CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

DETECÇÃO DE MOTIFS EM SÉRIES TEMPORAIS USANDO AUTOENCODERS: UMA ANÁLISE DO USO EM DADOS DE ECG

Luiz Cláudio Lemos de Oliveira

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Eduardo Soares Ogasawara

Rio de Janeiro,

15 de junho de 2026

Detecção de Motifs em Séries Temporais Usando Autoencoders: Uma Análise do Uso em Dados de ECG

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Luiz Cláudio Lemos de Oliveira

Aprovada por:

Presidente, Professor D.Sc. Eduardo Soares Ogasawara (CEFET/RJ)

Professora D.Sc. Laura Silva de Assis (CEFET/RJ)

Professora D.Sc. Helga Dolorico Balbi (CEFET/RJ)

D.Sc. Rebecca Pontes Salles (INRIA/FRA)

Rio de Janeiro,

15 de junho de 2026

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

O48 Oliveira, Luiz Cláudio Lemos de
Detecção de motifs em séries temporais usando autoencoders: uma análise do uso em dados de ECG / Luiz Cláudio Lemos de Oliveira – 2026.
81f : il. (algumas color.), enc.

Dissertação (Mestrado). Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2026.
Bibliografia : f. 73-81.
Orientador: Eduardo Soares Ogasawara.

1. Redes neurais. 2. Mineração de dados (Computação). 3. Eletrocardiografia — Processamento de dados. I. Ogasawara, Eduardo Soares (Orient.). II. Título.

CDD 006.3

Elaborada pelo bibliotecário Leandro Mota de Menezes – CRB/7 nº 5281

DEDICATÓRIA

À minha esposa Carolina e aos meus filhos Gabriel e Maria Luiza, fontes de motivação para trilhar este caminho. À minha mãe Gloria e meu pai Pedro, por me transmitirem o valor da educação. A todos que me apoiaram e acreditaram nesta jornada.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Eduardo Ogasawara, pelos direcionamentos, correções e auxílio permanente na condução deste trabalho. Aos membros do Grupo de Pesquisa Data Analytics LAB (DAL), pelas atividades colaborativas e pelas contribuições para o meu desenvolvimento profissional. Aos professores do PPCIC, pelos conhecimentos que fundamentaram minha jornada acadêmica e científica. À Petrobras, por meio de meus gestores e colegas, pelo incentivo e suporte que permitiram compatibilizar as atividades corporativas com os compromissos de pesquisa.

RESUMO

Detecção de Motifs em Séries Temporais Usando Autoencoders: Uma Análise do Uso em Dados de ECG

Luiz Cláudio Lemos de Oliveira

Orientadores:
Eduardo Soares Ogasawara

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

A descoberta de *motifs* em séries temporais biomédicas, como o eletrocardiograma (ECG), consiste na identificação de padrões recorrentes com potencial diagnóstico. Métodos tradicionais como o *Symbolic Aggregate approXimation* (SAX) apresentam limitações por dependerem de pressupostos estatísticos rígidos, inadequados para sinais fisiológicos complexos. Paralelamente, *autoencoders* demonstram capacidade superior de aprendizado de representações não lineares, mas sua aplicação na descoberta de *motifs* em ECG permanece inexplorada — lacuna que este trabalho endereça.

Este trabalho propõe o Latent Encoding for Motif Discovery (LEMOD), um método cuja contribuição central é a substituição da discretização estatística do SAX por codificação neural adaptativa via *autoencoders*, combinada com mecanismo de busca e ranqueamento de *motifs* por frequência de ocorrência e entropia de *Shannon*. A metodologia compreende três etapas: (*i*) validação da capacidade de reconstrução do *autoencoder*; (*ii*) treinamento de modelos; e (*iii*) detecção, análise de sensibilidade paramétrica e avaliação da generalização, com comparação sistemática ao SAX.

Os experimentos foram realizados com dados da MIT-BIH *Arrhythmia Database*. Na etapa de validação, observou-se fidelidade de reconstrução superior a 99%. Nos testes de detecção de *motifs*, o LEMOD apresentou ganho relativo de 92% em revocação e superioridade de 78% em F_1 frente ao SAX, evidenciando maior sensibilidade a padrões morfológicos sutis sem comprometer a especificidade. A entropia média dos *motifs* (0,44 bits *versus* 0,90 bits do SAX) foi menor, indicando maior homogeneidade morfológica interna dos padrões descobertos — evidência de que a codificação neural identifica estruturas recorrentes mais coerentes do que a discretização tradicional. O tempo de detecção foi 15% menor no LEMOD. O experimento de detecção cruzada confirmou generalização das representações latentes a séries distintas da mesma derivação.

Conclui-se que a substituição da discretização tradicional por codificação neural adaptativa é viável e proporciona ganhos quantitativos e qualitativos na descoberta de *motifs* em ECG, estabelecendo base metodológica para ferramentas de análise automatizada de sinais biomédicos.

Palavras-chave: Séries temporais, *Autoencoders*, eletrocardiograma (ECG)

Rio de Janeiro,

15 de junho de 2026

ABSTRACT

Detecção de Motifs em Séries Temporais Usando Autoencoders: Uma Análise do Uso em Dados de ECG

Luiz Cláudio Lemos de Oliveira

Advisors:
Eduardo Soares Ogasawara

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

The discovery of motifs in biomedical time series, such as electrocardiograms (ECG), consists in identifying recurrent patterns with diagnostic potential. Traditional methods such as SAX present limitations due to their dependence on rigid statistical assumptions, which are inadequate for complex physiological signals. In parallel, autoencoders demonstrate superior capacity for learning nonlinear representations, but their application to motif discovery in ECG remains unexplored — a gap that this work addresses.

This work proposes LEMOD, a method whose central contribution is the replacement of the statistical discretization of SAX by adaptive neural encoding via autoencoders, combined with a mechanism for searching and ranking motifs by frequency of occurrence and Shannon entropy. The methodology comprises three stages: (*i*) validation of the autoencoder's reconstruction capacity; (*ii*) model training; and (*iii*) detection, parametric sensitivity analysis and generalization evaluation, with systematic comparison to SAX.

Experiments were conducted using data from the MIT-BIH Arrhythmia Database. In the validation stage, reconstruction fidelity above 99% was observed. In detection tests, LEMOD achieved a relative gain of 92% in recall and 78% superiority in F_1 over SAX, evidencing greater sensitivity to subtle morphological patterns without compromising specificity. The average entropy of the motifs (0.44 bits *versus* 0.90 bits for SAX) was lower, indicating greater internal morphological homogeneity — evidence that neural encoding identifies more coherent recurrent structures than traditional discretization. Detection time was 15% lower in LEMOD. The cross-detection experiment confirmed generalization of the learned latent representations to distinct series of the same lead.

It is concluded that replacing traditional discretization with adaptive neural encoding is viable and provides quantitative and qualitative gains in motif discovery in ECG signals, establishing a methodological basis for automated biomedical signal analysis tools.
Keywords: Time series, Autoencoders, electrocardiogram (ECG)

Rio de Janeiro,

June 15, 2026

SUMÁRIO

I Introdução **1**

II Referencial Teórico **7**

II.1 Séries temporais 7

II.2 Pré-processamento de séries temporais 10

II.3 Detecção de *motifs* em séries temporais 13

II.4 Métricas de avaliação na detecção de *motifs* 14

II.5 Redes neurais e *autoencoders* 19

III Trabalhos Relacionados **22**

III.1 Métodos de Detecção de *Motifs* em Séries Temporais 22

III.2 Revisões Sistemáticas e Estudos Comparativos 27

III.3 Síntese de Trabalhos Relacionados 28

IV Metodologia **30**

IV.1 Visão Geral da Abordagem Proposta 30

IV.2 Caracterização do *Dataset* e Protocolo de Seleção 34

IV.3 Etapa I: Validação da Capacidade de Reconstrução 36

IV.4 Etapa II: Treinamento dos Modelos *Autoencoder* 40

IV.5 Etapa III: Comparação Experimental Sistemática 41

V Avaliação Experimental **47**

V.1 Configuração Experimental 47

V.2 Etapa I: Validação da Capacidade de Reconstrução 49

V.3 Etapa II: Treinamento dos Modelos *Autoencoder* 57

V.4 Etapa III: Comparação Experimental Sistemática 57

V.5 Síntese do Capítulo 67

VI Conclusão **69**

Referências **73**

LISTA DE FIGURAS

II.1 Série temporal univariada de eletrocardiograma do MIT-BIH *Arrhythmia Database*. 8

II.2 Série temporal estacionária: ECG com batimentos cardíacos normais. 9

II.3 Série temporal não estacionária: ECG com episódios de arritmia. 9

IV.1 Estrutura metodológica em três etapas experimentais interdependentes. 33

V.1 Distribuição de R^2 em função de $train_size$. 50

V.2 Dispersão relativa de R^2 em função de $train_size$. 50

V.3 Distribuição de NRMSE (%) em função de $train_size$. 51

V.4 Distribuição de MAPE (%) em função de $train_size$. 51

V.5 Distribuição de R^2 em função de sw_size. 52

V.6 Dispersão relativa de R^2 em função de sw_size. 52

V.7 Distribuição de NRMSE (%) em função de sw_size. 53

V.8 Distribuição de MAPE (%) em função de sw_size. 53

V.9 Distribuição de R^2 em função de $ratio$. 54

V.10 Dispersão relativa de R^2 em função de $ratio$. 55

V.11 Distribuição de NRMSE (%) em função de $ratio$. 55

V.12 Distribuição de MAPE (%) em função de $ratio$. 56

V.13 Comparação das métricas de desempenho entre o LEMOD e o SAX. 59

V.14 Entropia de *Shannon* dos *motifs* identificados e tempo de detecção medido. 60

V.15 Exemplo de detecções em segmento de ritmo sinusal regular da série `r100_MLII`. 60

V.16 Exemplo de detecções em segmento com arritmia da série `r100_MLII`. 61

V.17 Acurácia em função de $alphabet_size$, fixando $occurrences = 6$. 61

V.18 Precisão em função de $alphabet_size$, fixando $occurrences = 6$. 62

V.19 Revocação em função de $alphabet_size$, fixando $occurrences = 6$. 62

V.20 F_1 em função de $alphabet_size$, fixando $occurrences = 6$. 63

V.21 Acurácia em função de $occurrences$, fixando $alphabet_size = 10$. 63

V.22 Precisão em função de $occurrences$, fixando $alphabet_size = 10$. 64

V.23 Revocação em função de $occurrences$, fixando $alphabet_size = 10$. 64

V.24 F_1 em função de $occurrences$, fixando $alphabet_size = 10$. 64

LISTA DE TABELAS

II.1 Configuração da Matriz de Confusão da detecção de *motifs*. 15

II.2 Métricas básicas para avaliação da detecção de *motifs*. 15

III.1 Principais algoritmos clássicos de descoberta de *motifs* em séries temporais. 25

III.2 Principais algoritmos especializados de descoberta de *motifs* em séries temporais. 27

IV.1 Caracterização do MIT-BIH *Arrhythmia Database*. 35

V.1 Síntese das configurações experimentais implementadas nas três etapas avaliativas. 49

V.2 Estatísticas descritivas de R^2 (%) em função de $train_size$. 50

V.3 Estatísticas descritivas de R^2 (%) em função de sw_size. 52

V.4 Estatísticas descritivas de R^2 (%) em função de $ratio$. 54

V.5 Distribuição agregada das métricas de fidelidade (%) para a configuração de melhor balanço. 56

V.6 Métricas de desempenho agregadas para configuração principal. 58

V.7 Métricas de desempenho em detecção cruzada estratificadas por derivação. 65

LISTA DE ACRÔNIMOS

R^2	*Coeficiente de Determinação.*
AdaReLU	*Adaptive Rectified Linear Unit.*
aSAX	*adaptive* SAX.
CV	*Coeficiente de Variação.*
DAME	*Disk Aware Motif Enumeration.*
DNA	ácido desoxirribonucleico.
DTW	*Dynamic Time Warping.*
ECG	eletrocardiograma.
EEG	eletroencefalograma.
EMD	*Extended Motif Discovery.*
FFT	*Fast Fourier Transform.*
FN	*Falso Negativo.*
FP	*Falso Positivo.*
GPU	*Graphics Processing Unit.*
IQR	*Intervalo Interquartil.*
LEMOD	Latent Encoding for Motif Discovery.
LSTM	*Long Short-Term Memory.*
MAD	*Desvio Absoluto da Mediana.*
MAPE	*Erro Percentual Absoluto Médio.*
MDL	*Minimum Description Length.*
MDLats	*Motif Discovery for Large-scale Time Series.*
MIT-BIH	*Massachusetts Institute of Technology Beth Israel Hospital.*
MK	*Mueen-Keogh algorithm.*

MOEN	*MOtif ENumerator*.
MP	*Matrix Profile*.
MSE	*Erro Quadrático Médio*.
NRMSE	*Erro Quadrático Médio Normalizado*.
PAA	*Piecewise Aggregate Approximation*.
PCA	*Principal Component Analysis*.
RCV	*Coeficiente de Variação Robusto*.
ReLU	*Rectified Linear Unit*.
RMSE	*Raiz do Erro Quadrático Médio*.
SAX	*Symbolic Aggregate approXimation*.
SCRIMP++	Nome que faz alusão ao verbo inglês *scrimp* (economizar/ poupar), enfatizando o processamento econômico do algoritmo.
SoftED	*Soft Event Detection*.
STAMP	*Scalable Time Series Anytime Matrix Profile*.
STOMP	*Scalable Time Series Ordered-Search Matrix Profile*.
SWAMP	*Scalable Warping Aware Matrix Profile*.
TN	*Verdadeiro Negativo*.
TP	*Verdadeiro Positivo*.
VALMOD	*Variable Length Motif Discovery*.

Capítulo I Introdução

O cenário contemporâneo de geração massiva de dados, provenientes de fontes diversificadas, consolida as séries temporais como um dos tipos de dados mais prevalentes na comunidade científica nas últimas décadas, abrangendo praticamente todos os domínios científicos e sociais. Essa ubiquidade reflete a diversidade de aplicações que geram dados sequenciais e a importância crescente das técnicas de mineração de dados temporais para a análise de fenômenos complexos em múltiplas áreas do conhecimento [Linardi et al., 2020].

Esse interesse crescente configura oportunidades para o desenvolvimento de técnicas de detecção de padrões significativos por múltiplas comunidades de pesquisa, dado que a identificação desses padrões em séries temporais se traduz em informações valiosas para uma ampla variedade de domínios de aplicação [Chandola et al., 2009].

Séries temporais constituem conjuntos ordenados de variáveis de valor real, em que cada elemento da sequência representa uma observação temporal específica, sendo a preservação da ordem cronológica fundamental para qualquer procedimento analítico subsequente [Keogh and Lin, 2005]. Embora a dimensão que impõe a ordenação das séries seja geralmente o tempo, séries de dados também podem ser organizadas segundo outras medidas, como ângulos em perfis radiais em astronomia ou posições em sequências de genoma. Os termos série temporal, série de dados e sequência são utilizados de forma intercambiável na literatura [Linardi et al., 2020].

Motifs são definidos como sequências recorrentes em séries temporais, desempenhando papel fundamental na exploração não supervisionada de dados ao possibilitar a sumarização por meio da identificação de estruturas repetitivas que revelam comportamentos subjacentes relevantes [Mueen, 2013]. *Motifs* podem representar eventos críticos em diferentes contextos, como um padrão que precede uma convulsão em uma série temporal de eletroencefalograma (EEG) ou uma sequência de ácido desoxirribonucleico (DNA) preservada pela evolução [Castro and Azevedo, 2010]. A capacidade de revelar hierarquias estruturais e padrões relevantes, muitas vezes ocultos devido à complexidade e ao ruído nos dados, torna a descoberta de *motifs* uma primitiva importante na mineração de séries temporais e uma ferramenta analítica em campos como neurociência, medicina e biologia [Mueen, 2013; Zhu et al., 2018].

Em contextos biomédicos, exames de eletrocardiograma (ECG) constituem ferramenta consolidada na avaliação cardiovascular, permitindo identificar diversas patologias cardíacas, especialmente

distúrbios do ritmo [Roy et al., 2023]. A identificação de alterações rítmicas fundamenta-se na distinção entre padrões normais e anormais de atividade cardíaca, tornando o reconhecimento de regularidades nos registros eletrocardiográficos elemento central na construção de diagnósticos precisos [Roy et al., 2023]. No entanto, a análise desse tipo de dado apresenta dificuldades consideráveis, pois envolve características estruturais complexas e dependências de longo prazo determinantes para a identificação adequada de arritmias [Ding et al., 2023].

Apesar da relevância demonstrada, diversos desafios permeiam o problema de detecção de *motifs* em séries temporais, especialmente porque a natureza desses eventos é desconhecida, o que dificulta a escolha e a parametrização de métodos que dependem fundamentalmente das premissas assumidas sobre as propriedades estatísticas dos dados [Salles et al., 2020].

Poirier–Herbeck et al. [2022] identificam obstáculos a serem superados pelos algoritmos de detecção, como a dificuldade em detectar *motifs* de comprimentos desconhecidos, questões de eficiência computacional, a descoberta de padrões com variações morfológicas e restrições no processamento de dados em tempo real. Chandola et al. [2009] ampliam essa perspectiva ao destacar outros desafios, incluindo a dificuldade na demarcação clara de padrões regulares e irregulares, a natureza dinâmica dos perfis comportamentais ao longo do tempo e a escassez de conjuntos de dados previamente categorizados para o desenvolvimento e a avaliação de algoritmos de detecção.

Um aspecto crítico salientado por Chiu et al. [2003] é o impacto do ruído na detecção de *motifs*, pois perturbações mínimas podem comprometer severamente as métricas de similaridade, dificultando o reconhecimento de padrões semelhantes em conjuntos de dados. A determinação das dimensões temporais dos *motifs*, conforme observado por Linardi et al. [2020], constitui outro desafio, já que os comprimentos dos padrões procurados não são previamente conhecidos.

Além disso, a implementação de soluções computacionalmente eficientes continua sendo questão central, particularmente considerando as demandas por processamento em tempo real e a necessidade de adaptação às propriedades dinâmicas inerentes aos dados temporais [Rong et al., 2020]. Chiu et al. [2003] observam que implementações de força bruta para identificação de *motifs* apresentam limitações de escalabilidade, pois a complexidade computacional cresce de forma quadrática com as dimensões dos conjuntos analisados, inviabilizando sua aplicação em bases de grande volume.

Métodos aproximados baseados em discretização simbólica, como o SAX, reduzem o custo computacional por meio de agregação temporal e indexação, alcançando complexidade $O(n \log n)$ [Lin et al., 2003, 2007]. Essa estratégia viabiliza buscas eficientes e justifica sua adoção como *baseline* em descoberta de *motifs*. Entretanto, o SAX depende da suposição de normalidade estatística e de *breakpoints* fixos. Em sinais de eletrocardiograma (ECG), cuja morfologia é frequentemente não gaussiana e altamente assimétrica em contextos patológicos, *breakpoints* fixos tendem a agrupar padrões morfologicamente distintos sob o mesmo símbolo ou a fragmentar variações sutis clinicamente

relevantes [Butler and Kazakov, 2015; Castro and Azevedo, 2012].

Em contraste aos métodos aproximados baseados em discretização simbólica, métodos exatos baseados em comparação exaustiva apresentam complexidade computacional quadrática $O(n^2)$, o que compromete sua escalabilidade em bases reais [Mueen et al., 2009]. Essa complexidade computacional motiva o desenvolvimento de métodos mais eficientes, como o *Matrix Profile* (MP), uma estrutura de dados que armazena as métricas de similaridade mínimas para cada segmento temporal em relação ao seu correspondente mais similar na série analisada [Zhu et al., 2018]. O MP consolidou-se como ferramenta para identificação de *motifs* por meio de algoritmos especializados como *Scalable Time Series Anytime Matrix Profile* (STAMP) e *Scalable Time Series Ordered-Search Matrix Profile* (STOMP), embora a maioria dessas implementações empregue exclusivamente métricas baseadas na Distância Euclidiana [Zhu et al., 2018].

Do ponto de vista da eficiência computacional, embora o MP tenha introduzido otimizações baseadas em *Fast Fourier Transform* (FFT), métodos exatos, incluindo essas implementações otimizadas, permanecem fundamentalmente limitados pelo piso assintótico de $\Omega(n^2)$, uma vez que o problema exige a avaliação de distâncias entre todos os pares de subsequências. Esse limite inferior implica que, independentemente de otimizações de implementação, o custo computacional cresce rapidamente com o tamanho da série, restringindo sua aplicabilidade em cenários de grande escala ou análises exploratórias.

Esse tipo de abordagem de similaridade, apesar da eficiência computacional, apresenta restrições na análise de padrões com defasagens temporais ou variações de velocidade, pois a Distância Euclidiana pressupõe alinhamento temporal perfeito entre as sequências comparadas. Quando um padrão se manifesta em momentos diferentes ou é executado em ritmos distintos, essa métrica pode falhar em reconhecer similaridades morfológicas, levando a interpretações errôneas [Alaee et al., 2020].

O *Dynamic Time Warping* (DTW) representa uma estratégia de correspondência mais flexível, possibilitando o reconhecimento de padrões similares independentemente de variações temporais locais por meio de alinhamentos adaptativos [Alaee et al., 2020]. Entretanto, a integração entre técnicas de DTW e algoritmos de descoberta de *motifs* apresenta demandas computacionais consideráveis, pois ambas as estratégias requerem uso intensivo de recursos como tempo de processamento e memória, devido à complexidade envolvida na análise de grandes conjuntos de dados [Alaee et al., 2020].

Esta desafiadora complexidade, inerente à identificação de *motifs* em séries temporais, impulsiona a investigação de algoritmos mais sofisticados, incluindo abordagens de aprendizado profundo como *autoencoders* [Rong et al., 2020]. Estas arquiteturas neurais destacam-se no contexto de aprendizado não supervisionado por sua capacidade de extrair representações latentes dos dados

enquanto executam compressão dimensional, oferecendo soluções promissoras para a análise de conjuntos complexos [Berahmand et al., 2024]. Para dados sequenciais, os *autoencoders* demonstram eficácia na captura de características temporais relevantes, estabelecendo codificações compactas que preservam propriedades informacionais dos sinais originais [Liang et al., 2021].

Comparados com implementações convencionais, como métodos lineares de redução de dimensionalidade, modelos rasos baseados em engenharia manual de características e abordagens tradicionais de modelagem temporal, os modelos de aprendizado profundo apresentam vantagens na análise de dados temporais, incluindo extração automática de características, capacidade de aprender mapeamentos não lineares complexos e maior robustez no processamento de conjuntos de dados de grande escala [Berahmand et al., 2024]. Para dados biomédicos, particularmente sinais de ECG, técnicas de aprendizado profundo evidenciam capacidades superiores de extração de características e de desempenho classificatório em comparação com métodos tradicionais de aprendizado de máquina [Liang et al., 2021], representando uma direção metodológica promissora que pode superar fragilidades dos processos convencionais de análise [Ding et al., 2023].

Apesar desse potencial, revisões recentes indicam que ainda são escassas as abordagens que integram explicitamente *autoencoders* a *pipelines* clássicos de descoberta de *motifs* e inexistentes quando aplicadas em dados de ECG [Guerrmi et al., 2025]. A literatura evidencia um compromisso entre métodos exatos, que oferecem precisão à custa de complexidade quadrática, e métodos simbólicos aproximados, que escalam eficientemente mas impõem pressupostos estatísticos rígidos. Embora *autoencoders* demonstrem capacidade superior para aprender representações latentes expressivas em diversos domínios [Neloy and Turgeon, 2024], sua aplicação específica para detecção de *motifs* em séries temporais permanece pouco explorada.

Esse hiato é particularmente relevante, pois métodos tradicionais, como o SAX, frequentemente falham em identificar padrões sutis mas clinicamente significativos em dados fisiológicos, impondo representações fixas que podem não capturar a complexidade dos dados biomédicos [Castro and Azevedo, 2012]. A ausência de implementações que combinem a flexibilidade de representação dos *autoencoders* com técnicas robustas de detecção de *motifs* baseadas em entropia representa oportunidade para avanços metodológicos que podem impactar diretamente aplicações clínicas. Essas carências fundamentam a questão central desta pesquisa: o uso de *autoencoders* para representação de séries temporais resulta em melhor desempenho na detecção de *motifs* quando comparado a métodos tradicionais?

A hipótese baseia-se na capacidade demonstrada dos *autoencoders* de descobrir padrões de forma não supervisionada com qualidade comparável às abordagens clássicas, oferecendo vantagens na descoberta automática de padrões que cobrem a série temporal de forma mais abrangente [Noering et al., 2021]. Essa capacidade sugere que a integração de *autoencoders* pode oferecer alternativa mais

robusta e adaptável para a identificação de *motifs* em cenários de alta complexidade e variabilidade dos dados, justificando a presente investigação sobre essa integração.

Para preencher essa lacuna, a presente pesquisa propõe o **Latent Encoding for Motif Discovery (LEMOD)**, abordagem que substitui a discretização tradicional por *autoencoders*, explorando o potencial de representações latentes aprendidas automaticamente para melhorar a detecção de *motifs* em séries temporais usando dados de ECG. O objetivo geral desta investigação é desenvolver e validar um método que combine a capacidade de compressão não linear dos *autoencoders* com técnicas de descoberta de *motifs* baseadas em entropia, superando limitações identificadas por Rong et al. [2020] relacionadas à perda de propriedades de similaridade nos métodos tradicionais de redução de dimensionalidade, nas quais padrões originalmente similares podem ser identificados como dissimilares após a transformação. Esse tratamento visa preservar características essenciais dos padrões temporais enquanto mantém complexidade computacional viável ($O(n \log n)$ na fase de descoberta), permitindo detecção eficiente e precisa de estruturas repetitivas em sinais biomédicos complexos.

Os objetivos específicos desta pesquisa são: (*i*) avaliar a sensibilidade dos parâmetros presentes na implementação do *autoencoder*, de modo a verificar sua capacidade de reconstrução dos dados da série temporal após codificação e decodificação, (*ii*) comparar o LEMOD e o método SAX tradicional, utilizando métricas *soft* de avaliação (precisão, revocação e F_1 *score*) aplicadas à tarefa de detecção de *motifs*, e (*iii*) validar empiricamente a capacidade do método proposto em identificar padrões clinicamente relevantes no *Massachusetts Institute of Technology Beth Israel Hospital* (MIT-BIH) *Arrhythmia Database* [Moody and Mark, 1992, 2001], demonstrando sua aplicabilidade prática em cenários de análise de sinais biomédicos. Esses objetivos permitem a avaliação sistemática da eficácia desta abordagem em comparação com métodos consolidados na literatura.

As contribuições desta pesquisa abrangem aspectos metodológicos, aplicados e teóricos no campo da análise de séries temporais biomédicas. Os resultados são interpretados no contexto dos dados de ECG. Do ponto de vista metodológico, espera-se desenvolver proposta para detecção de *motifs* que preserve as propriedades de similaridade dos dados originais por meio do uso de representações latentes aprendidas automaticamente, superando limitações dos métodos de discretização linear identificadas na literatura.

No âmbito aplicado, prevê-se demonstrar a eficácia do tratamento proposto na identificação de padrões regulares e arrítmicos em dados de ECG, contribuindo para o avanço das técnicas de análise automatizada de sinais biomédicos e fornecendo ferramenta complementar aos métodos tradicionais de diagnóstico cardiovascular.

Do ponto de vista teórico, a pesquisa visa contribuir para o entendimento de como *autoencoders* podem ser integrados em *pipelines* de mineração de séries temporais, estabelecendo princípios e

diretrizes para investigações futuras que combinem aprendizado profundo com técnicas clássicas de descoberta de padrões em dados temporais complexos.

Este trabalho está organizado em cinco capítulos: após esta Introdução, o Capítulo II aborda o Referencial Teórico sobre séries temporais e *autoencoders*. O Capítulo III revisa os Trabalhos Relacionados à detecção de *motifs* em séries temporais. Em seguida, o Capítulo IV detalha a Metodologia proposta para o LEMOD na detecção de *motifs*. O Capítulo V apresenta a Avaliação Experimental e os resultados obtidos e o Capítulo VI expõe a Conclusão, destacando as contribuições da presente pesquisa e as oportunidades de trabalhos futuros.

Capítulo II Referencial Teórico

Este capítulo apresenta o referencial teórico sobre séries temporais e detecção de *motifs* utilizando *autoencoders*. A Seção II.1 aborda aspectos elementares de séries temporais e desafios centrais de sua análise. A Seção II.2 discute técnicas de pré-processamento aplicadas a esse tipo de dado. A Seção II.3 explora o conceito de *motifs* em séries temporais, examinando métodos para sua detecção. A Seção II.4 apresenta métricas de avaliação de algoritmos utilizadas na materialização dos resultados desta pesquisa. Por fim, a Seção II.5 trata de redes neurais e *autoencoders*, avaliando sua aplicação em compressão e reconstrução de séries temporais, aspectos essenciais para a abordagem metodológica proposta.

II.1 Séries temporais

Uma série temporal constitui uma sequência ordenada de observações, em que cada elemento representa uma medição em um instante específico, preservando a dependência sequencial inerente aos dados temporais [Esling and Agon, 2012; Fu, 2011; Ogasawara et al., 2025; Shumway and Stoffer, 2025]. Pode ser expressa como na Equação II.1, onde X representa a série temporal, x_i denota cada observação individual, n é o número total de observações e i é o índice temporal da observação.

$$X = (x_1, x_2, \ldots, x_n), \quad x_i \in \mathbb{R} \tag{II.1}$$

Essa estrutura apresenta propriedades matemáticas que incluem indexação temporal explícita, ordem cronológica obrigatória e dependência entre observações consecutivas, distinguindo-se de conjuntos de dados convencionais pela necessidade de preservar a sequencialidade para análises significativas [Keogh and Kasetty, 2003; Shumway and Stoffer, 2025]. As séries temporais podem ser univariadas, quando representam a evolução de uma única variável ao longo do tempo, ou multivariadas, quando múltiplas variáveis são observadas simultaneamente em cada instante [Esling and Agon, 2012; Fu, 2011; Ogasawara et al., 2025]. Assume-se sincronização entre as medições e aplicam-se técnicas de interpolação ou imputação quando houver dados faltantes [Ogasawara et al., 2025].

Para facilitar o processamento e a aplicação de algoritmos de mineração de dados temporais, essas estruturas organizam-se em representação matricial $X \in \mathbb{R}^{n \times d}$, em que n denota o compri-

mento temporal da série e d representa a dimensionalidade dos atributos observados. Para o caso univariado, adota-se $d = 1$ e $X \in \mathbb{R}^{n \times 1}$. A Figura II.1 ilustra uma série temporal univariada.

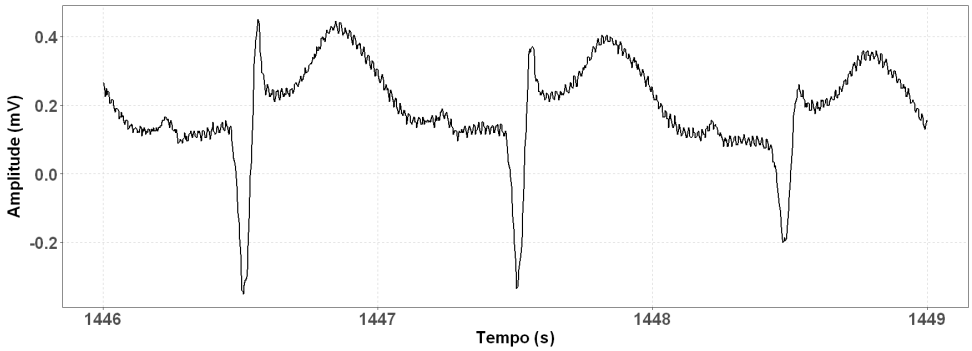


Figura II.1: Série temporal de ECG do MIT-BIH *Arrhythmia Database* [Moody and Mark, 1992]: amplitudes do sinal elétrico cardíaco em (mV) no eixo y e tempo em (s) no eixo x; taxa de amostragem de 360 Hz.

As propriedades estatísticas das séries temporais manifestam-se por componentes estruturais que determinam seu comportamento [Esling and Agon, 2012; Fu, 2011; Ogasawara et al., 2025; Shumway and Stoffer, 2025]. A tendência representa o movimento de longo prazo da série e pode ser detectada por regressão linear ou filtros de suavização. A sazonalidade caracteriza padrões cíclicos que se repetem em intervalos fixos e pode ser analisada por decomposição aditiva ou multiplicativa, conforme a interação entre os componentes. As flutuações não explicadas pela tendência ou sazonalidade compõem o termo de ruído, tratado como processo estocástico, geralmente assumido como ruído branco gaussiano com média zero e variância finita [Ogasawara et al., 2025]. A decomposição aditiva pode ser expressa pela Equação II.2, onde τ, π e ω representam, respectivamente, as componentes de tendência, sazonalidade e ruído aleatório.

$$x_t = \tau_t + \pi_t + \omega_t \tag{II.2}$$

A autocorrelação quantifica a dependência temporal entre observações por meio da função de autocovariância, conceito estabelecido na literatura de séries temporais [Esling and Agon, 2012; Shumway and Stoffer, 2025], expressa na Equação II.3, onde $\gamma_x(s,t)$ representa a função de autocovariância, s e t são instantes temporais específicos, x_s e x_t denotam as observações nos respectivos instantes, μ_s e μ_t são as médias das observações nos instantes correspondentes, e $E[\cdot]$ é o operador esperança matemática.

$$\gamma_x(s,t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)] \tag{II.3}$$

Essa função fornece informações sobre a estrutura de dependência temporal e fundamenta a detecção de padrões repetitivos.

A estacionariedade constitui conceito central na análise de séries temporais [Salles et al., 2019; Shumway and Stoffer, 2025]. De modo geral, uma série é considerada estacionária quando suas propriedades estatísticas permanecem invariantes ao longo do tempo. Distinguem-se, nesse contexto, duas noções complementares: **Estacionariedade fraca** que caracteriza-se pela constância do valor esperado e pela invariância da função de autocovariância, a qual depende exclusivamente da defasagem entre observações, e não do instante temporal específico. Trata-se da forma mais utilizada na prática, sendo suficiente para a aplicação da maioria dos modelos lineares clássicos. A **Estacionariedade forte** define uma condição mais restritiva, segundo a qual toda a distribuição de probabilidade do processo deve permanecer inalterada sob translações temporais. Assim, todas as suas propriedades estatísticas devem ser invariantes no tempo. A suposição de estacionariedade é central para a validade de diversos procedimentos inferenciais e algoritmos de modelagem, razão pela qual a ausência dessa propriedade demanda transformações apropriadas, como diferenciação de primeira ordem ou transformações de potência (por exemplo, Box–Cox) [Shumway and Stoffer, 2025].

A avaliação formal da estacionariedade, por meio de testes estatísticos adequados, constitui etapa indispensável na preparação dos dados, sobretudo em aplicações que envolvem detecção de *motifs* e processos de validação estatística. A Figura II.2 apresenta exemplo de série estacionária, enquanto a Figura II.3 mostra série não estacionária.

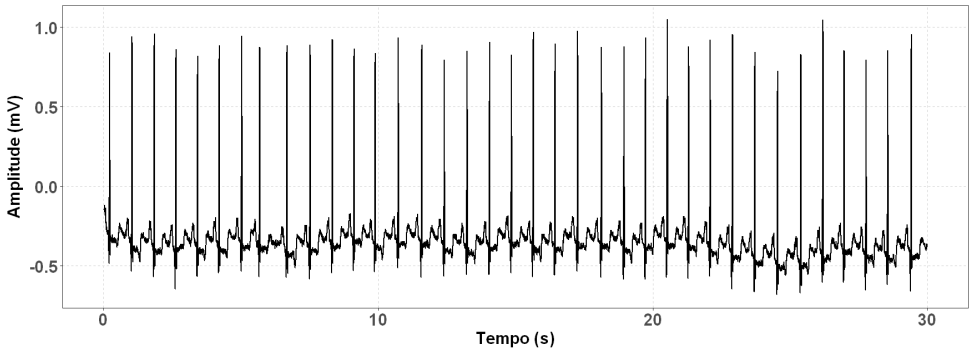


Figura II.2: Série temporal estacionária: ECG evidenciando batimentos cardíacos normais.

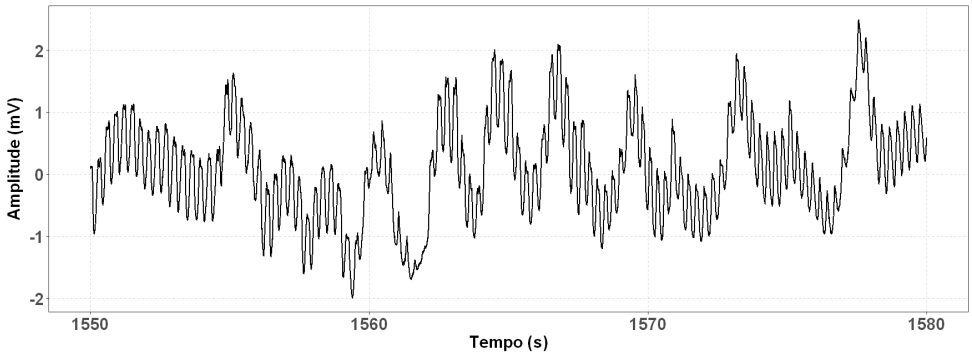


Figura II.3: Série temporal não estacionária: ECG caracterizado por episódios de arritmia.

As considerações sobre similaridade temporal evidenciam a necessidade de métricas que considerem a natureza sequencial dos dados [Esling and Agon, 2012; Fu, 2011; Keogh and Kasetty, 2003; Ogasawara et al., 2025]. Os desafios de alinhamento surgem da comparação entre subsequências possivelmente deslocadas no tempo, exigindo técnicas capazes de identificar padrões similares independentemente da posição temporal. A invariância a transformações constitui outro aspecto crucial, pois padrões podem manifestar-se com diferentes escalas, amplitudes ou orientações, demandando métricas robustas que preservem a similaridade estrutural.

A análise para detecção de eventos enfrenta desafios computacionais e metodológicos que impactam a eficácia dos algoritmos e a qualidade dos resultados. Esses obstáculos decorrem de características dos dados temporais e de limitações de métodos tradicionais, exigindo abordagens adequadas. Um desafio está relacionado à maldição da dimensionalidade, que se manifesta em séries temporais multivariadas; a crescente dimensionalidade compromete a viabilidade dos métodos de detecção, requerendo técnicas de pré-processamento para reduzir a complexidade. Ademais, a não estacionariedade pode impactar a modelagem, levando a inferências incorretas e exigindo técnicas específicas de estabilização [Ogasawara et al., 2025].

A escalabilidade computacional é crítica na descoberta de *motifs*. Mueen et al. [2009] observam que o algoritmo direto de extração de *motifs* tem complexidade quadrática no comprimento da série. Embora existam técnicas de aproximação para contornar essa limitação, a complexidade pode ser gerenciada por otimização de algoritmos, processamento paralelo e redução de dados [Ogasawara et al., 2025].

II.2 Pré-processamento de séries temporais

O pré-processamento constitui etapa fundamental na detecção de eventos, sobretudo na descoberta de *motifs*. Trata a variabilidade de escalas e a complexidade computacional por métodos diversos. No escopo deste trabalho, empregam-se normalização, segmentação por janelas deslizantes e representação simbólica, conceituadas a seguir.

A normalização possibilita comparar séries com diferentes deslocamentos e amplitudes [Lin et al., 2003]. Essa operação ajusta as escalas dos dados, eliminando diferenças de magnitude que podem mascarar padrões relevantes. A normalização *Z-Score* transforma os dados em distribuições com média zero e desvio padrão unitário. Ogasawara et al. [2010] definem que, nesse caso, os valores de uma série temporal X são normalizados de acordo com sua média $\mu(X)$ e desvio padrão $\sigma(X)$, sendo um valor $x_i \in X$ normalizado para $\hat{x}_i$ pela Equação II.4.

$$\hat{x}_i = \frac{x_i - \mu(X)}{\sigma(X)} \tag{II.4}$$

O método Min-Max mapeia os valores para um intervalo predefinido, tipicamente $[0, 1]$ ou $[-1, 1]$. Os valores são ajustados conforme mínimo e máximo da série temporal analisada, convertendo $x_i \in X$ para $\hat{x}_i$ no intervalo definido pela Equação II.5, onde *high* e *low* são os limites superior e inferior do intervalo de destino, e $\min(X)$ e $\max(X)$ representam os valores mínimo e máximo da série temporal [Ogasawara et al., 2009].

$$\hat{x}_i = (\text{high} - \text{low}) \cdot \frac{x_i - \min(X)}{\max(X) - \min(X)} + \text{low} \tag{II.5}$$

A normalização adaptativa surge como alternativa para superar limitações de métodos tradicionais, especialmente em contextos não estacionários e para dar suporte a métodos de aprendizado de máquina. Ogasawara et al. [2010] propõem um método que transforma a série em sequência de dados da qual as propriedades estatísticas são calculadas e consideradas na normalização. O procedimento compreende quatro passos: (i) calcular a média móvel para cada janela deslizante, (ii) obter o resíduo de cada janela em relação à sua média móvel, (iii) calcular a distribuição de resíduos para todas as janelas removendo as que contêm *outliers* e, por fim, (iv) normalizar todas as janelas usando Min-Max extraído da distribuição de resíduos limpa.

A segmentação por janelas deslizantes divide as séries em subsequências de tamanho fixo e permite a análise de propriedades locais [Lin et al., 2003]. No contexto de detecção de *motifs*, possibilita a comparação direta de segmentos de igual comprimento e oferece suporte a métodos de aprendizado de máquina [Ogasawara et al., 2025]. A subsequência é um subconjunto contínuo dos valores da série com comprimento fixo iniciando em uma posição específica. Yeh et al. [2018] formalizam como mostrado na Equação II.6, onde $X_{i,sw}$ representa a subsequência, x_i denota cada elemento da série X, sw é o comprimento da janela, i é a posição inicial e n é o tamanho total da série:

$$X_{i,sw} = x_i, x_{i+1}, \ldots, x_{i+sw-1}, \quad 1 \leq i \leq n - sw + 1 \tag{II.6}$$

O conjunto de todas as subsequências possíveis W é obtido deslizando uma janela de comprimento sw através de X, resultando no conjunto definido na Equação II.7 adaptada de [Yeh et al., 2018].

$$W = \{X_{1,sw}, X_{2,sw}, ..., X_{n-sw+1,sw}\} \tag{II.7}$$

Assume-se passo de deslizamento unitário ($s = 1$), gerando sobreposição máxima entre janelas consecutivas, embora seja possível adotar outros valores de passo $s \in \mathbb{N}$ com $1 \leq s \leq sw$. Janelas incompletas nas extremidades são descartadas, sempre garantindo subsequências com comprimento sw.

As estratégias de sobreposição são inerentes às janelas deslizantes, nas quais cada subsequência compartilha $sw - 1$ elementos com a subsequência adjacente. Isso permite detectar padrões que cruzam limites de janelas [Lin et al., 2003].

A representação simbólica transforma dados numéricos contínuos em sequências discretas de símbolos de alfabeto finito. Essa abordagem viabiliza algoritmos de processamento de *strings* e técnicas de mineração de textos, oferecendo redução dimensional, eficiência computacional e capacidade de indexação [Lin et al., 2007].

Piecewise Aggregate Approximation (PAA) constitui uma etapa intermediária na simbolização, promovendo a redução do número de observações por agregação temporal [Ogasawara et al., 2025]. Lin et al. [2003] definem que uma série X de comprimento n pode ser representada em espaço w-dimensional pelo vetor $\bar{X} = \bar{x}_1, ..., \bar{x}_w$, calculado pela Equação II.8, onde $\bar{x}_i$ representa o i-ésimo elemento agregado, w é a dimensionalidade do espaço reduzido, n é o comprimento da série original e x_j denota cada observação individual.

$$\bar{x}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} x_j \tag{II.8}$$

Essa agregação pode ser vista como aproximação da série original por combinação linear de funções de segmentação retangulares, exemplo de redução de dimensionalidade intuitiva [Lin et al., 2003].

SAX representa a extensão simbólica da PAA, mapeando coeficientes agregados para símbolos alfabéticos por particionamento estatístico. Lin et al. [2003, 2007] mostram que séries normalizadas tendem à distribuição gaussiana, permitindo determinar *breakpoints* que produzem áreas de igual tamanho sob a curva. O SAX particiona a série em intervalos segundo essas quebras, cada um associado a um símbolo particular [Lin et al., 2007]. O tamanho do alfabeto determina o número de partições para o domínio, e todos os valores são substituídos por seus símbolos associados [Ogasawara et al., 2025]. A Equação II.9 transforma uma série com n observações e um alfabeto de tamanho α na série baseada em SAX, onde y_i representa o símbolo resultante, x_i é cada observação da série original, $(a_1, \ldots, a_\alpha)$ constitui o alfabeto simbólico e $\beta_1, \ldots, \beta_{\alpha-1}$ são os *breakpoints* determinados pela distribuição gaussiana [Ogasawara et al., 2025].

$$y_i = \begin{cases} a_1, & \text{se } x_i \in [-\infty, \beta_1] \\ a_2, & \text{se } x_i \in [\beta_1, \beta_2] \\ \vdots & \vdots \\ a_\alpha, & \text{se } x_i \in [\beta_{\alpha-1}, \infty] \end{cases} \tag{II.9}$$

A agregação temporal fornecida por PAA diminui a influência do componente residual e, em seguida, a aplicação do SAX permite a descoberta de *motifs* mais relacionados aos componentes de tendência e sazonalidade mais representativos de um fenômeno [Ogasawara et al., 2025]. Essas ideias introduzem as abordagens específicas de descoberta de *motifs*.

II.3 Detecção de *motifs* em séries temporais

Motifs em séries temporais são subsequências frequentes que apresentam similaridade morfológica [Maletzke et al., 2014], podendo também ser entendidos como pares de subsequências individuais ou segmentos de uma série mais longa que exibem semelhança entre si [Mueen et al., 2009]. A finalidade da descoberta é identificar padrões desconhecidos sem conhecimento prévio sobre quando ocorrem ou que forma assumem [Torkamani and Lohweg, 2017]. Os *motifs* encerram informações sobre o sistema e oferecem percepções sobre o fenômeno estudado [Torkamani and Lohweg, 2017].

Seja q uma sequência de tamanho sw e W o conjunto de janelas deslizantes de tamanho sw obtidas de uma série X. O conjunto de ocorrências de q em X é o conjunto de janelas w_i de W similares a q, conforme a Equação II.10 [Ogasawara et al., 2025].

$$occurs(q, X) = \{w_i\} \mid \forall w_i \in W, w_i \approx q \tag{II.10}$$

A Equação II.11 mostra que uma sequência q é um *motif* em X com suporte ν se e somente se q ocorre em X pelo menos ν vezes [Ogasawara et al., 2025]:

$$motif(q, X) \Leftrightarrow |occurs(q, X)| \geq \nu \tag{II.11}$$

A literatura distingue métodos exatos e aproximados para descoberta de *motifs*. Os métodos exatos computam a medida de distância entre sequências, e os aproximados baseiam-se em indexação [Ogasawara et al., 2025]. Correspondências triviais constituem aspecto crítico: ocorrem quando duas subsequências compartilham sobreposição substancial ou quando uma subsequência corresponde a si mesma; tais casos devem ser excluídos para garantir padrões informativos [Lin et al., 2002].

A abordagem *top-k motifs* identifica os k *motifs* mais significativos segundo critérios de ordenação e desconsiderando correspondências triviais, enquanto métodos baseados em *threshold* detectam todos os *motifs* que satisfazem limiar predefinido [Torkamani and Lohweg, 2017]. Mueen et al. [2009] formalizam o k-ésimo *motif* como o k-ésimo par mais similar na base, estabelecendo hierarquia de relevância.

A avaliação da qualidade dos *motifs* descobertos representa desafio metodológico. Caso um *motif* q ocorra com frequência excessiva, descaracterizando seu desconhecimento prévio, ou não agregue informação, ele pode ser descartado ou classificado como falso [Ogasawara et al., 2025].

A significância estatística pode ser calculada por testes de hipótese que comparam a frequência observada com expectativas [Castro and Azevedo, 2011].

A seguir, apresentam-se técnicas de detecção de *motifs*. Essas técnicas vão de algoritmos de força bruta a estratégias de otimização, com *trade-offs* entre precisão, eficiência e escalabilidade. A força bruta representa a estratégia direta, que realiza a comparação exaustiva entre todas as subsequências da série temporal; embora essa técnica garanta completude na busca, apresenta complexidade $O(n^2)$ [Lin et al., 2002; Mueen et al., 2009; Yeh et al., 2018]. As limitações de escalabilidade desses métodos exatos motivaram o desenvolvimento de algoritmos aproximados [Mueen et al., 2009]. Técnicas de aceleração, como *early abandoning* e *pruning*, que interrompem cálculos de distância assim que um limiar é excedido, permitem reduzir a complexidade prática mantendo garantias de exatidão [Yankov et al., 2007].

O método baseado em indexação surge como alternativa eficiente à busca exaustiva, introduzindo tolerância controlada por meio de técnicas de discretização simbólica como PAA e SAX [Ogasawara et al., 2025]. Essa abordagem, considerada aproximada, viabiliza a descoberta de *motifs* com significativa redução do custo computacional, alcançando complexidade $O(n \cdot \log n)$ por meio de estruturas de indexação que evitam comparações redundantes [Ogasawara et al., 2025].

II.4 Métricas de avaliação na detecção de *motifs*

No contexto de detecção de eventos, a medição do desempenho de um algoritmo traduz-se na avaliação da aderência de um método a uma aplicação específica [Ogasawara et al., 2025]. Essa avaliação ocorre pela comparação do resultado do detector, positivo (P) ou negativo (N), com os eventos marcados e conhecidos previamente [Han et al., 2022]. Na detecção de *motifs*, os resultados positivos indicam a presença de um *motif* e as negativos indicam observações típicas [Ogasawara et al., 2025].

Dessa comparação, surgem os elementos usados no cálculo das métricas: *Verdadeiro Positivo* (TP), evento marcado previamente na série e corretamente identificado; *Verdadeiro Negativo* (TN), observação típica sem identificação de evento; *Falso Positivo* (FP), observação típica com identificação equivocada; e *Falso Negativo* (FN), evento marcado previamente não identificado [Han et al., 2022]. Esses elementos são resumidos em uma matriz de confusão, útil para analisar o desempenho do método. TP e TN indicam detecção correta, enquanto FP e FN indicam erros [Han et al., 2022]. A matriz de confusão contabiliza as classificações segundo a unidade de análise definida, que pode ser por observação individual, por janela deslizante ou por evento detectado, dependendo da abordagem metodológica. A Tabela II.1 [Han et al., 2022] apresenta a matriz de confusão. No melhor cenário, os elementos fora da diagonal principal devem assumir valor zero ou próximos de zero.

Com os elementos da matriz de confusão, calcula-se as métricas para medir o desempenho do

Tabela II.1: Configuração da Matriz de Confusão da detecção de *motifs*.

		Detectado	
		Positivo	Negativo
Real	Positivo	TP	FN
	Negativo	FP	TN

algoritmo. A acurácia representa a porcentagem de acertos, com ou sem *motifs*. A precisão mede a exatidão, indicando a porcentagem de *motifs* identificados que realmente o são. A revocação mede a completude, mostrando a porcentagem de *motifs* corretamente identificados. Precisão perfeita (1,0) significa que todos os *motifs* detectados eram *motifs*, mas não informa sobre *motifs* perdidos. Revocação perfeita (1,0) indica que todos os *motifs* foram identificados, mas não revela quantas observações típicas foram incorretamente rotuladas como *motifs*. Há relação inversa entre precisão e revocação, sendo possível aumentar uma à custa da outra. A medida F_1 combina precisão e revocação em métrica única, oferecendo equilíbrio entre ambas, pois é a média harmônica entre as duas. A Tabela II.2 mostra as métricas utilizadas no escopo deste trabalho [Han et al., 2022].

Tabela II.2: Métricas básicas para avaliação da detecção de *motifs*.

Métrica	**Fórmula**
Acurácia	$\frac{TP+TN}{P+N}$
Precisão	$\frac{TP}{TP+FP}$
Revocação	$\frac{TP}{TP+FN}$
F_1	$\frac{2 \times \text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$

Considere um exemplo de detecção de *motifs* em série com 10 observações, em que o algoritmo produziu 3 detecções nos tempos 2, 5 e 8, sendo os *motifs* verdadeiros localizados nos tempos 5, 8 e 9. Aplicando pareamento exato, obtém-se: $TP = 2$, $TN = 6$, $FP = 1$ e $FN = 1$, produzindo acurácia de 0,8, precisão de 0,67, revocação de 0,67 e F_1 de 0,67.

Embora as métricas tradicionais forneçam avaliação de desempenho, elas apresentam limitações no contexto de séries temporais, em que detecções temporalmente próximas aos eventos reais podem surgir devido a efeitos antecedentes ou subsequentes do fenômeno observado [Salles et al., 2024]. Apesar de detecções vizinhas possuírem valor prático, as métricas tradicionais adotam abordagem binária que desconsidera a proximidade temporal, tratando qualquer detecção não exata como erro

absoluto [Salles et al., 2024]. A demanda por métricas que considerem tempo e tolerância torna-se evidente e motiva a adoção das métricas *Soft Event Detection* (SoftED), que suavizam a avaliação ao incorporar essa tolerância.

As métricas SoftED redefinem os elementos da matriz de confusão. Cada detecção d_i recebe um *score soft* $d_s(d_i)$ baseado em sua proximidade temporal aos eventos reais, calculado por uma função de pertinência triangular definida na Equação II.12, onde $\mu_{e_j}(t_{d_i})$ representa o grau de pertinência da detecção ao evento, t_{d_i} é o instante da detecção, t_{e_j} é o instante do evento e k é o parâmetro de tolerância que define a janela simétrica de $\pm k$ observações ao redor do evento [Salles et al., 2024].

$$\mu_{e_j}(t_{d_i}) = \max\left(\min\left(\frac{t_{d_i} - (t_{e_j} - k)}{k}, \frac{(t_{e_j} + k) - t_{d_i}}{k}\right), 0\right) \tag{II.12}$$

O *score* final $d_s(d_i)$ é determinado atribuindo-se cada detecção ao evento temporalmente mais próximo e aplicando-se restrições para manter integridade com as métricas tradicionais, em que apenas a melhor detecção para cada evento recebe *score* não zero, conforme a Equação II.13 [Salles et al., 2024].

$$d_s(d_i) = \begin{cases} \max_{e_j} \mu_{e_j}(t_{d_i}), & \text{se } d_i \text{ é a detecção representativa do evento mais próximo} \\ 0, & \text{caso contrário} \end{cases} \tag{II.13}$$

Os TP_s, FN_s, FP_s e TN_s, em versões *soft*, são apresentados nas Equações II.14, II.15, II.16 e II.17 [Salles et al., 2024]. Essas métricas *soft* permitem que detecções próximas aos eventos contribuam proporcionalmente, oferecendo avaliação mais realista da capacidade de detecção de *motifs*.

$$TP_s = \sum_{i=1}^{\delta} d_s(d_i) \tag{II.14}$$

$$FN_s = m - TP_s \tag{II.15}$$

$$FP_s = \sum_{i=1}^{\delta} (1 - d_s(d_i)) \tag{II.16}$$

$$TN_s = (|X| - m) - FP_s \tag{II.17}$$

onde δ é o número de detecções, m é o número de eventos reais e $|X|$ é o comprimento da série.

Outro aspecto complementar à qualidade dos *motifs* detectados diz respeito à entropia infor-

macional, que oferece perspectiva quantitativa baseada na teoria da informação para distinguir estruturas significativas de ruído. A entropia de *Shannon*, aplicada a sequências simbólicas, quantifica esse valor por meio da distribuição de probabilidade dos símbolos e apresenta-se mais elevada em sequências com maior diversidade [Cover and Thomas, 2012]. Para uma variável aleatória discreta com função de massa de probabilidade e alfabeto específico, a entropia é definida na Equação II.18, onde $H(\xi)$ representa a entropia da variável aleatória ξ, $p(x)$ é a função de massa de probabilidade e $\mathcal{A}$ denota o alfabeto ou conjunto de valores possíveis [Cover and Thomas, 2012].

$$H(\xi) = -\sum_{x \in \mathcal{A}} p(x) \log_2 p(x) \tag{II.18}$$

A relação entre entropia e qualidade dos *motifs* manifesta-se pelo conceito de que *motifs* representam regiões de estrutura informacional específica, em que a entropia serve como indicador de complexidade e relevância [Borges et al., 2020]. O uso da entropia para ranqueamento e filtragem de candidatos constitui estratégia de priorização baseada em critérios informacionais [Borges et al., 2020]. Essa metodologia identifica padrões com maior potencial analítico, priorizando *motifs* mais ricos e relevantes em detrimento de estruturas triviais, contribuindo para a eficiência e a qualidade do processo de descoberta.

No uso de *autoencoders* para detecção de eventos, é fundamental avaliar a qualidade da representação latente (espaço de menor dimensionalidade para onde são mapeados os dados de entrada). Métricas quantificam a capacidade de reconstrução pela comparação entre a série original e sua versão reconstruída, mensurando o quanto das características originais é preservado. A função de perda típica em *autoencoders*, baseada no erro de reconstrução $L(x, g(f(x)))$, relaciona-se diretamente com essas métricas, em que g e f representam *decoder* e *encoder*, respectivamente [Goodfellow et al., 2016]. A minimização do erro quadrático médio durante o treinamento busca garantir que a representação latente preserve informações suficientes para reconstrução fidedigna [Goodfellow et al., 2016]. Para cumprir essa validação no escopo desta pesquisa, foram selecionadas as métricas descritas a seguir.

O *Coeficiente de Determinação* (R^2), coeficiente de determinação, avalia a qualidade de predição e reconstrução, representando a proporção da variância na variável dependente explicável pelo modelo [Chicco et al., 2021; James et al., 2021]. Valores próximos de 1.0 indicam alta capacidade reconstrutiva, sendo matematicamente expresso na Equação II.19, onde R^2 representa o coeficiente de determinação, y_i são os valores observados, $\hat{y}_i$ são os valores preditos pelo modelo, $\bar{y}$ é a média dos valores observados e η é o número de observações [James et al., 2021]:

$$R^2 = 1 - \frac{\sum_{i=1}^{\eta}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{\eta}(y_i - \bar{y})^2} \tag{II.19}$$

A *Raiz do Erro Quadrático Médio* (RMSE), quantifica o erro entre valores preditos e observados, fornecendo uma medida na mesma unidade dos dados [Chicco et al., 2021; James et al., 2021]. O RMSE deriva do *Erro Quadrático Médio* (MSE) pela raiz quadrada e penaliza erros de grande magnitude, sendo definido na Equação II.20.

$$RMSE = \sqrt{\frac{1}{\eta}\sum_{i=1}^{\eta}(y_i - \hat{y}_i)^2} \tag{II.20}$$

O *Erro Percentual Absoluto Médio* (MAPE), expressa o erro como percentual dos valores observados, oferecendo interpretabilidade em termos de erro relativo [Chicco et al., 2021]. Embora útil para comparação entre séries de diferentes escalas, apresenta limitações quando aplicado a dados com valores próximos de zero, sendo definido na Equação II.21.

$$MAPE = \frac{1}{\eta}\sum_{i=1}^{\eta}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{II.21}$$

A avaliação da estabilidade das métricas de reconstrução sob diferentes configurações experimentais requer quantificação de dispersão relativa. O *Coeficiente de Variação* (CV), definido na Equação II.22, expressa a razão entre desvio padrão (σ) e média (μ).

$$\text{CV} = \frac{\sigma}{\mu} \tag{II.22}$$

Como alternativa robusta a *outliers*, o *Coeficiente de Variação Robusto* (RCV), definido na Equação II.23, substitui essas estatísticas paramétricas por suas contrapartes não paramétricas: o *Desvio Absoluto da Mediana* (MAD) (Equação II.24) em lugar do desvio padrão e a mediana ($\tilde{x}$) em lugar da média.

$$\text{RCV} = \frac{\text{MAD}}{\tilde{x}} \tag{II.23}$$

$$\text{MAD} = 1{,}4826 \times \text{mediana}(|x_i - \tilde{x}|) \tag{II.24}$$

onde o fator 1,4826 garante comparabilidade com o desvio padrão em distribuições normais.

Valores baixos de CV e RCV indicam estabilidade relativa, confirmando robustez da configuração testada independentemente de particularidades das séries analisadas. A comparação entre CV e RCV permite identificar se a dispersão observada decorre de variabilidade generalizada ou de valores atípicos isolados [Everitt and Skrondal, 2010].

II.5 Redes neurais e *autoencoders*

As redes neurais artificiais consistem em modelos paramétricos não lineares que transformam um conjunto de variáveis de entrada em um conjunto de variáveis de saída por funções controladas por parâmetros ajustáveis [Bishop, 2006]. Essas arquiteturas simulam o sistema nervoso biológico, em que o processamento ocorre por unidades interconectadas denominadas neurônios, que recebem entradas por conexões ponderadas e transmitem saídas após computações locais [Aggarwal, 2017]. A combinação dessas operações permite aproximar funções contínuas.

Fundamentadas nesses princípios, as redes neurais artificiais constituem a base dos *autoencoders*, viabilizando o aprendizado de representações complexas a partir de dados. Esses fundamentos são essenciais para compreender como *autoencoders* processam e codificam informações temporais, estabelecendo bases para sua aplicação em detecção de *motifs* e análise de séries temporais.

O *perceptron* representa a forma básica de rede neural, sendo um modelo de regressão linear com duas camadas de nós na entrada e um único nó de saída, em que o número de nós de entrada é igual à dimensão dos dados [Aggarwal, 2017]. Expandindo essa arquitetura, o *perceptron* multicamadas caracteriza-se pela presença de camadas ocultas entre as camadas de entrada e saída, em que cada unidade oculta computa combinação linear ponderada de suas entradas, transformada por função de ativação não linear [Bishop, 2006]. Essa sucessão de transformações constitui a arquitetura fundamental das redes neurais modernas [Goodfellow et al., 2016].

Matematicamente, cada neurônio em uma camada calcula sua ativação de acordo com as saídas da camada anterior [Bishop, 2006]. A Equação II.25 estende esses conceitos para redes com múltiplas camadas ocultas, onde $a_i^{(l)}$ representa a ativação do neurônio i na camada l, $W_{ij}^{(l-1)}$ são os pesos conectando o neurônio j da camada $l-1$ ao neurônio i da camada l, $b_i^{(l)}$ é o *bias* do neurônio i na camada l, $f(\cdot)$ é a função de ativação e θ é o número de neurônios na camada anterior.

$$a_i^{(l)} = f\left(\sum_{j=1}^{\theta} W_{ij}^{(l-1)} a_j^{(l-1)} + b_i^{(l)}\right) \tag{II.25}$$

As funções de ativação introduzem não linearidades necessárias ao aprendizado de padrões complexos [Bishop, 2006; Aggarwal, 2017], permitindo que redes multicamadas aproximem funções complexas [Goodfellow et al., 2016; Sakurada and Yairi, 2014]. Entre as mais utilizadas, destacam-se a sigmoide logística e a tangente hiperbólica, expressas nas Equações II.26 e II.27, onde z representa a entrada pré-ativação da unidade neural [Bishop, 2006].

$$S(z) = \frac{1}{1+\exp(-z)} \tag{II.26}$$

$$\tanh(z) = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)} \tag{II.27}$$

O algoritmo *backpropagation* representa o mecanismo de treinamento de redes neurais multicamadas, baseado na aplicação da regra da cadeia para calcular gradientes da função de erro em relação aos parâmetros da rede [Bishop, 2006; Goodfellow et al., 2016; Hinton and Salakhutdinov, 2006]. O processo envolve duas fases por instância de treinamento: propagação direta das ativações da entrada para a saída e propagação reversa dos erros da saída para a entrada [Aggarwal, 2017].

Com esses fundamentos, os *autoencoders* constituem uma classe específica de redes projetadas para aprender representações eficientes por codificação e decodificação, em que a rede é treinada para reconstruir sua entrada na saída [Goodfellow et al., 2016]. A arquitetura *encoder-decoder* representa o núcleo conceitual, com duas componentes interconectadas por representação latente compacta [Goodfellow et al., 2016; Hinton and Salakhutdinov, 2006]. O *encoder*, $h = f(x)$, mapeia os dados de entrada para espaço de menor dimensionalidade, enquanto o *decoder* produz reconstrução $r = g(h)$ a partir dessa representação [Goodfellow et al., 2016; Bascol et al., 2016; Roy et al., 2023]. Nesse processo, a rede preserva características salientes dos dados, resultando em representação latente que captura a estrutura essencial da informação original.

As funções de perda em *autoencoders* baseiam-se na minimização do erro de reconstrução entre a entrada original e a saída decodificada [Goodfellow et al., 2016; Roy et al., 2023; Noering et al., 2021]. A função de custo típica é $L(x, g(f(x)))$, em que L representa uma medida de dissimilaridade, comumente o erro quadrático médio para dados contínuos [An and Cho, 2015]. O processo de otimização ajusta iterativamente os parâmetros das redes *encoder* e *decoder*, minimizando a discrepância entre entrada e reconstrução [Hinton and Salakhutdinov, 2006].

A redução de dimensionalidade com *autoencoders* oferece vantagens sobre métodos lineares como *Principal Component Analysis* (PCA), permitindo representações não lineares mais expressivas [Hinton and Salakhutdinov, 2006; Goodfellow et al., 2016]. Quando regularizados, *autoencoders* aprendem representações hierárquicas e abstratas que capturam variações complexas nos dados [Bascol et al., 2016; Noering et al., 2021]. Em séries temporais, essa capacidade é valiosa, pois padrões complexos podem ser comprimidos em representações latentes que preservam informações críticas [Roy et al., 2023]. O *trade-off* entre compressão e fidelidade manifesta-se na necessidade de reduzir a dimensionalidade preservando a qualidade da reconstrução [Hinton and Salakhutdinov, 2006; Goodfellow et al., 2016]. A avaliação dessa qualidade é crucial para a utilidade prática dos *autoencoders* [Sakurada and Yairi, 2014; Goodfellow et al., 2016].

O erro de reconstrução, medido por métricas como o erro quadrático médio, constitui simultaneamente a função objetivo de treinamento e a métrica de avaliação da capacidade do modelo [Hinton and Salakhutdinov, 2006]. Em detecção de anomalias, valores elevados de erro indicam padrões que diferem dos dados de treinamento [Roy et al., 2023; An and Cho, 2015]. Além disso, a interpretabilidade da representação latente e sua aplicabilidade para tarefas subsequentes constituem aspectos

relevantes [Bascol et al., 2016; Noering et al., 2021]. Representações bem aprendidas podem servir como características para classificação, *clustering* ou detecção de *motifs*, oferecendo vantagens computacionais por redução dimensional [Hinton and Salakhutdinov, 2006; Sakurada and Yairi, 2014]. Em séries temporais, a capacidade de extrair padrões interpretáveis é especialmente valiosa, pois a representação comprimida pode revelar estruturas latentes e facilitar a descoberta automática de *motifs*, consolidando os *autoencoders* como ferramentas para análise de dados temporais complexos [Bascol et al., 2016; Noering et al., 2021].

Capítulo III Trabalhos Relacionados

Com o objetivo de fornecer base sólida para este trabalho, foi realizado um mapeamento sistemático da literatura sobre descoberta de *motifs* em séries temporais, com ênfase no uso de *autoencoders*. Buscou-se estabelecer um panorama atualizado da temática, a fim de identificar oportunidades de investigação e posicionar a proposta no cenário científico contemporâneo.

A estratégia de busca foi executada pela última vez em agosto de 2025 e foi empregada na base Scopus como fonte primária, com o filtro inicial *TITLE-ABS-KEY ("time series motif" OR ("autoencoder" AND "time series" AND "motif") OR ("autoencoder" AND "ECG") OR ("SAX" AND "motif"))*, resultando em 689 publicações em inglês. Após avaliação exploratória de aderência ao tema, eliminou-se cerca de 74% dos trabalhos. A análise dos resumos dos 180 estudos restantes levou à seleção de 45 trabalhos, posteriormente expandida para 53 publicações através da utilização da técnica de *snowballing*.

A organização do capítulo contempla três seções. A Seção III.1 examina 43 trabalhos sobre detecção de *motifs* em séries temporais, abrangendo diferentes paradigmas e abordagens. A Seção III.2 analisa 13 estudos que incluem revisões sistemáticas, avaliações comparativas e *frameworks* de avaliação. A Seção III.3 sintetiza as principais tendências, limitações e direções promissoras que orientam a metodologia proposta nesta dissertação.

III.1 Métodos de Detecção de *Motifs* em Séries Temporais

A detecção de *motifs* em séries temporais evoluiu significativamente desde suas formulações iniciais, constituindo um campo ativo e em constante desenvolvimento. Revisões abrangentes como [Torkamani and Lohweg, 2017], [Keogh, 2025] e [Guerrmi et al., 2025] documentam essa progressão, evidenciando a transição dos algoritmos fundacionais de força bruta para arquiteturas contemporâneas baseadas em *deep learning*. Para uma exposição estruturada, os métodos são apresentados por paradigma e em ordem cronológica, dos fundamentos históricos às técnicas baseadas em *deep learning*, destacando avanços conceituais e ganhos de eficiência computacional.

Brute-Force Discovery

O conceito de *motifs* em séries temporais é introduzido por Patel et al. [2002] e Lin et al. [2002], em analogia aos *motifs* discretos da biologia computacional, para definir padrões recorrentes e previamente desconhecidos. *Motifs* caracterizam-se como subsequências altamente similares dentro de uma série temporal, distinguindo-se de buscas por padrões conhecidos. Esses trabalhos inaugurais motivam a descoberta automática de estruturas repetitivas em grandes bases, com aplicações em sumarização, visualização e como subrotina para regras de associação, agrupamento e classificação [Patel et al., 2002; Lin et al., 2002]. O algoritmo fundacional apresenta custo computacional $O(n^2)$ devido à natureza exaustiva do procedimento.

Index-Based Discovery

Algoritmos desse paradigma buscam eficiência ao organizar subsequências em estruturas indexadas, reduzindo comparações. Um avanço central é o *Symbolic Aggregate approXimation* (SAX) [Lin et al., 2003, 2007], que combina *Piecewise Aggregate Approximation* (PAA) com discretização por *breakpoints* gaussianos, produzindo representações simbólicas com propriedade de *lower bound* para medidas de distância. Considerado *index-based*, o SAX introduz tolerância via indexação, evitando o cálculo exato de distâncias e alcançando assim complexidade $O(n \cdot \log n)$. Sua eficiência e simplicidade sustentam ampla adoção e inspiram algoritmos subsequentes.

Limitações decorrem da suposição de gaussianidade. Butler and Kazakov [2015] mostram que a PAA altera a distribuição original, tornando símbolos não equiprováveis. A distorção cresce com o tamanho do segmento e com a autocorrelação, afetando séries com baixa correlação temporal. O *adaptive* SAX (aSAX) [Pham et al., 2010] responde com *breakpoints* adaptativos via *k-means*, mas ainda depende de suposições e conhecimento *a priori*, o que restringe a aplicabilidade em contextos complexos, como dados biomédicos.

Para escalabilidade em *datasets* massivos, Mueen et al. [2011] propõem o *Disk Aware Motif Enumeration* (DAME), primeiro algoritmo otimizado para processamento em disco, com acesso sequencial que minimiza operações de entrada e saída, viabilizando bases com centenas de GB e aplicações como EEG. Ademais, Yin et al. [2014] tratam de *motifs* de comprimento variável com representação simbólica aplicada à ECG, usando os *motifs* como características discriminativas em tarefas de classificação.

Random-Projection Discovery

Em paralelo aos métodos simbólicos, Chiu et al. [2003] adaptam projeções aleatórias [Buhler and Tompa, 2001], propondo uma abordagem *anytime* que produz candidatos imediatamente e refina

resultados progressivamente. A técnica incorpora seções *don't care*, que são posições dentro da sequência ou padrão procurado, onde qualquer valor é permitido, funcionando como curingas que aumentam a robustez ao ruído, superando a rigidez de representações discretas.

Early-Abandon Discovery

Mueen et al. [2009] desenvolveram o primeiro algoritmo exato e viável, o *Mueen-Keogh algorithm* (MK), com *pruning* e *early abandoning*, técnicas de aceleração que interrompem cálculos de distância assim que um limiar é excedido. O método obtém reduções de até três ordens de magnitude no tempo em relação à força bruta, sem comprometer a exatidão, estabelecendo novo patamar de escalabilidade. Ainda assim, métodos exatos permanecem fundamentalmente limitados pelo limite inferior assintótico $\Omega(n^2)$ inerente à comparação entre pares de subsequências, o que restringe sua aplicabilidade em cenários de larga escala ou análise exploratória.

Matrix Profile

O *Matrix Profile* (MP) [Yeh et al., 2016] reformula a descoberta de *motifs* como busca de similaridade entre todos os pares de subsequências, armazenando a distância de cada subsequência ao vizinho mais próximo em estrutura única. O *Scalable Time Series Anytime Matrix Profile* (STAMP) oferece capacidades *anytime* e paralelização, enquanto o *Scalable Time Series Ordered-Search Matrix Profile* (STOMP), com *Graphics Processing Unit* (GPU), viabiliza o processamento de *datasets* com até 100 milhões de subsequências [Zhu et al., 2016]. Extensões abrangem séries multidimensionais [Yeh et al., 2017]. O SCRIMP++ acelera a convergência mantendo propriedades *anytime* [Zhu et al., 2018]. O *Variable Length Motif Discovery* (VALMOD) resolve a descoberta exata de comprimento variável com *lower bounding* que reutiliza computações [Linardi et al., 2018].

Embora a Distância Euclidiana apresente simplicidade conceitual, ela é altamente sensível a variações de amplitude e desalinhamentos temporais. O DTW amplia a flexibilidade da correspondência ao permitir alinhamentos elásticos, porém sua integração aos algoritmos de descoberta de *motifs* acarreta custos computacionais ainda mais elevados, limitando sua aplicação prática. A integração com DTW via *Scalable Warping Aware Matrix Profile* (SWAMP) torna viável a descoberta exata sob DTW, podando até 99,99% das computações [Alaee et al., 2020, 2021]. Métodos adicionais para comprimento variável incluem abordagem voltada à indústria [Van Onsem et al., 2023] e o MP MAD [Linardi et al., 2020].

No entanto, mesmo com otimizações baseadas em *Fast Fourier Transform* (FFT) e paralelização em GPU, os métodos baseados em MP permanecem sujeitos à complexidade quadrática, o que restringe sua aplicabilidade em cenários exploratórios ou de larga escala, especialmente em dados biomédicos com séries longas. A Tabela III.1 sintetiza os principais algoritmos clássicos de

descoberta de *motifs* em séries temporais, organizando-os por trabalho de referência e paradigma metodológico. Essa visão consolidada permite comparar a evolução histórica das abordagens.

Tabela III.1: Principais algoritmos clássicos de descoberta de *motifs* em séries temporais.

Algoritmo	**Trabalho(s)**	**Método**
Find-1-Motif	Patel et al. [2002] e Lin et al. [2002]	*Brute-Force*
SAX	Lin et al. [2003]	*Index-Based*
Random Projection Motif Discovery	Chiu et al. [2003]	*Random-Projection*
MK	Mueen et al. [2009]	*Early-Abandon*
aSAX	Pham et al. [2010]	*Index-Based*
DAME	Mueen et al. [2011]	*Index-Based*
Variable Length Motif Discovery	Yin et al. [2014]	*Index-Based*
STAMP	Yeh et al. [2016]	*Matrix Profile*
STOMP	Zhu et al. [2016]	*Matrix Profile*
SCRIMP++	Zhu et al. [2018]	*Matrix Profile*
VALMOD	Linardi et al. [2018]	*Matrix Profile*
MP MAD	Linardi et al. [2020]	*Matrix Profile*
SWAMP	Alaee et al. [2021]	*Matrix Profile*
MP Variable Length	Van Onsem et al. [2023]	*Matrix Profile*

Embora consolidadas, as abordagens clássicas inspiram extensões, introdução de técnicas de otimização e métodos especializados que exploram requisitos de domínio ou novas formulações.

Abordagens baseadas em *Minimum Description Length* (MDL) e redução de dimensionalidade incluem o *Extended Motif Discovery* (EMD), que utiliza PCA para dados multidimensionais e admite comprimentos variáveis [Tanaka et al., 2005]. A enumeração eficiente para múltiplos comprimentos aparece no *MOtif ENumerator* (MOEN), que reutiliza *lower bounds* ao longo de comprimentos sucessivos e reduz a execução de algoritmos separados [Mueen, 2013]. Os *k-Motiflets* substituem o limiar de distância r pelo número desejado de ocorrências k, tornando a parametrização mais intuitiva [Schäfer and Leser, 2022]. O LoCoMotif supera limitações de cardinalidade, comprimento fixo, variabilidade temporal e univariância, lidando com séries multivariadas e variações temporais não lineares [Van Wesenbeeck et al., 2024].

Em otimização, o SWARMMOTIF estrutura o problema como busca multimodal por *particle swarm*, com qualidade comparável a métodos exatos e menor tempo [Serrà and Arcos, 2016]. Algoritmos genéticos multiobjetivo equilibram compreensibilidade, confiança e interesse, com validação no MIT-BIH *Arrhythmia Database* e alta acurácia diagnóstica [Ramanujam and Padmavathi, 2016]. Em *constrained discovery*, MCFull e MCInc incorporam conhecimento de domínio por restrições de localização, acelerando a busca sem perda de acurácia [Mohammad and Nishida, 2009]. Para uso clínico, o *Motif Discovery for Large-scale Time Series* (MDLats) foi implantado para classificação de ECG, com eficiência e altos índices de precisão e revocação [Liu et al., 2015].

Arquiteturas profundas emergem como alternativa não linear para a descoberta de padrões em séries temporais. Bascol et al. [2016] propõem *autoencoders* convolucionais interpretáveis nos quais filtros convolucionais 1D funcionam como *templates* deslizantes que geram mapas de ativação temporal. A detecção de *motifs* ocorre pela identificação de máximos locais nesses mapas: selecionam-se picos de ativação acima de um limiar, extraem-se subsequências correspondentes aos campos receptivos dos filtros e aplicam-se técnicas de supressão de máximos não locais. As subsequências resultantes são agrupadas ou promediadas, formando *templates* representativos interpretados como *motifs*. A abordagem incorpora *Adaptive Rectified Linear Unit* (AdaReLU) e regularização *group-lasso* para promover interpretabilidade.

O FastM [Rong et al., 2020] integra *autoencoders* a funções de perda orientadas ao agrupamento, visando preservar similaridade e reduzir dimensionalidade. O método segmenta a série em janelas deslizantes, treina o *autoencoder* minimizando o erro de reconstrução, coleta os vetores latentes e aplica algoritmo de agrupamento (por exemplo, *k-means*), em que cada *cluster* corresponde a um candidato a *motif*. Para cada *cluster*, obtém-se o *motif* representativo calculando a média ou um *medoid* das janelas associadas. O método demonstra eficiência superior em três *datasets* reais. Noering et al. [2021] comparam *autoencoders* a métodos clássicos (incluindo DTW, discretização e MP), indicando qualidade similar na descoberta de padrões.

Entretanto, essas abordagens baseadas em *autoencoders* concentram-se em bases industriais ou genéricas e realizam a descoberta de *motifs* por meio de agrupamento ou cálculos de distância no espaço latente. Tais estratégias introduzem custos computacionais adicionais: o agrupamento via *k-means* apresenta complexidade típica de $O(n \cdot k \cdot d \cdot i)$, onde n é o número de janelas, k é o número de *clusters*, d a dimensão latente e i o número de iterações, resultando em complexidade pelo menos superlinear e frequentemente próxima a quadrática em implementações práticas. Essa limitação compromete a escalabilidade que caracteriza métodos simbólicos tradicionais, particularmente em aplicações biomédicas com séries longas.

Aplicações de *autoencoders* em ECG também incluem classificação supervisionada de arritmias com *stacked denoising autoencoders* e *active learning* [Rahhal et al., 2016], *sparse autoencoders* com regularização L1 e divergência KL [Liang et al., 2021], arquiteturas híbridas com *transformers* e *stacked autoencoders* [Ding et al., 2023], e *Long Short-Term Memory* (LSTM) *autoencoders* para classificação multiclasse [Guerra et al., 2025]. Mecanismos de *self-attention* podem melhorar a identificação de regiões relevantes [Shah et al., 2025]. Para detecção de anomalias, destacam-se LSTM *autoencoders* [Roy et al., 2023], combinações com redes adversariais para bases desbalanceadas [Nawaz et al., 2024], e abordagens *patient-specific* com LSTM bidirecional [Zhong and Sun, 2023]. Em processamento técnico, há compressão e *denoising* com LSTM [Dasan and Panneerselvam, 2021], e detecção de picos R com arquiteturas 1D [Dogan et al., 2023].

A Tabela III.2 consolida os principais algoritmos especializados de descoberta de *motifs* discutidos nesta seção, incluindo extensões baseadas em MDL, otimização, restrições de domínio e abordagens baseadas em *deep learning*. A organização evidencia a diversidade metodológica do campo e reforça a lacuna identificada na integração entre *autoencoders* e descoberta de *motifs* em séries de ECG.

Tabela III.2: Principais algoritmos especializados de descoberta de *motifs* em séries temporais.

Algoritmo	**Trabalho(s)**	**Método**
EMD	[Tanaka et al., 2005]	*MDL+PCA*
MOEN	[Mueen, 2013]	*Lower Bounds Cascading*
k-Motiflets	[Schäfer and Leser, 2022]	*Occurrence-based Discovery*
LoCoMotif	[Van Wesenbeeck et al., 2024]	*Multivariate Time Warping*
SWARMMOTIF	[Serrà and Arcos, 2016]	*Particle Swarm Optimization*
Multi-objective GA	[Ramanujam and Padmavathi, 2016]	*Multi-objective Optimization*
MCFull/MCInc	[Mohammad and Nishida, 2009]	*Constraint-based Discovery*
MDLats	[Liu et al., 2015]	*Adaptive Discovery*
Autoencoder Convolucional	[Bascol et al., 2016]	*Deep Learning*
FastM	[Rong et al., 2020]	*Deep Learning*

III.2 Revisões Sistemáticas e Estudos Comparativos

A diversidade metodológica descrita demanda estudos comparativos que orientem a seleção de técnicas para aplicações específicas. Apesar de avanços na descoberta de *motifs*, avaliações comparativas rigorosas ainda são limitadas, sobretudo em aplicações biomédicas. Esta seção organiza os trabalhos em *surveys* e revisões, estudos comparativos experimentais, *frameworks* e métricas, e limitações dos trabalhos existentes.

Torkamani and Lohweg [2017] apresentam um survey abrangente com glossário e taxonomia por medidas de similaridade e propriedades de representação, destacando aplicações em múltiplos domínios, inclusive medicina. Em *machine learning*, Berahmand et al. [2024] revisam *autoencoders* e aplicações, com foco em detecção de anomalias. Guerrmi et al. [2025] combinam revisão ampla com avaliação experimental de 11 algoritmos em 8 conjuntos reais e *datasets* sintéticos, propondo diretrizes de escolha. Keogh [2025] discutem tornar problemas difíceis tratáveis via algoritmos *anytime* e restrições adicionais nos padrões retornados, com exemplos biomédicos.

No que tange a comparações experimentais, Fodil et al. [2024] avaliam SAX, *1d-SAX* e *fABBA*, propõe *UniformSAX*. Eles mostram que representações simbólicas funcionam melhor para comprimentos fixos e perdem eficácia para comprimentos variáveis. Noering et al. [2021] comparam *autoencoders* a métodos clássicos (DTW, discretização e MP) e observam qualidade similar. Neloy and Turgeon [2024] avaliam 11 arquiteturas de *autoencoders* em detecção de anomalias, discutindo *trade-offs*. Goldstein and Uchida [2016] comparam 19 algoritmos em 10 bases multivariadas. Schmidl et al. [2022] implementam 71 algoritmos em 976 conjuntos, avaliando eficácia, eficiência e robustez em mais de 69.000 experimentos. Embora extensos, muitos estudos focam em anomalias, não *motifs*.

Quanto a *benchmarks*, Wu and Keogh [2023] criticam conjuntos amplamente usados, apontando trivialidades, densidade irreal de anomalias, rotulagem incorreta e viés, e propõem o *UCR Time Series Anomaly Archive*. Essa crítica reforça a necessidade de avaliações rigorosas e *datasets* confiáveis.

Finalmente, no que concerne a ferramentas e métricas, o Harbinger [Salles et al., 2020] integra detectores, avaliação e visualização para eventos diversos, incluindo *motifs*. As métricas *Soft Event Detection* (SoftED) [Salles et al., 2024] incorporam tolerância temporal, suavizando precisão, revocação e F_1 por funções de pertinência triangulares, e ampliam a capacidade avaliativa em relação às métricas tradicionais. Críticas metodológicas clássicas, como apresentado em [Keogh and Kasetty, 2003], permanecem atuais e corroboram limitações sistemáticas apontadas recentemente.

III.3 Síntese de Trabalhos Relacionados

A revisão sistemática abrange métodos fundacionais, técnicas contemporâneas de *deep learning*, estudos comparativos e *frameworks*. Observam-se três eixos: evolução metodológica, lacunas críticas e oportunidades para contribuições.

A evolução progride de Patel et al. [2002]; Lin et al. [2002] ao LoCoMotif [Van Wesenbeeck et al., 2024], com marcos como SAX [Lin et al., 2003], MK [Mueen et al., 2009], MP [Yeh et al., 2016] e integração com DTW [Alaee et al., 2020, 2021]. A escalabilidade avança de milhares de pontos para 100 milhões de subsequências com GPU [Zhu et al., 2016]. O SAX consolida-se como representação simbólica dominante [Lin et al., 2003, 2007; Pham et al., 2010] e o MP como paradigma unificador [Yeh et al., 2016; Zhu et al., 2016; Linardi et al., 2018, 2020; Alaee et al., 2021; Van Onsem et al., 2023]. As aplicações biomédicas validam viabilidade técnica e relevância clínica [Mueen et al., 2011; Yin et al., 2014; Liu et al., 2015; Ramanujam and Padmavathi, 2016].

A literatura revisada evidencia um compromisso entre métodos exatos de descoberta de *motifs*, que oferecem precisão à custa de complexidade quadrática ($\Omega(n^2)$), e métodos simbólicos aproximados, que escalam eficientemente ($O(n \cdot \log n)$) mas impõem pressupostos estatísticos rígidos.

Embora existam aplicações de *autoencoders* em ECG para classificação e detecção de anomalias [Rahhal et al., 2016; Liang et al., 2021; Guerra et al., 2025; Roy et al., 2023; Zhong and Sun, 2023; Dasan and Panneerselvam, 2021; Dogan et al., 2023] e métodos de descoberta de *motifs* em contextos biomédicos [Liu et al., 2015; Ramanujam and Padmavathi, 2016; Yin et al., 2014], a interseção dessas duas linhas permanece inexplorada. *Surveys* recentes sobre *autoencoders* [Berahmand et al., 2024] não registram aplicações específicas para descoberta de *motifs* em séries de ECG.

As abordagens existentes que integram *autoencoders* à descoberta de *motifs* [Bascol et al., 2016; Rong et al., 2020; Noering et al., 2021] realizam a busca de padrões por agrupamento ou cálculos de distância no espaço latente, introduzindo custos computacionais superlineares ($O(n \cdot k \cdot d \cdot i)$ para *k-means*) que podem aproximar-se de $O(n^2)$ em implementações práticas. Essas abordagens concentram-se predominantemente em bases industriais ou genéricas, sem validações em séries biomédicas. Combinadas às limitações de pressupostos gaussianos documentadas para o SAX [Butler and Kazakov, 2015; Fodil et al., 2024], essas insuficiências evidenciam lacuna metodológica: a ausência de abordagens que preservem a complexidade linear da busca simbólica aplicando-a diretamente sobre representações latentes aprendidas, evitando tanto os pressupostos estatísticos rígidos quanto os custos quadráticos do agrupamento no espaço latente. Essa lacuna é especialmente crítica em aplicações biomédicas com séries longas, nas quais a eficiência computacional constitui requisito operacional fundamental.

O próximo capítulo detalha a metodologia proposta, fundamentada nos achados e críticas aqui sintetizados.

Capítulo IV Metodologia

A descoberta de padrões recorrentes em séries temporais biomédicas representa um desafio metodológico que combina eficiência computacional, capacidade de generalização e interpretabilidade clínica. Este capítulo apresenta uma abordagem para detecção de *motifs* em séries de eletrocardiograma (ECG), fundamentada na substituição de técnicas convencionais de discretização simbólica por arquiteturas de aprendizado profundo. A proposta central consiste em investigar se representações latentes aprendidas automaticamente por *autoencoders* superam as limitações de métodos baseados em discretização simbólica tradicional, preservando características discriminativas essenciais e capturando a complexidade morfológica de sinais cardíacos.

A organização metodológica estrutura-se em cinco seções: duas de fundamentação conceitual e três correspondentes aos eixos experimentais da pesquisa. A Seção IV.1 estabelece a visão geral da abordagem proposta, articulando a lacuna científica identificada na revisão de literatura (Capítulo III) com a estratégia adotada, formalizando as hipóteses de pesquisa e descrevendo o desenho das três etapas experimentais. A Seção IV.2 caracteriza o *MIT-BIH Arrhythmia Database*. As três seções seguintes detalham o **protocolo** de cada etapa experimental: a Seção IV.3 descreve o desenho da Etapa I, voltada à validação da capacidade de reconstrução do *autoencoder* e identificação da melhor configuração de hiperparâmetros; a Seção IV.4 documenta o procedimento de treinamento dos 96 modelos *autoencoder* (Etapa II); e a Seção IV.5 estabelece o protocolo comparativo entre *autoencoder* e SAX (Etapa III), especificando as métricas de avaliação aplicadas à detecção de *motifs*. Os **resultados** obtidos pela aplicação desta metodologia são apresentados e discutidos no Capítulo V.

IV.1 Visão Geral da Abordagem Proposta

A identificação de *motifs* em séries temporais de ECG constitui um problema de mineração de dados com implicações diretas para diagnóstico cardiovascular assistido por computador. Métodos tradicionais abordam essa questão por meio de transformações que convertem sinais contínuos em sequências discretas de símbolos, viabilizando a aplicação de algoritmos de busca de padrões sobre representações compactas. O *Symbolic Aggregate approXimation* (SAX) consolidou-se como técnica predominante nesse contexto ao combinar agregação temporal por média de segmentos com

discretização baseada em quantis da distribuição normal [Lin et al., 2003, 2007]. Entretanto, a imposição de normalidade estatística como premissa fundamental introduz fragilidades metodológicas documentadas empiricamente: a transformação por *Piecewise Aggregate Approximation* altera as propriedades distribucionais dos dados originais, resultando em partições simbólicas com probabilidades desiguais e comprometendo a equiprobabilidade esperada dos símbolos alfabéticos [Butler and Kazakov, 2015].

Em séries como ECG, essas limitações manifestam-se de forma particularmente crítica. A morfologia do batimento cardíaco exibe variabilidade intra e interindividual que não se conforma necessariamente a distribuições gaussianas, especialmente em contextos patológicos como arritmias ventriculares ou distúrbios de condução. Discretizações lineares tendem a agrupar padrões morfologicamente distintos sob o mesmo símbolo, enquanto variações sutis, porém clinicamente relevantes, podem ser fragmentadas em categorias diferentes. Essa inadequação motivou investigações sobre representações alternativas capazes de capturar não linearidades e preservar informação discriminativa. Arquiteturas de aprendizado profundo, particularmente *autoencoders*, demonstraram capacidade de aprender mapeamentos complexos de forma não supervisionada, oferecendo uma alternativa promissora para codificação adaptativa de séries temporais sem pressupor estruturas estatísticas específicas [Hinton and Salakhutdinov, 2006; Goodfellow et al., 2016].

A revisão da literatura científica conduzida no Capítulo III revela o desenvolvimento paralelo de duas linhas de pesquisa. De um lado, múltiplos estudos aplicam *autoencoders* a ECG para tarefas de classificação supervisionada de arritmias [Liang et al., 2021; Ding et al., 2023], detecção de anomalias [Roy et al., 2023; Zhong and Sun, 2023] e compressão de sinais [Dasan and Panneerselvam, 2021]. Essas investigações confirmam que representações latentes neurais preservam características relevantes dos dados originais. Por outro lado, a descoberta de *motifs* em séries temporais avançou com algoritmos baseados no paradigma de discretização simbólica iniciado pelo SAX. No entanto, a literatura não registra aplicações de *autoencoders* como substituto direto dessa discretização tradicional na descoberta de *motifs*.

Essa lacuna metodológica fundamenta a **hipótese central** investigada nesta pesquisa: representações latentes aprendidas por *autoencoders* preservam propriedades discriminativas de sinais de ECG com fidelidade superior à obtida por discretização linear, resultando em descoberta de *motifs* de maior conteúdo informacional sem comprometimento da qualidade de detecção. Formalmente, a hipótese desdobra-se em três predições empiricamente testáveis. Predição 1 (Fidelidade Reconstrutiva): modelos *autoencoder* treinados com configurações refinadas de hiperparâmetros alcançarão $R^2 > 0{,}95$ na reconstrução das séries, demonstrando que a compressão dimensional preserva propriedades essenciais do sinal original. Predição 2 (Eficácia de Detecção): métricas de desempenho em detecção de eventos, como precisão, revocação e F_1, bem como o desempenho computacional, per-

manecerão em níveis equivalentes ou superiores aos obtidos pelo método de referência, validando a viabilidade prática da substituição metodológica. Predição 3 (Complexidade Informacional): *motifs* identificados a partir de representações latentes neurais apresentarão características informacionais distintas daquelas observadas em *motifs* derivados de discretização SAX, refletindo diferenças na captura de padrões morfológicos.

O protocolo experimental desenvolvido para testar essas predições organiza-se em três etapas sequenciais e logicamente interdependentes, cada uma endereçando aspectos complementares da validação científica:

Etapa I - Validação da Capacidade de Reconstrução (Seção IV.3): concentra-se exclusivamente na caracterização técnica da capacidade reconstrutiva de arquiteturas *autoencoder* aplicadas a sinais de eletrocardiograma. Por meio de busca exaustiva em grade (*exhaustive grid search*) sobre espaço paramétrico multidimensional, essa fase identifica configurações que maximizam a fidelidade de reconstrução segundo métricas quantitativas estabelecidas (*Coeficiente de Determinação* (R^2), *Erro Quadrático Médio Normalizado* (NRMSE) e *Erro Percentual Absoluto Médio* (MAPE)). A configuração resultante dessa exploração sistemática será adotada universalmente nas Etapas II e III, assegurando que comparações metodológicas ocorram sob as melhores condições técnicas.

Etapa II – Treinamento dos Modelos *Autoencoder* (Seção IV.4): operacionaliza o treinamento dos modelos *autoencoder* que serão posteriormente empregados na fase de descoberta de *motifs*. Utilizando a configuração de melhor custo-benefício obtida na Etapa I, treina-se um modelo específico para cada série do MIT-BIH *database*. Cada modelo executa a fase `fit` sobre o conjunto de treinamento de sua respectiva série, persistindo os resultados em disco, viabilizando reutilização nas execuções de detecção subsequentes. Essa separação temporal entre treinamento (Etapa II) e detecção (Etapa III) reflete um *workflow* realista: em aplicações práticas, modelos são treinados *offline* uma única vez e aplicados repetidamente a novos dados ou sob diferentes configurações de descoberta. O método SAX, por não requerer fase de treinamento análoga ao *autoencoder* (suas transformações são determinísticas, baseadas em estatísticas calculadas diretamente dos dados de entrada), executa operação equivalente ao `fit` (estimação de média e desvio padrão para normalização) de forma integrada à fase de detecção na Etapa III.

Etapa III - Comparação Experimental Sistemática (Seção IV.5): executa confrontação empírica direta entre a abordagem proposta e o método SAX, aplicando ambos à descoberta de *motifs* em todas as séries presentes no MIT-BIH *database*. O protocolo experimental emprega configuração de parâmetros de descoberta fundamentada em robustez estatística, validada por análise de sensibilidade local que explora diferentes configurações em eixos ortogonais de granularidade simbólica (*alphabet_size*) e critério de recorrência (*occurrences*). O desenho assegura equidade

comparativa: ambos os métodos processam as mesmas séries sob condições idênticas, diferindo apenas na etapa de transformação — codificação neural adaptativa *versus* discretização linear por quantis gaussianos — isolando o efeito da representação como única variável experimental manipulada. Complementarmente, conduz-se experimento de detecção cruzada (*cross-detection*) para avaliar a capacidade de generalização dos modelos *autoencoder*: um modelo treinado sobre série de referência é aplicado às demais séries, comparando-se seu desempenho com os modelos específicos. Esse experimento investiga se representações latentes aprendidas capturam estruturas morfológicas transferíveis entre registros distintos. A avaliação quantitativa registra métricas de acurácia (*soft*), entropia informacional e eficiência computacional.

A Figura IV.1 sintetiza visualmente a arquitetura metodológica tripartite, explicitando relações de dependência entre as etapas experimentais. A aplicação de cada um dos métodos testados, SAX e *Autoencoder*, às séries temporais (fase `fit`) produz uma representação transformada que é submetida ao algoritmo de descoberta de *motifs* baseado em frequência e entropia, assegurando que diferenças observadas nos resultados finais de detecção decorram exclusivamente da etapa de transformação, não de vieses introduzidos por lógicas de busca distintas. A fase `evaluate` consolida métricas de desempenho segundo múltiplas dimensões, permitindo caracterização abrangente do desempenho comparativo.

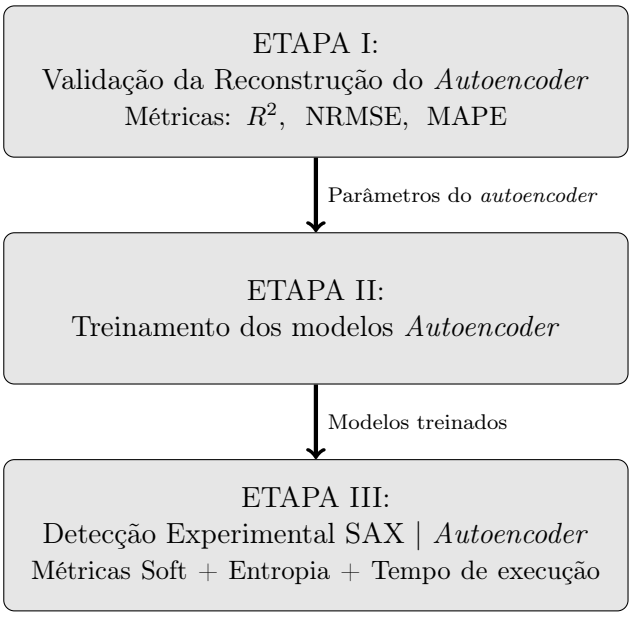


Figura IV.1: Estrutura metodológica em três etapas experimentais interdependentes. Etapa I (superior): Validação da capacidade de reconstrução do *Autoencoder*, buscando o ajustamento dos parâmetros. Etapa II (central): Treinamento dos modelos segundo os parâmetros obtidos. Etapa III (inferior): Detecção experimental e comparação dos dois métodos segundo métricas soft, entropia informacional e tempo de execução.

IV.2 Caracterização do *Dataset* e Protocolo de Seleção

A validação empírica de métodos para detecção de *motifs* em séries de eletrocardiograma requer um conjunto de dados que combine três propriedades essenciais: (*i*) representatividade de condições fisiológicas diversas, (*ii*) anotações de eventos que permitam avaliação objetiva de desempenho de algoritmos de detecção e (*iii*) adoção consolidada pela comunidade científica. O MIT-BIH *Arrhythmia Database*[1] [Moody and Mark, 2001] constitui recurso de referência internacional que satisfaz simultaneamente esses critérios, tendo sido empregado em mais de 500 estudos desde sua publicação em 1980. Este conjunto de dados originou-se do esforço colaborativo entre o *Massachusetts Institute of Technology* e o *Beth Israel Hospital* de Boston, desenvolvido ao longo da década de 1970 com o objetivo de estabelecer um padrão para avaliação de detectores de arritmias cardíacas. É disponibilizado atualmente pela plataforma *Physionet (Research Resource for Complex Physiologic Signals)*, mantida pelos *National Institutes of Health* (NIH) desde 1999, oferecendo acesso gratuito a dados fisiológicos e clínicos para pesquisa biomédica [Goldberger et al., 2000].

O conjunto é composto por 48 registros ambulatoriais, cada um com duração aproximada de 30 minutos, totalizando em torno de 110000 batimentos cardíacos individualmente anotados. Os dados foram coletados entre 1975 e 1979 de 47 pacientes distintos (um paciente contribuiu com dois registros em momentos diferentes), sendo 25 indivíduos do sexo masculino com idades entre 32 e 89 anos e 22 do sexo feminino com idades entre 23 e 89 anos. A população estudada reflete distribuição mista entre pacientes internados (aproximadamente 60%) e ambulatoriais (40%), proporcionando diversidade representativa de condições clínicas encontradas na prática cardiológica. A estratificação do conjunto seguiu protocolo deliberado: 23 registros foram selecionados aleatoriamente de um arquivo contendo 4000 gravações de Holter de 24 horas, representando amostra probabilística da população; os 25 registros remanescentes foram escolhidos intencionalmente por conterem arritmias raras, porém clinicamente significativas, que não seriam adequadamente representadas em amostragem puramente aleatória.

As especificações técnicas de aquisição dos sinais refletem padrões tecnológicos da época, mas permanecem adequadas para investigações metodológicas contemporâneas. Cada registro foi digitalizado a taxa de amostragem de 360 Hz, resolução que proporciona margem acima do critério de Nyquist para captura de componentes espectrais relevantes do ECG (tipicamente até 150 Hz), com resolução de amplitude de 11 bits sobre faixa dinâmica de ± 5 mV. Cada registro compreende duas derivações (*lead*) simultâneas, sendo a primeira invariavelmente MLII (derivação bipolar de membro modificada) e a segunda selecionada entre as derivações precordiais V1, V2, V4 ou V5, conforme critérios de qualidade de sinal e conteúdo informativo. As anotações foram produzidas por cardi-

[1] Disponível em https://physionet.org/content/mitdb/1.0.0/

ologistas especializados de forma independente e posteriormente reconciliadas, classificando cada batimento segundo taxonomia padronizada que distingue batimentos normais, ectópicos ventriculares, ectópicos supraventriculares, batimentos de marcapasso e artefatos, entre outras categorias clinicamente relevantes. Essa configuração de duas derivações simultâneas viabiliza análises que exploram redundância espacial ou discordância entre perspectivas distintas do campo elétrico cardíaco.

Para os propósitos desta investigação, o *dataset* foi organizado como conjunto de 96 séries temporais independentes, derivadas dos 48 registros × 2 derivações por registro. Cada série possui duração aproximada de 30 minutos, adquirida a 360 Hz, totalizando 650.000 amostras por série. A Tabela IV.1 consolida as características técnicas do MIT-BIH *Arrhythmia Database* relevantes para o desenho experimental.

Tabela IV.1: Caracterização do MIT-BIH *Arrhythmia Database*.

Característica	**Especificação**
Origem institucional	*MIT e Beth Israel Hospital (Boston, EUA)*
Período de coleta	1975–1979
Pacientes	47 (25 homens: 32–89 anos; 22 mulheres: 23–89 anos)
Estratificação	23 aleatórios + 25 arritmias raras
Total de Séries	96 (48 registros × 2 derivações)
Derivações	MLII + V1/V2/V4/V5 (sempre duas por registro)
Taxa de amostragem	360 Hz
Total de amostras de cada série	650000 (∼30 min)
Resolução de amplitude	11 bits (±5 mV)
Batimentos anotados	∼110000
Seleção para Etapa I	40 séries V1
Seleção para Etapas II-III	96 séries (ambas derivações)

É importante mencionar que o MIT-BIH *Arrhythmia Database* constitui conjunto de dados público, anonimizado e disponibilizado para pesquisa científica desde 1980. A presente pesquisa não envolveu coleta primária de dados, contato com pacientes ou acesso a informações identificáveis, caracterizando-se como análise secundária de dados previamente coletados e desidentificados.

A divisão dos dados em conjuntos de treinamento e teste preservou rigorosamente a ordem temporal, prática essencial em análise de séries temporais. Na Etapa I de validação da capacidade reconstrutiva (Seção IV.3), explorou-se sistematicamente proporções de dados de treinamento variando de 10% a 50% das amostras iniciais de cada série.

Na Etapa II (Seção IV.4), a configuração refinada obtida na Etapa I foi adotada universalmente para o treinamento dos 96 modelos *autoencoder*: os pesos de cada rede neural foram ajustados exclusivamente sobre a porção inicial da respectiva série. Na Etapa III de comparação experimental (Seção IV.5), ambos os métodos processaram as séries completas, assegurando paridade comparativa. Para o *autoencoder*, modelos pré-treinados (cujos pesos foram fixados após a Etapa II) foram aplicados via método `detect` à integralidade de cada série, gerando representações latentes e descobrindo *motifs* em todas as janelas temporais, independentemente de terem sido observadas durante treinamento. Para o SAX, que não requer fase de treinamento prévia, executou-se sequência integrada `fit–detect` sobre a série completa, sem distinção entre fases de ajuste de parâmetros e aplicação. Essa estratégia assegura que ambas as abordagens executem descoberta de *motifs* sobre conjuntos idênticos de dados, diferindo apenas na etapa de discretização e simbolização.

Complementarmente ao experimento principal de detecção, no qual modelos *autoencoder* são treinados e aplicados sobre a mesma série, conduziu-se investigação adicional de detecção cruzada (*cross-detection*) para avaliar capacidade de generalização das representações latentes aprendidas. Nesse experimento, um único modelo *autoencoder* foi treinado sobre a primeira série do conjunto de dados, tomada como referência. Esse modelo foi então aplicado às 95 séries remanescentes, executando a descoberta de *motifs* da mesma forma que no experimento principal. A comparação de desempenho entre modelos específicos (treinados na própria série-alvo) e modelo generalista (treinado externamente) permite quantificar degradação introduzida pela generalização e verificar se representações latentes capturam estruturas morfológicas transferíveis entre registros distintos. Esse protocolo adicional, detalhado na Seção IV.5, complementa a validação metodológica ao testar robustez da abordagem proposta em cenário operacional relevante: análise de novos dados sem re-treinamento.

IV.3 Etapa I: Validação da Capacidade de Reconstrução

A integração de *autoencoders* em *pipelines* de descoberta de *motifs* fundamenta-se na premissa de que a compressão dimensional executada pelo codificador preserva características temporais essenciais das séries originais. Essa premissa não constitui garantia teórica universal, mas uma propriedade emergente que depende criticamente de escolhas arquiteturais e configurações de treinamento. A validação empírica dessa capacidade reconstrutiva representa, portanto, etapa metodológica incontornável, que antecede logicamente a aplicação do método proposto à tarefa de detecção. Esta seção documenta o protocolo de exploração sistemática do espaço de hiperparâmetros, conduzido com objetivo duplo: (*i*) estabelecer metodologia rigorosa para quantificar fidelidade reconstrutiva de diferentes configurações de *autoencoder*; (*ii*) criar base empírica para seleção de parâmetros adequados que serão empregados nas etapas subsequentes.

A arquitetura neural empregada segue o paradigma *vanilla autoencoder*, caracterizado por topologia totalmente conectada, sem incorporação de estruturas especializadas como convoluções ou recorrências. Essa escolha deliberadamente conservadora prioriza simplicidade e interpretabilidade sobre sofisticação arquitetural, permitindo isolar o efeito da representação latente não linear sem confundi-lo com capacidades indutivas específicas de arquiteturas mais elaboradas. A implementação utiliza a classe `autoenc_ed` do pacote `daltoolboxdp`, que encapsula modelo `PyTorch` acessado via interface `reticulate`.

A rede organiza-se em três componentes funcionais dispostos sequencialmente. O *autoencoder* recebe como entrada a série temporal previamente normalizada, transformada em uma matriz por meio da aplicação de uma janela deslizante de tamanho equivalente ao parâmetro `sw_size`. Cada linha dessa matriz corresponde a um segmento temporal contíguo da série de comprimento `sw_size`, processado pelo *encoder* por meio de sucessivas transformações lineares e não lineares, com dimensionalidade progressivamente reduzida, até atingir o gargalo dimensional representado pela camada latente. Esta, por sua vez, constitui representação compacta de dimensão resultante da aplicação da taxa de compressão (*ratio*) sobre a dimensão original ($encoding_size = \lfloor sw_size \times (1 - ratio) \rfloor$, onde $ratio \in (0, 1)$ e $\lfloor \cdot \rfloor$ denota a função piso —*floor function*). O *decoder* reconstrói uma aproximação do sinal original por expansão simétrica, partindo do vetor latente e aplicando camadas densas com dimensionalidades crescentes até produzir saída de dimensão original `sw_size`.

A não linearidade essencial ao poder representacional da rede é introduzida por funções de ativação aplicadas após transformações lineares em cada camada oculta. A classe `autoenc_ed` do pacote `daltoolboxdp` emprega a função *Rectified Linear Unit* (ReLU), definida como $f(x) = \max(0, x)$, que preserva valores positivos inalterados enquanto anula valores negativos. A camada de saída utiliza ativação linear (função identidade), apropriada para problemas de regressão nos quais o objetivo consiste em predizer valores contínuos sem restrições de faixa, permitindo que a rede reconstrua sinais normalizados em toda a extensão de variação observada nos dados.

O espaço de configurações investigado abrange três dimensões paramétricas principais que influenciam diretamente a capacidade de reconstrução:

- **Proporção de treinamento** (`train_size`): define fração inicial da série temporal reservada para ajuste de pesos da rede, testada em $\{10\%, 20\%, 30\%, 40\%, 50\%\}$, refletindo *trade-off* entre quantidade de exemplos disponíveis para aprendizado e tempo computacional dispensado para o treinamento do modelo.

- **Dimensão de entrada** (`sw_size`): equivale ao comprimento temporal da janela deslizante aplicada à série, com valores explorados em $\{10, 20, 30, 40, 50, 60, 70, 80, 90\}$ amostras.

- **Taxa de compressão** (`ratio`): especifica fração da dimensionalidade eliminada na camada

latente, variando em $\{10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%\}$. Valores próximos a zero implicam compressão leve (espaço latente quase tão grande quanto a entrada), enquanto valores próximos a um impõem compressão agressiva (gargalo dimensional severo).

A combinação cartesiana dessas três dimensões principais — 5 proporções de treinamento $\times$ 9 tamanhos de janela $\times$ 9 taxas de compressão — totaliza 405 configurações únicas por série temporal; multiplicando-se pelas 40 séries selecionadas do MIT-BIH para esta fase exploratória (correspondentes à derivação V1), obtém-se universo experimental de **16200 execuções independentes**.

O protocolo de processamento de cada configuração seguiu sequência padronizada de operações. Inicialmente, a série temporal bruta foi submetida a normalização *z-score*, transformando amplitudes para distribuição com média zero e variância unitária. Em seguida, aplicou-se segmentação por janelas deslizantes de comprimento `sw_size` com passo unitário, gerando conjunto de segmentos sobrepostos. Esse conjunto foi particionado temporalmente: a fração inicial da série correspondente ao parâmetro `train_size` forma o conjunto de treinamento. Um modelo *autoencoder* com arquitetura determinada pelos parâmetros `sw_size` e `ratio` foi instanciado por meio da classe `autoenc_ed`, ajustado e aplicado à série normalizada, utilizando os mecanismos `fit` e `transform` da biblioteca `daltoolbox`. Finalmente, foram calculadas três métricas quantificando fidelidade reconstrutiva, detalhadas a seguir.

- *Coeficiente de Determinação* (R^2) quantifica a proporção da variância dos dados originais explicada pelas reconstruções, fornecendo indicador adimensional de qualidade de ajuste. A interpretação é direta: $R^2 = 1$ indica reconstrução perfeita, enquanto $R^2 = 0$ corresponde ao desempenho de um modelo que não supera predição constante pela média. No contexto desta investigação, R^2 serve como métrica primária de seleção de configurações, com expectativa de valores superiores a 95% para validar suficiência da representação latente.

- *Erro Quadrático Médio Normalizado* (NRMSE) expressa a magnitude típica do erro de reconstrução em relação à amplitude do sinal, proporcionando medida de erro relativo independente de escala absoluta. Valores baixos de NRMSE (tipicamente $< 5\%$) indicam desvios pequenos em relação à variação natural do sinal, sugerindo reconstrução de alta fidelidade.

- *Erro Percentual Absoluto Médio* (MAPE) quantifica o desvio relativo médio das predições em relação aos valores reais, expresso como percentual. Diferentemente do NRMSE, que normaliza globalmente pela faixa de variação, o MAPE normaliza ponto a ponto pelo valor observado, penalizando erros proporcionalmente à magnitude local. O MAPE é particularmente sensível a valores próximos de zero no denominador, podendo inflar artificialmente em regiões de baixa amplitude; não obstante, permanece útil como indicador complementar de erro relativo, especialmente quando a distribuição de amplitudes apresenta assimetrias.

A identificação da melhor configuração de hiperparâmetros procedeu por estratégia de análise sequencial em três estágios, cada um refinando sucessivamente o espaço de busca por eliminação de dimensões já otimizadas. Essa abordagem de otimização em cascata reduz progressivamente a cardinalidade do conjunto de candidatos, facilitando visualização e interpretação de tendências em cada dimensão paramétrica isoladamente. O R^2 foi estabelecido como métrica primária de seleção, com NRMSE e MAPE servindo como indicadores complementares de validação cruzada: configurações que maximizam R^2 devem simultaneamente minimizar as métricas de erro, confirmando consistência da escolha.

No primeiro estágio, analisou-se a dimensão `train_size` sobre o conjunto completo de 16200 execuções. Para cada um dos cinco valores testados (10%, 20%, 30%, 40%, 50%), foram calculadas as estatísticas descritivas das três métricas agregadas sobre todas as combinações de `sw_size` e `ratio`: *Média Aritmética*, *Mediana*, *Variância*, *Desvio Padrão*, *Desvio Absoluto da Mediana* (MAD), *Intervalo Interquartil* (IQR), *Coeficiente de Variação* (CV) e *Coeficiente de Variação Robusto* (RCV). Foram geradas as visualizações de tendência central (R^2 vs. `train_size`, NRMSE vs. `train_size`, MAPE vs. `train_size`) e dispersão relativa (CV vs. `train_size`, RCV vs. `train_size`).

No segundo estágio, fixou-se `train_size` no ponto de balanço entre o tamanho da amostra de treinamento e o melhor valor das métricas de reconstrução identificado no primeiro estágio. Então analisou-se a dimensão `sw_size` sobre o subconjunto filtrado de 3240 execuções (9 tamanhos de janela $\times$ 9 taxas de compressão $\times$ 40 séries). Para cada tamanho testado (`sw_size` $\in \{10, 20, \ldots, 90\}$), foram computadas as mesmas estatísticas descritivas e medidas de dispersão, agregando sobre todas as taxas de compressão. Foram plotados gráficos análogos ao estágio anterior.

No terceiro estágio, foram fixados `train_size` e sw_size, nos valores de melhor custo-benefício identificados, analisando-se a dimensão `ratio` sobre subconjunto final de 360 execuções (9 taxas de compressão $\times$ 40 séries). Para cada taxa testada (`ratio` $\in \{10\%, 20\%, \ldots, 90\%\}$), foram obtidas as mesmas estatísticas descritivas e medidas de dispersão dos estágios anteriores. Novamente foram plotados gráficos análogos aos estágios anteriores, que revelam o *trade-off* entre compactação dimensional e fidelidade de reconstrução das séries.

Esta estratégia de otimização sequencial assume independência entre dimensões paramétricas e a configuração de melhor custo-benefício resultante dessa exploração sistemática foi adotada universalmente nas Etapas II e III, garantindo que comparações metodológicas reflitam exclusivamente diferenças entre esquemas de representação. Os resultados detalhados dessa análise, incluindo valores numéricos de métricas para cada dimensão, gráficos de sensibilidade, tabelas de estatísticas descritivas e identificação da configuração adotada com justificativa quantitativa, são apresentados no Capítulo V.

IV.4 Etapa II: Treinamento dos Modelos *Autoencoder*

A Etapa II operacionaliza o treinamento dos modelos *autoencoder* que serão posteriormente empregados na fase de descoberta de *motifs*, aplicando a configuração de hiperparâmetros identificada na exploração sistemática da Etapa I. Esta seção documenta o procedimento de treinamento dos 96 modelos específicos — um por série do conjunto de dados — e o esquema de persistência que viabilizará sua aplicação na Etapa III, onde serão integrados ao *framework* `Harbinger` para comparação sistemática entre métodos.

O treinamento dos modelos empregou a arquitetura *vanilla autoencoder* validada na Etapa I, cuja estrutura topológica foi detalhadamente especificada na Seção IV.3. A implementação realizou-se pela construção do algoritmo `hmo_autoencoder`, que materializa a interface padronizada `fit` e `detect`, por meio do uso de funções dos pacotes `daltoolbox` (normalização), `tspredit` (transformações de série, como janelas deslizantes e construção do conjunto de treinamento) e `daltoolboxdp` (criação de modelos usando *autoencoder*).

O treinamento dos 96 modelos foi conduzido sequencialmente através dos Algoritmos 1, 2 e 3, que, respectivamente, criam o objeto *autoencoder* com a configuração definida na Etapa I, treinam o modelo (fase `fit`) e persistem o modelo resultante. Esse fluxo é executado sequencialmente para cada uma das 96 séries do MIT-BIH, garantindo independência entre os modelos treinados.

Algorithm 1 Autoencoder Object Creation

```
1: function CreateAutoencoderObject(sw_size, ratio, train_size)
2:     obj ← CreateBaseObject
3:     if ratio ≤ 0 or ratio > 1 then
4:         stop "Ratio must be between 0 and 1"
5:     end if
6:     if train_size ≤ 0 or train_size > 1 then
7:         stop "Train size must be between 0 and 1"
8:     end if
9:     obj.sw_size ← sw_size
10:    obj.ratio ← ratio
11:    obj.train_size ← train_size
12:    obj.encoding_size ← ComputeEncodingSize(sw_size, ratio)
13:    SetClass(obj, "AutoencoderObject")
14:    return obj
15: end function
```

Após o treinamento de cada modelo, o esquema de persistência emprega formato dual: pesos da rede neural são salvos em arquivo `.pth` (formato `PyTorch` *checkpoint*) via `torch.save`, enquanto metadados contextuais (estatísticas de normalização, hiperparâmetros, histórico de convergência) são salvos em arquivo `.rds` (formato nativo do `R`) via `saveRDS`. Os artefatos são organizados hierarquicamente em diretório dedicado, indexados por identificador de série, o que permite recuperação determinística de qualquer modelo sem necessidade de re-treinamento.

Algorithm 2 Autoencoder Fitting

```
function FitAutoencoder(obj, serie)
    if serie is null then
        stop "No data was provided"
    end if
    obj ← StoreReferences(obj, serie)
    normModel ← CreateNormalizer
    normData ← Normalize(normModel, serie)
    windowedData ← SlidingWindow(normData, obj.sw_size)
    if obj.train_size == 1 then
        test_size ← 1
    else
        test_size ← ComputeTestSize(windowedData, obj.train_size)
    end if
    (trainData, testData) ← Split(windowedData, test_size)
    autoencoder ← InitAutoencoder(obj.sw_size, obj.encoding_size)
    obj.model ← Train(autoencoder, trainData)
    return obj
end function
```

Algorithm 3 Save Autoencoder Model

```
function SaveModel(obj)
    params ← ExtractParameters(obj)
    weights ← ExtractWeights(obj)
    SaveParameters(params, paramsFile)
    SaveWeights(weights, weightsFile)
    return (paramsFile, weightsFile)
end function
```

Ao término da Etapa II, dispõe-se de 96 modelos treinados e persistidos em disco. Esses modelos constituem insumo essencial para a Etapa III (Seção IV.5), onde serão carregados e aplicados repetidamente sob diferentes configurações de descoberta (variando `alphabet_size` e `occurrences`), amortizando o custo fixo de treinamento sobre múltiplas execuções de detecção. Adicionalmente, modelos de referência por tipo de derivação serão aplicados a séries distintas no experimento de detecção cruzada (*cross-detection*), avaliando capacidade de generalização de representações latentes aprendidas. Os resultados quantitativos do processo de treinamento serão apresentados no Capítulo V.

IV.5 Etapa III: Comparação Experimental Sistemática

A Etapa III constitui o núcleo avaliativo desta investigação, operacionalizando a comparação sistemática entre o método proposto — que substitui a discretização estatística do SAX por codificação neural via *autoencoder* — e a abordagem de referência baseada exclusivamente em SAX. Esta seção documenta o protocolo experimental que permite quantificar, sob condições controladas, as diferenças de desempenho entre as duas estratégias de descoberta de *motifs* em séries temporais de

ECG. A metodologia subdivide-se em dois experimentos complementares: o experimento principal de detecção, no qual ambos os métodos processam as 96 séries do conjunto de dados completo, e o experimento de detecção cruzada (*cross-detection*), que avalia a capacidade de generalização das representações latentes aprendidas. Os resultados obtidos pela aplicação deste protocolo também são apresentados no Capítulo V.

A implementação dos métodos comparados adere rigorosamente à interface padronizada estabelecida pelo *framework* `Harbinger` [Salles et al., 2020], que define protocolo para detectores de eventos em séries temporais baseado em linhas de experimento: fase `fit` para treinamento de modelos, fase `detect` para detecção dos *motifs* e fase `evaluate` para o cálculo de métricas de desempenho mediante confronto com anotações de referência.

O método SAX, na forma como implementado no pacote `Harbinger` e utilizado neste experimento, não requer fase `fit` de treinamento de modelo, ao contrário do método proposto com *autoencoder*. As transformações do SAX são determinísticas, calculadas diretamente dos dados de entrada durante a própria fase de detecção. Em razão desta diferença, o método baseado em *autoencoder* operacionaliza o carregamento dos modelos pré-treinados e persistidos em disco (Etapa II) para proceder com a transformação da série completa em representações latentes. Esta tarefa é realizada por meio do Algoritmo 4.

Algorithm 4 Load Autoencoder Model

```
1: function LoadModel(paramsFile, weightsFile)
2:     params ← LoadParameters(paramsFile)
3:     model ← CreateAutoencoderObject
4:     AssignParameters(model, params)
5:     model.network ← InitializeNetwork(params)
6:     LoadWeights(model.network, weightsFile)
7:     AssignTrainingMetrics(model, params)
8:     return model
9: end function
```

Fase *Detect*: Simbolização e Descoberta

A fase `detect` de ambos os métodos organiza-se em duas etapas sequenciais.

A etapa de **simbolização** transforma a série temporal contínua em sequências simbólicas. No método SAX, essa transformação ocorre pela normalização *z-score* da série, seguida de agrupamento dos dados por quantis e mapeamento para um alfabeto de tamanho predefinido `alphabet_size`, finalizado com a transformação da série utilizando janelas deslizantes de tamanho `encoding_size`. Na abordagem usando *autoencoder*, esta etapa também ocorre pela aplicação da normalização *z-score* à série, transformando-a em seguida pela utilização de janelas deslizantes de tamanho `sw_size`, o que permite a geração da representação latente de dimensão `encoding_size` utilizando os modelos pré-treinados e carregados. A partir daí, há a discretização da série resultante por particionamento

uniforme (*equal-width binning*) em intervalos de mesmo tamanho, mapeando cada componente latente para símbolo do alfabeto de tamanho `alphabet_size`. A utilização dos mesmos parâmetros em ambos os métodos assegura que as representações simbólicas produzidas, submetidas à etapa posterior de detecção, tenham a mesma estrutura. A discretização e simbolização realizadas na abordagem usando *autoencoder* são apresentadas no Algoritmo 5.

Algorithm 5 Symbolic Encoding via Autoencoder

```
1: function EncodeToSymbols(model, serie, alphabet_size)
2:     ValidateAlphabet(alphabet_size)
3:     norm ← CreateNormalizer
4:     normSerie ← Normalize(norm, serie)
5:     windowedSerie ← SlidingWindow(normSerie, model.sw_size)
6:     latent ← EncodeWithAutoencoder(model, windowedSerie)
7:     vector ← Vectorize(latent)
8:     bins ← Discretize(vector, alphabet_size)
9:     symbols ← MapBinsToAlphabet(bins)
10:    sequences ← CreateSymbolMatrix(symbols, model.encoding_size)
11:    return sequences
12: end function
```

A etapa de **descoberta**, adotada de forma idêntica nos dois métodos, detecta subsequências simbólicas que se repetem com frequência superior a um limiar estabelecido e ordena-as segundo uma medida de conteúdo informacional, selecionando padrões que combinam alta frequência com alta distintividade.

Para cada sequência simbólica produzida na etapa de simbolização em ambos os métodos, calcula-se a entropia de *Shannon* segundo a distribuição empírica de símbolos componentes. As sequências são então agrupadas por identidade simbólica, filtradas pelo limiar de recorrência `occurrences` e ordenadas de forma decrescente, primeiramente por frequência de ocorrência e, em seguida, por entropia, de modo que padrões mais frequentes e informativamente mais ricos sejam priorizados. Para cada candidato na lista ordenada, verificam-se suas ocorrências na série e, quando o suporte mínimo é atingido, o padrão é registrado como *motif*; uma vizinhança de tamanho equivalente à metade de `encoding_size` em torno de cada posição identificada é então excluída da busca subsequente, prevenindo correspondências triviais por sobreposição. O processo retorna o objeto de detecção contendo *motifs*, sequências e entropias associadas.

Este mecanismo de ranqueamento por conteúdo informacional — que prioriza *motifs* segundo entropia de *Shannon* — constitui contribuição metodológica central do LEMOD: ao aplicar um critério baseado em teoria da informação diretamente sobre representações latentes, o método seleciona padrões estruturalmente mais coerentes e morfologicamente específicos. *Motifs* compostos por símbolos uniformemente distribuídos apresentam entropia máxima, indicando maior complexidade estrutural; *motifs* dominados por símbolo único exibem entropia próxima de zero, sugerindo padrões

morfologicamente simples ou degenerados. O Algoritmo 6 formaliza este mecanismo.

Algorithm 6 Motif Discovery

```
1: function DiscoverMotifs(sequences, ocurrences, encoding_size)
2:     for i = 1 to |sequences| do
3:         entropy_i ← −Σ_{a∈Σ} p(a) log_2 p(a)                ▷ Shannon
4:     end for
5:     data ← CreateDataFrame(indices, sequences, entropies)
6:     gp ← GroupBySequence(data)
7:     filtered ← Filter(gp, count ≥ occurrences)
8:     candidates ← SortBy(filtered, desc(count, entropy))
9:     Δ ← ⌊encoding_size/2⌋
10:    motifs ← ∅
11:    for word ∈ candidates do
12:        support ← FindOccurrences(data, word.sequence)
13:        if |support| ≥ occurrences then
14:            motifs ← motifs ∪ support
15:            data ← ExcludeNeighborhood(data, support, Δ)
16:        end if
17:    end for
18:    return BuildObject(motifs, sequences, entropies)
19: end function
```

Fase *Evaluate*: Métricas de Avaliação

A fase `evaluate`, também idêntica em ambos os métodos, utiliza as métricas de avaliação SoftED [Salles et al., 2024] implementadas no `Harbinger` [Salles et al., 2020]. Trata-se de extensão das métricas clássicas de classificação adaptada ao contexto de séries temporais mediante introdução de tolerância temporal. Diferentemente de métricas rígidas que exigem coincidência exata de posições, essas métricas contabilizam um evento como verdadeiro positivo se ocorrer dentro de uma janela temporal de $\pm k$ amostras em torno de evento anotado. Essa flexibilização reconhece a realidade prática de que eventos em séries temporais não possuem delimitação temporal precisa, especialmente em séries de ECG sujeitas a variabilidade fisiológica e ruído de aquisição.

São calculadas quatro métricas fundamentais. A **acurácia** quantifica proporção global de classificações corretas; a **precisão** mede fração de detecções que correspondem a *motifs* reais; a **revocação** quantifica fração de *motifs* verdadeiros efetivamente detectados; a medida $\mathbf{F_1}$ harmoniza precisão e revocação via média harmônica, constituindo indicador balanceado que penaliza simultaneamente excesso de falsos positivos e falsos negativos.

Nesta investigação, empregou-se uma janela de tolerância de $k = 100$ amostras, equivalente a aproximadamente 278 ms na taxa de amostragem de 360 Hz das séries de ECG do MIT-BIH *database*. Esse valor reflete uma margem temporal compatível com a variabilidade natural de eventos cardíacos, que não ocorrem com precisão absoluta entre batimentos sucessivos. Pequenos deslocamentos entre o instante anotado e o detectado, na escala de centenas de milissegundos, podem ser clinicamente aceitáveis e não caracterizam erros relevantes. A escolha de $k = 100$

busca, portanto, equilibrar rigor avaliativo e realismo fisiológico, evitando penalizações decorrentes de desalinhamentos mínimos, porém sem ampliar excessivamente o critério de tolerância.

Complementarmente às métricas de detecção, a medida de entropia de *Shannon* empregada no ranqueamento dos *motifs* também é aproveitada como indicador de avaliação. Como o processo de descoberta já estima a entropia de cada subsequência candidata para ordenar padrões segundo seu conteúdo informacional, esses mesmos valores são utilizados para quantificar o grau de complexidade estrutural capturado por cada método. Para cada série, calcula-se a entropia média dos *motifs* identificados, ponderada por suas frequências de ocorrência, produzindo uma métrica agregada de diversidade simbólica. Adicionalmente, mensura-se o tempo de processamento da fase de detecção em ambos os métodos, permitindo avaliar a eficiência computacional associada à descoberta de *motifs*.

Experimento de Detecção Cruzada

O experimento de detecção cruzada (*cross-detection*) investiga capacidade de generalização das representações latentes aprendidas por *autoencoders*, testando se um modelo treinado em série específica mantém desempenho adequado quando aplicado a séries distintas da mesma derivação. Esta avaliação possui relevância metodológica — verificar se representações latentes capturam estruturas morfológicas transferíveis, para além de memorização de características idiossincráticas da série de treinamento — e relevância prática, ou seja, viabilizar cenários operacionais nos quais seria impraticável treinar um modelo específico para cada nova série analisada.

O protocolo experimental selecionou uma série de referência por tipo de derivação disponível no MIT-BIH *Arrhythmia Database*: `r100_MLII` para derivação MLII, `r101_V1` para derivação V1, `r102_V2` para V2 e `r104_V5` para V5. A derivação V4 não foi testada por possuir apenas uma série no banco de dados (`r124_V4`), impossibilitando avaliação de transferência. Para cada derivação, treinou-se um modelo autoencoder sobre a série de referência com a configuração identificada na Etapa I, e esse modelo foi aplicado às séries remanescentes da mesma derivação, executando descoberta de *motifs* com os parâmetros da configuração principal. A comparação de desempenho procede por confronto direto entre duas condições experimentais para cada série-alvo: (*i*) detecção com modelo treinado na própria série, conforme experimento principal; (*ii*) detecção com modelo generalista treinado externamente. A degradação de desempenho, quantificada por diferenças nas métricas de acurácia, precisão, revocação e F_1, estima a penalidade introduzida pela generalização.

Organização de Artefatos e Reprodutibilidade

A execução dos experimentos produziu conjunto estruturado de artefatos persistidos em disco, organizados hierarquicamente por etapa metodológica e série processada. Modelos *autoencoder* treinados (96 arquivos `.pth` contendo pesos neurais e 96 arquivos `.rds` contendo metadados) foram armazenados em diretório dedicado, indexados por identificador de série. Objetos de detecção

retornados pelos métodos `detect` de ambas as abordagens foram persistidos em arquivos `.RData` individuais, preservando estruturas completas de *motifs* (posições temporais, composições simbólicas, frequências de ocorrência, entropias individuais) para análises posteriores; métricas consolidadas foram exportadas em formato tabular `.csv`, facilitando ingestão por ferramentas de análise estatística e visualização.

A metodologia descrita nesta seção estabelece um protocolo reprodutível para comparação sistemática de métodos de descoberta de *motifs*, isolando fontes de variação não relacionadas ao objeto de investigação e empregando métricas estabelecidas na literatura de detecção de eventos. A adesão à interface `Harbinger` assegura extensibilidade futura, permitindo incorporação de métodos adicionais sem necessidade de reimplementação do *pipeline* avaliativo. Os resultados empíricos obtidos pela aplicação deste protocolo são apresentados no Capítulo V.

Capítulo V Avaliação Experimental

Este capítulo apresenta os resultados da avaliação experimental do LEMOD para detecção de *motifs* em séries temporais de ECG utilizando *autoencoders*. O objetivo central é responder à questão de pesquisa que orienta esta investigação: **o uso de *autoencoders* para representação de séries temporais proporciona melhor desempenho na detecção de *motifs* quando comparado ao método tradicional SAX?**

O protocolo experimental, detalhado no Capítulo IV, organizou-se em três etapas interdependentes. A Seção V.1 apresenta a síntese das configurações executadas nas três etapas, consolidada na Tabela V.1. As seções subsequentes apresentam os resultados de cada etapa: a Seção V.2 reporta os resultados da investigação de hiperparâmetros do *autoencoder* e a configuração ideal identificada; a Seção V.3 apresenta os resultados do treinamento dos 96 modelos autoencoder; a Seção V.4 documenta os resultados da comparação experimental sistemática entre o LEMOD e o SAX, incluindo análise de sensibilidade paramétrica e experimento de detecção cruzada; e a Seção V.5 integra os achados das três etapas, valida as predições de pesquisa e discute limitações do estudo.

V.1 Configuração Experimental

Os experimentos foram conduzidos seguindo o protocolo metodológico detalhado no Capítulo IV, aplicado ao MIT-BIH *Arrhythmia Database* [Moody and Mark, 2001], cujas características técnicas foram apresentadas na Seção IV.2.

As execuções ocorreram em infraestrutura computacional de alto desempenho equipada com processador AMD EPYC 9354 (32 núcleos físicos, 64 *threads*, frequência base 3,25 GHz, cache L3 de 256 MB), aceleradora gráfica NVIDIA H100 80GB (arquitetura Hopper, 16.896 núcleos CUDA, memória HBM3 com largura de banda de 3 TB/s) e 512 GB de memória RAM DDR5. O sistema operacional empregado foi Fedora Linux 39, ambiente que suporta nativamente tecnologias de conteinerização utilizadas para isolamento de dependências.

Todo o código-fonte produzido foi desenvolvido primariamente em linguagem `R` (versão 4.4.2), privilegiando as seguintes bibliotecas de código aberto com comunidades ativas de desenvolvimento e manutenção:

- `daltoolbox` e `daltoolboxdp` para componentes de pré-processamento, transformações de sé-

ries temporais e abstração de modelos de aprendizado profundo;

- `tspredit` para manipulação especializada de estruturas de séries temporais;
- `Harbinger` [Salles et al., 2020] para detecção de eventos, cálculo de métricas de avaliação e comparação sistemática de métodos.

A integração ao `Harbinger` é particularmente relevante por garantir interfaces padronizadas (`fit`, `detect`, `evaluate`) que facilitam extensão futura e incorporação de métodos adicionais sem necessidade de reimplementação de *pipelines* de avaliação.

A interoperabilidade com o ecossistema `Python`, especificamente o *framework* `PyTorch` (versão 2.0.1) para implementação de arquiteturas de redes neurais, foi estabelecida via pacote `reticulate` (versão 1.28), que permite invocação bidirecional de objetos e funções `Python` a partir de sessões `R`. Todas as fontes de aleatoriedade durante treinamento foram controladas por definição explícita de sementes geradoras fixas (seed=42), garantindo reprodutibilidade determinística das execuções em ambiente idêntico.

O *workflow* computacional foi estruturado para maximizar rastreabilidade e viabilizar recuperação de execuções interrompidas. Resultados intermediários são persistidos em formatos estruturados ao término de cada fase:

- Modelos *autoencoder* treinados são salvos em arquivos `.pth` (formato nativo `PyTorch` *checkpoint*, contendo dicionários de tensores de pesos e metadados de arquitetura);
- Objetos `R` contendo estruturas de detecção são persistidos em arquivos `.RData` (formato binário comprimido padrão de `R`);
- Tabelas de métricas consolidadas são exportadas em formato `.csv` (valores separados por vírgula), facilitando interoperabilidade com ferramentas de análise estatística externas e possibilitando ingestão em *pipelines* de visualização ou bases de dados relacionais.

O código-fonte completo — incluindo scripts de execução dos experimentos e funções auxiliares de processamento e análise — foi versionado e disponibilizado em repositório público no GitHub[1], assegurando a plena reprodutibilidade dos experimentos em alinhamento com princípios de ciência aberta e reprodutibilidade computacional.

A Tabela V.1 consolida os parâmetros e volumes experimentais implementados nas três etapas avaliativas, servindo como referência para as seções subsequentes.

[1]Disponível em: https://github.com/cefet-rj-dal-members/luiz-oliveira

Tabela V.1: Síntese das configurações experimentais implementadas nas três etapas avaliativas. Parâmetros metodológicos fundamentados no Capítulo IV.

Etapa	**Componente**	**Configuração**
I	*Dataset*	40 segmentos das séries de derivação V1 (10–90 s)
	Parâmetros	$train_size \in \{10\%, 20\%, 30\%, 40\%, 50\%\}$ (5 níveis)
		$sw_size \in \{10, 20, \ldots, 90\}$ (9 níveis)
		$ratio \in \{10\%, 20\%, \ldots, 90\%\}$ (9 níveis)
	Métricas	R^2, NRMSE, MAPE
	Experimentos	16200 ($40 \times 9 \times 9 \times 5$)
II	*Dataset*	96 séries completas (48 registros $\times$ 2 derivações)
	Configuração	$train_size = 30\%$, $sw_size = 30$, $ratio = 60\%$,
		$encoding_size = 12$ dimensões latentes
	Modelos	96 treinados (`.pth` + `.rds`)
III	*Dataset*	96 séries completas (48 registros $\times$ 2 derivações)
	Métodos	SAX vs. LEMOD
	Configuração principal	$alphabet_size = 10$, $occurrences = 6$
	Análise de sensibilidade	Dimensão 1: $alphabet_size \in \{14, 18, 22, 26\}$
		Dimensão 2: $occurrences \in \{2, 4, 8, 10\}$
	Avaliação	SoftED (tolerância $k = 100$ amostras)
	Experimentos	1728 ($192 + 768 + 768$)
Total		17928 experimentos

V.2 Etapa I: Validação da Capacidade de Reconstrução

Esta seção apresenta os resultados da exploração sistemática do espaço de hiperparâmetros do *autoencoder*, conduzida conforme o protocolo descrito na Seção IV.3, com o objetivo de identificar a configuração que maximiza a fidelidade de reconstrução das séries de ECG. Os resultados são organizados segundo a estratégia de otimização sequencial em três estágios, cada um refinando uma dimensão paramétrica a partir dos resultados do estágio anterior, garantindo que a configuração resultante reflita o comportamento conjunto dos hiperparâmetros.

Estágio 1 — Proporção de Treinamento (`train_size`)

A proporção de dados destinada ao treinamento governa diretamente a capacidade de generalização do *autoencoder*: valores muito baixos limitam o aprendizado das estruturas morfológicas da série, enquanto valores elevados reduzem a disponibilidade de amostras para avaliação da qualidade reconstrutiva. A Figura V.1 apresenta a distribuição de R^2 e a Figura V.2 as métricas de dispersão em função de $train_size$. A Tabela V.2 consolida as estatísticas descritivas correspondentes.

Tabela V.2: Estatísticas descritivas de R^2 (%) em função de *train_size* com destaque para o valor de melhor balanço identificado.

train_size	Média	Mediana	SD	MAD	CV	RCV	IQR
10%	91,04	95,22	11,12	3,52	12,22	3,70	9,74
20%	94,15	97,42	9,36	1,96	9,95	2,02	5,58
30%	**95,62**	**98,53**	**8,39**	**1,08**	**8,78**	**1,10**	**3,48**
40%	93,96	97,96	9,67	1,62	10,29	1,66	5,77
50%	94,02	97,68	8,79	1,88	9,35	1,92	6,22

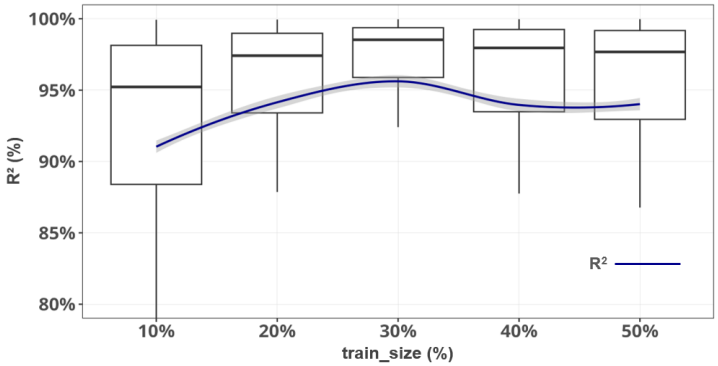


Figura V.1: Distribuição de R^2 em função de *train_size*. Cada *boxplot* agrega 3240 experimentos (40 séries × 9 níveis de *sw_size* × 9 níveis de *ratio*).

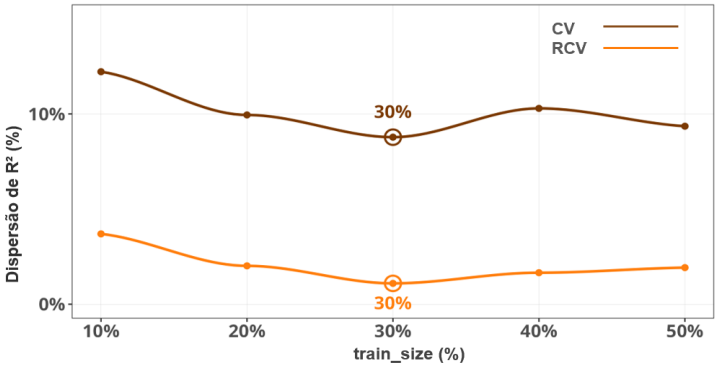


Figura V.2: Dispersão relativa de R^2 em função de *train_size*, confirmando a escolha de *train_size*=30%.

Observa-se tendência crescente da mediana de R^2 até *train_size* = 30%, seguida de estabilização e queda para proporções superiores. A configuração *train_size* = 10% apresentou mediana de 95,22% com dispersão relativa elevada (CV=12,22%, RCV=3,70%, *Intervalo Interquartil* (IQR)=9.74%), indicando instabilidade na capacidade de generalização quando o modelo dispõe de poucos exemplos de treinamento. O patamar de melhor balanço foi atingido em *train_size* = 30%, que alcançou a maior média (95,62%) e mediana (98,53%), com menor dispersão relativa (CV=8,78%, RCV=1,10%, IQR=3,48%), equilibrando quantidade de dados para aprendizado e disponibilidade de amostras para avaliação da capacidade de generalização.

As métricas de erro apresentaram comportamento complementar ao R^2, com mínimos coincidindo com o máximo em $train_size = 30\%$. A Figura V.3 apresenta a distribuição de NRMSE e a Figura V.4 de MAPE em função de $train_size$.

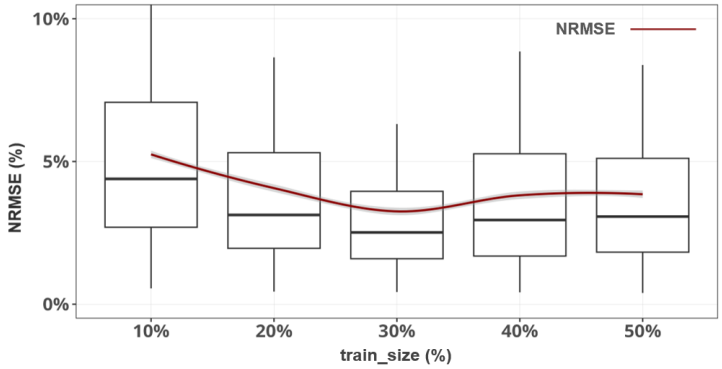


Figura V.3: Distribuição de NRMSE (%) em função de $train_size$.

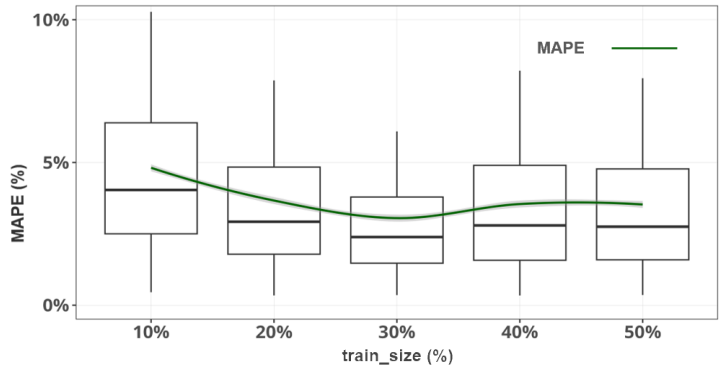


Figura V.4: Distribuição de MAPE (%) em função de $train_size$.

O NRMSE decresceu de 4,39% ($train_size = 10\%$) para 2,51% ($train_size = 30\%$), enquanto o MAPE reduziu-se de 4,04% para 2,39%. Configurações com $train_size \geq 40\%$ apresentaram deterioração das métricas, indicando que uma maior proporção de treinamento não representa, necessariamente, ganho de qualidade reconstrutiva.

Estágio 2 — Tamanho da Janela Deslizante (`sw_size`)

O tamanho da janela deslizante determina o contexto temporal capturado em cada amostra de entrada do *autoencoder*: janelas curtas reduzem a complexidade da tarefa de reconstrução, mas podem ser insuficientes para representar a morfologia completa de um batimento cardíaco; janelas longas ampliam o contexto temporal disponível, porém aumentam a dimensionalidade de entrada e exigem maior capacidade da camada latente para preservar as características essenciais do sinal.

Fixando $train_size = 30\%$, analisou-se o impacto do tamanho da janela deslizante sobre a fidelidade de reconstrução. A Tabela V.3 consolida as estatísticas descritivas de R^2 em função de sw_size. A Figura V.5 apresenta a distribuição de R^2 e a Figura V.6 apresenta as métricas de dispersão de R^2 em função de sw_size.

Tabela V.3: Estatísticas descritivas de R^2 (%) em função de sw_size, fixando $train_size = 30\%$, com destaque para o valor de melhor balanço identificado.

sw_size	Média	Mediana	SD	MAD	CV	RCV	IQR
10	97,65	99,56	4,86	0,33	4,97	0,33	1,58
20	98,12	99,34	3,44	0,47	3,51	0,47	1,47
30	**97,90**	**99,09**	**3,40**	**0,63**	**3,47**	**0,63**	**2,00**
40	97,02	98,68	5,11	0,95	5,26	0,96	2,52
50	96,13	98,34	6,70	1,13	6,97	1,15	3,62
60	94,93	98,08	8,70	1,33	9,17	1,36	3,89
70	93,72	97,73	10,71	1,49	11,42	1,52	4,62
80	93,00	97,20	11,26	1,88	12,11	1,94	5,58
90	92,12	96,75	12,62	2,18	13,70	2,25	5,73

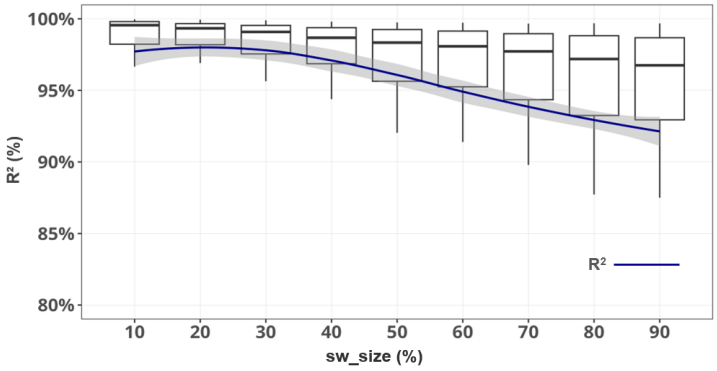


Figura V.5: Distribuição de R^2 em função de sw_size, fixando $train_size = 30\%$. Cada *boxplot* agrega 360 experimentos (40 séries × 9 níveis de $ratio$).

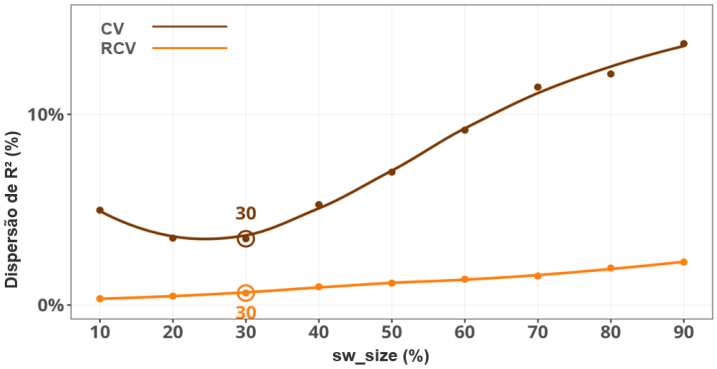


Figura V.6: Dispersão relativa de R^2 (CV e RCV) em função de sw_size, confirmando a escolha de $sw_size = 30$.

Janelas curtas ($sw_size \leq 20$) alcançaram medianas superiores a 99%, refletindo a facilidade da tarefa reconstrutiva quando o contexto temporal é limitado. A escolha de $sw_size = 30$ fundamentou-se no equilíbrio entre dois critérios: (i) preservação de alta fidelidade ($R^2 > 99\%$) e (ii) complexidade computacional gerenciável, pois janelas maiores ampliam a dimensionalidade

de entrada da rede neural, exigindo maior capacidade da camada latente para preservar informação.

Contudo, a partir de $sw_size = 30$, observa-se degradação gradual: a mediana de R^2 decresce monotonicamente de 99,09% ($sw_size = 30$) para 96,75% ($sw_size = 90$), enquanto a dispersão aumenta progressivamente (CV cresce de 3,47% para 13,70%, RCV de 0,63% para 2,25%, IQR de 2,00 para 5,73).

Embora $sw_size = 20$ apresente fidelidade marginalmente superior, a configuração $sw_size = 30$ captura contexto temporal morfológico mais rico — possibilita a descoberta de *motifs* maiores — ao mesmo tempo em que mantém uma dispersão mínima antes do crescimento observado para janelas maiores, conforme confirmado pelos já mencionados valores de CV, RCV e IQR na Figura V.6. Esse comportamento é corroborado pelas métricas de erro absolutas. As Figuras V.7 e V.8 apresentam as distribuições de NRMSE e MAPE em função de sw_size.

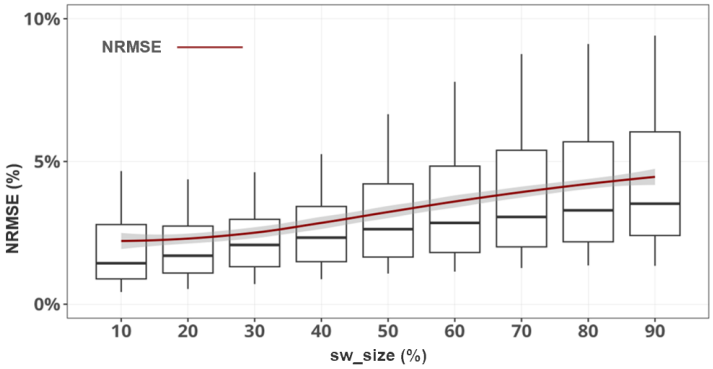


Figura V.7: Distribuição de NRMSE (%) em função de sw_size, fixando $train_size = 30\%$. Cada *boxplot* agrega 360 experimentos (40 séries × 9 níveis de *ratio*).

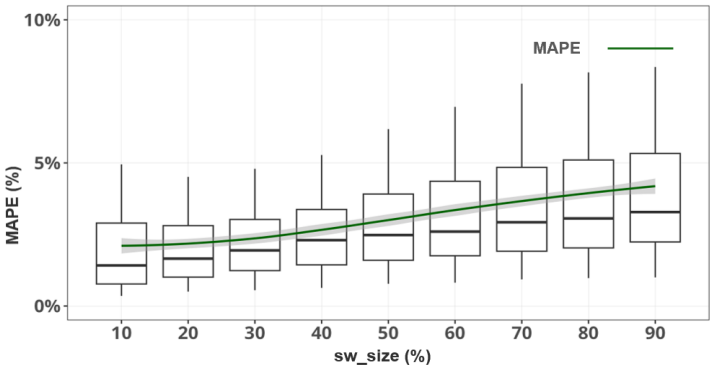


Figura V.8: Distribuição de MAPE (%) em função de sw_size, fixando $train_size = 30\%$. Cada *boxplot* agrega 360 experimentos (40 séries × 9 níveis de *ratio*).

Ambas as métricas reforçam a escolha: a dispersão relativa de R^2 atinge seu valor mínimo em $sw_size = 30$ (CV=3,47%), imediatamente antes do crescimento observado para janelas maiores — o que evidencia que $sw_size = 30$ é o ponto de maior estabilidade entre os candidatos com R^2 mediano superior a 99%. A partir de $sw_size = 40$, os erros crescem continuamente: o NRMSE sobe de 2,08% para 3,52% e o MAPE de 1,91% para 3,28% ao ampliar a janela até 90 amostras.

Estágio 3 — Taxa de Compressão (`ratio`)

A taxa de compressão dimensional determina o tamanho da camada latente do *autoencoder*: valores baixos preservam mais informação, mas produzem representações extensas que comprometem a eficiência da etapa de simbolização; valores elevados geram representações compactas, porém podem descartar características morfológicas essenciais do sinal. O parâmetro *ratio* governa diretamente o *encoding_size* resultante, de modo que sua escolha condiciona tanto a qualidade reconstrutiva quanto a viabilidade prática da descoberta de *motifs* nas etapas subsequentes. Com $train_size = 30\%$ e $sw_size = 30$, fixados a partir dos estágios anteriores, explorou-se o *trade-off* entre compressão dimensional e fidelidade de reconstrução. A Tabela V.4 quantifica esse comportamento.

Tabela V.4: Estatísticas descritivas de R^2 (%) em função de *ratio*, fixando $train_size = 30\%$ e $sw_size = 30$, com destaque para o valor de melhor balanço identificado.

ratio	Média	Mediana	SD	MAD	CV	RCV	IQR
10%	98,57	99,40	2,45	0,36	2,49	0,36	1,46
20%	98,52	99,35	2,14	0,40	2,18	0,40	1,50
30%	98,59	99,29	1,66	0,44	1,69	0,45	1,54
40%	98,52	99,33	2,00	0,39	2,03	0,39	1,61
50%	98,44	99,26	1,84	0,47	1,86	0,48	1,58
60%	**98,45**	**99,22**	**1,96**	**0,45**	**1,99**	**0,45**	**1,81**
70%	97,95	99,10	2,88	0,58	2,94	0,58	2,10
80%	97,38	98,52	3,21	1,05	3,30	1,06	2,53
90%	94,65	97,02	7,06	1,92	7,46	1,98	4,88

A Figura V.9 apresenta a distribuição de R^2 em função de *ratio*, evidenciando o comportamento não linear da fidelidade de reconstrução ao longo da faixa explorada. A Figura V.10 apresenta as métricas de dispersão correspondentes.

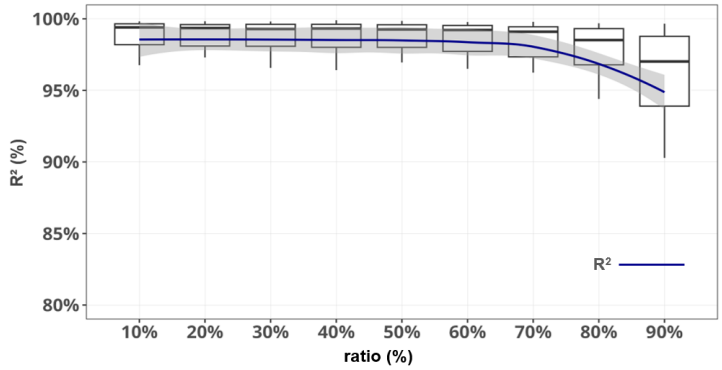


Figura V.9: Distribuição de R^2 em função de *ratio*, fixando $train_size = 30\%$ e $sw_size = 30$. Cada *boxplot* agrega 40 experimentos (40 séries da derivação V1).

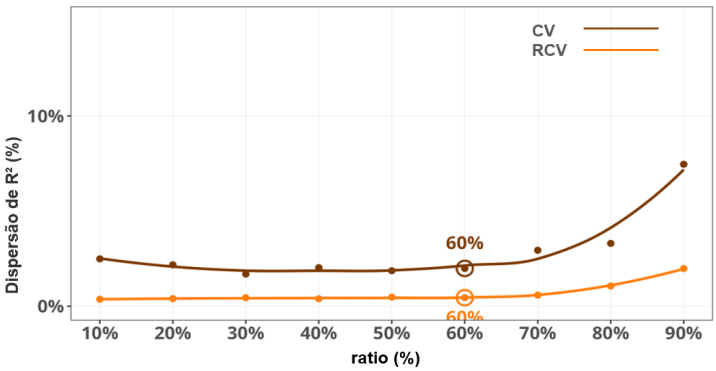


Figura V.10: Dispersão relativa de R^2 (CV e RCV) em função de *ratio*, confirmando a escolha de $ratio = 60\%$.

Taxas de compressão moderadas ($ratio \leq 60\%$) mantêm R^2 acima de 98%, enquanto compressões agressivas ($ratio \geq 70\%$) induzem degradação acelerada. A escolha de $ratio = 60\%$ representa o ponto de melhor balanço entre compressão dimensional e fidelidade de reconstrução: essa configuração reduz a dimensionalidade em 60% (de 30 para $encoding_size = 12$ dimensões latentes), maximiza a compressão viabilizando representações simbólicas compactas nas etapas subsequentes, não compromete a qualidade reconstrutiva, ao mesmo tempo em que preserva R^2 superior a 98%. Trata-se de um compromisso entre custo computacional e qualidade de solução — há configurações com maior fidelidade absoluta ($ratio \leq 20\%$), mas a compressão adicional oferecida por $ratio = 60\%$ é essencial para viabilizar a etapa de simbolização com $encoding_size = 12$ dimensões, sem degradação substancial. A Figura V.10 confirma esse comportamento: para $ratio \leq 60\%$, os valores de CV, RCV e IQR permanecem baixos e estáveis, ao passo que compressões superiores ($ratio \geq 70\%$) induzem crescimento exponencial dessas métricas, comprometendo a estabilidade e a reprodutibilidade da representação latente entre séries distintas.

Esse comportamento é corroborado pelas métricas de erro absolutas, que confirmaram o ponto de inflexão em $ratio = 60\%$. As Figuras V.11 e V.12 apresentam as distribuições de NRMSE e MAPE em função de *ratio*, respectivamente.

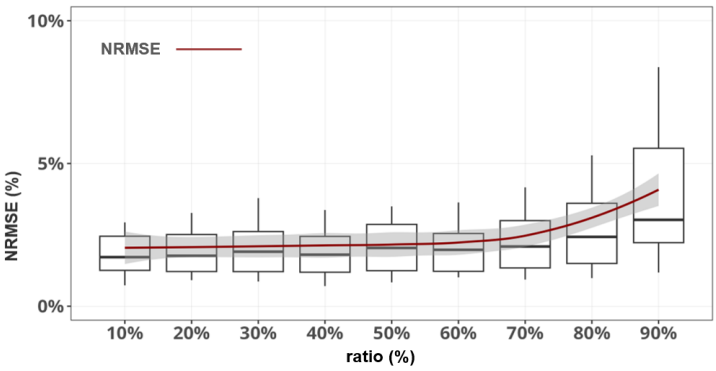


Figura V.11: Distribuição de NRMSE (%) em função de *ratio*, fixando $train_size = 30\%$ e $sw_size = 30$. Cada *boxplot* agrega 40 experimentos (40 séries da derivação V1).

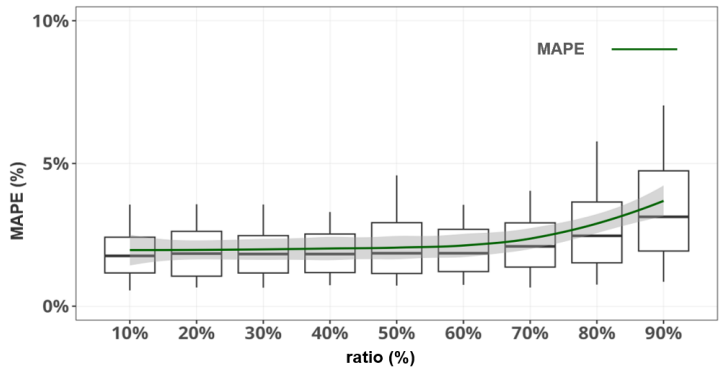


Figura V.12: Distribuição de MAPE (%) em função de *ratio*, fixando $train_size = 30\%$ e $sw_size = 30$. Cada *boxplot* agrega 40 experimentos (40 séries da derivação V1).

O NRMSE manteve-se em 1,98% para $ratio = 60\%$, elevando-se para 3,03% em $ratio = 90\%$. O MAPE seguiu trajetória similar: 1,85% para $ratio = 60\%$, crescendo para 3,13% em $ratio = 90\%$. Tomados em conjunto, os resultados dos três estágios convergem para a configuração $train_size = 30\%$, $sw_size = 30$ e $ratio = 60\%$, resultando em $encoding_size = 12$ dimensões latentes.

A Tabela V.5 consolida a distribuição das três métricas de fidelidade calculadas sobre as 40 séries usando a configuração identificada como de melhor balanço. Esses indicadores validam empiricamente que *autoencoders* com arquitetura adequadamente parametrizada preservam características morfológicas essenciais de séries temporais de ECG: 75% das séries alcançaram $R^2 \geq 99\%$, com mediana de 99,22%; o NRMSE situou-se abaixo de 2% para 75% das séries e o MAPE apresentou mediana de 1,85%, indicando desvios típicos inferiores a 2% dos valores observados.

Tabela V.5: Distribuição agregada das métricas de fidelidade (%) para a configuração de melhor balanço. ($train_size = 30\%$, $sw_size = 30$, $ratio = 60\%$) aplicada às 40 séries da derivação V1.

Métrica	Média	Mediana	Desvio Padrão
R^2	98,45%	99,22%	1,96%
NRMSE	2,19%	1,98%	1,17%
MAPE	2,07%	1,85%	1,13%

Os resultados desta Etapa I demonstraram que a exploração sistemática do espaço de hiperparâmetros identificou configurações que equilibram fidelidade de reconstrução, compressão dimensional e viabilidade computacional. A configuração resultante ($train_size = 30\%$, $sw_size = 30\%$, $ratio = 60\%$, resultando em $encoding_size = 12$) será adotada nas Etapas II para treinamento dos modelos e sustentará as comparações metodológicas da Etapa III. As representações latentes de 12 dimensões preservam mais de 99% da variância das séries originais, com erros de reconstrução tipicamente inferiores a 2%, estabelecendo base sólida para aplicação em descoberta de *motifs* nas etapas subsequentes.

V.3 Etapa II: Treinamento dos Modelos *Autoencoder*

A Etapa II operacionalizou o treinamento de 96 modelos *autoencoder* independentes — um para cada série do MIT-BIH *Arrhythmia Database* — aplicando exclusivamente a configuração de hiperparâmetros identificada como a de melhor balanço estabelecida na exploração sistemática da Etapa I ($train_size = 30\%$, $sw_size = 30$, $ratio = 60\%$, resultando em $encoding_size = 12$ dimensões latentes).

O protocolo de treinamento, detalhado na Seção IV.4, foi aplicado sequencialmente às 96 séries. Cada modelo convergiu de forma independente, com os pesos neurais otimizados até estabilização da função de perda. Ao término de cada execução, os artefatos foram persistidos em formato dual: arquivo `.pth` contendo os pesos finais da rede neural e arquivo `.rds` armazenando os metadados contextuais — parâmetros de normalização *z-score*, histórico de *loss* e configuração arquitetural.

Os 96 modelos assim produzidos constituem artefatos estáveis que viabilizam a execução eficiente da Etapa III, eliminando a necessidade de re-treinamento a cada configuração de descoberta. Essa separação entre fase de treinamento (*offline*) e fase de detecção (*online*) reflete um *workflow* realista para aplicações práticas com grandes volumes de dados: o custo fixo de treinamento é amortizado sobre as 1.728 execuções comparativas da Etapa III, em que cada modelo é carregado e aplicado repetidamente sob variações de $alphabet_size$ e $occurrences$. Essa estratégia assegura que a comparação metodológica entre o LEMOD e o SAX na Etapa III ocorra sob condições experimentais rigorosamente controladas, isolando o efeito da representação latente como única variável manipulada.

É importante destacar o contraste arquitetural com o método SAX: por ser determinístico e baseado em estatísticas estimadas diretamente dos dados de entrada, o SAX não requer fase de treinamento análoga — sua operação equivalente ao *fit* (estimação de média e desvio padrão para normalização *z-score*) é executada de forma integrada à fase de detecção na Etapa III. Essa assimetria não constitui desvantagem do LEMOD, mas sim a manifestação do custo de aprendizado adaptativo: o treinamento *offline* é o mecanismo pelo qual o modelo aprende representações morfológicas específicas das séries, conferindo a sensibilidade superior a padrões sutis observada nos resultados da Etapa III (Seção V.4).

V.4 Etapa III: Comparação Experimental Sistemática

A Etapa III constitui o núcleo avaliativo desta investigação, operacionalizando a comparação sistemática entre o LEMOD — que substitui a discretização estatística do SAX por codificação neural via *autoencoder* — e a abordagem de referência baseada exclusivamente em SAX. Esta seção documenta os resultados obtidos em três experimentos complementares: (i) detecção sob

configuração principal dos parâmetros de descoberta (*alphabet_size* e *occurrences*), (*ii*) análise de sensibilidade paramétrica em duas dimensões ortogonais, totalizando 1728 execuções de detecção; e (*iii*) experimento de detecção cruzada (*cross-detection*), avaliando a capacidade de generalização das representações latentes aprendidas quando modelos treinados em séries específicas são aplicados a séries distintas da mesma derivação. As métricas de avaliação abrangem desempenho classificatório (acurácia, precisão, revocação e F_1), complexidade informacional (entropia de *Shannon* dos *motifs* identificados) e eficiência computacional (tempo de execução da fase de detecção por série).

A configuração principal adotou $alphabet_size = 10$ e $occurrences = 6$. O tamanho da janela foi fixado em $sw_size = 30$ para ambos os métodos, e o $word_size$ do SAX foi igualado ao $encoding_size = 12$ do *autoencoder*, assegurando comparabilidade estrutural das representações simbólicas produzidas. A avaliação empregou as métricas SoftED [Salles et al., 2024] com tolerância temporal $k = 100$ amostras — equivalente a aproximadamente 278 ms na taxa de amostragem de 360 Hz do MIT-BIH — valor escolhido de modo a acomodar, de forma conservadora, a variabilidade fisiológica natural no tempo de ocorrência dos eventos cardíacos, que não se repetem com precisão absoluta entre batimentos sucessivos [Moody and Mark, 2001].

Configuração principal

A Tabela V.6 consolida as métricas de desempenho agregadas sobre as 96 séries para ambos os métodos sob a configuração principal.

Tabela V.6: Métricas de desempenho agregadas para configuração principal ($alphabet_size = 10$, $occurrences = 6$, tolerância $k = 100$). N=96 séries. Tempo médio de detecção por série. A cor atribuída aos valores da última linha "Diferença"reflete o método de melhor performance.

Método	Acurácia	Precisão	Revocação	F_1	Entropia (bits)	Tempo (s)
LEMOD	95,2%	95,8%	9,8%	17,3%	0,44	39,5
SAX	93,6%	93,3%	5,1%	9,7%	0,90	46,3
Diferença	+1,6%	+2,5%	+4,7%	+7,6%	-0,46	-6,8

O LEMOD alcançou acurácia média de 95,2%, superando o SAX (93,6%). A diferença de 1,6 pontos percentuais, embora modesta, indica vantagem consistente da codificação neural na capacidade de discriminação entre observações típicas e *motifs* verdadeiros. A precisão média foi de 95,8% para o LEMOD contra 93,3% para o SAX, refletindo menor taxa de falsos positivos na abordagem proposta: dos *motifs* identificados pelo LEMOD, 95,8% correspondem efetivamente a padrões recorrentes verdadeiros, comparativamente a 93,3% no método tradicional.

A revocação apresentou diferença acentuada: 9,8% para o LEMOD versus 5,1% para o SAX, representando ganho relativo de 92%. Essa diferença indica que a codificação neural identifica praticamente o dobro de *motifs* verdadeiros em comparação com a discretização tradicional, evi-

denciando maior sensibilidade a padrões morfologicamente sutis que a representação simbólica fixa do SAX não captura adequadamente. A medida F_1, que harmoniza precisão e revocação via média harmônica, consolidou a superioridade do método proposto: 17,3% contra 9,7%, diferença de 7,6% que reflete simultaneamente menor taxa de falsos positivos e maior taxa de verdadeiros positivos.

A Figura V.13 apresenta visualização comparativa das quatro métricas de desempenho classificatório. A superioridade do LEMOD em acurácia e precisão é mais discreta, enquanto a diferença em revocação e F_1 é mais acentuada, evidenciando que a principal vantagem da codificação neural reside na capacidade de identificar *motifs* verdadeiros sem comprometer significativamente a especificidade. Essa característica é particularmente relevante em aplicações biomédicas, onde a sensibilidade diagnóstica possui maior valor clínico do que a especificidade absoluta.

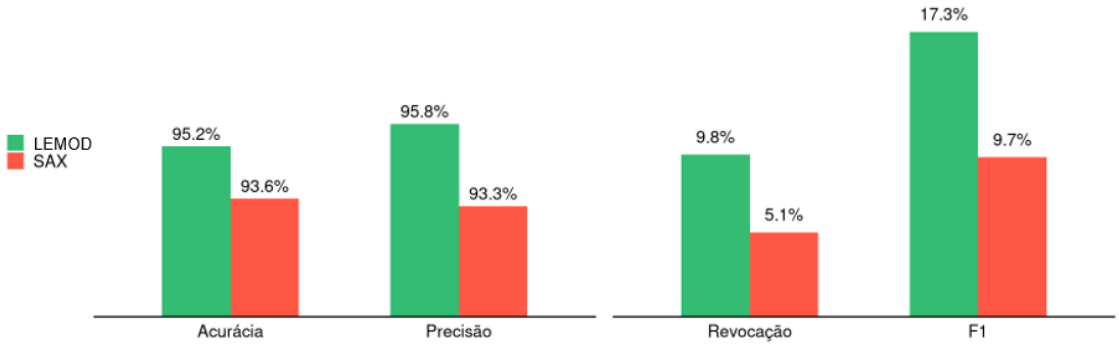


Figura V.13: Comparação das métricas de desempenho (acurácia, precisão, revocação e F_1) entre o LEMOD (verde) e SAX (vermelho). Valores médios agregados sobre as 96 séries.

A entropia média dos *motifs* identificados apresentou comportamento inverso: 0,44 bits para o LEMOD versus 0,90 bits para o SAX. A entropia de *Shannon*, calculada sobre a distribuição de símbolos componentes de cada padrão, quantifica a diversidade simbólica interna dos *motifs* descobertos: valores menores indicam padrões morfologicamente mais homogêneos, com distribuição de símbolos mais concentrada, enquanto valores maiores refletem maior variabilidade interna. O resultado obtido — entropia 51% menor para o LEMOD — indica que a codificação neural tende a identificar padrões estruturalmente mais coerentes e específicos, enquanto a discretização tradicional do SAX tende a agrupar subsequências morfologicamente heterogêneas sob o mesmo *motif*, comprometendo a precisão morfológica da descoberta. Essa maior homogeneidade é, portanto, interpretada como vantagem qualitativa do LEMOD: os *motifs* descobertos são mais representativos de estruturas recorrentes genuínas.

O tempo médio de execução da fase de detecção por série foi de 39,5 segundos para o LEMOD contra 46,3 segundos para o SAX, representando redução de aproximadamente 15%. Essa vantagem decorre da eficiência da transformação neural pré-treinada: uma vez carregados os pesos do modelo, a codificação das janelas temporais em representações latentes é uma operação matricial vetorizada

de alta eficiência, enquanto o SAX requer cálculos iterativos de agregação PAA e discretização por *breakpoints* para cada janela. Considerando as 96 séries, o tempo total de detecção foi de aproximadamente 63 minutos para o LEMOD e 74 minutos para o SAX, consolidando a viabilidade computacional da abordagem proposta. A Figura V.14 apresenta comparativamente as entropias médias e os tempos de detecção dos dois métodos, evidenciando que o LEMOD é simultaneamente mais eficiente em tempo de execução e identifica *motifs* de maior homogeneidade morfológica.

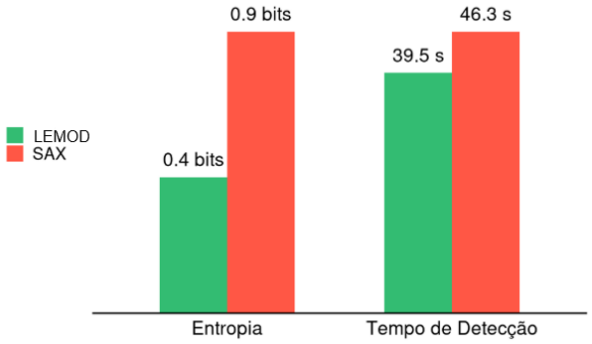


Figura V.14: Entropia de *Shannon* (bits) dos *motifs* identificados e o tempo medido na detecção para ambos os métodos: LEMOD (verde) e SAX (vermelho).

Para ilustrar qualitativamente o comportamento diferencial dos métodos experimentados, a Figura V.15 apresenta as detecções realizadas por ambos em segmento de ritmo sinusal regular da série `r100_MLII`.

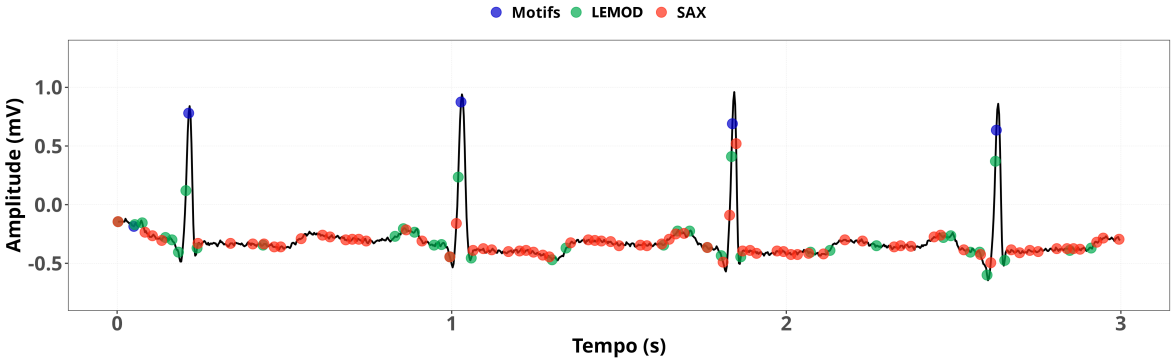


Figura V.15: Exemplo de detecções em segmento de ritmo sinusal regular da série `r100_MLII` (intervalo 0-3.5 s). Círculos azuis: *motifs* verdadeiros (anotações de referência). Marcações verdes: detecções do LEMOD. Marcações vermelhas: detecções do SAX. Configuração: $alphabet_size = 10$, $occurrences = 6$.

Em ritmo regular, o SAX produz detecções densas ao longo da linha de base, refletindo alta recorrência morfológica de segmentos de baixa variabilidade, enquanto o LEMOD concentra detecções nos batimentos regulares, demonstrando maior seletividade. A Figura V.16 apresenta segmento correspondente da mesma série em contexto arrítmico, permitindo contrastar o comportamento dos métodos sob condições morfológicas distintas.

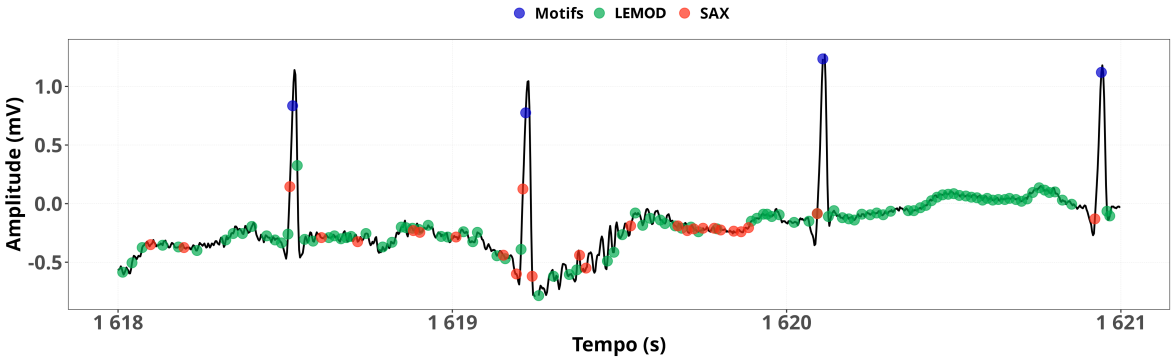


Figura V.16: Exemplo de detecções em segmento com arritmia da série `r100_MLII` (intervalo 1618-1621 s). Círculos azuis: *motifs* verdadeiros (anotações de referência). Marcações verdes: detecções do LEMOD. Marcações vermelhas: detecções do SAX. Configuração: $alphabet_size = 10$, $occurrences = 6$.

Em contexto arrítmico, o LEMOD amplifica a densidade de detecções, capturando padrões recorrentes sutis associados à variabilidade morfológica, enquanto o SAX apresenta distribuição mais esparsa. Esse comportamento diferencial sugere que a codificação neural oferece maior sensibilidade a contextos patológicos, potencialmente relevante para aplicações diagnósticas que privilegiam a detecção de eventos anômalos.

Análise de sensibilidade paramétrica

Para avaliar a robustez das conclusões frente a variações nos parâmetros de descoberta, ambos os métodos foram submetidos à análise de sensibilidade paramétrica em duas dimensões ortogonais. A primeira dimensão explorou o tamanho do alfabeto simbólico, fixando $occurrences = 6$ e variando $alphabet_size \in \{10, 14, 18, 22, 26\}$, totalizando 960 execuções (5 configurações $\times$ 96 séries $\times$ 2 métodos). A segunda dimensão explorou o limiar de recorrência, fixando $alphabet_size = 10$ e variando $occurrences \in \{2, 4, 6, 8, 10\}$, totalizando 960 execuções adicionais. A Etapa III somou 1728 experimentos de detecção comparativa.

Dimensão 1 — variação de *alphabet_size*:

A Figura V.17 apresenta a evolução da acurácia em função de *alphabet_size*.

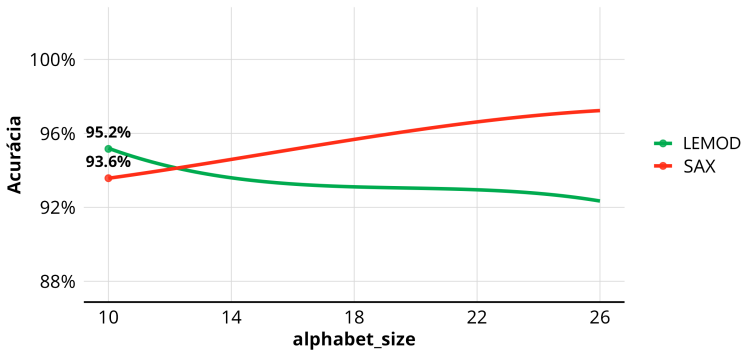


Figura V.17: Acurácia em função de *alphabet_size*, fixando $occurrences = 6$.

O LEMOD mantém acurácia decrescente entre 95% e 92% ao longo de toda a faixa testada, enquanto o SAX apresenta crescimento gradual de 93,6% ($alphabet_size = 10$) para aproximadamente 97,5% ($alphabet_size = 26$), sugerindo que a granularidade simbólica crescente beneficia o método tradicional na discriminação global. A Figura V.18 revela o mecanismo subjacente a esse comportamento.

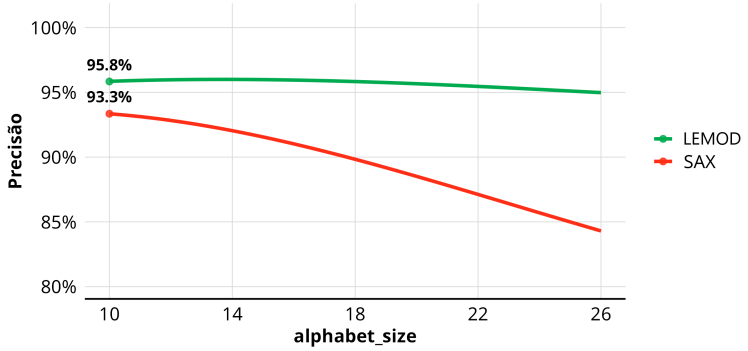


Figura V.18: Precisão em função de $alphabet_size$, fixando $occurrences = 6$.

A precisão do LEMOD permanece estável em aproximadamente 95,8%, enquanto o SAX apresenta decrescimento acentuado de 93,3% para aproximadamente 84% em $alphabet_size = 26$, indicando que o aumento de granularidade amplifica a taxa de falsos positivos no método tradicional. Essa estabilidade da precisão do LEMOD evidencia que a codificação neural mantém especificidade independentemente do tamanho do alfabeto, consolidando vantagem metodológica à medida que a granularidade aumenta.A Figura V.19 complementa essa análise com a evolução da revocação.

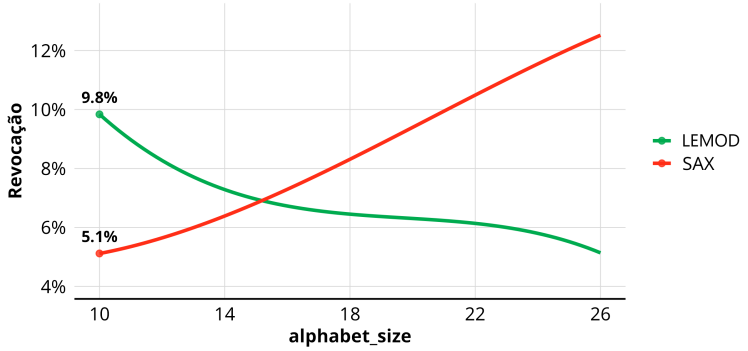


Figura V.19: Revocação em função de $alphabet_size$, fixando $occurrences = 6$.

O LEMOD exibe decrescimento gradual de 9,8% para aproximadamente 5% ($alphabet_size = 26$), enquanto o SAX apresenta crescimento substancial de 5,1% para aproximadamente 12,5% ($alphabet_size = 26$), com as curvas cruzando-se em $alphabet_size \approx 15$. Esse comportamento deve ser interpretado em conjunto com a degradação de precisão documentada: o SAX identifica mais *motifs* verdadeiros para alfabetos maiores, mas ao custo de identificar também um número muito superior de falsos positivos. A Figura V.20 sintetiza esse *trade-off* via F_1.

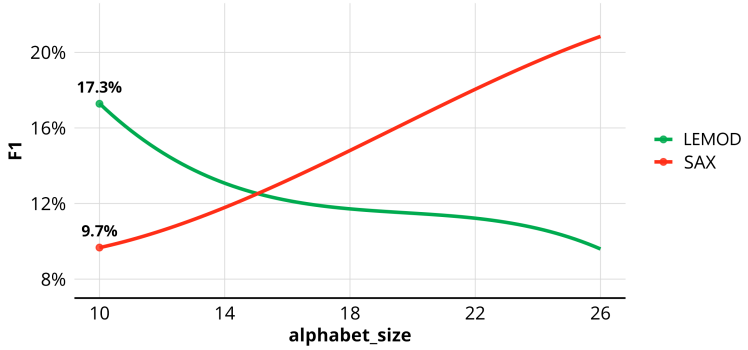


Figura V.20: F_1 em função do tamanho de $alphabet_size$, fixando $occurrences = 6$.

O LEMOD decresce de 17,3% para aproximadamente 9,5%, enquanto o SAX cresce de 9,7% para aproximadamente 20%, superando o LEMOD para $alphabet_size \geq 15$. Esse resultado deve ser interpretado com cautela: o ganho nominal de F_1 do SAX em alfabetos grandes decorre primariamente de aumento de revocação acompanhado por degradação de precisão de 93% para 84%, comprometendo utilidade prática em contextos biomédicos sensíveis a falsos positivos. A configuração principal ($alphabet_size = 10$) situa-se no regime em que o LEMOD mantém superioridade em F_1, sustentando a escolha adotada.

Dimensão 2 — variação de *occurrences*:

A Figura V.21 apresenta a evolução da acurácia em função de *occurrences*, avaliando se o critério de seletividade afeta a capacidade global de discriminação entre observações típicas e *motifs* verdadeiros.

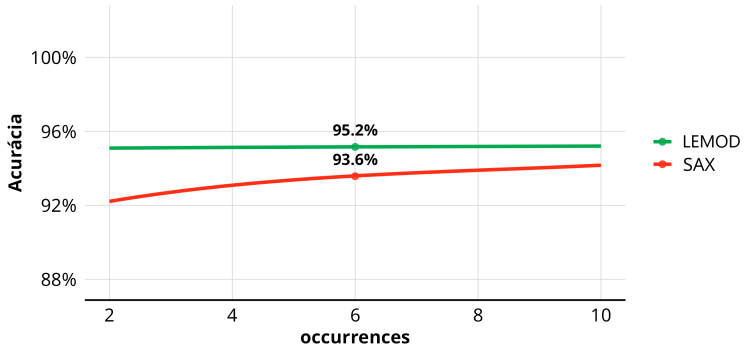


Figura V.21: Acurácia em função de *occurrences*, fixando $alphabet_size = 10$.

Ambos os métodos exibem comportamento estável, com o LEMOD mantendo-se em aproximadamente 95,2% e o SAX em 93,6%. Esse padrão confirma que a vantagem marginal do LEMOD em discriminação global é robusta frente a variações em *occurrences*, persistindo independentemente do critério de seletividade adotado. A Figura V.22 apresenta a evolução da precisão em função de *occurrences*, verificando se este limiar afeta a especificidade dos métodos na identificação de *motifs* verdadeiros.

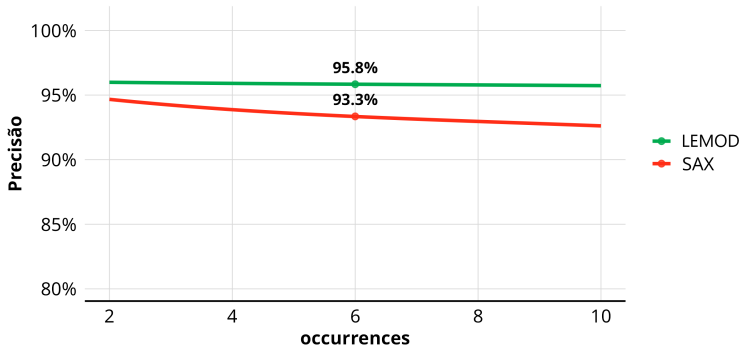


Figura V.22: Precisão em função de *occurrences*, fixando $alphabet_size = 10$.

Os métodos apresentam precisão estável ao longo da faixa testada, com o LEMOD preservando vantagem marginal consistente de 2 a 3 pontos percentuais, indicando menor taxa de falsos positivos independentemente do critério de seletividade. A Figura V.23 apresenta a evolução da revocação.

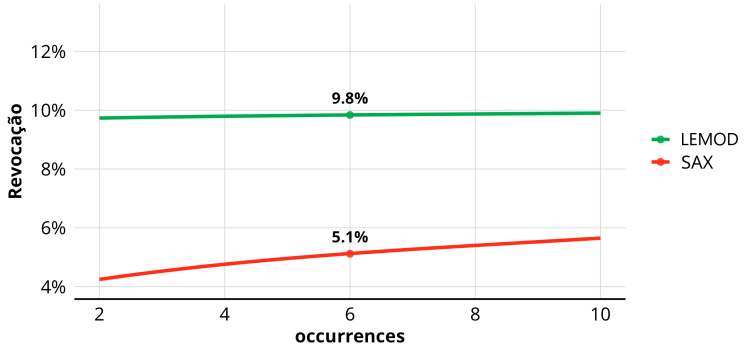


Figura V.23: Revocação em função de *occurrences*, fixando $alphabet_size = 10$.

O LEMOD mantém revocação estável em aproximadamente 9,8%, enquanto o SAX apresenta crescimento gradual de 4,2% ($occurrences = 2$) para 5,8% ($occurrences = 10$), com superioridade do LEMOD de 4 a 5 pontos. A Figura V.24 sintetiza o desempenho comparativo via F_1.

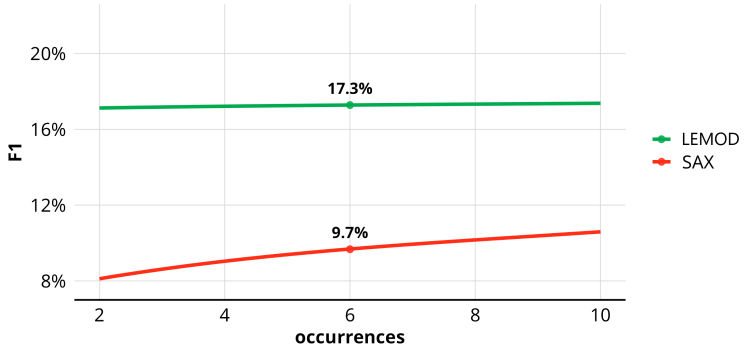


Figura V.24: F_1 em função de *occurrences*, fixando $alphabet_size = 10$.

O LEMOD preserva superioridade consistente em F_1 de 6 a 8 pontos percentuais ao longo de toda a faixa, validando a robustez da vantagem metodológica frente a diferentes critérios de seletividade.

Em conjunto, as duas dimensões da análise de sensibilidade demonstram que a vantagem do LEMOD sobre o SAX não é restrita à configuração principal, mas constitui propriedade estável do método: a codificação neural preserva equilíbrio superior entre especificidade e sensibilidade independentemente das escolhas de *alphabet_size* e *occurrences*, desde que se opere no regime em que a precisão não é sacrificada em favor de revocação. A investigação complementar a seguir examina uma dimensão adicional da robustez do método: sua capacidade de generalização.

Experimento de detecção cruzada

O experimento de detecção cruzada (*cross-detection*) investigou a capacidade de generalização das representações latentes, testando se modelos treinados em séries específica mantêm desempenho adequado quando aplicado a séries distintas da mesma derivação. Esse cenário possui relevância metodológica ao verificar se as representações latentes capturam estruturas morfológicas transferíveis, bem como relevância prática, ao viabilizar aplicações em que seria impraticável treinar um modelo específico para cada nova série analisada.

O protocolo selecionou uma série de referência por tipo de derivação disponível no MIT-BIH *Arrhythmia Database*: `r100_MLII` para MLII, `r101_V1` para V1, `r102_V2` para V2 e `r104_V5` para V5. A derivação V4 não foi testada por possuir apenas uma série no banco de dados (`r124_V4`), impossibilitando avaliação de transferência. Para cada derivação, o modelo treinado sobre a série de referência foi aplicado às séries remanescentes da mesma derivação com os parâmetros da configuração principal ($alphabet_size = 10$, $occurrences = 6$). A Tabela V.7 consolida os resultados estratificados por derivação.

Tabela V.7: Métricas de desempenho em detecção cruzada estratificadas por derivação. Modelo de referência aplicado a séries distintas da mesma derivação. Configuração: $alphabet_size = 10$, $occurrences = 6$, tolerância $k = 100$.

Lead	N séries-alvo	Método	Acurácia	Precisão	Revocação	F_1
MLII	45	LEMOD cross-detection	**95,7%**	86,6%	**6,8%**	**12,6%**
		SAX	93,2%	**94,1%**	4,8%	9,2%
V1	39	LEMOD cross-detection	**98,6%**	48,4%	**12,7%**	**20,0%**
		SAX	93,9%	**92,8%**	5,7%	10,6%
V2	3	LEMOD cross-detection	**95,8%**	79,6%	**5,5%**	**10,3%**
		SAX	93,3%	**94,2%**	4,1%	7,9%
V5	4	LEMOD cross-detection	91,5%	92,2%	3,2%	6,1%
		SAX	**93,2%**	**94,0%**	**4,1%**	**7,8%**

Para a derivação MLII, o modelo `r100_MLII` aplicado a 45 séries-alvo alcançou acurácia de 95,7%, superando o SAX (93,2%). A precisão foi de 86,6%, inferior aos 94,1% do SAX, indicando taxa

mais elevada de falsos positivos no cenário de generalização. Contudo, a revocação de 6,8% (contra 4,8% do SAX) e o F_1 de 12,6% (contra 9,2%) demonstram que, mesmo aplicando um único modelo a 45 séries distintas, o LEMOD mantém capacidade superior de identificar *motifs* verdadeiros—evidência de transferibilidade das representações latentes dentro da mesma derivação.

Para a derivação V1, o modelo `r101_V1` aplicado a 39 séries-alvo apresentou acurácia de 98,6%, a mais elevada entre as derivações testadas, com revocação de 12,7% (substancialmente superior aos 5,7% do SAX) e F_1 de 20,0% (praticamente o dobro dos 10,6% do método tradicional. Contudo, a precisão foi de apenas 48,4%, indicando taxa elevada de falsos positivos. Esse *trade-off* — alta revocação à custa de precisão reduzida — captura padrões morfológicos sutis com alta sensibilidade, identificando genuínas recorrências, mas também gerando detecções espúrias em subsequências que compartilham características latentes similares sem serem efetivamente recorrentes. A superioridade do F_1 (20,0% *versus* 10,6%) indica que, nesse cenário, o ganho em revocação supera a perda de precisão, posicionando a derivação V1 como contexto particularmente favorável à generalização de representações latentes.

As derivações V2 e V5 apresentam limitações amostrais (3 e 4 séries-alvo, respectivamente), restringindo inferências estatísticas robustas. Para V2, o LEMOD preserva vantagem em F_1 (10,3% *versus* 7,9%). Para V5, observou-se inversão: o SAX alcançou F_1 de 7,8% contra 6,1% do LEMOD, sendo este o único contexto em que o método tradicional superou a codificação neural. Esse resultado pode refletir características morfológicas particulares da derivação V5 não adequadamente capturadas pelo modelo treinado na série $r104_V5$, ou instabilidade decorrente do tamanho amostral reduzido.

Os resultados da Etapa III demonstram que a substituição da discretização estatística do SAX por codificação neural adaptativa via *autoencoder* proporciona ganhos consistentes em desempenho classificatório, particularmente em sensibilidade (revocação), sem comprometer substancialmente a especificidade (precisão). O ganho relativo de 92% em revocação sob a configuração principal (9,8% *versus* 5,1%) sustenta que representações latentes capturam padrões morfológicos sutis que a discretização linear do SAX não identifica adequadamente. A análise de sensibilidade paramétrica evidencia que o LEMOD preserva estabilidade de precisão e equilíbrio superior entre especificidade e sensibilidade ao longo de toda a faixa de $alphabet_size$ testada, ainda que o SAX supere o LEMOD em F_1 para alfabetos grandes, fazendo-o via degradação de precisão clinicamente problemática. O experimento de detecção cruzada fornece evidências de generalização das representações latentes dentro da mesma derivação, com desempenho competitivo quando modelos treinados em séries específicas são aplicados a séries distintas. A derivação V1 emergiu como contexto particularmente favorável, com acurácia de 98,6% e F_1 de 20,0% em generalização. A redução de aproximadamente 15% no tempo de detecção por série (39,5s *versus* 46,3s) complementa as vantagens metodológicas

com ganho de eficiência computacional, consolidando a viabilidade operacional da abordagem proposta.

V.5 Síntese do Capítulo

Este capítulo apresentou avaliação experimental sistemática do LEMOD, respondendo à questão de pesquisa central desta investigação: representações latentes aprendidas por *autoencoders* proporcionam desempenho superior na descoberta de *motifs* em séries temporais de ECG quando comparadas à discretização simbólica do SAX.

Os resultados obtidos nas três etapas experimentais permitem responder afirmativamente a essa questão, com as seguintes evidências quantitativas:

- **Etapa I** validou que *autoencoders* adequadamente parametrizados ($train_size = 30\%$, $sw_size = 30$, $ratio = 60\%$, $encoding_size = 12$) preservam mais de 99% da variância dos sinais originais (R^2 mediano de 99,22%), com erros de reconstrução tipicamente inferiores a 2%, confirmando a Predição 1 (Fidelidade Reconstrutiva).

- **Etapa II** operacionalizou o treinamento de 96 modelos independentes, consolidando artefatos estáveis que sustentam a Etapa III sob condições experimentais rigorosamente controladas.

- **Etapa III** demonstrou que o LEMOD alcança ganho relativo de 92% em revocação (9,8% *versus* 5,1% do SAX) e superioridade de 78% em F1 (17,3% *versus* 9,7%), com precisão superior (95,8% *versus* 93,3%) e redução de 15% no tempo de detecção, confirmando a Predição 2 (Eficácia de Detecção).

A Predição 3 (Complexidade Informacional) foi confirmada: a entropia média dos *motifs* identificados pelo LEMOD foi de 0,44 bits, contra 0,90 bits do SAX, evidenciando características informacionais distintas entre os dois métodos. Conforme discutido na Seção V.4, a menor entropia do LEMOD reflete maior homogeneidade morfológica interna dos padrões descobertos — indicativo de que a codificação neural identifica estruturas recorrentes genuínas com maior coerência simbólica, enquanto o SAX tende a agrupar subsequências heterogêneas sob o mesmo *motif*.

A análise de sensibilidade paramétrica e o experimento de detecção cruzada reforçam a robustez dessas conclusões: a vantagem do LEMOD em precisão e equilíbrio entre especificidade e sensibilidade mantém-se ao longo de toda a faixa de *occurrences* testada e na configuração principal de $alphabet_size$, e as representações latentes demonstram capacidade de transferência dentro da mesma derivação, com a derivação V1 emergindo como contexto particularmente favorável ($F1 = 20,0\%$ em generalização *versus* $10,6\%$ do SAX).

Os resultados devem ser contextualizados por limitações que circunscrevem o alcance das conclusões. Primeiro, a sensibilidade ao parâmetro *alphabet_size* revela um *trade-off* complexo: para alfabetos grandes (≥ 15), o SAX supera o LEMOD em F_1, porém via degradação de precisão (de 93% para 84%), o que compromete sua utilidade prática em contextos biomédicos sensíveis a falsos positivos. Segundo, a revocação absoluta de ambos os métodos permanece inferior a 13% mesmo nas configurações mais favoráveis, sugerindo que a descoberta de *motifs* baseada exclusivamente em frequência de ocorrência e entropia pode não capturar todos os padrões clinicamente relevantes, motivando a investigação de critérios complementares de seleção. Terceiro, a validação empírica limitou-se ao MIT-BIH *Arrhythmia Database*, conjunto consolidado, porém datado, cuja população e protocolos de aquisição podem não refletir plenamente a diversidade morfológica de bases contemporâneas de maior escala. Quarto, o experimento de detecção cruzada apresenta limitações amostrais para as derivações V2 e V5 (3 e 4 séries-alvo, respectivamente), restringindo inferências estatísticas robustas para esses contextos.

A combinação de alta fidelidade de reconstrução, convergência estável de treinamento e desempenho superior em detecção estabelece base empírica sólida para as conclusões apresentadas no Capítulo VI. As limitações identificadas apontam direções concretas para trabalhos futuros, discutidas nas considerações finais.

Capítulo VI Conclusão

A descoberta de *motifs* em séries temporais biomédicas constitui desafio metodológico relevante, pois combina complexidade computacional, necessidade de representações compactas e preservação de características morfológicas essenciais para identificação de padrões recorrentes clinicamente significativos. Em sinais de eletrocardiograma, a capacidade de identificar automaticamente estruturas repetitivas oferece potencial para suporte à decisão clínica, detecção precoce de arritmias e descoberta de biomarcadores cardíacos previamente imperceptíveis. O avanço das técnicas de aprendizado profundo nas últimas décadas consolidou arquiteturas neurais como ferramentas capazes de aprender representações não lineares complexas de forma não supervisionada, oferecendo alternativa promissora aos métodos convencionais de análise de séries temporais que dependem de pressupostos estatísticos rígidos sobre as propriedades dos dados.

Métodos tradicionais de descoberta de *motifs*, particularmente aqueles baseados em discretização simbólica como o *Symbolic Aggregate approXimation* (SAX), apresentam limitações metodológicas documentadas na literatura. A imposição de normalidade estatística como premissa fundamental introduz fragilidades: a transformação por *Piecewise Aggregate Approximation* (PAA) altera as propriedades distribucionais dos dados originais, resultando em partições simbólicas com probabilidades desiguais, comprometendo a equiprobabilidade esperada dos símbolos alfabéticos. Em séries temporais de ECG, essas limitações manifestam-se de forma particularmente crítica, uma vez que a morfologia do batimento cardíaco exibe variabilidade que não se conforma necessariamente a distribuições gaussianas, especialmente em contextos patológicos. Discretizações lineares tendem a agrupar padrões morfologicamente distintos sob o mesmo símbolo, enquanto variações sutis, mas clinicamente relevantes, podem ser fragmentadas em categorias diferentes, motivando a investigação de representações alternativas capazes de capturar não linearidades sem pressupor estruturas estatísticas específicas.

Esta pesquisa propôs e validou empiricamente o LEMOD, um método que substitui a discretização estatística do SAX por codificação neural adaptativa via *autoencoder*, preservando a estrutura consolidada de descoberta de *motifs* baseada em frequência de ocorrência e entropia de *Shannon*. A abordagem mantém o *pipeline* estabelecido — segmentação por janelas deslizantes, transformação para representação simbólica compacta, busca de subsequências recorrentes, ranqueamento por conteúdo informacional — substituindo exclusivamente a etapa crítica de discretização linear por

representações latentes aprendidas automaticamente a partir dos dados. Essa integração permite explorar a capacidade dos *autoencoders* de extrair características morfológicas relevantes sem impor suposições estatísticas restritivas, ao mesmo tempo em que preserva a eficiência computacional e a interpretabilidade dos métodos baseados em indexação simbólica.

A validação empírica, documentada em detalhe no Capítulo V, confirmou as três predições formuladas na Seção IV.1. A Predição 1 (Fidelidade Reconstrutiva) foi confirmada: a configuração de melhor balanço identificada ($train_size = 30\%$, $sw_size = 30$, $ratio = 60\%$, $encoding_size = 12$) preserva mais de 99% da variância dos sinais originais com erros de reconstrução tipicamente inferiores a 2%, demonstrando que a compressão dimensional preserva as propriedades morfológicas essenciais dos sinais de ECG. A Predição 2 (Eficácia de Detecção) foi confirmada com destaque: sob a configuração principal ($alphabet_size = 10$, $occurrences = 6$), o LEMOD alcançou ganho relativo de 92% em revocação (9,8% *versus* 5,1% do SAX) e superioridade de 78% em F_1 (17,3% *versus* 9,7%), mantendo precisão superior (95,8% *versus* 93,3%) e reduzindo o tempo de detecção em aproximadamente 15% (39,5 s *versus* 46,3 s por série). A Predição 3 (Complexidade Informacional) foi confirmada: a entropia média dos *motifs* identificados pelo LEMOD foi de 0,44 bits contra 0,90 bits do SAX, evidenciando que a codificação neural identifica padrões morfologicamente mais homogêneos e coerentes, enquanto a discretização tradicional tende a agrupar subsequências heterogêneas sob o mesmo *motif*. A análise de sensibilidade paramétrica e o experimento de detecção cruzada reforçaram a robustez dessas conclusões: a vantagem do LEMOD em precisão e equilíbrio entre especificidade e sensibilidade mantém-se ao longo de toda a faixa de *occurrences* testada e na configuração principal de $alphabet_size$, e as representações latentes demonstraram capacidade de transferência dentro da mesma derivação, com destaque para a derivação V1 ($F_1 = 20,0\%$ em generalização *versus* 10,6% do SAX).

As contribuições científicas desta pesquisa abrangem dimensões metodológica, empírica e prática no campo da análise de séries temporais biomédicas. Do ponto de vista metodológico, o trabalho estabelece protocolo para integração de *autoencoders* em *frameworks* de descoberta de *motifs*, demonstrando que representações latentes não lineares superam discretizações estatísticas lineares na preservação de características morfológicas essenciais, oferecendo alternativa a métodos que dependem de suposições rígidas sobre as propriedades dos dados. A identificação da melhor configuração de hiperparâmetros ($train_size = 30\%$, $sw_size = 30$, $ratio = 60\%$) por exploração sistemática de 16200 experimentos fornece ponto de partida reproduzível para investigações futuras, enquanto a validação de alta fidelidade de reconstrução ($R^2 > 99\%$) estabelece critério de qualidade para aplicações da arquitetura proposta. Do ponto de vista empírico, a validação sobre 96 séries do MIT-BIH *Arrhythmia Database* por meio de 1728 experimentos de detecção, incluindo análise de sensibilidade paramétrica e avaliação de capacidade de generalização, indica superioridade consistente do método

proposto em revocação e F_1, com ganhos em sensibilidade a padrões morfológicos sutis. A demonstração de viabilidade computacional e as evidências de generalização (modelos únicos por derivação mantêm desempenho competitivo) consolidam as contribuições práticas, oferecendo base para desenvolvimento de ferramentas de análise automatizada de sinais de ECG com potencial aplicação em sistemas de monitoramento contínuo e triagem diagnóstica.

As conclusões devem ser contextualizadas por quatro limitações que circunscrevem o alcance dos resultados: (*i*) sensibilidade ao parâmetro *alphabet_size*, cujo aumento favorece o SAX em F_1 via degradação de precisão de 93% para 84%, comprometendo utilidade prática em contextos biomédicos sensíveis a falsos positivos; (*ii*) revocação absoluta inferior a 13% em ambos os métodos, sugerindo que a descoberta de *motifs* baseada exclusivamente em frequência de ocorrência e entropia pode não capturar todos os padrões clinicamente relevantes, motivando a investigação de critérios complementares de seleção; (*iii*) validação empírica restrita ao MIT-BIH, conjunto consolidado, porém datado, cuja população e protocolos de aquisição podem não refletir plenamente a diversidade morfológica de bases contemporâneas de maior escala; e (*iv*) limitações amostrais no experimento de detecção cruzada para as derivações V2 e V5 (3 e 4 séries-alvo, respectivamente), restringindo inferências estatísticas robustas nesses contextos.

Trabalhos futuros podem explorar múltiplas direções de aprofundamento e extensão desta pesquisa. A investigação de arquiteturas alternativas de *autoencoders*, como *Variational Autoencoders* (VAE), que incorporam estrutura probabilística no espaço latente, ou arquiteturas baseadas em mecanismos de atenção (*transformer-based autoencoders*), que capturam dependências temporais de longo alcance, pode oferecer ganhos adicionais de qualidade na representação de padrões cardíacos complexos. A otimização conjunta de hiperparâmetros por técnicas de busca mais sofisticadas, como otimização bayesiana ou algoritmos evolutivos, pode identificar configurações superiores às obtidas por busca em grade, particularmente ao explorar interações não lineares entre dimensões paramétricas. A validação em bancos de dados adicionais e contemporâneos, como PTB-XL[1] [Wagner et al., 2020] ou INCART 12-Lead[2], permitiria avaliar robustez da abordagem frente a protocolos de aquisição distintos, populações diversificadas e condições patológicas mais amplas. A extensão da metodologia a outros sinais biomédicos, como eletroencefalograma (EEG) para detecção de padrões epileptiformes, permitiria avaliar a generalidade da estratégia de codificação neural adaptativa além do domínio cardiovascular. A integração com sistemas de suporte à decisão clínica, desenvolvendo interfaces que traduzam *motifs* identificados em evidências interpretáveis por profissionais de saúde, pode consolidar aplicabilidade prática da pesquisa. Por fim, a análise de interpretabilidade do espaço latente aprendido, investigando se dimensões específicas capturam características morfológicas conhecidas, pode oferecer evidências sobre os mecanismos pelos quais a codificação neural supera

[1]Disponível em https://physionet.org/content/ptb-xl/1.0.3/
[2]Disponível em https://physionet.org/content/incartdb/1.0.0/

a discretização tradicional, informando o desenvolvimento de arquiteturas mais especializadas para sinais cardíacos.

Os resultados consolidados nesta dissertação sustentam a viabilidade metodológica e prática da substituição de técnicas convencionais de discretização por codificação neural adaptativa na descoberta de *motifs* em séries temporais de ECG. A demonstração de alta fidelidade de reconstrução (Etapa I), convergência estável de treinamento (Etapa II) e desempenho superior em detecção (Etapa III) confirma que representações latentes aprendidas por *autoencoders* capturam características morfológicas essenciais com qualidade superior à obtida por métodos baseados em suposições estatísticas rígidas. A contribuição central desta investigação reside em demonstrar empiricamente que técnicas de aprendizado profundo, tradicionalmente empregadas em tarefas supervisionadas de classificação ou detecção de anomalias, podem ser integradas de forma eficaz a *frameworks* não supervisionados de mineração de padrões, oferecendo ganhos mensuráveis em sensibilidade e especificidade. Os achados estabelecem base metodológica para o desenvolvimento de ferramentas de análise automatizada de sinais biomédicos, com potencial para impactar sistemas de monitoramento cardiovascular contínuo, triagem diagnóstica em larga escala e descoberta de biomarcadores cardíacos em cardiologia digital.

Referências

Aggarwal, C. C. *Outlier Analysis*. Springer International Publishing, Cham, 2017.

Alaee, S., Kamgar, K., and Keogh, E. Matrix profile XXII: Exact discovery of time series motifs under DTW. In *Proc. IEEE Int. Conf. Data Min. ICDM*, volume 2020-November, pages 900–905. Institute of Electrical and Electronics Engineers Inc., 2020. https://doi.org/10.1109/ICDM50108.2020.00099

Alaee, S., Mercer, R., Kamgar, K., and Keogh, E. Time series motifs discovery under DTW allows more robust discovery of conserved structure. *Data Mining and Knowledge Discovery*, 35(3):863 – 910, 2021. https://doi.org/10.1007/s10618-021-00740-0

An, J. and Cho, S. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. *Special lecture on IE*, 2(1):1–18, 2015.

Bascol, K., Emonet, R., Fromont, E., and Odobez, J.-M. Unsupervised Interpretable Pattern Discovery in Time Series Using Autoencoders. In Robles-Kelly, A., Loog, M., Biggio, B., Escolano, F., and Wilson, R., editors, *Structural, Syntactic, and Statistical Pattern Recognition*, pages 427–438, Cham. Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-49055-7_38

Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., and Xu, Y. Autoencoders and their applications in machine learning: a survey. *Artificial Intelligence Review*, 57(2):28, 2024. https://doi.org/10.1007/s10462-023-10662-6

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer New York, 2006.

Borges, H., Dutra, M., Bazaz, A., Coutinho, R., Perosi, F., Porto, F., Masseglia, F., Pacitti, E., and Ogasawara, E. Spatial-time motifs discovery. *Intelligent Data Analysis*, 24(5):1121–1140, 2020. https://doi.org/10.3233/IDA-194759

Buhler, J. and Tompa, M. Finding motifs using random projections. In *Proceedings of the fifth annual international conference on Computational biology*, volume 9, pages 69–76, Montreal Quebec Canada. ACM, 2001. https://doi.org/10.1145/369133.369172

Butler, M. and Kazakov, D. SAX Discretization Does Not Guarantee Equiprobable Symbols. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):1162–1166, 2015. https://doi.org/10.1109/TKDE.2014.2382882

Castro, N. and Azevedo, P. Multiresolution Motif Discovery in Time Series. In *Proceedings of the 2010 SIAM International Conference on Data Mining (SDM)*, Proceedings, pages 665–676. Society for Industrial and Applied Mathematics, 2010. https://doi.org/10.1137/1.9781611972801.73

Castro, N. and Azevedo, P. J. Time Series Motifs Statistical Significance. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 687–698. Society for Industrial and Applied Mathematics, 2011. https://doi.org/10.1137/1.9781611972818.59

Castro, N. C. and Azevedo, P. J. Significant motifs in time series. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(1):35–53, 2012. https://doi.org/10.1002/sam.11134

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 2009. https://doi.org/10.1145/1541880.1541882

Chicco, D., Warrens, M. J., and Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7:1 – 24, 2021. https://doi.org/10.7717/PEERJ-CS.623

Chiu, B., Keogh, E., and Lonardi, S. Probabilistic discovery of time series motifs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 493 – 498, 2003. https://doi.org/10.1145/956750.956808

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2012.

Dasan, E. and Panneerselvam, I. A novel dimensionality reduction approach for ECG signal via convolutional denoising autoencoder with LSTM. *Biomedical Signal Processing and Control*, 63:102225, 2021. https://doi.org/10.1016/j.bspc.2020.102225

Ding, C., Wang, S., Jin, X., Wang, Z., and Wang, J. A novel transformer-based ECG dimensionality reduction stacked auto-encoders for arrhythmia beat detection. *Medical Physics*, 50(9):5897–5912, 2023. https://doi.org/10.1002/mp.16534

Dogan, R. O., Dogan, H., and Kayikcioglu, T. An Efficient 1D Autoencoder-Based Approach for R-Peaks Detection in Electrocardiogram Signals. In *2023 10th International Conference on Electrical and Electronics Engineering (ICEEE)*, pages 106–110. Institute of Electrical and Electronics Engineers Inc., 2023. https://doi.org/10.1109/ICEEE59925.2023.00027

Esling, P. and Agon, C. Time-series data mining. *ACM Comput. Surv.*, 45(1):12:1–12:34, 2012. https://doi.org/10.1145/2379776.2379788

Everitt, B. S. and Skrondal, A. *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge, UK, 3rd edition, 2010.

Fodil, N., Olivier, D., and Tranouez, P. Towards Better Motif Detection: Comparative Analysis of Several Symbolic Methods:. In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods*, pages 311–319, Rome, Italy. SCITEPRESS - Science and Technology Publications, 2024. https://doi.org/10.5220/0012368500003654

Fu, T.-c. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011. https://doi.org/10.1016/j.engappai.2010.09.007

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., and Stanley, H. E. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. https://doi.org/10.1161/01.CIR.101.23.e215

Goldstein, M. and Uchida, S. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE*, 11(4):e0152173, 2016. https://doi.org/10.1371/journal.pone.0152173

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Guerra, R. d. T., Yamaguchi, C. K., Stefenon, S. F., Coelho, L. d. S., and Mariani, V. C. Deep Learning Approach for Automatic Heartbeat Classification. *Sensors*, 25(5):1400, 2025. https://doi.org/10.3390/s25051400

Guerrmi, V., Germain, T., Truong, C., Oudre, L., and Boniol, P. Time Series Motif Discovery: A Comprehensive Evaluation. *Proc. VLDB Endow.*, 18(7):2226–2239, 2025. https://doi.org/10.14778/3734839.3734857

Han, J., Pei, J., and Tong, H. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Cambridge, MA, 4th edition edition, 2022.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, 2006. https://doi.org/10.1126/science.1127647

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021.

Keogh, E. Finding Repeated Structures in Time Series: Algorithms and Applications: A Unifying View of Time Series Motifs/Shapelets/Chains and Similar Primitives. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 6054–6062, Toronto ON Canada. ACM, 2025. https://doi.org/10.1145/3711896.3736551

Keogh, E. and Kasetty, S. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003. https://doi.org/10.1023/A:1024988512476

Keogh, E. and Lin, J. Clustering of time-series subsequences is meaningless: Implications for previous and future research. *Knowledge and Information Systems*, 8(2):154 – 177, 2005. https://doi.org/10.1007/s10115-004-0172-7

Liang, M., Jiang, A., Liu, X., Kwan, H. K., and Zhu, Y. Arrhythmia classification algorithm based on sparse autoencoder. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1333–1337, 2021.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, DMKD '03, pages 2–11, New York, NY, USA. Association for Computing Machinery, 2003. https://doi.org/10.1145/882082.882086

Lin, J., Keogh, E., Lonardi, S., and Patel, P. Finding motifs in time series. *Proceedings of the Second Workshop on Temporal Data Mining*, 2002.

Lin, J., Keogh, E., Wei, L., and Lonardi, S. Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144, 2007. https://doi.org/10.1007/s10618-007-0064-z

Linardi, M., Zhu, Y., Palpanas, T., and Keogh, E. Matrix Profile X: VALMOD - Scalable Discovery of Variable-Length Motifs in Data Series. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pages 1053–1066, New York, NY, USA. Association for Computing Machinery, 2018. https://doi.org/10.1145/3183713.3183744

Linardi, M., Zhu, Y., Palpanas, T., and Keogh, E. Matrix profile goes MAD: variable-length motif and discord discovery in data series. *Data Mining and Knowledge Discovery*, 34(4):1022–1071, 2020. https://doi.org/10.1007/s10618-020-00685-w

Liu, B., Li, J., Chen, C., Tan, W., Chen, Q., and Zhou, M. Efficient Motif Discovery for Large-Scale Time Series in Healthcare. *IEEE Transactions on Industrial Informatics*, 11(3):583–590, 2015. https://doi.org/10.1109/TII.2015.2411226

Maletzke, A. G., Lee, H. D., Enrique, G., Batista, A. P. A., Coy, C. S. R., Fagundes, J. J., and Chung, W. F. Time Series Classification with Motifs and Characteristics. In Espin, R., Pérez, R. B., Cobo, A., Marx, J., and Valdés, A. R., editors, *Soft Computing for Business Intelligence*, pages 125–138. Springer, Berlin, Heidelberg, 2014.

Mohammad, Y. and Nishida, T. Constrained Motif Discovery in Time Series. *New Generation Computing*, 27(4):319–346, 2009. https://doi.org/10.1007/s00354-009-0068-x

Moody, G. and Mark, R. The impact of the MIT-BIH Arrhythmia Database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001. https://doi.org/10.1109/51.932724

Moody, G. B. and Mark, R. G. MIT-BIH Arrhythmia Database, 1992.

Mueen, A. Enumeration of Time Series Motifs of All Lengths. In *2013 IEEE 13th International Conference on Data Mining*, pages 547–556, 2013. https://doi.org/10.1109/ICDM.2013.27

Mueen, A., Keogh, E., Zhu, Q., Cash, S., and Westover, B. Exact discovery of time series motifs. In *Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining 2009, Proceedings in Applied Mathematics*, volume 1, pages 469 – 480, 2009.

Mueen, A., Keogh, E., Zhu, Q., Cash, S. S., Westover, M. B., and Bigdely-Shamlo, N. A disk-aware algorithm for time series motif discovery. *Data Mining and Knowledge Discovery*, 22(1-2):73 – 105, 2011. https://doi.org/10.1007/s10618-010-0176-8

Nawaz, A., Umar, M. A., Shuaib, K., Ahmad, A., and Belkacem, A. N. Autoencoder-based Arrhythmia Detection using Synthetic ECG Generation Technique. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7. Institute of Electrical and Electronics Engineers Inc., 2024. https://doi.org/10.1109/EMBC53108.2024.10781537

Neloy, A. A. and Turgeon, M. A comprehensive study of auto-encoders for anomaly detection: Efficiency and trade-offs. *Machine Learning with Applications*, 17:100572, 2024. https://doi.org/10.1016/j.mlwa.2024.100572

Noering, F. K.-D., Schroeder, Y., Jonas, K., and Klawonn, F. Pattern discovery in time series using autoencoder in comparison to nonlearning approaches. *Integrated Computer-Aided Engineering*, 28(3):237–256, 2021. https://doi.org/10.3233/ica-210650

Ogasawara, E., Martinez, L. C., de Oliveira, D., Zimbrão, G., Pappa, G. L., and Mattoso, M. Adaptive Normalization: A novel data normalization approach for non-stationary time series. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2010. https://doi.org/10.1109/IJCNN.2010.5596746

Ogasawara, E., Murta, L., Zimbrão, G., and Mattoso, M. Neural networks cartridges for data mining on time series. In *Proceedings of the International Joint Conference on Neural Networks*, pages 2302 – 2309, 2009. https://doi.org/10.1109/IJCNN.2009.5178615

Ogasawara, E., Salles, R., Porto, F., and Pacitti, E. *Event Detection in Time Series.* Synthesis Lectures on Data Management. Springer Nature Switzerland, Cham, 2025.

Patel, P., Keogh, E., Lin, J., and Lonardi, S. Mining motifs in massive time series databases. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 370–377, 2002. https://doi.org/10.1109/ICDM.2002.1183925

Pham, N. D., Le, Q. L., and Dang, T. K. Two Novel Adaptive Symbolic Representations for Similarity Search in Time Series Databases. In *2010 12th International Asia-Pacific Web Conference*, pages 181–187, 2010. https://doi.org/10.1109/APWeb.2010.23

Poirier–Herbeck, L., Lahalle, E., Saurel, N., and Marcos, S. Unknown-length motif discovery methods in environmental monitoring time series. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–5. Institute of Electrical and Electronics Engineers Inc., 2022. https://doi.org/10.1109/ICECET55527.2022.9873093

Rahhal, M. M. A., Bazi, Y., AlHichri, H., Alajlan, N., Melgani, F., and Yager, R. R. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345:340–354, 2016. https://doi.org/10.1016/j.ins.2016.01.082

Ramanujam, E. and Padmavathi, S. Multi-objective genetic motif discovery technique for time series classification. *International Journal of Business Intelligence and Data Mining*, 11(4):318–337, 2016. https://doi.org/10.1504/IJBIDM.2016.082214

Rong, C., Chen, Z., Lin, C., and Wang, J. Motif Discovery Using Similarity-Constraints Deep Neural Networks. In Nah, Y., Cui, B., Lee, S.-W., Yu, J. X., Moon, Y.-S., and Whang, S. E., editors, *Database Systems for Advanced Applications*, pages 587–603, Cham. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-59410-7_39

Roy, M., Majumder, S., Halder, A., and Biswas, U. ECG-NET: A deep LSTM autoencoder for detecting anomalous ECG. *Engineering Applications of Artificial Intelligence*, 124:106484, 2023. https://doi.org/10.1016/j.engappai.2023.106484

Sakurada, M. and Yairi, T. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, MLSDA'14, pages 4–11, New York, NY, USA. Association for Computing Machinery, 2014. https://doi.org/10.1145/2689746.2689747

Salles, R., Belloze, K., Porto, F., Gonzalez, P. H., and Ogasawara, E. Nonstationary time series transformation methods: An experimental review. *Knowledge-Based Systems*, 164:274–291, 2019. https://doi.org/10.1016/j.knosys.2018.10.041

Salles, R., Escobar, L., Baroni, L., Zorrilla, R., Ziviani, A., Kreischer, V., Delicato, F., Pires, P. F., Maia, L., Coutinho, R., Assis, L., and Ogasawara, E. Harbinger: Um framework para integração e análise de métodos de detecção de eventos em séries temporais. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 73–84. SBC, 2020. https://doi.org/10.5753/sbbd.2020.13626

Salles, R., Lima, J., Reis, M., Coutinho, R., Pacitti, E., Masseglia, F., Akbarinia, R., Chen, C., Garibaldi, J., Porto, F., and Ogasawara, E. SoftED: Metrics for soft evaluation of time series event detection. *Computers and Industrial Engineering*, 198, 2024. https://doi.org/10.1016/j.cie.2024.110728

Schmidl, S., Wenig, P., and Papenbrock, T. Anomaly detection in time series: a comprehensive evaluation. *Proc. VLDB Endow.*, 15(9):1779–1797, 2022. https://doi.org/10.14778/3538598.3538602

Schäfer, P. and Leser, U. Motiflets: Simple and Accurate Detection of Motifs in Time Series. *Proc. VLDB Endow.*, 16(4):725–737, 2022. https://doi.org/10.14778/3574245.3574257

Serrà, J. and Arcos, J. L. Particle swarm optimization for time series motif discovery. *Knowledge-Based Systems*, 92:127 – 137, 2016. https://doi.org/10.1016/j.knosys.2015.10.021

Shah, A., Singh, D., Mohamed, H. G., Bharany, S., Rehman, A. U., and Hussen, S. Electrocardiogram analysis for cardiac arrhythmia classification and prediction through self attention based auto encoder. *Scientific Reports*, 15(1):9230, 2025. https://doi.org/10.1038/s41598-025-93906-5

Shumway, R. H. and Stoffer, D. S. *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer Nature Switzerland, Cham, 2025.

Tanaka, Y., Iwamoto, K., and Uehara, K. Discovery of Time-Series Motif from Multi-Dimensional Data Based on MDL Principle. *Machine Learning*, 58:269–300, 2005. https://doi.org/10.1007/s10994-005-5829-2

Torkamani, S. and Lohweg, V. Survey on time series motif discovery. *WIREs Data Mining and Knowledge Discovery*, 7(2):e1199, 2017. https://doi.org/10.1002/widm.1199

Van Onsem, M., Ledoux, V., Mélange, W., Dreesen, D., and Van Hoecke, S. Variable Length Motif Discovery in Time Series Data. *IEEE Access*, 11:73754–73766, 2023. https://doi.org/10.1109/ACCESS.2023.3295995

Van Wesenbeeck, D., Yurtman, A., Meert, W., and Blockeel, H. LoCoMotif: discovering time-warped motifs in time series. *Data Mining and Knowledge Discovery*, 38(4):2276–2305, 2024. https://doi.org/10.1007/s10618-024-01032-z

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., and Schaeffter, T. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. https://doi.org/10.1038/s41597-020-0495-6

Wu, R. and Keogh, E. J. Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2421–2429, 2023. https://doi.org/10.1109/TKDE.2021.3112126

Yankov, D., Keogh, E., Medina, J., Chiu, B., and Zordan, V. Detecting time series motifs under uniform scaling. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 844 – 853, 2007. https://doi.org/10.1145/1281192.1281282

Yeh, C.-C. M., Kavantzas, N., and Keogh, E. Matrix Profile VI: Meaningful Multidimensional Motif Discovery. In *2017 IEEE International Conference on Data Mining (ICDM)*, volume 2017-November, pages 565–574, 2017. https://doi.org/10.1109/ICDM.2017.66

Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Silva, D. F., Mueen, A., and Keogh, E. Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1317–1322. Institute of Electrical and Electronics Engineers Inc., 2016. https://doi.org/10.1109/ICDM.2016.0179

Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Zimmerman, Z., Silva, D. F., Mueen, A., and Keogh, E. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. *Data Mining and Knowledge Discovery*, 32(1):83–123, 2018. https://doi.org/10.1007/s10618-017-0519-9

Yin, M. S., Tangsripairoj, S., and Pupacdi, B. Variable Length Motif-Based Time Series Classification. In Boonkrong, S., Unger, H., and Meesad, P., editors, *Recent Advances in Information and Communication Technology*, pages 73–82, Cham. Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-06538-0_8

Zhong, Z. and Sun, L. PerDetect: A Personalized Arrhythmia Detection System Based on Unsupervised Autoencoder. In *2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT)*, pages 914–919. Institute of Electrical and Electronics Engineers Inc., 2023. https://doi.org/10.1109/ACAIT60137.2023.10528650

Zhu, Y., Yeh, C.-C. M., Zimmerman, Z., Kamgar, K., and Keogh, E. Matrix Profile XI: SCRIMP++: Time Series Motif Discovery at Interactive Speeds. In *2018 IEEE International Conference on Data Mining (ICDM)*, volume 2018-November, pages 837–846, 2018. https://doi.org/10.1109/ICDM.2018.00099

Zhu, Y., Zimmerman, Z., Senobari, N. S., Yeh, C.-C. M., Funning, G., Mueen, A., Brisk, P., and Keogh, E. Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 739–748. Institute of Electrical and Electronics Engineers Inc., 2016. https://doi.org/10.1109/ICDM.2016.0085

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO/DIVULGAÇÃO DE DOCUMENTO ELETRÔNICO NO CATÁLOGO DO SISTEMA DE BIBLIOTECAS DO CEFET/RJ E NO REPOSITÓRIO INSTITUCIONAL

Na qualidade de titular dos direitos de autor da publicação abaixo citada, de acordo com a Lei nº 9610/98, autorizo, a partir da presente data, o CEFET/RJ, a disponibilizar gratuitamente, sem ressarcimento de direitos autorais, conforme permissões assinadas abaixo, no Catálogo Online do Sistema de Bibliotecas e no Repositório Institucional do Cefet/RJ, ou qualquer outro formato de disseminação da informação que a instituição implemente, no formato PDF, para fins de leitura, impressão e/ou download pela internet, a título de divulgação da produção científica gerada por esta Instituição.

1- Tipo de trabalho: [ ] Projeto Final (Graduação) [ ] Especialização
[X] Dissertação [ ] Tese

2- Identificação do trabalho/autor(es):

Curso: Mestrado em Ciência da Computação
Título: DETECÇÃO DE MOTIFS EM SÉRIES TEMPORAIS USANDO AUTOENCODERS: UMA ANÁLISE DO USO EM DADOS DE ECG
Autor(es): Luiz Cláudio Lemos de Oliveira
Orientador: Eduardo Soares Ogasawara
Co-orientador:_____

3- Informações de acesso ao trabalho:

Este trabalho é confidencial? [ ] Sim [X] Não
Pode ser liberado para publicação? [X] Total [ ] Parcial [ ] Não
Justifique:_____
_____
Em caso de publicação parcial, assinale as permissões:
[ ] Sumário
[ ] Capítulos. Especifique: _____
[ ] Bibliografia
[ ] Outras partes do trabalho: _____

Rio de Janeiro, 02 / 07 / 2026

LUIZ CLAUDIO LEMOS DE OLIVEIRA:07960228795
Assinado digitalmente por LUIZ CLAUDIO LEMOS DE OLIVEIRA:07960228795
DN: cn=LUIZ CLAUDIO LEMOS DE OLIVEIRA:07960228795, c=BR, o=ICPEdu, ou=CEFET-RJ - Centro Fed. Educ. Tecnol. Celso S. Fonseca, email=luiz.oliveira.7@aluno.cefet-rj.br
Motivo: Sou o autor deste documento
Data: 2026.07.02 23:07:37 -03'00'

_____
Assinatura do(s) autor(es)

gov.br
Documento assinado digitalmente
EDUARDO SOARES OGASAWARA
Data: 03/07/2026 09:08:01-0300
Verifique em https://validar.iti.gov.br

_____
Assinatura do orientador

_____
Assinatura do co-orientador