROZZIELLO, A.; REDMAN, C. W. G.; PAPAGEORGHIOU, A. T.; JORDANOV, I.; GEORGIEVA, A. Multimodal convolutional neural networks to detect fetal compromise during labor and delivery. **IEEE Access**, v. 7, p. 112026–112036, 2019.

PIRI, J.; MOHAPATRA, P. Exploring fetal health status using an association based classification approach. In: **2019 International Conference on Information Technology (ICIT)**. [*S. l.*: *s. n.*], 2019. p. 166–171.

PIRI, J.; MOHAPATRA, P.; DEY, R. Fetal health status classification using moga - cd based feature selection approach. In: **2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)**. [*S. l.*: *s. n.*], 2020. p. 1–6.

PORTER, P.; MUIRHEAD, F.; BRISBANE, J.; SCHNEIDER, B.; CHOVEAUX, J.; BEAR, N.; CARSON, J.; JONES, K.; SILVA, D.; NEPPE, C. Accuracy, clinical utility, and usability of a wireless self-guided fetal heart rate monitor. **Obstetrics & Gynecology**, v. 137, n. 4, p. 673–681, 2021. ISSN 0029-7844.

POTHARAJU, S. P.; SREEDEVI, M.; ANDE, V. K.; TIRANDASU, R. K. Data mining approach for accelerating the classification accuracy of cardiotocography. **Clinical Epidemiology and Global Health**, v. 7, n. 2, p. 160–164, 2019. ISSN 2213-3984. Disponível em: https://www.sciencedirect.com/science/article/pii/S2213398418300277.

RAMANUJAM, E.; CHANDRAKUMAR, T.; NANDHANA, K.; LAAXMI, N. T. Prediction of fetal distress using linear and non-linear features of ctg signals. In: SMYS, S.; TAVARES, J. M. R. S.; BALAS, V. E.; ILIYASU, A. M. (Ed.). **Computational Vision and Bio-Inspired Computing**. Cham: Springer International Publishing, 2020. p. 40–47. ISBN 978-3-030-37218-7.

RICCIARDI, C.; IMPROTA, G.; AMATO, F.; CESARELLI, G.; ROMANO, M. Classifying the type of delivery from cardiotocographic signals: A machine learning approach. **Computer Methods and Programs in Biomedicine**, v. 196, p. 12, 2020. ISSN 0169-2607. Disponível em: http://www.sciencedirect.com/science/article/pii/S0169260720315455.

ROMANO, M.; BIFULCO, P.; RUFFO, M.; IMPROTA, G.; CLEMENTE, F.; CESARELLI, M. Software for computerised analysis of cardiotocographic traces. **Computer Methods and Programs in Biomedicine**, 124, p. 121–137, FEB 2016. ISSN 0169-2607.

ROTARIU, C.; PASARICA, A.; COSTIN, H.; NEMESCU, D. Spectral analysis of fetal heart rate variability associated with fetal acidosis and base deficit values. In: **2014 International Conference on Development and Application Systems (DAS)**. [*S. l.*: *s. n.*], 2014. p. 210–213.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, v. 323, n. 6088, p. 533–536, Oct 1986. ISSN 1476-4687. Disponível em: https://doi.org/10.1038/323533a0.

SHAFFER, F.; GINSBERG, J. P. An overview of heart rate variability metrics and norms. **Frontiers in public health**, Frontiers Media S.A., v. 5, p. 258–270, Sep 2017. ISSN 2296-2565. Disponível em: https://pubmed.ncbi.nlm.nih.gov/29034226.

SHAH, S. A. A.; AZIZ, W.; ARIF, M.; NADEEM, M. S. A. Decision trees based classification of cardiotocograms using bagging approach. In: **2015 13th International Conference on Frontiers of Information Technology (FIT)**. [*S. l.*: *s. n.*], 2015. p. 12–17.

SHESKIN, D. J. **Handbook of Parametric and Nonparametric Statistical Procedures**. 4. ed. London, United Kingdom: Chapman & Hall/CRC, 2007. 972 p. ISBN 1584888148, 9781584888147.

SIGNORINI, M. G.; MAGENES, G. Advanced signal processing techniques for ctg analysis. In: KYRIACOU, E.; CHRISTOFIDES, S.; PATTICHIS, C. S. (Ed.). **XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016**. Cham: Springer International Publishing, 2016. p. 1205–1210. ISBN 978-3-319-32703-7.

SIGNORINI, M. G.; MAGENES, G.; CERUTTI, S.; ARDUINI, D. Linear and nonlinear parameters for the analysisof fetal heart rate signal from cardiotocographic recordings. **IEEE Transactions on Biomedical Engineering**, v. 50, n. 3, p. 365–374, 2003.

SIGNORINI, M. G.; PINI, N.; MALOVINI, A.; BELLAZZI, R.; MAGENES, G. Dataset on linear and non-linear indices for discriminating healthy and iugr fetuses. **Data in Brief**, v. 29, p. 10, 2020. ISSN 2352-3409. Disponível em: https://www.sciencedirect.com/science/article/pii/S2352340920300585.

SIGNORINI, M. G.; PINI, N.; MALOVINI, A.; BELLAZZI, R.; MAGENES, G. Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring. **Computer Methods and Programs in Biomedicine**, v. 185, p. 105015, 2020. ISSN 0169-2607. Disponível em: https://www.sciencedirect.com/science/article/pii/S0169260719308107.

SILVA NETO, M. G. da; GOMES, D. G. **Sysmap Supplementary Data Repository**. 2021. Disponível em: http://siswebfree.alwaysdata.net/biosysmap/. Acesso em: 01 Mar. 2021.

SILVA NETO, M. G. da; GOMES, D. G.; SOARES, J. M. Credibility on crowdsensing data acquisition: A systematic mapping study. **Journal of Communication and Information Systems**, v. 34, p. 248–269, 2019. ISSN 1980-6604. Disponível em: https://doi.org/10.14209/jcis.2019.26.

SONTAKKE, S. A.; LOHOKARE, J.; DANI, R.; SHIVAGAJE, P. Classification of cardiotocography signals using machine learning. In: **Intelligent Systems and Applications**. Cham: Springer International Publishing, 2019. p. 439–450. ISBN 978-3-030-01057-7.

SPILKA, J.; CHUDÁČEK, V.; HUPTYCH, M.; LEONARDUZZI, R.; ABRY, P.; DORET, M. Intrapartum fetal heart rate classification: Cross-database evaluation. In: KYRIACOU, E.; CHRISTOFIDES, S.; PATTICHIS, C. S. (Ed.). **XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016**. Cham: Springer International Publishing, 2016. p. 1199–1204. ISBN 978-3-319-32703-7.

SPILKA, J.; CHUDáčEK, V.; KOUCKý, M.; LHOTSKá, L.; HUPTYCH, M.; JANKů, P.; GEORGOULAS, G.; STYLIOS, C. Using nonlinear features for fetal heart rate classification. **Biomedical Signal Processing and Control**, v. 7, n. 4, p. 350 – 357, 2012. ISSN 1746-8094. Disponível em: http://www.sciencedirect.com/science/article/pii/S1746809411000619.

SPILKA, J.; FRECON, J.; LEONARDUZZI, R.; PUSTELNIK, N.; ABRY, P.; DORET, M. Sparse support vector machine for intrapartum fetal heart rate classification. **IEEE Journal of Biomedical and Health Informatics**, v. 21, n. 3, p. 664–671, May 2017. ISSN 2168-2194.

STROUX, L.; REDMAN, C. W.; GEORGIEVA, A.; PAYNE, S. J.; CLIFFORD, G. D. Doppler-based fetal heart rate analysis markers for the detection of early intrauterine growth restriction. **Acta Obstetricia et Gynecologica Scandinavica**, v. 96, n. 11, p. 1322–1329, 2017. Disponível em: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/aogs.13228.

SUBASI, A.; KADASA, B.; KREMIC, E. Classification of the cardiotocogram data for anticipation of fetal risks using bagging ensemble classifier. **Procedia Computer Science**, v. 168, p. 34–39, 2020.

SUPRATAK, A.; WU, C.; DONG, H.; SUN, K.; GUO, Y. Survey on feature extraction and applications of biosignals. In: **Machine Learning for Health Informatics: State-of-the-Art and Future Challenges**. Cham: Springer International Publishing, 2016, (Lecture Notes in Artificial Intelligence). p. 161–182. ISBN 978-3-319-50478-0. Disponível em: https://doi.org/10.1007/978-3-319-50478-0_8.

TAN, J. H.; HAGIWARA, Y.; PANG, W.; LIM, I.; OH, S. L.; ADAM, M.; TAN, R. S.; CHEN, M.; ACHARYA, U. R. Application of stacked convolutional and long short-term memory network for accurate identification of cad ecg signals. **Computers in Biology and Medicine**, v. 94, p. 19 – 26, 2018. ISSN 0010-4825. Disponível em: http://www.sciencedirect.com/science/article/pii/S0010482517304201.

TANG, H.; WANG, T.; LI, M.; YANG, X. The design and implementation of cardiotocography signals classification algorithm based on neural network. **Computational and Mathematical Methods in Medicine**, Hindawi, v. 2018, p. 17, Dec 2018. ISSN 1748-670X. Disponível em: https://doi.org/10.1155/2018/8568617.

ULLAH, F. U. M.; ULLAH, A.; HAQ, I. U.; RHO, S.; BAIK, S. W. Short-term prediction of residential power energy consumption via cnn and multi-layer bi-directional lstm networks. **IEEE Access**, v. 8, p. 369–380, 2020. ISSN 2169-3536.

WANG, C.; LIU, M.; WANG, X.; LU, Y. A novel method for nonlinear dynamics analysis of fetal heart rate in fetal distress using visibility graph. In: **Proceedings of the Fourth International Conference on Biological Information and Biomedical Engineering**. New York, NY, USA: Association for Computing Machinery, 2020. (BIBE2020). ISBN 9781450377096. Disponível em: https://doi-org.ez11.periodicos.capes.gov.br/10.1145/3403782.3403810.

WARMERDAM, G. J. J.; VULLINGS, R.; LAAR, J. O. E. H. V.; JAGT, M. B. V. der Hout-Van der; BERGMANS, J. W. M.; SCHMITT, L.; OEI, S. G. Detection rate of fetal distress using contraction-dependent fetal heart rate variability analysis. **Physiological Measurement**, IOP Publishing, v. 39, n. 2, p. 025008, feb 2018. Disponível em: https://doi.org/10.1088/1361-6579/aaa925.

WILCOXON, F. Individual comparisons by ranking methods. **Biometrics Bulletin**, [International Biometric Society, Wiley], v. 1, n. 6, p. 80–83, 1945. ISSN 00994987. Disponível em: http://www.jstor.org/stable/3001968.

WOHLIN, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: **Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering**. New York, NY, USA: ACM, 2014. (EASE '14), p. 38:1–38:10. ISBN 978-1-4503-2476-2. Disponível em: http://doi.acm.org/10.1145/2601248.2601268.

WOLF, H.; BRUIN, C.; DOBBE, J. G.; GORDIJN, S. J.; GANZEVOORT, W. Computerized fetal cardiotocography analysis in early preterm fetal growth restriction – a quantitative comparison of two applications. **Journal of Perinatal Medicine**, v. 47, n. 4, p. 439–447, 2019. Disponível em: https://doi.org/10.1515/jpm-2018-0412.

WU, W.; ZHANG, Y.; LV, Y.; YU, W.; LIN, Y. Shape pattern based sinusoidal fetal heart rate detection from scanned ctg records. In: **2019 IEEE 15th International Conference on Control and Automation (ICCA)**. [*S. l.*: *s. n.*], 2019. p. 1320–1325.

XUE, G. The application of machine learning models in fetal state auto-classification based on cardiotocograms. **IOP Conference Series: Earth and Environmental Science**, IOP Publishing, v. 310, p. 052007, sep 2019. Disponível em: https://doi.org/10.1088/1755-1315/310/5/052007.

YILMAZ, E. Fetal State Assessment from Cardiotocogram Data Using Artificial Neural Networks. **Journal of Medical and Biological Engineering**, 36, n. 6, SI, p. 820–832, DEC 2016. ISSN 1609-0985.

ZARMEHRI, M. N.; CASTRO, L.; SANTOS, J.; BERNARDES, J.; COSTA, A.; SANTOS, C. C. On the prediction of foetal acidaemia: A spectral analysis-based approach. **Computers in Biology and Medicine**, v. 109, p. 235–241, 2019. ISSN 0010-4825. Disponível em: https://www.sciencedirect.com/science/article/pii/S0010482519301490.

ZENG, R.; LU, Y.; LONG, S.; WANG, C.; BAI, J. Cardiotocography signal abnormality classification using time-frequency features and ensemble cost-sensitive svm classifier. **Computers in Biology and Medicine**, v. 130, p. 104218, 2021. ISSN 0010-4825. Disponível em: https://www.sciencedirect.com/science/article/pii/S0010482521000123.

ZHANG, Y.; ZHAO, Z. Fetal state assessment based on cardiotocography parameters using pca and adaboost. In: **2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)**. [*S. l.*: *s. n.*], 2017. p. 1–6.

ZHAO, Z.; DENG, Y.; ZHANG, Y.; ZHANG, Y.; ZHANG, X.; SHAO, L. Deepfhr: intelligent prediction of fetal acidemia using fetal heart rate signals based on convolutional neural network. **BMC Medical Informatics and Decision Making**, v. 19, n. 1, p. 286, Dec 2019. ISSN 1472-6947. Disponível em: https://doi.org/10.1186/s12911-019-1007-5.

ZHAO, Z.; ZHANG, Y.; COMERT, Z.; DENG, Y. Computer-aided diagnosis system of fetal hypoxia incorporating recurrence plot with convolutional neural network. **Frontiers in Physiology**, 10, MAR 12 2019. ISSN 1664-042X.

ZHAO, Z.; ZHANG, Y.; DENG, Y. A comprehensive feature analysis of the fetal heart rate signal for the intelligent assessment of fetal state. **Journal of Clinical Medicine**, 7, n. 8, AUG 2018. ISSN 2077-0383.

ZHOU, D.; BOUSQUET, O.; LAL, T. N.; WESTON, J.; SCHOLKOPF, B. Learning with local and global consistency. In: **Proceedings of the 16th International Conference on Neural Information Processing Systems**. Cambridge, MA, USA: MIT Press, 2004. (NIPS'03), p. 321–328.

APÊNDICE A – MAPEAMENTO SISTEMÁTICO - ESTUDOS SELECIONADOS

Tabela 45 – Lista de Publicações Selecionadas. Fonte: Elaboração própria

P.ID/Ref.	RQ1		RQ2	RQ3
	Tipo	Canal	Estágios	Base de dados
S001 (ZHAO *et al.*, 2018)	OT	J	DP/DT/MC	CTU-UHB
S002 (BOUDET *et al.*, 2020)	OT	J	DP/DT	UTSB
S003 (LU *et al.*, 2019)	OT	J	DP/DT	*private*
S004 (WANG *et al.*, 2020)	T	C	DP/DT	CTU-UHB
S005 (COMERT; KOCAMAZ, 2017)	OT	C	DP/DT/MC	CTU-UHB
S006 (HUDDAR; SONTAKKE, 2019)	M	C	MC	UCI-CTG
S007 (SIGNORINI; MAGENES, 2016)	M	C	DT/MC	*private*
S008 (FUENTEALBA *et al.*, 2017)	T	J	DP/DT	CTU-UHB
S009 (KANNAN *et al.*, 2021)	C	J	MC	UCI-CTG
S010 (MARQUES *et al.*, 2019)	T	J	DP/DT	*private*
S011 (MOLLA *et al.*, 2021)	M	C	MC	UCI-CTG
S012 (DAS *et al.*, 2020c)	T	J	DP/DT/MC	CTU-UHB
S013 (FUENTEALBA *et al.*, 2019a)	T	J	DP/DT/MC	CTU-UHB
S014 (AGRAWAL; MOHAN, 2019)	C	C	MC	UCI-CTG
S015 (BATRA *et al.*, 2017)	C	C	MC	UCI-CTG
S016 (ZENG *et al.*, 2021)	M	J	DP/DT/MC	CTU-UHB
S017 (SONTAKKE *et al.*, 2019)	M	C	MC	UCI-CTG
S018 (ABBAS *et al.*, 2018)	C	C	DP/DT/MC	CTU-UHB
S019 (SUBASI *et al.*, 2020)	C	J	MC	UCI-CTG

continua na próxima pagina

Tabela 45 – *continuação*

P.ID/Ref.	RQ1		RQ2	RQ3
	Tipo	Canal	Estágios	Base de dados
S020 (JEZEWSKI *et al.*, 2016)	M	J	DT/MC	UCI-CTG CTU-UHB
S021 (CZABANSKI *et al.*, 2020)	M	C	DT/MC	CTU-UHB
S022 (ZHAO *et al.*, 2019)	T	J	DP/DT/MC	CTU-UHB
S023 (GIULIANO *et al.*, 2017)	V	J	DP/DT	*private*
S024 (LU *et al.*, 2018)	OT	C	DP/DT	*private*
S025 (GEORGIEVA *et al.*, 2017)	V	J	DP/DT	*private*
S026 (WOLF *et al.*, 2019)	C	J	DP/DT	*private*
S027 (POTHARAJU *et al.*, 2019)	M	J	MC	UCI-CTG
S028 (SHAH *et al.*, 2015)	M	C	MC	UCI-CTG
S029 (ZHAO *et al.*, 2019)	M	J	DP/DT/MC	CTU-UHB
S030 (ANISHA *et al.*, 2021)	M	J	DP/DT/MC	NIFECGDB ADFECGDB PCCDB
S031 (WARMERDAM *et al.*, 2018)	M	J	DP/DT/MC	*private*
S032 (STROUX *et al.*, 2017)	M	J	DP/DT/MC	*private*
S033 (KADHIM; ABED, 2020)	M	C	MC	UCI-CTG
S034 (DAS *et al.*, 2020a)	M	J	MC	UCI-CTG
S035 (INTAN *et al.*, 2019)	C	W	DP/DT/MC	CTU-UHB
S036 (NAGENDRA *et al.*, 2017)	M	C	MC	UCI-CTG

continua na próxima pagina

Tabela 45 – *continuação*

P.ID/Ref.	RQ1		RQ2	RQ3
	Tipo	Canal	Estágios	Base de dados
S037 (PIRI; MOHAPATRA, 2019)	M	C	MC	UCI-CTG
S038 (ALSAYYARI, 2019)	M	J	MC	UCI-CTG
S039 (PIRI *et al.*, 2020)	M	C	MC	UCI-CTG
S040 (GAO; LU, 2019)	M	C	DP/DT/MC	*private*
S041 (BARQUERO-PEREZ *et al.*, 2017)	T	J	DP/DT/MC	HUFA
S042 (BOUDET *et al.*, 2019a)	T	J	DP/DT	UTSB
S043 (AFRIDI *et al.*, 2019)	M	J	MC	UCI-CTG
S044 (COMERT; KOCAMAZ, 2019)	M	J	DP/DT/MC	CTU-UHB
S045 (KEDDACHI; THELJANI, 2016)	M	C	MC	UCI-CTG
S046 (ZHANG; ZHAO, 2017)	M	C	MC	UCI-CTG
S047 (YILMAZ, 2016)	C	J	MC	UCI-CTG
S048 (CHAMIDAH; WASITO, 2015)	M	C	MC	UCI-CTG
S049 (KIM *et al.*, 2017)	T	J	DP/DT/MC	CTU-UHB
S050 (CZABANSKI *et al.*, 2016)	M	J	DP/DT/MC	*private*
S051 (JEZEWSKI *et al.*, 2019)	T	J	DT/MC	CTU-UHB
S052 (CHEN *et al.*, 2019)	M	C	MC	UCI-CTG
S053 (MA'SUM *et al.*, 2019)	M	W	DP/DT/MC	CTU-UHB
S054 (FUENTEALBA *et al.*, 2019b)	M	J	DP/DT/MC	CTU-UHB
S055 (SIGNORINI *et al.*, 2020b)	C	J	DT/MC	DNL-IUGR

continua na próxima pagina

Tabela 45 – *continuação*

P.ID/Ref.	RQ1		RQ2	RQ3
	Tipo	Canal	Estágios	Base de dados
S056 (HUANG *et al.*, 2020)	M	C	MC	UCI-CTG
S057 (ITO *et al.*, 2021)	T	J	DP/DT	*private*
S058 (FERGUS *et al.*, 2020)	M	J	DP/DT/MC	CTU-UHB
S059 (PETROZZIELLO *et al.*, 2019)	M	J	DP/DT/MC	*private* CTU-UHB SpaM
S060 (ZARMEHRI *et al.*, 2019)	T	J	DP/DT/MC	*private* SpaM CTU-UHB
S061 (COMERT; KOCAMAZ, 2018)	OT	J	DP/DT/MC	CTU-UHB
S062 (KAUR *et al.*, 2019)	M	J	MC	UCI-CTG
S063 (DAS *et al.*, 2020b)	M	J	DP/DT/MC	CTU-UHB
S064 (ALSAGGAF *et al.*, 2020)	T	J	DP/DT/MC	CTU-UHB
S065 (RAMANUJAM *et al.*, 2020)	M	C	DP/DT/MC	CTU-UHB
S066 (IRAJI, 2019)	M	J	MC	UCI-CTG
S067 (COMERT *et al.*, 2019)	M	J	DP/DT/MC	CTU-UHB
S068 (COMERT *et al.*, 2018)	M	J	DP/DT/MC	CTU-UHB
S069 (DAS *et al.*, 2019)	M	C	MC	UCI-CTG
S070 (WU *et al.*, 2019)	M	C	DP/DT/MC	*private*
S071 (ROMANO *et al.*, 2016)	OT	J	DP/DT	*private*
S072 (FENG *et al.*, 2018)	M	C	DP/DT/MC	CTU-UHB
S073 (XUE, 2019)	M	J	MC	UCI-CTG
S074 (HOODBHOY *et al.*, 2019)	C	J	MC	UCI-CTG

continua na próxima pagina

Tabela 45 – *continuação*

P.ID/Ref.	RQ1		RQ2	RQ3
	Tipo	Canal	Estágios	Base de dados
S075 (GYLLENCREUTZ *et al.*, 2018)	V	J	DP/DT	*private*