



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS REITOR JOÃO DAVID FERREIRA LIMA
CURSO DE GRADUAÇÃO EM ENGENHARIA ELETRÔNICA

Mateus Ribeiro de Sousa

**Estimativa dos Níveis de Glicose através do Processamento de Sinais de
Eletrocardiograma e Modelos de Aprendizado de Máquina**

Florianópolis
2025

Mateus Ribeiro de Sousa

**Estimativa dos Níveis de Glicose através do Processamento de Sinais de
Eletrocardiograma e Modelos de Aprendizado de Máquina**

Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia Eletrônica do Campus Reitor João David Ferreira Lima da Universidade Federal de Santa Catarina para a obtenção do título de bacharel em Engenharia Eletrônica.
Orientador: Prof. Cesar Ramos Rodrigues, Dr.

Florianópolis
2025

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Sousa, Mateus Ribeiro de
Estimativa dos Níveis de Glicose através do
Processamento de Sinais de Eletrocardiograma e Modelos de
Aprendizado de Máquina / Mateus Ribeiro de Sousa ;
orientador, Cesar Ramos Rodrigues, 2025.
69 p.

Trabalho de Conclusão de Curso (graduação) -
Universidade Federal de Santa Catarina, Centro Tecnológico,
Graduação em Engenharia Eletrônica, Florianópolis, 2025.

Inclui referências.

1. Engenharia Eletrônica. 2. Eletrocardiograma. 3.
Glicose. 4. Aprendizado de máquina. 5. Processamento de
sinais. I. Rodrigues, Cesar Ramos. II. Universidade
Federal de Santa Catarina. Graduação em Engenharia
Eletrônica. III. Título.

Mateus Ribeiro de Sousa

Estimativa dos Níveis de Glicose através do Processamento de Sinais de Eletrocardiograma e Modelos de Aprendizado de Máquina

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de “bacharel em Engenharia Eletrônica” e aprovado em sua forma final pelo Curso de Graduação em Engenharia Eletrônica.

Florianópolis, 8 de dezembro de 2025.

 Documento assinado digitalmente
Fernando Rangel de Sousa
Data: 16/12/2025 13:36:18-0300
CPF: ***.649.114-**
Verifique as assinaturas em <https://v.ufsc.br>

Prof. Fernando Rangel de Sousa, Dr.
Subcoordenador do Curso

Banca Examinadora:

 Documento assinado digitalmente
Cesar Ramos Rodrigues
Data: 16/12/2025 14:06:45-0300
CPF: ***.557.990-**
Verifique as assinaturas em <https://v.ufsc.br>

Prof. Cesar Ramos Rodrigues, Dr.
Orientador

 Documento assinado digitalmente
Jefferson Luiz Brum Marques
Data: 16/12/2025 10:39:02-0300
CPF: ***.624.060-**
Verifique as assinaturas em <https://v.ufsc.br>

Prof. Jefferson Luiz Brum Marques, Dr.
Avaliador

Universidade Federal de Santa Catarina

 Documento assinado digitalmente
SOFIA PINHEIRO KLAUTAU
Data: 16/12/2025 11:58:00-0300
CPF: ***.254.552-**
Verifique as assinaturas em <https://v.ufsc.br>

Sofia Pinheiro Klautau, Eng. Biomed.
Avaliadora
Universidade Federal de Santa Catarina

Aos que não desistem, mesmo quando tudo parece impossível.

AGRADECIMENTOS

Gostaria de expressar minha sincera gratidão à minha família, especialmente ao meu pai, à minha mãe e à minha irmã, que foram fundamentais durante toda a minha trajetória acadêmica. Agradeço imensamente pelo amor, paciência e incentivo constantes, que sempre me motivaram a seguir em frente, mesmo nos momentos mais desafiadores. Também sou grato ao meu orientador, Prof. Dr. Cesar Ramos Rodrigues, por ter despertado em mim o interesse por seguir os estudos nesta área. Sua orientação excepcional e seus ensinamentos ao longo deste trabalho foram essenciais para o meu crescimento acadêmico e pessoal. Agradeço também à colega Sofia Pinheiro Klautau, pelo apoio e direcionamento durante todo o desenvolvimento deste projeto. Por fim, agradeço a todos os amigos que estiveram ao meu lado ao longo dessa jornada.

“A curiosidade é o motor da descoberta.”

RESUMO

O monitoramento dos níveis de glicose é essencial para prevenir doenças e sintomas associados à hiperglicemia e à hipoglicemia. O diabetes, por exemplo, está entre as doenças que mais causam mortes no mundo, tornando o acompanhamento contínuo da glicose sanguínea uma prática indispensável para pessoas diagnosticadas com a condição. No entanto, os métodos tradicionais de medição são, em sua maioria, invasivos, causam desconforto, dor e exigem a coleta de amostras de sangue, o que pode dificultar o monitoramento frequente. Diante disso, cresce o interesse em pesquisas que buscam estimar os níveis de glicose de forma não invasiva. Este estudo propõe o uso de sinais de eletrocardiograma (ECG), obtidos por meio de um dispositivo vestível, para estimar a glicose sanguínea. O conjunto de dados foi desenvolvido especificamente para este projeto e envolve etapas de pré-processamento e filtragem dos sinais, extração de características relevantes e treinamento de modelos de aprendizado de máquina, com o objetivo de desenvolver uma abordagem precisa, prática e menos desconfortável para o monitoramento contínuo da glicose. Os modelos com melhor performance foram *Random Forest*, com erro RMSE de 11.295 e MAPE de 8,63% e 93% das estimativas na Zona A do Gráfico de Clarke, e o modelo *Gradient Boosting*, com erro RMSE de 11.768 e MAPE de 9,09% e 84% das estimativas na Zona A. O modelo *CatBoost* também chegou próximo, com RMSE e MAPE iguais a 12.512 e 9,53% e 81% das estimativas na Zona A. Portanto, os resultados deste trabalho reforçam que variações nos níveis glicêmicos afetam características morfológicas do sinal de eletrocardiograma.

Palavras-chave: Eletrocardiograma, Glicose, Aprendizado de Máquina, Processamento de Sinais.

ABSTRACT

Monitoring glucose levels is essential to prevent diseases and symptoms associated with hyperglycemia and hypoglycemia. Diabetes, for instance, is among the leading causes of death worldwide, making continuous glucose monitoring indispensable for diagnosed individuals. However, traditional measurement methods are mostly invasive, cause discomfort and pain, and require blood sample collection, which can hinder frequent monitoring. Consequently, there is growing interest in research aimed at estimating glucose levels non-invasively. This study proposes the use of electrocardiogram (ECG) signals, obtained through a wearable device, to estimate blood glucose levels. The dataset was specifically developed for this project and involves signal preprocessing and filtering, extraction of relevant features, and training of machine learning models, with the goal of developing an accurate, practical, and less discomforting approach for continuous glucose monitoring. The best-performing models were *Random Forest*, with an RMSE of 11.295 and MAPE of 8.63% and 93% of estimates in Clarke's Zone A, and *Gradient Boosting*, with an RMSE of 11.768 and MAPE of 9.09% and 84% of estimates in Zone A. The *CatBoost* model also performed closely, with RMSE and MAPE of 12.512 and 9.53%, and 81% of estimates in Zone A. Therefore, the results of this work reinforce that variations in glucose levels affect the morphological characteristics of the electrocardiogram signal.

Keywords: Electrocardiogram, Glucose, Machine Learning, Signal Processing.

LISTA DE FIGURAS

Figura 1 – Forma de onda do ECG.	19
Figura 2 – Exemplo de <i>Smartwatch</i> com tecnologia de coleta de ECG	20
Figura 3 – Filtro Savitzky-Golay.	22
Figura 4 – Resposta em magnitude de um filtro Butterworth ideal de ordem N. . .	23
Figura 5 – Gráfico de Clarke.	26
Figura 6 – Exemplo de sinal de ECG.	31
Figura 7 – Exemplo de sinal de ECG com picos de ruído.	33
Figura 8 – Exemplo de filtragem com Savitzky-Golay.	34
Figura 9 – Exemplo de sinal com os picos detectados.	36
Figura 10 – Exemplo de sinal com qualidade alta (qualidade mínima de 0.9192). . .	39
Figura 11 – Exemplo de sinal com baixa qualidade (qualidade mínima de 0.0578). .	39
Figura 12 – Efeito da qualidade sobre o número de janelas, sinais e diferentes níveis de glicose.	40
Figura 13 – Gráfico do número de janelas por nível de glicose para todas as janelas.	41
Figura 14 – Gráfico do número de janelas por nível de glicose para qualidade acima de 0.9.	41
Figura 15 – Gráfico do número de janelas por nível de glicose para qualidade acima de 0.95.	42
Figura 16 – Glicose Real x Predita para o modelo CatBoost para todas as janelas. .	44
Figura 17 – Glicose Real x Predita para o modelo CatBoost para janelas com qua- lidade acima de 0.2.	44
Figura 18 – Glicose Real x Predita para o modelo CatBoost para janelas com qua- lidade acima de 0.7.	45
Figura 19 – Glicose Real x Predita para o modelo CatBoost para janelas com qua- lidade acima de 0.9.	45
Figura 20 – Gráfico de Clarke para todas as janelas.	46
Figura 21 – Gráfico de Clarke para janelas de qualidade acima de 0.2.	47
Figura 22 – Gráfico de Clarke para janelas de qualidade acima de 0.7.	47
Figura 23 – Gráfico de Clarke para janelas de qualidade acima de 0.9.	48
Figura 24 – K-Fold Cross-Validation.	50
Figura 25 – Gráfico de Clarke para o modelo <i>Random Forest</i>	53
Figura 26 – Gráfico de Clarke para o modelo <i>Gradient Boosting</i>	54
Figura 27 – Gráfico de Clarke para o modelo <i>CatBoost</i>	54
Figura 28 – Correlação das <i>Features</i> com a Glicose.	55
Figura 29 – DNN (<i>Deep Neural Network</i>).	56
Figura 30 – MAE treino x validação.	57
Figura 31 – MAE de validação para diferentes otimizadores	58

Figura 32 – MAE de validação para diferentes <i>Learning rates</i>	58
Figura 33 – MAE de validação para diferentes Batch Sizes	59
Figura 34 – MAE treino x validação	59
Figura 35 – Gráfico de Clarke para a Rede Neural	60

LISTA DE QUADROS

Quadro 1 – Hiperparâmetros otimizados.	49
Quadro 2 – Descrição dos Hiperparâmetros otimizados.	68

LISTA DE TABELAS

Tabela 1 – Comparação de desempenho dos modelos para todas as janelas.	43
Tabela 2 – Comparação de desempenho para janelas com qualidade acima de 0.2.	43
Tabela 3 – Comparação de desempenho para janelas com qualidade acima de 0.7.	43
Tabela 4 – Comparação de desempenho para janelas com qualidade acima de 0.9.	43
Tabela 5 – Resultado dos modelos após otimização de hiperparâmetros	52
Tabela 6 – Melhores hiperparâmetros encontrados para cada modelo	53
Tabela 7 – Resultados dos diferentes conjuntos de entrada na rede neural	56
Tabela 8 – Métricas do treinamento final com os melhores parâmetros	60

LISTA DE ABREVIATURAS E SIGLAS

AUC	<i>Area Under the Curve</i>
AverageQRS	Método de Análise de Qualidade de Sinal
BGL	<i>Blood Glucose Level</i> (Nível de Glicose do Sangue)
CAAE	Certificado de Apresentação para Apreciação Ética
CatBoost	Algoritmo de Aprendizado de Máquina
CEGA	<i>Clarke Error Grid Analysis</i>
CNN	<i>Convolutional Neural Network</i> (Rede Neural Convolucional)
D1NAMO	Banco de Dados
DM	Diabetes Mellitus
DNN	<i>Deep Neural Network</i>
DWT	Transformada Discreta de Wavelet
ECG	Eletrocardiograma
EMI	<i>Electromagnetic Interference</i> (Interferência Eletromagnética)
GBDTs	<i>Gradient Boosted Decision Trees</i>
GPR	<i>Exponential Gaussian Process Regression</i>
GridSearch	Técnica de otimização de hiperparâmetros
HR	<i>Heart Rate</i> (Frequência Cardíaca)
HRV	Variabilidade da Frequência Cardíaca
IEB	Instituto de Engenharia Biomédica
KNN	<i>K-Nearest Neighbors</i>
LDA	Análise Discriminante Linear
MAE	<i>Mean Absolute Error</i> (Erro Médio Absoluto)
MAPE	Erro Percentual Absoluto Médio
ME	Erro Médio
Min-Max	Normalização de Dados
ML	<i>Machine Learning</i>
MLP	Perceptron de Múltiplas Camadas
NeuroKit2	Biblioteca de Processamento de Dados Biomédicos
Optuna	Técnica de otimização de hiperparâmetros
PhysioNet	Plataforma Pública de Banco de Dados Fisiológicos
PLI	<i>Power Line Interference</i> (Interferência da Rede Elétrica)
PPG	Fotopletismografia
QT-dataset	Banco de Dados
RMSE	Erro Quadrático Médio
R ²	Coefficiente de Determinação
SVM	<i>Support Vector Machine</i>
templatematch	Método de Análise de Qualidade de Sinal
UFSC	Universidade Federal de Santa Catarina

XGBoost
Zhao2018

Algoritmo de Aprendizado de Máquina
Método de Análise de Qualidade de Sinal

SUMÁRIO

	Lista de quadros	11
	Lista de tabelas	12
1	INTRODUÇÃO	16
1.1	OBJETIVOS	17
1.2	OBJETIVOS ESPECÍFICOS	17
2	FUNDAMENTAÇÃO TEÓRICA	19
2.1	ELETROCARDIOGRAMA	19
2.2	PROCESSAMENTO DE SINAIS	21
2.2.1	Filtro Savitzky-Golay	21
2.2.2	Filtro Butterworth	22
2.2.3	Normalização	23
2.2.4	Aprendizado de máquina	24
2.2.5	Gráfico de Clarke	25
3	REVISÃO BIBLIOGRÁFICA	27
4	METODOLOGIA	30
4.1	AMBIENTE DE DESENVOLVIMENTO	30
4.2	CONJUNTO DE DADOS	30
4.3	PRÉ-PROCESSAMENTO	32
4.3.1	Filtragem	33
4.4	JANELAMENTO DOS SINAIS	34
4.5	NORMALIZAÇÃO	35
4.6	EXTRAÇÃO DE ATRIBUTOS	35
4.7	MODELOS DE APRENDIZADO DE MÁQUINA	36
4.8	MÉTRICAS DE AVALIAÇÃO	38
4.9	ANÁLISE DA QUALIDADE DO SINAL	39
4.9.1	Treinamento de modelos para diferentes valores de qualidade.	42
4.10	OTIMIZAÇÃO DE HIPERPARÂMETROS	48
4.10.1	K-Fold Cross-Validation	49
4.11	DIVISÃO DOS DADOS	50
5	RESULTADOS	52
5.1	REDE NEURAL	55
6	DISCUSSÃO	61
7	CONCLUSÃO	63
	REFERÊNCIAS	64
	APÊNDICE A –	68

1 INTRODUÇÃO

O Diabetes Mellitus (DM) é uma condição crônica caracterizada por alterações na capacidade do organismo em metabolizar a glicose, resultando em níveis elevados de glicemia. Essa desregulação pode estar associada à deficiência na produção de insulina ou à resistência à sua ação, ocasionando episódios persistentes de hiperglicemia. Como ainda não existe cura para o diabetes, os pacientes precisam monitorar continuamente os níveis de glicose no sangue e utilizar insulina ou outros tratamentos prescritos, mantendo a glicemia dentro de uma faixa segura, geralmente entre 70 e 180 mg/dL. Dessa forma, o manejo clínico da doença busca equilibrar os níveis de glicose, corrigindo a hiperglicemia e prevenindo a hipoglicemia, reduzindo, assim, os riscos e complicações associados (ARBI et al., 2023).

Estudos recentes têm buscado o desenvolvimento de métodos não invasivos para o monitoramento da glicose sanguínea, motivados principalmente pelas limitações dos métodos tradicionais. As técnicas clinicamente aceitas exigem a coleta de amostras de sangue, o que pode causar desconforto e dor aos pacientes, além de apresentar riscos potenciais de infecção. Nesse contexto, surgem alternativas promissoras, como o uso de sinais biomédicos combinados a técnicas de processamento de sinais e aprendizado de máquina, que apresentam potencial para viabilizar a detecção contínua e não invasiva da hiperglicemia (CORONADO-REYES et al., 2025).

O diabetes é uma doença que atualmente afeta cerca de 537 milhões de adultos em todo o mundo, e estima-se que esse número possa chegar a 783 milhões até 2045 (SATTER et al., 2024). O monitoramento contínuo da glicose é essencial para o manejo da doença e técnicas não invasivas para esse monitoramento vem sendo estudadas. Nesse contexto, a fotopletismografia (PPG) surge como uma alternativa promissora, permitindo a estimativa não invasiva dos níveis de glicose por meio da análise de variações de volume sanguíneo na circulação periférica, tornando o PPG uma solução viável e menos invasiva para o acompanhamento da glicemia em pacientes com diabetes (SATTER et al., 2024).

Além das técnicas ópticas para monitoramento da glicose, sinais de eletrocardiograma têm se mostrado uma ferramenta promissora para auxiliar na detecção de alterações nos níveis de glicose. O eletrocardiograma (ECG) registra a atividade elétrica do coração, sendo sensível a variações fisiológicas. Estudos indicam que alterações nos níveis de glicose podem influenciar parâmetros como frequência cardíaca e intervalos do ECG, permitindo que essas informações sejam utilizadas para monitoramento contínuo e não invasivo (ARBI et al., 2023).

Segundo Sanai et al. (2021), estudos recentes também vêm explorando a bioimpedância como alternativa para estimar o *Blood Glucose Level* (BGL). Embora a relação biológica ainda não seja totalmente compreendida, acredita-se que mudanças na membrana das hemácias e nos fluidos intracelulares e extracelulares influenciem a impedância

elétrica, refletindo variações na glicose plasmática.

Este cenário evidencia a necessidade de soluções que possibilitem o acompanhamento remoto, o controle à distância e o suporte a tecnologias de autogestão do diabetes, promovendo, ao mesmo tempo, uma abordagem integrada e sustentável. A hiperglicemia, caracterizada pela elevação dos níveis de glicose no sangue, está associada a complicações crônicas, incluindo doenças cardiovasculares, neuropatia e retinopatia diabéticas. Por outro lado, a hipoglicemia pode causar efeitos imediatos e graves, como confusão mental, tremores, sudorese intensa, perda de consciência, coma e até óbito.

O diagnóstico precoce do diabetes é essencial, pois, quando realizado em estágios iniciais, permite o tratamento antes que a doença evolua para condições críticas. Na ausência de detecção e intervenção precoce, a progressão da doença representa um risco considerável à saúde. Dessa forma, o monitoramento contínuo e rigoroso da glicemia torna-se fundamental, tanto para indivíduos com pré-diabetes quanto para pacientes diagnosticados com a doença (FELLAH ARBI; SOULIMANE; SAFFIH, 2023).

1.1 OBJETIVOS

Este trabalho tem como objetivo investigar o uso de sinais de eletrocardiograma (ECG) como ferramenta para a detecção de alterações nos níveis de glicose sanguínea. A proposta consiste em analisar como variações glicêmicas podem influenciar características específicas do sinal cardíaco, tais como variabilidade da frequência cardíaca, morfologia das ondas e parâmetros temporais e espectrais. Com base nessa análise, busca-se identificar padrões e correlações que permitam inferir mudanças no estado metabólico do indivíduo. Ao explorar exclusivamente o ECG, pretende-se avaliar seu potencial como um método não invasivo e de baixo custo para o monitoramento da glicose, contribuindo para o avanço de soluções alternativas às técnicas invasivas tradicionais utilizadas no controle do diabetes.

1.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos deste trabalho incluem:

1. Analisar e organizar uma base de dados composto por sinais de ECG de participantes com diferentes níveis de glicose, estabelecendo a associação desses sinais com os valores glicêmicos obtidos por métodos tradicionais de medição.
2. Realizar o pré-processamento dos sinais, incluindo remoção de ruído, segmentação de ciclos e normalização dos dados.
3. Extrair características temporais e morfológicas relevantes dos sinais, como intervalos RR, variabilidade da frequência cardíaca (HRV) e morfologia das ondas P, T e complexo QRS do ECG.

4. Realizar análise da qualidade dos sinais.
5. Desenvolver e treinar modelos clássicos e de redes neurais profundas a partir dos sinais processados.
6. Avaliar o desempenho dos modelos treinados utilizando métricas adequadas.
7. Comparar o desempenho dos diferentes modelos e arquiteturas de redes neurais utilizadas e discutir os resultados obtidos em relação a métodos mais tradicionais.
8. Investigar possíveis correlações entre alterações nos sinais fisiológicos e os níveis de glicose, com o objetivo de identificar padrões relevantes para a estimativa não invasiva da glicemia.

2 FUNDAMENTAÇÃO TEÓRICA

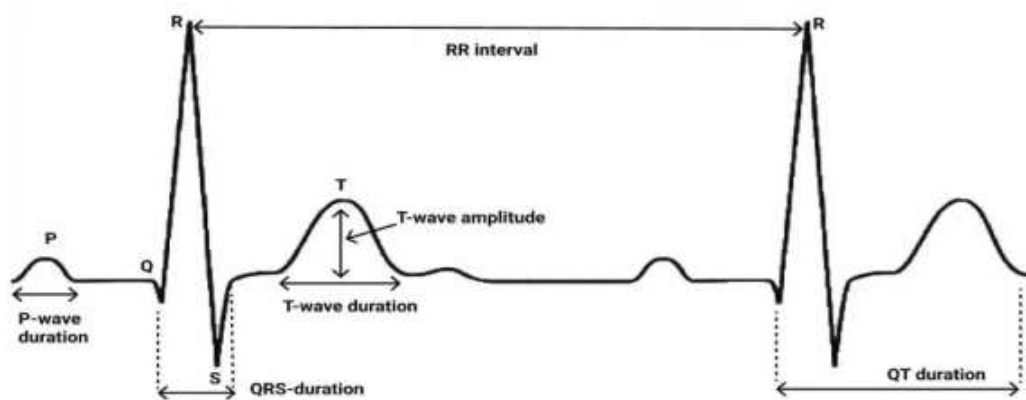
Neste capítulo, serão apresentados e discutidos os principais conceitos teóricos que embasam o desenvolvimento deste trabalho. Serão abordados os fundamentos relacionados ao sinal fisiológico utilizado, o ECG, bem como as etapas de processamento desses dados, incluindo filtragem, segmentação e extração de características relevantes para a análise. Além disso, serão introduzidos os princípios do aprendizado de máquina aplicados à interpretação de sinais biológicos.

2.1 ELETROCARDIOGRAMA

O eletrocardiograma é um exame utilizado para registrar a atividade elétrica do coração ao longo do tempo. Trata-se de uma ferramenta não invasiva e eficaz, amplamente empregada em diversas aplicações biomédicas, como a avaliação da frequência cardíaca, a análise do ritmo dos batimentos cardíacos e a identificação de anomalias cardíacas, entre outras utilidades clínicas e científicas. A análise de sinais de ECG tem ganhado destaque por sua capacidade de fornecer informações detalhadas sobre o estado de saúde cardiovascular de forma prática, contínua e acessível (KAPLAN BERKAYA et al., 2018).

Por se tratar de sinais com amplitudes bastante reduzidas, geralmente entre 0,4 mV e 3 mV, torna-se essencial o uso de técnicas eficazes de filtragem antes de realizar qualquer tipo de análise biológica. Na aquisição desses sinais, empregam-se comumente eletrodos de superfície, os quais podem ser configurados de três maneiras distintas: bipolar, unipolar e unipolar aumentada (APPATHURAI et al., 2019).

Figura 1 – Forma de onda do ECG.



Fonte: Adaptado de (ARBI et al., 2023).

Diversos tipos de interferências podem comprometer a qualidade do sinal de ECG. Entre os principais ruídos estão a flutuação da linha de base (conhecida também como ruído de baixa frequência), a *power line interference* (PLI), o ruído muscular e a *electromagnetic interference* (EMI) gerada por equipamentos eletrônicos próximos. Os sinais de ECG geralmente operam em uma faixa de frequência de 0,5 Hz a 100 Hz, e a presença de artefatos representa um desafio considerável para a correta interpretação dos dados. A forma de onda típica de um ECG é composta pelas letras P, Q, R, S e T, sendo que os componentes P e T correspondem a baixas frequências, enquanto o complexo QRS concentra as informações de maior frequência (APPATHURAI et al., 2019).

Além da extração tradicional do sinal de ECG por meio de eletrodos fixados à pele, atualmente é possível realizar a coleta desses sinais de forma não invasiva por meio de dispositivos vestíveis, como smartwatches. Esses dispositivos, amplamente disponíveis no mercado de consumo, permitem o monitoramento contínuo e portátil da atividade elétrica cardíaca, favorecendo o acompanhamento remoto de pacientes com condições crônicas ou assintomáticas, que antes dependiam exclusivamente de tecnologias como o monitor Holter. O avanço dos dispositivos com sensores de ECG integrados representa uma mudança significativa na forma de se realizar o monitoramento da saúde cardiovascular, possibilitando, além da coleta de dados em tempo real, a aplicação de algoritmos inteligentes para diagnóstico precoce e prevenção de doenças cardíacas (WANG et al., 2024).

Figura 2 – Exemplo de *Smartwatch* com tecnologia de coleta de ECG .



Apple Watch

Fonte: Adaptado de (WANG et al., 2024).

2.2 PROCESSAMENTO DE SINAIS

O processamento dos sinais de eletrocardiograma é uma etapa fundamental para garantir a precisão das análises fisiológicas e dos diagnósticos clínicos. Esses sinais, por sua natureza biológica e pela sensibilidade dos sensores utilizados, estão sujeitos à presença de diversos tipos de ruído que podem comprometer a extração de informações relevantes. As principais fontes de ruído incluem movimentos corporais durante a aquisição do sinal, variações na posição dos sensores, instabilidade dos equipamentos, interferências elétricas e até mesmo o ciclo respiratório do paciente. Para minimizar esses efeitos e preservar as características originais do sinal, é essencial aplicar técnicas adequadas de filtragem. Entre os métodos mais utilizados, destacam-se o filtro Savitzky-Golay, eficiente na suavização de sinais sem perda significativa de componentes importantes, e o filtro Butterworth, amplamente empregado por sua resposta em frequência suave e capacidade de atenuar ruídos indesejados, mantendo a forma do sinal fisiológico.

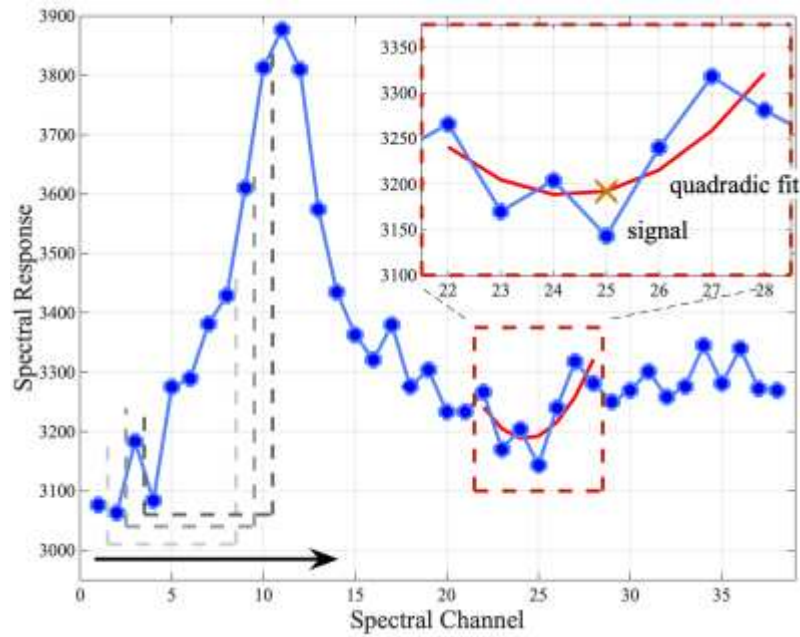
2.2.1 Filtro Savitzky-Golay

Segundo Gallagher (2020), o filtro de Savitzky-Golay é amplamente utilizado como etapa de pré-processamento em aplicações de espectroscopia e processamento de sinais, por sua capacidade de suavizar dados e reduzir ruídos de alta frequência sem distorcer significativamente a forma original do sinal. Além disso, o método permite atenuar componentes de baixa frequência por meio da diferenciação, tornando-se útil tanto para a limpeza quanto para o realce de detalhes sutis nos dados.

De acordo com Dombi e Dineva (2018), o funcionamento do filtro baseia-se no ajuste de um polinômio de baixa ordem, por meio do método dos mínimos quadrados, sobre um conjunto de pontos dentro de uma janela deslizante ao longo do tempo. O valor suavizado correspondente ao ponto central dessa janela é obtido por convolução, resultando em uma filtragem eficaz e computacionalmente simples.

Um dos parâmetros mais importantes do filtro é a ordem do polinômio. A escolha dessa ordem influencia diretamente o grau de suavização do sinal. Polinômios de ordem mais baixa tendem a gerar uma suavização mais intensa, podendo atenuar detalhes importantes, enquanto ordens mais altas permitem capturar melhor as variações locais, resultando em menor suavização. Dessa forma, é necessário encontrar um equilíbrio adequado: a ordem deve ser suficiente para reduzir o ruído sem perder características relevantes do sinal, como picos e vales.

Figura 3 – Filtro Savitzky-Golay.



Fonte: Adaptado de (Gallagher 2020).

2.2.2 Filtro Butterworth

Os filtros dessa classe são definidos por dois parâmetros principais: a frequência de corte e a ordem do filtro. Sua resposta em frequência é suave e contínua, sem oscilações na banda passante, e a transição entre a banda de passagem e a de rejeição torna-se mais acentuada conforme aumenta a ordem do filtro. A resposta em frequência do filtro Butterworth é uniforme na banda passante, mantendo o mesmo ganho para todas as frequências úteis, e diminui gradualmente até zero na banda de rejeição. Um filtro de primeira ordem apresenta uma inclinação de cerca de 6 dB por oitava; um de segunda ordem, de 12 dB por oitava; um de terceira ordem, de 18 dB por oitava; e assim sucessivamente.

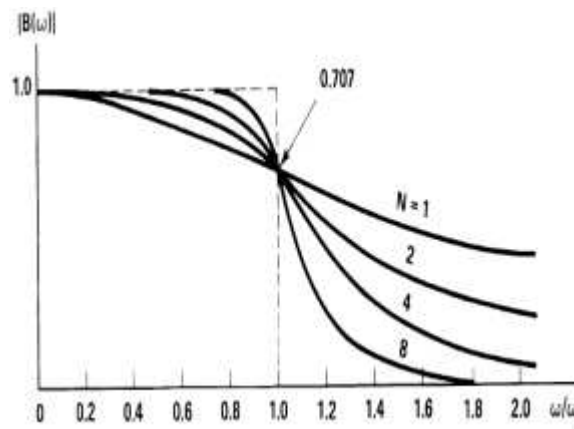
Resposta em frequência de um filtro Butterworth:

$$|H(j\omega)| = \frac{1}{\sqrt{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}}}$$

- $|H(j\omega)|$ — Magnitude da resposta em frequência.
- ω_c — Frequência de corte.
- n — Ordem do filtro.

O filtro Butterworth é o único que mantém essa forma característica mesmo para ordens mais altas, enquanto outros tipos, como os filtros de Bessel, Chebyshev e elíptico, apresentam comportamentos distintos. Em comparação com os filtros Chebyshev do tipo I ou II, o Butterworth possui uma transição mais suave entre as bandas, exigindo uma ordem maior para atingir a mesma atenuação. Em contrapartida, oferece uma resposta de fase mais linear na banda passante, o que é desejável em aplicações como o processamento de sinais de ECG (PANDEY; TIWARI, 2014).

Figura 4 – Resposta em magnitude de um filtro Butterworth ideal de ordem N .



Fonte: Adaptado de (PANDEY; TIWARI, 2014).

2.2.3 Normalização

A normalização Min-Max é uma técnica utilizada para reescalar valores de um conjunto de dados numéricos para um intervalo predefinido, geralmente entre 0 e 1. O processo consiste em subtrair de cada valor o mínimo presente na coluna e dividir pela diferença entre o valor máximo e o mínimo dessa mesma coluna. Dessa forma, o menor valor passa a ser 0, o maior passa a ser 1, e todos os demais são ajustados proporcionalmente dentro desse intervalo. Esse método é amplamente utilizado na preparação de dados para modelos de aprendizado de máquina, pois garante que todas as variáveis tenham escala semelhante, evitando que atributos com valores numericamente maiores exerçam influência desproporcional durante o treinamento do modelo.

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

- x — valor original do dado.
- $x_{\min}$ — menor valor presente no conjunto de dados.
- $x_{\max}$ — maior valor presente no conjunto de dados.
- x_{norm} — valor normalizado, reescalado para o intervalo desejado (geralmente entre 0 e 1).

2.2.4 Aprendizado de máquina

Segundo Zhou, Tang e Zheng (2020), o aprendizado de máquina ou *machine learning* (ML), vem sendo amplamente reconhecido como uma ferramenta poderosa para pesquisadores que lidam com problemas complexos em processamento de sinais e imagens. Essa tecnologia permite a criação de métodos automáticos de aprendizado capazes de identificar padrões recorrentes em grandes volumes de dados. Com base nesses padrões, é possível realizar decisões inteligentes que seriam muito difíceis de serem determinadas manualmente. Assim, o aprendizado de máquina oferece soluções que combinam eficiência e precisão, tornando-se essencial em contextos nos quais a complexidade e a variabilidade dos dados impedem abordagens tradicionais.

O aprendizado de máquina desempenha um papel fundamental na área da saúde ao permitir a extração automática de informações relevantes provenientes de diferentes tipos de dados, como sinais fisiológicos, imagens médicas e registros clínicos. A partir desses dados, algoritmos de aprendizado de máquina conseguem identificar padrões complexos e construir modelos preditivos. Na prática, essas técnicas têm sido aplicadas para apoiar profissionais no processo de diagnóstico, auxiliar na tomada de decisão clínica e oferecer ferramentas mais confiáveis para o acompanhamento da evolução do paciente. Além disso, o uso de aprendizado de máquina possibilita o monitoramento contínuo e em tempo real, a detecção precoce de anomalias e a análise automatizada de grandes volumes de informação, contribuindo para intervenções mais rápidas, personalizadas e eficientes.

Além disso, técnicas avançadas de ML têm sido empregadas para tarefas específicas como a identificação de similaridades entre diferentes modalidades de imagem, a localização precisa de órgãos em exames médicos, a detecção de alterações anatômicas ao longo do tempo, a classificação de tecidos e o desenvolvimento de sistemas de diagnóstico assistido por computador. Conforme Zhou, Tang e Zheng (2020), essas aplicações demonstram que o aprendizado de máquina não apenas otimiza a análise de dados médicos, mas

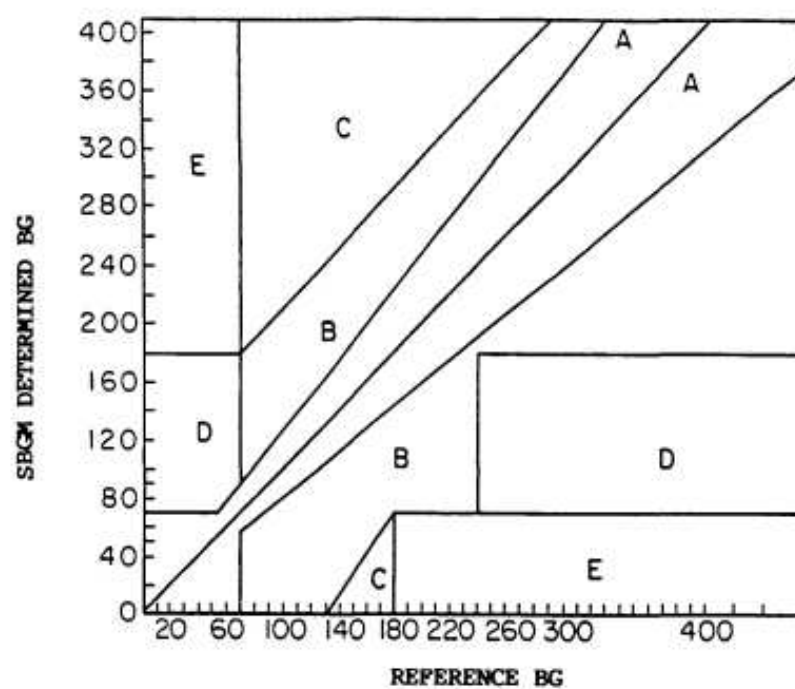
também potencializa a capacidade de médicos e pesquisadores em interpretar informações complexas, reduzir erros e aprimorar tratamentos, evidenciando seu papel estratégico na evolução da medicina baseada em dados.

2.2.5 Gráfico de Clarke

O Gráfico de Clarke, também conhecido como Clarke Error Grid Analysis (CEGA), é um método desenvolvido para avaliar a precisão clínica dos sistemas de automonitoramento da glicose no sangue. Diferente das análises estatísticas tradicionais, o gráfico de Clarke considera o impacto clínico dos erros de medição. Ele compara os valores de glicose obtidos pelo paciente (no eixo Y) com os valores de referência laboratoriais (no eixo X), dividindo o gráfico em cinco zonas (A a E) que representam diferentes graus de concordância e suas possíveis consequências clínicas. As zonas A e B indicam erros clinicamente aceitáveis, enquanto as zonas C, D e E representam erros que podem levar a tratamentos inadequados ou perigosos.

Assim, o gráfico de Clarke fornece uma forma prática e visual de avaliar se um dispositivo de medição de glicose é confiável para uso clínico, destacando não apenas a exatidão numérica, mas também a relevância terapêutica dos resultados. Essa abordagem é amplamente utilizada na validação de novos medidores e sensores de glicose, pois permite compreender se os erros de leitura podem comprometer a segurança do paciente (CLARKE et al., 1987).

Figura 5 – Gráfico de Clarke.



Fonte: Adaptado de (Clarke et al., 1987).

3 REVISÃO BIBLIOGRÁFICA

Nos últimos anos, os estudos na área têm apresentado avanços significativos impulsionados pelo desenvolvimento das técnicas de aprendizado de máquina. O aprimoramento dos algoritmos, aliado ao aumento da capacidade computacional e à disponibilidade de grandes volumes de dados, tem possibilitado a criação de modelos mais precisos e eficientes na análise e interpretação de sinais. Essa evolução tem promovido o surgimento de novas abordagens e metodologias, ampliando o campo de pesquisa e contribuindo para o desenvolvimento de soluções práticas.

O estudo de Fella Arbi et al. (2023) apresentou uma abordagem para a estimativa não invasiva da concentração de glicose utilizando sinais de eletrocardiograma. Os autores propõem que dispositivos portáteis de ECG, amplamente disponíveis, acessíveis e não invasivos, podem servir como sensores vestíveis para integrar sistemas de pâncreas artificial.

O estudo foi dividido em duas etapas principais: pré-processamento e extração de características do ECG, incluindo filtragem, segmentação e detecção de pontos característicos como Q, R, S e T, e aplicação de modelos de regressão para estimar a BGL com base em seis parâmetros extraídos do ECG. Foram utilizados dois conjuntos de dados abertos: o QT-dataset da plataforma pública de banco de dados fisiológicos PhysioNet, para treinamento da segmentação do ECG, e o D1NAMO, para estimativa da BGL em pacientes diabéticos.

Para a segmentação do ECG, foi empregada uma Rede Neural Convolucional (CNN), alcançando uma acurácia de 94%. Em seguida, cinco modelos de regressão foram avaliados, sendo o *Exponential Gaussian Process Regression* (GPR) o mais eficaz, com erro quadrático médio (RMSE) de 0,32 a 0,67 e coeficiente de determinação (R^2) variando de 70% a 98% entre os pacientes. O estudo também utilizou métricas como o Erro Médio Absoluto (MAE) para comparar a BGL estimada com a medida.

Os resultados demonstram que sinais de ECG contêm informações para estimar a glicemia em tempo real, mostrando a viabilidade de monitoramento contínuo não invasivo da glicose por meio de dispositivos vestíveis. Este trabalho evidencia o potencial das técnicas de aprendizado de máquina aplicadas a sinais biomédicos para monitoramento de doenças metabólicas, abrindo caminho para soluções acessíveis e confortáveis para pacientes com diabetes.

Igbe Tobore et al. (2019), realizaram análise estatística e espectral de diferentes segmentos do ECG, incluindo HRV e frequência cardíaca (HR). O objetivo do estudo foi identificar quais parâmetros do ECG poderiam refletir variações na glicose sanguínea, possibilitando estimativas sem a necessidade de métodos invasivos.

No experimento, 16 participantes foram submetidos a dois protocolos: no primeiro, consumiram 75g de uma solução de glicose, no segundo, não ingeriram glicose. Os resultados mostraram que segmentos como HRV e HR apresentaram padrões consistentes

com alterações nos níveis de glicose, enquanto outros segmentos, como QT e ST, também exibiram variações significativas.

Alexakis et al. (2003) investigaram a extração de características e a classificação de sinais de eletrocardiograma (ECG) com o objetivo de identificar alterações relacionadas à hipoglicemia em pacientes com diabetes tipo 1. O estudo utilizou técnicas automáticas para extrair características temporais e morfológicas do ECG, incluindo intervalos de tempo como RR e RTc, bem como parâmetros da onda T, como amplitude, assimetria (*skewness*) e achatamento (*kurtosis*). Essas características foram escolhidas por refletirem alterações na repolarização ventricular, que pode ser afetada por episódios de hipoglicemia. Os dados foram obtidos de onze pacientes com diabetes tipo 1 durante duas noites consecutivas, combinando medições de ECG com níveis de glicose sanguínea.

Para a classificação dos sinais de ECG, os autores compararam métodos lineares e não lineares, utilizando Análise Discriminante Linear (LDA) e Perceptron de Múltiplas Camadas (MLP), sendo que o MLP apresentou melhor desempenho. Os resultados indicaram que a extração automática de características morfológicas e temporais do ECG pode fornecer informações relevantes para detectar alterações cardíacas induzidas por hipoglicemia. Os autores ressaltam que ajustes adicionais nos parâmetros da rede neural e a exploração de novas combinações de características podem melhorar ainda mais a acurácia do sistema. O estudo também sugere a aplicação futura dessas técnicas em populações jovens, onde a incidência de morte súbita relacionada à hipoglicemia é mais elevada.

O estudo de Coronado-Reyes et al. (2025) analisou a possibilidade de detecção não invasiva da hiperglicemia a partir de sinais de eletrocardiograma, utilizando a Transformada Discreta de Wavelet (DWT) para extração de características e técnicas de aprendizado de máquina para classificação. A pesquisa baseou-se em evidências de que o diabetes mellitus provoca alterações em parâmetros cardiovasculares, como a frequência cardíaca e a variabilidade da frequência cardíaca. O processamento dos sinais com a DWT permitiu decompor o ECG em diferentes faixas de frequência, possibilitando a identificação de padrões sutis relacionados às variações glicêmicas.

Foram analisados registros de ECG de 210 indivíduos, incluindo participantes saudáveis e diabéticos, com idades entre 18 e 70 anos. Após a extração das features, os dados foram utilizados como entrada em classificadores *Support Vector Machine* (SVM) e *K-Nearest Neighbors* (KNN). Os resultados obtidos demonstraram altos níveis de acurácia, alcançando 97% na classificação de níveis normais de glicose e 93% na detecção de hiperglicemia. Esses resultados confirmam que alterações no sinal de ECG podem refletir estados metabólicos distintos, portanto, fornecem uma base para futuras aplicações clínicas e para o aprimoramento de algoritmos voltados ao monitoramento contínuo da glicose por meio de sinais fisiológicos.

Já o estudo de Hosseinzadehketilath, Adami e Karimian (2024) realizou uma nova abordagem para monitoramento não invasivo de hiperglicemia utilizando sinais de

ECG provenientes de um banco de dados com 1119 participantes. Para superar limitações de generalização presentes em trabalhos anteriores, os autores desenvolveram uma rede neural profunda capaz de extrair características relevantes em diferentes regiões do sinal e analisar dependências internas entre essas características. O ECG de cada indivíduo foi segmentado em ciclos cardíacos, permitindo processamento mais rápido e eficiente. O modelo foi treinado com 727 indivíduos, validado com 168 e testado em 224 indivíduos nunca vistos, totalizando 9.000 segmentos. A metodologia demonstrou alto desempenho na detecção de hiperglicemia, alcançando AUC de 91,60%, sensibilidade de 81,05% e especificidade de 85,54%, reforçando o potencial do ECG como ferramenta viável para estimativa não invasiva de glicemia.

Diferentemente dos estudos revisados, que em sua maioria utilizam bases de dados públicas, sinais adquiridos em ambientes controlados ou equipamentos médicos dedicados, este trabalho se destaca pela utilização de um dispositivo vestível comercial, especificamente um relógio inteligente, como plataforma de aquisição do sinal de ECG. Essa abordagem aproxima a pesquisa de uma aplicação real e cotidiana, pois avalia o desempenho dos modelos em um cenário mais próximo do uso clínico e doméstico, considerando ruídos, variações de movimento e diversas limitações. Além disso, o uso de um smartwatch amplia o potencial de escalabilidade da solução, tornando viável o monitoramento contínuo da glicose de forma não invasiva, acessível e confortável ao usuário, sem a necessidade de equipamentos especializados. Dessa forma, o trabalho contribui não apenas do ponto de vista científico, mas também prático e tecnológico, fortalecendo a transição das técnicas de aprendizado de máquina aplicadas ao ECG do ambiente experimental para aplicações reais.

4 METODOLOGIA

Este capítulo apresenta o detalhamento dos procedimentos e métodos empregados no desenvolvimento deste trabalho. São descritas as etapas realizadas sobre o conjunto de dados, desde o pré-processamento até o treinamento dos modelos. O objetivo é fornecer uma visão clara e estruturada do processo adotado, de modo a garantir a reprodutibilidade dos resultados e a compreensão das escolhas metodológicas realizadas.

4.1 AMBIENTE DE DESENVOLVIMENTO

Para a execução dos códigos e experimentos deste trabalho, foi utilizado o ambiente Google Colab, por se tratar de uma plataforma online que oferece recursos computacionais gratuitos e compatibilidade com bibliotecas de aprendizado de máquina. O Google Drive foi empregado para o armazenamento e acesso ao conjunto de dados, facilitando a integração com o Google Colab e garantindo maior praticidade no gerenciamento dos arquivos. Toda a implementação foi realizada na linguagem de programação Python, escolhida por sua ampla utilização na área de ciência de dados e pela disponibilidade de diversas bibliotecas voltadas ao processamento, análise e modelagem de dados.

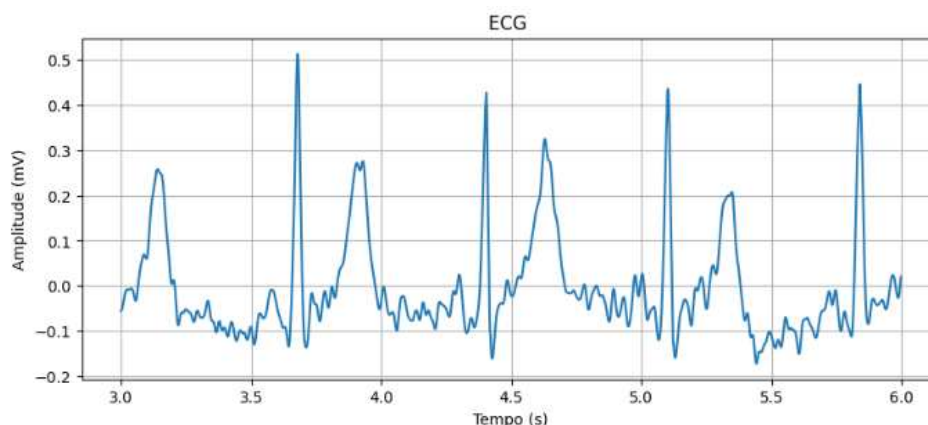
4.2 CONJUNTO DE DADOS

O conjunto de dados utilizado neste trabalho foi desenvolvido em colaboração com o projeto atualmente conduzido por Sofia Klautau no Instituto de Engenharia Biomédica (IEB-UFSC) da Universidade Federal de Santa Catarina (UFSC). Esse conjunto é composto por registros de 9 participantes, para os quais foram coletados sinais de eletrocardiograma (ECG), fotopletismografia (PPG) e bioimpedância, sendo que neste trabalho serão explorados apenas os sinais de ECG. Cada paciente participou de diversas sessões experimentais, correspondentes a diferentes níveis de glicose no sangue, totalizando um banco de dados com 165 sessões, ou seja, para este trabalho, foi obtido 165 registros de ECG, associados a medições específicas de glicose.

Este projeto foi aprovado pelo Comitê de Ética em Pesquisa com Seres Humanos da UFSC, sob o número do CAAE: 84846024.9.0000.0121

Durante as coletas, em alguns casos, os participantes foram orientados a ingerir determinados alimentos ou bebidas com o objetivo de elevar os níveis de glicose, permitindo posteriormente analisar de que forma essa variação influenciava os sinais fisiológicos registrados.

Figura 6 – Exemplo de sinal de ECG.



Fonte: Autor.

Neste trabalho, optou-se por utilizar exclusivamente os sinais de ECG, coletados por meio de um relógio inteligente Galaxy Watch 6, enquanto as medições de glicose foram obtidas com um glicosímetro da marca G-Tech, modelo Free ou com os sensores intersticiais (GCM) Freestyle Libre e Smart MedLevensohn.

A diferença entre os dois métodos é que o glicosímetro tradicional coleta uma gota de sangue da ponta do dedo, exigindo perfuração da pele. Já os sensores GCM medem a glicose no fluido intersticial sob a pele, fornecendo leituras contínuas ao longo do dia com atraso de alguns minutos em relação ao sangue. Enquanto o glicosímetro é mais simples e de baixo custo, o sensor intersticial permite acompanhar tendências glicêmicas, detectar episódios de hipoglicemia ou hiperglicemia e reduzir a necessidade de múltiplas punções diárias.

Cada registro de ECG possui 30 segundos de duração e uma taxa de amostragem de 500 Hz, resultando em 15.000 pontos por sessão para cada participante. Em estudos que envolvem técnicas de aprendizado de máquina, o conjunto de dados exerce um papel central e determinante para o desempenho e a validade dos modelos desenvolvidos. A qualidade, a representatividade e o balanceamento dos dados influenciam diretamente na capacidade do algoritmo de aprender padrões relevantes e generalizar para novos casos. Um conjunto de dados bem estruturado, com variabilidade adequada e medições consistentes, garante que o modelo não apenas alcance bons resultados em fase de treinamento, mas também apresente robustez em situações reais.

O desempenho dos modelos de ML depende diretamente da qualidade dos conjuntos de dados utilizados no treinamento e na avaliação. Conjuntos de dados bem construídos devem refletir com precisão os fenômenos do mundo real, ser abrangentes e livres de vieses, pois qualquer inconsistência ou erro pode comprometer a eficácia do modelo. Assim, compreender os diferentes aspectos que afetam a qualidade dos dados, como integridade, consistência, representatividade e ausência de viés, é essencial para garantir resultados

confiáveis (GONG et al., 2023).

Um elemento central na manutenção da qualidade dos dados é a identificação e correção de problemas, como valores ausentes, erros de medição ou registros corrompidos. Processos de limpeza, pré-processamento e, quando necessário, aumento de dados (*data augmentation*) contribuem para melhorar a robustez do conjunto. Dessa forma, pesquisas sobre qualidade de dados, assim como este estudo se mostram fundamentais para otimizar o desempenho de tarefas de aprendizado de máquina e orientar profissionais na criação de datasets confiáveis e eficientes.

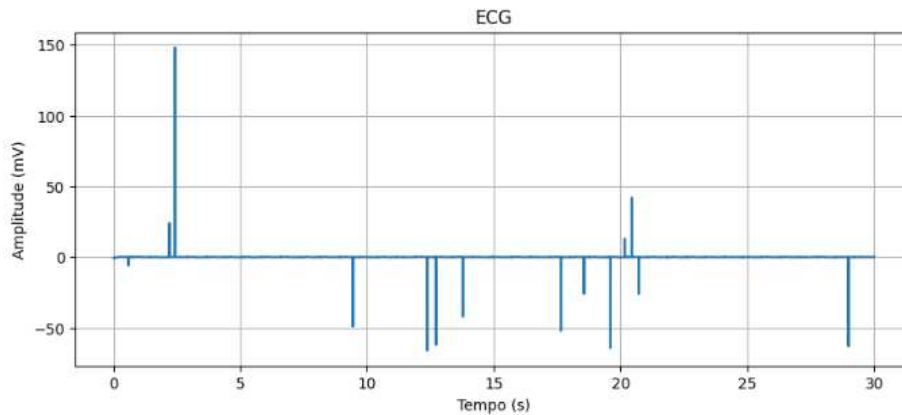
4.3 PRÉ-PROCESSAMENTO

Neste trabalho, foi dada ênfase a um robusto pré-processamento dos dados, devido ao fato de que os sinais de ECG foram coletados por um dispositivo atualmente não considerado altamente preciso para medições clínicas rigorosas. O ECG obtido pelo relógio apresentou, na maior parte dos participantes, níveis significativos de ruído e interferências. O ECG obtido através do relógio, é um ECG de uma única derivação e os eletrodos estão presentes na parte de trás do relógio, em contato com o pulso e no botão na parte lateral em contato com o dedo da mão. Por este motivo o sinal é mais ruidoso que um ECG clínico, em que os eletrodos são afixados em cima do peito e possuem gel condutor em sua composição.

No entanto, o objetivo deste estudo era justamente a obtenção do ECG por meio de um método prático e não invasivo, como um relógio, de forma que o sinal pudesse ser coletado de maneira simples e contínua ao longo do dia pelo próprio paciente. Por esse motivo, foram aplicadas técnicas específicas de filtragem, voltadas à remoção de ruídos de alta e baixa frequência, bem como procedimentos de janelamento, para segmentação do sinal. Além disso, foi realizada uma análise da qualidade do sinal, com o objetivo de identificar trechos inadequados para processamento ou que pudessem comprometer a extração de características relevantes. Esse cuidado no pré-processamento é essencial para garantir que os dados utilizados no treinamento e avaliação dos modelos sejam confiáveis e representativos, minimizando o impacto de ruídos e artefatos na precisão dos algoritmos de *machine learning* aplicados posteriormente.

A Figura 6, por exemplo, mostra um problema que aconteceu em alguns dos sinais, em que picos de ruído de alta amplitude afetavam partes do sinal, e ao plotar o gráfico do sinal completo, devido a escala, não é possível se observar o sinal de ECG que tem amplitude muito menor.

Figura 7 – Exemplo de sinal de ECG com picos de ruído.



Fonte: Autor.

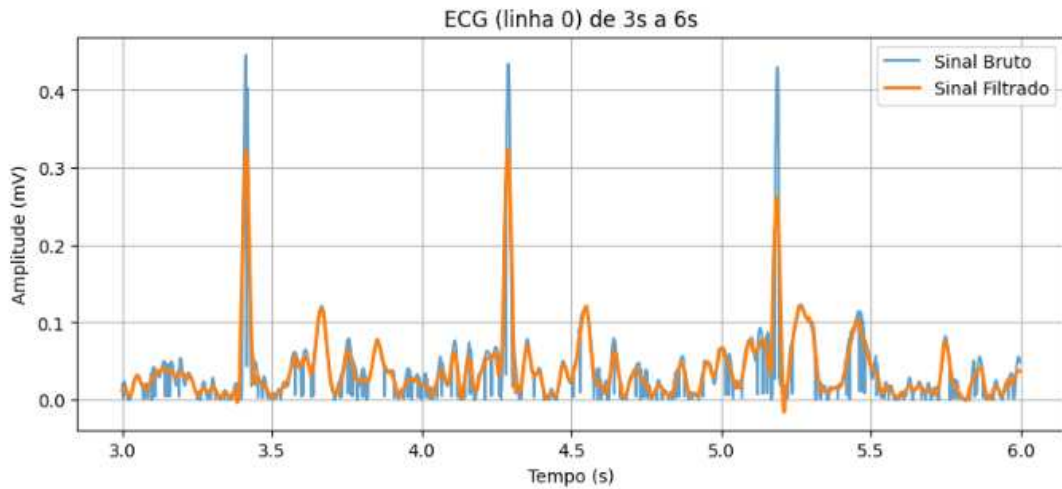
4.3.1 Filtragem

O primeiro filtro aplicado ao sinal foi o filtro de Savitzky-Golay. Esse filtro foi escolhido por ser amplamente utilizado na suavização de sinais biomédicos, como o eletrocardiograma, devido à sua capacidade de reduzir o ruído, preservando, ao mesmo tempo, as características morfológicas do sinal, como picos e curvaturas relevantes. Neste filtro, foram utilizados os parâmetros *frame* (ou *window length*) igual a 23 e *order* (ordem do polinômio) igual a 3, valores que demonstraram desempenho satisfatório na suavização da maioria dos sinais analisados.

O parâmetro *frame* define o tamanho da janela usada para o ajuste do polinômio local em torno de cada ponto do sinal. Essa janela deve possuir comprimento ímpar e ser maior que a ordem do polinômio, garantindo que haja pontos suficientes para o ajuste e que o filtro possa operar de forma simétrica em torno de cada amostra. Já o parâmetro *order*, representa a ordem do polinômio ajustado dentro de cada janela. Ordens mais baixas produzem uma suavização mais intensa, o que é útil para reduzir ruídos de alta frequência, porém podem resultar em perda de detalhes sutis. Por outro lado, ordens mais elevadas preservam melhor as curvaturas e picos característicos do sinal, embora apresentem menor efeito de filtragem.

Dessa forma, a escolha dos parâmetros buscou um equilíbrio entre suavização e preservação das características fisiológicas do ECG, permitindo reduzir o ruído sem comprometer a integridade das informações necessárias para as etapas posteriores de processamento e extração de características.

Figura 8 – Exemplo de filtragem com Savitzky-Golay.



Fonte: Autor.

A segunda etapa de filtragem foi realizada por meio do processamento do sinal utilizando a biblioteca NeuroKit2, desenvolvida para análise e pré-processamento de sinais fisiológicos. Nessa etapa, foi aplicado um filtro digital Butterworth passa-alta, com frequência de corte de 0,5 Hz e ordem igual a 5, conforme configuração padrão da função da biblioteca. O objetivo desse procedimento foi eliminar componentes de baixa frequência presentes no sinal, geralmente associadas a movimentos do corpo, respiração e variações lentas, que podem comprometer a interpretação morfológica das ondas do ECG. O filtro Butterworth foi escolhido por apresentar uma resposta em frequência suave, sem ondulações na banda passante, preservando assim a forma original do sinal e evitando distorções nas ondas P, QRS e T.

4.4 JANELAMENTO DOS SINAIS

Os sinais de ECG, com uma duração total de 30 segundos, foram segmentados em janelas de 2 segundos com o objetivo de aumentar a quantidade de dados disponíveis para análise. Essa abordagem permite gerar um número maior de dados, o que pode ajudar no desempenho dos modelos. A escolha de 2 segundos garante que cada janela contenha pelo menos dois batimentos cardíacos, o que é essencial para a extração da característica do intervalo RR além de possibilitar o cálculo da média das *features*.

4.5 NORMALIZAÇÃO

A normalização de dados em *machine learning* é um processo fundamental para garantir que todas as variáveis de um conjunto de dados contribuam de forma equilibrada para o modelo. Como diferentes atributos podem ter escalas numéricas muito distintas, variáveis com valores maiores podem dominar os cálculos e distorcer os resultados. Ao normalizar, ajustamos os dados para uma mesma escala. Isso melhora o desempenho de algoritmos que dependem de medidas de distância ou gradiente, tornando o aprendizado mais preciso. Dessa forma, a normalização ajuda o modelo a interpretar corretamente os padrões e relações existentes nos dados.

Portanto, nesta etapa foi aplicada a normalização Min-Max, ajustando os valores das amplitudes do sinal para o intervalo entre 0 e 1. É importante destacar que essa normalização foi realizada após o janelamento do sinal, uma vez que o sinal completo de ECG apresentava, em alguns casos, picos de ruído muito elevados, em comparação com a amplitude do sinal de ECG, o que poderia distorcer a escala e prejudicar a representação das demais amostras. Ao normalizar cada janela individualmente, os efeitos desses picos ficam restritos apenas às janelas em que ocorrem, preservando assim a qualidade e a consistência da escala nas demais partes do sinal.

4.6 EXTRAÇÃO DE ATRIBUTOS

Essa fase tem como objetivo identificar e selecionar informações relevantes do sinal, reduzindo sua complexidade e destacando padrões significativos que representem adequadamente o comportamento do fenômeno estudado. A correta extração de características permite que o modelo aprenda de forma mais eficiente, aumentando sua capacidade de generalização e precisão na classificação ou predição de eventos. Dessa forma, essa etapa garante que os dados fornecidos ao modelo sejam representativos e informativos.

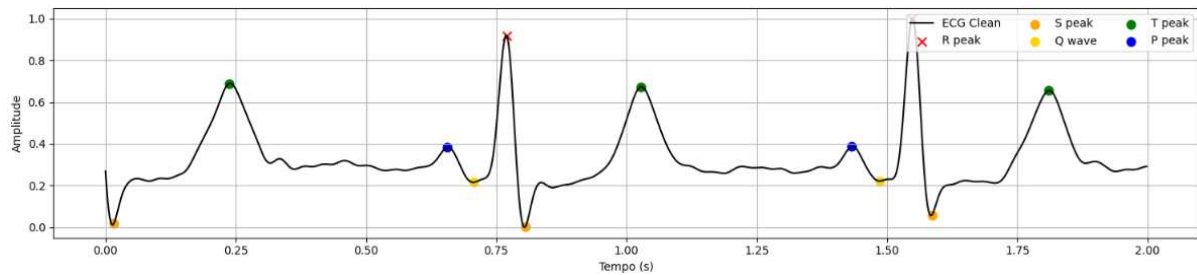
Com o sinal devidamente processado e filtrado, realizou-se a etapa de extração de características (*features*), na qual foram detectados os pontos Q, R, S, P e T. A partir desses pontos, foram extraídas as seguintes características do sinal:

1. QRS_durations: distância entre os picos das ondas Q e S.
2. RR_intervals: distância entre dois picos R consecutivos.
3. QT_segments: distância entre os picos das ondas Q e T.
4. ST_segments: distância entre os picos das ondas S e T.
5. RT_segments: distância entre os picos das ondas R e T.
6. R_amplitudes: amplitude do pulso R.

7. T_amplitudes: amplitude do pulso T.
8. P_amplitudes: amplitude do pulso P.
9. PR_segments: distância entre os picos das ondas P e R.
10. PT_segments: distância entre os picos das ondas P e T.

Após a extração das *features*, foi calculada a média das características em cada janela. Por exemplo, caso uma determinada janela contenha dois valores da *feature* QT_segment, é realizada a média entre esses valores.

Figura 9 – Exemplo de sinal com os picos detectados.



Fonte: Autor.

4.7 MODELOS DE APRENDIZADO DE MÁQUINA

Nesta etapa do trabalho, foram aplicados diversos modelos de aprendizado de máquina com o objetivo de comparar o desempenho de diferentes abordagens na estimativa dos níveis de glicose a partir dos sinais de ECG. Os modelos utilizados foram *Support Vector Machine* (SVM), Regressão Linear, *Random Forest*, *Gradient Boosting*, *K-Nearest Neighbors* (KNN), *CatBoost* e *XGBoost*.

Segundo William S. Noble (2006), o SVM é um algoritmo de aprendizado de máquina supervisionado que aprende a classificar dados a partir de exemplos, separando diferentes classes por meio de um limite matemático chamado hiperplano. O objetivo do método é encontrar o hiperplano que melhor divide os dados, garantindo a maior distância possível entre as classes, o que proporciona uma separação mais precisa. Quando os dados não podem ser separados de forma linear, o SVM utiliza funções conhecidas como *kernels*, que transformam os dados em um espaço de maior dimensão, tornando a separação viável. Essa técnica é amplamente aplicada em reconhecimento de padrões, análise de imagens

e, especialmente, em contextos biomédicos, como na classificação de perfis de expressão gênica para auxiliar no diagnóstico de doenças, incluindo diferentes tipos de leucemia.

Segundo Xiaogang Su, Xin Yan e Chih-Ling Tsai (2012), a regressão linear tem um papel fundamental na modelagem estatística, servindo como uma das abordagens mais simples e interpretáveis para relacionar variáveis preditoras e respostas contínuas. Além de sua base teórica bem consolidada, a regressão linear também é amplamente utilizada por sua eficiência computacional, sendo particularmente eficaz quando o conjunto de dados é limitado.

De acordo com Liu, Wang e Zhang (2011), o modelo proposto por Leo Breiman, combina múltiplas árvores de decisão que votam de forma independente, resultando em classificações mais robustas e precisas. As *Random Forests* destacam-se por sua alta acurácia, resistência a ruídos e *outliers*, e baixa propensão ao *overfitting*, tornando-se uma das abordagens mais empregadas em aplicações biológicas.

De acordo com Friedman (2002), o *Gradient Boosting* é uma técnica de aprendizado de máquina que cria modelos de regressão de forma sequencial, ajustando pequenas funções chamadas *base learners* aos erros do modelo anterior. A cada etapa, o método busca reduzir esses erros, melhorando gradualmente a precisão do modelo. Para tornar o processo mais rápido e evitar sobreajuste, o algoritmo pode usar apenas uma parte aleatória dos dados em cada iteração, o que também aumenta sua robustez.

De acordo com Gongde Guo et al. (2003), o método KNN é uma técnica de classificação não paramétrica amplamente utilizada por sua simplicidade e eficiência. Ele funciona identificando os k vizinhos mais próximos de uma amostra e determinando sua classe com base em votação majoritária. O desempenho do KNN, entretanto, depende fortemente da escolha do valor de k , o que pode influenciar significativamente os resultados. Para reduzir essa dependência, foi proposta uma abordagem baseada em probabilidade contextual que considera múltiplos conjuntos de vizinhos, proporcionando classificações mais estáveis e precisas, embora com maior custo computacional.

Os métodos baseados em Gradient Boosted Decision Trees (GBDTs) constituem uma das abordagens mais eficazes para problemas de classificação e regressão em ambientes de *Big Data*. Entre essas técnicas, o algoritmo *CatBoost* se destaca por lidar de maneira eficiente com dados categóricos e heterogêneos, além de apresentar alto desempenho em diferentes contextos. No entanto, a eficácia do *CatBoost* depende fortemente do ajuste adequado de seus hiperparâmetros, o que exige um processo cuidadoso de calibração. Assim, compreender suas vantagens e limitações é essencial para aplicá-lo de forma apropriada em problemas de aprendizado de máquina (HANCOCK; KHOSHGOFTAAR, 2020).

O *XGBoost* é um algoritmo de aprendizado de máquina baseado em *boosting*, que combina vários modelos simples (geralmente árvores de decisão) para formar um modelo forte e mais preciso. Ele funciona de forma sequencial, onde cada nova árvore é treinada para corrigir os erros das anteriores. A cada etapa, o *XGBoost* calcula os gradientes e o

hessiano da função de perda (usando derivadas) para ajustar os pesos das amostras mais difíceis. Ele também utiliza regularização para evitar *overfitting* e otimizações internas que tornam o treinamento rápido e eficiente. O resultado é um modelo robusto, com alto desempenho em tarefas de regressão e classificação (LI, 2019).

4.8 MÉTRICAS DE AVALIAÇÃO

A escolha das métricas de avaliação é uma etapa fundamental no desenvolvimento e na análise de modelos de aprendizado de máquina, especialmente em problemas de regressão. As métricas são responsáveis por quantificar o desempenho do modelo, permitindo verificar se ele está realmente aprendendo os padrões presentes nos dados e se suas previsões possuem qualidade suficiente para aplicação prática.

Além disso, o uso de métricas corretas possibilita realizar comparações justas entre diferentes modelos e abordagens. Sem critérios de avaliação bem definidos, não é possível determinar de forma objetiva qual modelo apresenta o melhor desempenho. Métricas consistentes também garantem a reprodutibilidade dos resultados e a credibilidade das conclusões obtidas.

Em problemas de regressão, o objetivo é estimar valores numéricos contínuos, e para avaliar o desempenho dos modelos são utilizadas métricas que medem o quanto as previsões se aproximam dos valores reais. Entre as métricas mais comuns estão o Erro Médio (ME), o Erro Absoluto Médio (MAE), a Raiz do Erro Quadrático Médio (RMSE), o Coeficiente de Determinação (R^2) e o Erro Percentual Absoluto Médio (MAPE).

O Erro Médio (Mean Error – ME) calcula a média simples das diferenças entre os valores previstos e os valores reais. Essa métrica indica se o modelo tende a superestimar ou subestimar as previsões, sendo positiva quando o modelo tende a prever valores maiores que os observados e negativa quando tende a prever valores menores.

O Erro Absoluto Médio (Mean Absolute Error – MAE) mede a média dos valores absolutos das diferenças entre as previsões e os valores observados. Diferentemente do ME, o MAE considera apenas a magnitude dos erros, ignorando o sinal.

A Raiz do Erro Quadrático Médio (Root Mean Squared Error – RMSE) é a raiz quadrada da média dos quadrados das diferenças entre os valores previstos e reais. Essa métrica penaliza erros grandes de forma mais severa, mas mantém o resultado na mesma escala da variável de saída, permitindo uma interpretação mais intuitiva.

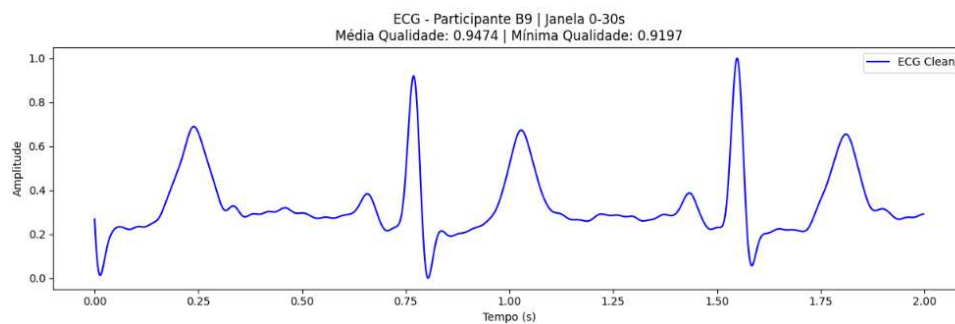
O Coeficiente de Determinação (R^2) mede o quanto da variabilidade dos dados é explicada pelo modelo. Seu valor varia de 0 a 1, sendo que valores mais próximos de 1 indicam um melhor ajuste.

Por fim, o Erro Percentual Absoluto Médio (Mean Absolute Percentage Error – MAPE) representa a média dos erros absolutos em termos percentuais em relação aos valores reais.

4.9 ANÁLISE DA QUALIDADE DO SINAL

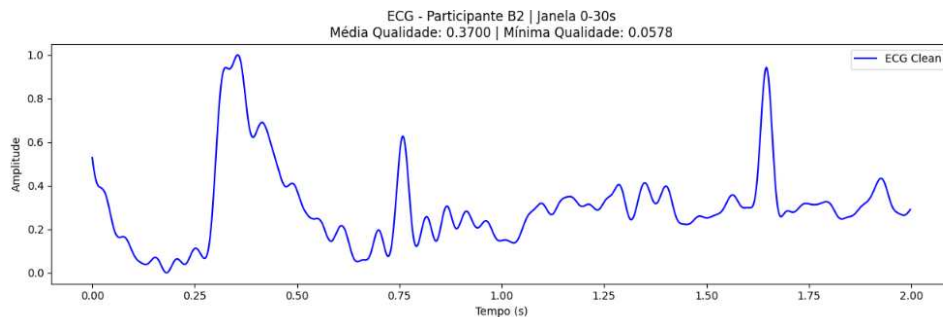
Após a etapa de janelamento, foi realizada uma análise de qualidade das janelas utilizando a biblioteca de processamento de dados biomédicos (NeuroKit2). Essa biblioteca gera um vetor de valores entre 0 e 1 para cada ponto do sinal, indicando a qualidade do sinal de ECG, sendo valores próximos de 1 associados a melhor qualidade. Inicialmente, foram calculadas, para cada janela, a média e o valor mínimo dessas notas. Contudo, observou-se que a média podia permanecer elevada mesmo em casos onde a janela continha tanto batimentos de boa qualidade quanto de baixa qualidade, mascarando a presença de trechos ruidosos. Diante disso, optou-se por utilizar o menor valor de qualidade dentro de cada janela como métrica principal, garantindo uma avaliação mais rigorosa e representativa das regiões com pior qualidade no sinal.

Figura 10 – Exemplo de sinal com qualidade alta (qualidade mínima de 0.9192).



Fonte: Autor.

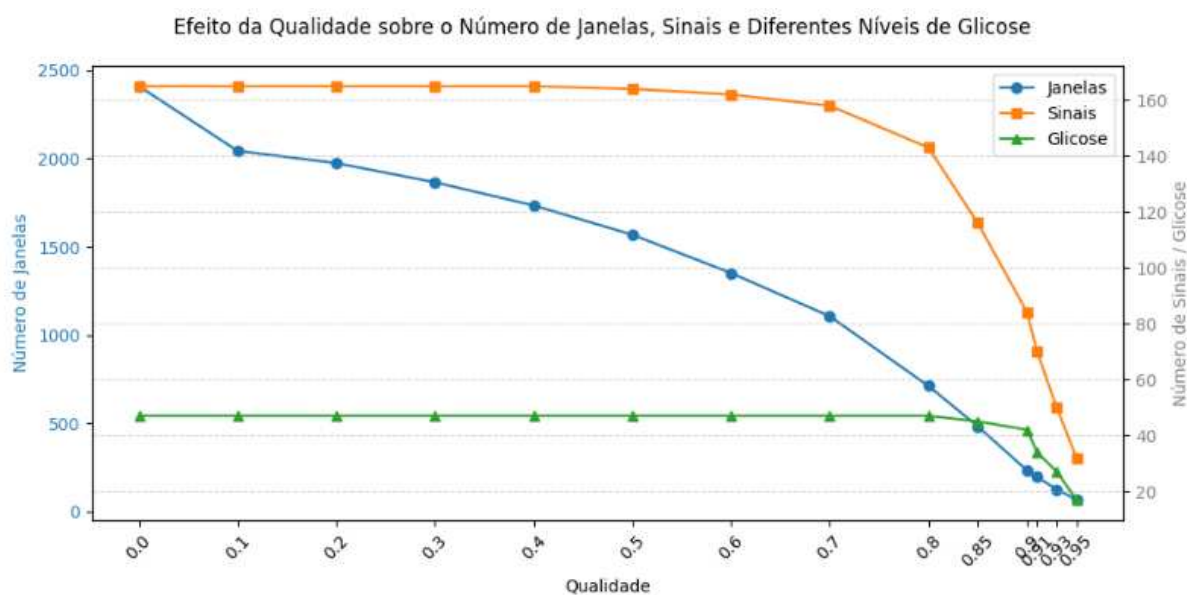
Figura 11 – Exemplo de sinal com baixa qualidade (qualidade mínima de 0.0578).



Fonte: Autor.

Entretanto, há dois aspectos a serem considerados. Quanto maior for a nota mínima escolhida para filtrar as janelas, melhor tende a ser a qualidade do sinal selecionado. Por outro lado, esse critério mais rigoroso reduz a quantidade total de dados disponíveis para análise. Assim, com o objetivo de avaliar a relação entre a nota de qualidade e o número de janelas resultantes, foi elaborado o gráfico apresentado a seguir.

Figura 12 – Efeito da qualidade sobre o número de janelas, sinais e diferentes níveis de glicose.

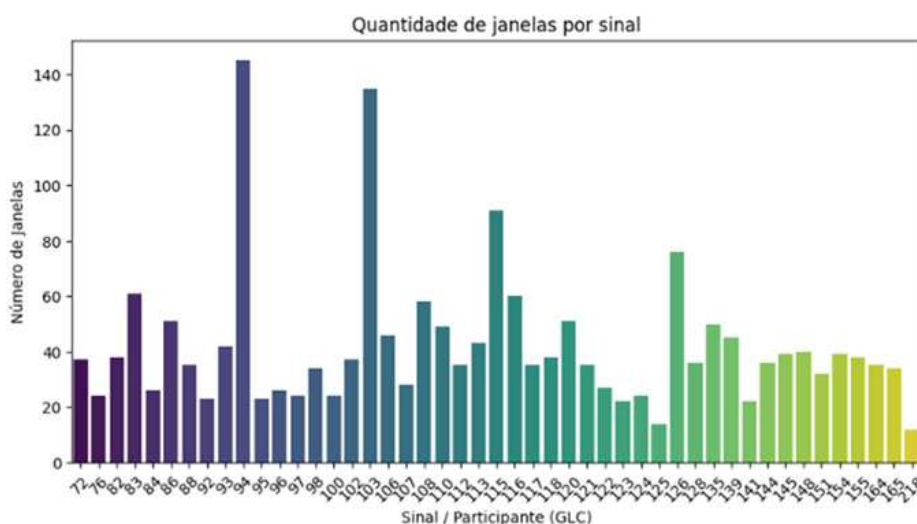


Fonte: Autor.

No gráfico da figura 12, observa-se que para qualidade 0, ou seja, sem filtragem, estão presentes todos os sinais e janelas do conjunto, totalizando 165 sinais e 2475 janelas, distribuídos em 47 níveis distintos de glicose. À medida que o critério de qualidade aumenta, ocorre inicialmente uma redução no número de janelas, enquanto o número de sinais permanece constante. Isso acontece porque apenas parte das janelas de cada sinal é removida, e não o sinal inteiro. A partir da qualidade 0.7, nota-se uma diminuição no número de sinais, o que indica que, em alguns casos, todas as janelas de um determinado sinal foram eliminadas. Mesmo assim, a quantidade de níveis de glicose se mantém, pois existem várias sessões correspondentes a cada nível. Assim, foi escolhido o valor de qualidade mínima igual a 0.9, que garante uma boa qualidade dos sinais de ECG e, apesar da redução significativa no número de janelas e sinais, ainda preserva a variabilidade dos níveis de glicose.

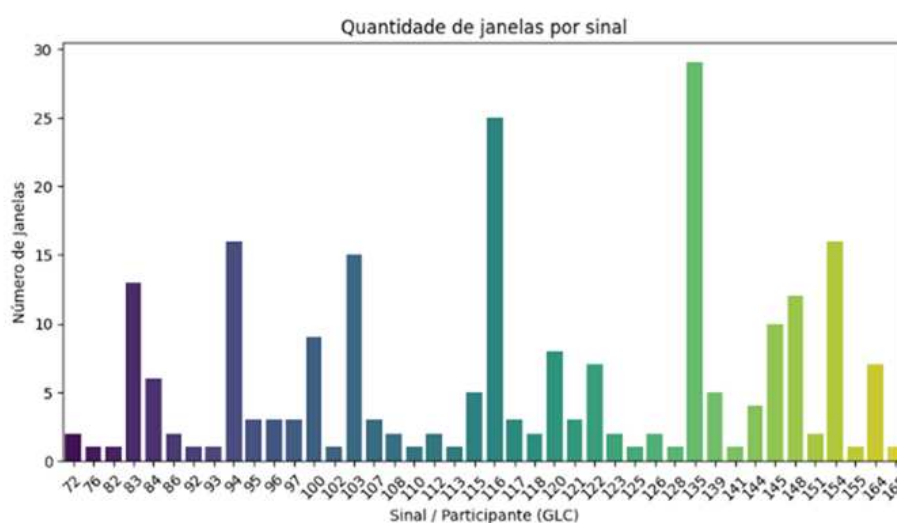
As Figuras 13, 14 e 15 apresentam a quantidade de janelas para cada nível de glicose. Observa-se, na Figura 15, que para valores de qualidade das janelas acima de 0.95 o número de diferentes níveis de glicose é significativamente reduzido. Por outro lado, para valores de qualidade acima de 0.90, embora haja uma diminuição na quantidade de janelas, ainda se preserva grande parte da diversidade dos níveis de glicose.

Figura 13 – Gráfico do número de janelas por nível de glicose para todas as janelas.



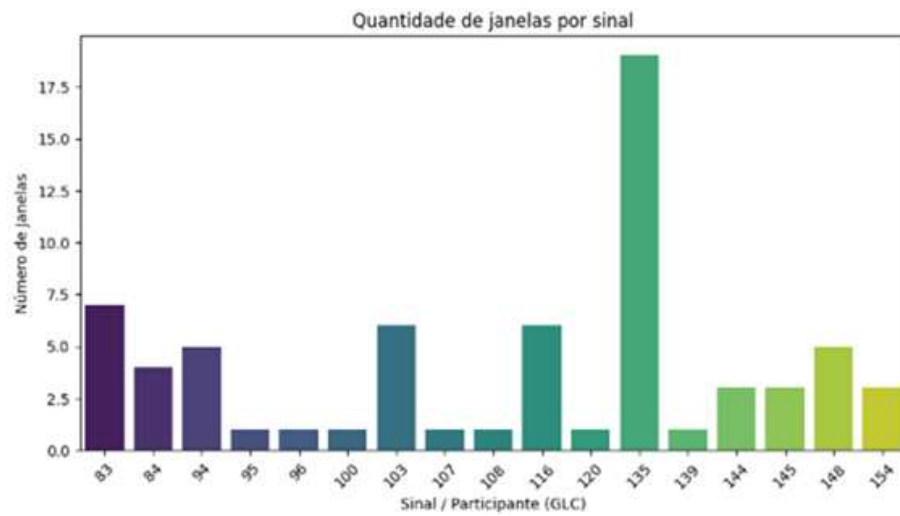
Fonte: Autor.

Figura 14 – Gráfico do número de janelas por nível de glicose para qualidade acima de 0.9.



Fonte: Autor.

Figura 15 – Gráfico do número de janelas por nível de glicose para qualidade acima de 0.95.



Fonte: Autor.

Após a análise de qualidade dos sinais, foram também, excluídas as janelas em que geraram dados faltantes (*NaN*), para alguma das *features* extraídas, uma vez que pode ter ocorrido falhas na detecção dos picos.

Para validação prática, foram realizados testes envolvendo o treinamento e a avaliação do modelo em diferentes faixas de qualidade de sinal. Inicialmente, foram realizados testes com todas as janelas, sem nenhuma filtragem pela qualidade do sinal. Posteriormente, foram consideradas janelas com valores de qualidade acima de 0.2, de modo a excluir apenas as amostras com notas extremamente baixas e preservar um grande volume de dados. Também foram testados limiares mais restritivos, como acima de 0.7, ponto em que, segundo o gráfico da Figura 12, acima disso ocorre uma queda na quantidade de sinais. Por fim, foram testadas janelas com qualidade acima de 0.9, valor máximo que ainda mantém diversidade nos níveis de glicose, permitindo avaliar o impacto da qualidade do sinal no desempenho do modelo.

4.9.1 Treinamento de modelos para diferentes valores de qualidade.

As tabelas a seguir, mostram a comparação dos resultados para cada modelo, para o treinamento e avaliação com cada valor de nota de qualidade:

Tabela 1 – Comparação de desempenho dos modelos para todas as janelas.

Modelo	ME (mg/dL)	MAE (mg/dL)	MAPE (%)	RMSE (mg/dL)	R ²
SVM	2.395	22.520	19.31	32.810	-0.360
Regressão Linear	1.644	20.680	17.70	28.129	0.001
Random Forest	2.580	19.069	16.10	28.219	-0.006
Gradient Boosting	2.308	19.414	16.45	28.448	-0.022
KNN Regressor	2.519	20.473	17.45	29.004	-0.062
CatBoost	2.772	19.407	16.35	28.562	-0.030
XGBoost	3.005	19.479	16.41	28.652	-0.037

Tabela 2 – Comparação de desempenho para janelas com qualidade acima de 0.2.

Modelo	ME (mg/dL)	MAE (mg/dL)	MAPE (%)	RMSE (mg/dL)	R ²
SVM	1.068	22.044	18.98	32.426	-0.339
Regressão Linear	1.088	20.399	17.54	27.876	0.010
Random Forest Regressor	2.111	18.693	15.77	27.913	0.008
Gradient Boosting Regressor	1.883	18.946	16.04	28.245	-0.016
KNN Regressor	1.482	20.481	17.55	29.059	-0.075
CatBoost Regressor	2.044	18.956	16.00	28.250	-0.016
XGBoost Regressor	2.331	19.113	16.19	28.315	-0.021

Tabela 3 – Comparação de desempenho para janelas com qualidade acima de 0.7.

Modelo	ME (mg/dL)	MAE (mg/dL)	MAPE (%)	RMSE (mg/dL)	R ²
SVM	-3.068	15.937	14.48	20.075	-0.085
Regressão Linear	-1.557	15.970	14.33	19.580	-0.033
Random Forest Regressor	-1.653	12.335	11.31	16.320	0.283
Gradient Boosting Regressor	-1.812	14.009	12.74	17.687	0.157
KNN Regressor	-1.627	14.310	12.99	18.326	0.095
CatBoost Regressor	-1.874	13.088	11.92	16.383	0.277
XGBoost Regressor	-0.801	12.980	11.81	16.756	0.244

Tabela 4 – Comparação de desempenho para janelas com qualidade acima de 0.9.

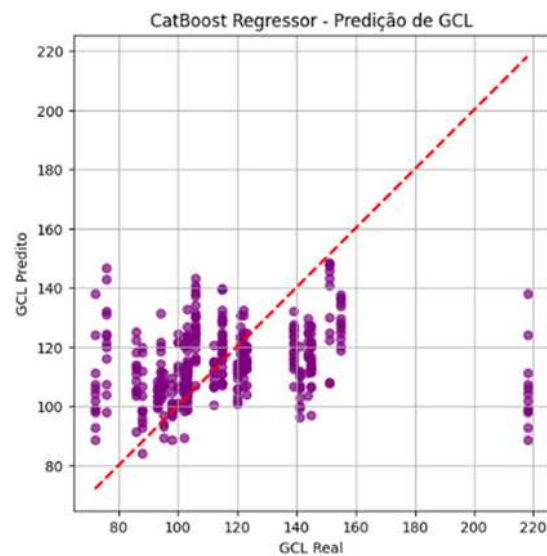
Modelo	ME (mg/dL)	MAE (mg/dL)	MAPE (%)	RMSE (mg/dL)	R ²
SVM	-5.992	17.044	16.99	26.925	-1.035
Regressão Linear	-8.080	12.006	11.55	14.981	0.370
Random Forest Regressor	-3.606	9.157	8.95	11.912	0.602
Gradient Boosting Regressor	-3.667	10.344	10.27	14.269	0.428
KNN Regressor	-6.033	12.851	12.45	16.447	0.241
CatBoost Regressor	-4.891	9.797	9.58	12.613	0.553
XGBoost Regressor	-5.416	11.288	11.10	14.100	0.442

Portanto, observamos que como previsto, uma qualidade de janelas acima de 0.9, se fez mais eficiente, um cenário em que se tem uma boa qualidade de sinal e apesar de

perder uma quantidade significativa de janelas, ainda temos quase toda a diversidade de níveis de glicose.

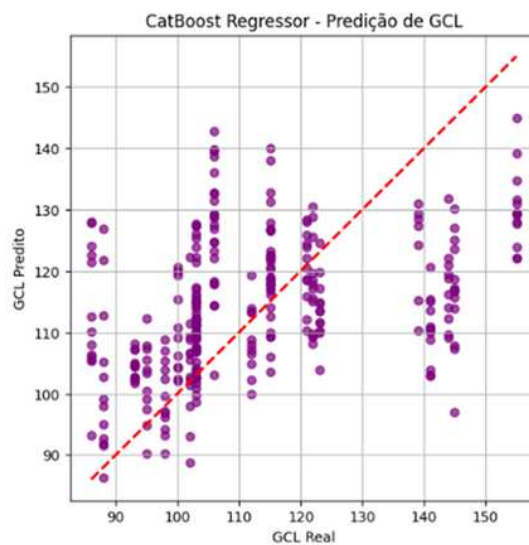
Os gráficos a seguir, mostram a glicose real x glicose predita, para o modelo CatBoost, em que temos a reta em vermelho, mostrando o local ideal de onde os pontos deveriam estar, ou seja, prevendo exatamente o valor correto de glicose.

Figura 16 – Glicose Real x Predita para o modelo CatBoost para todas as janelas.



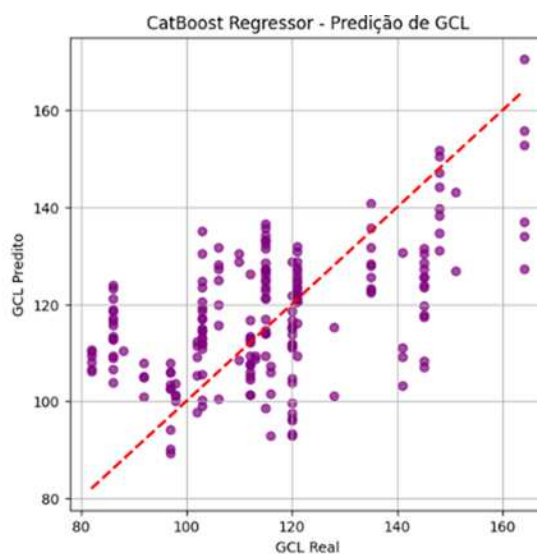
Fonte: Autor.

Figura 17 – Glicose Real x Predita para o modelo CatBoost para janelas com qualidade acima de 0.2.



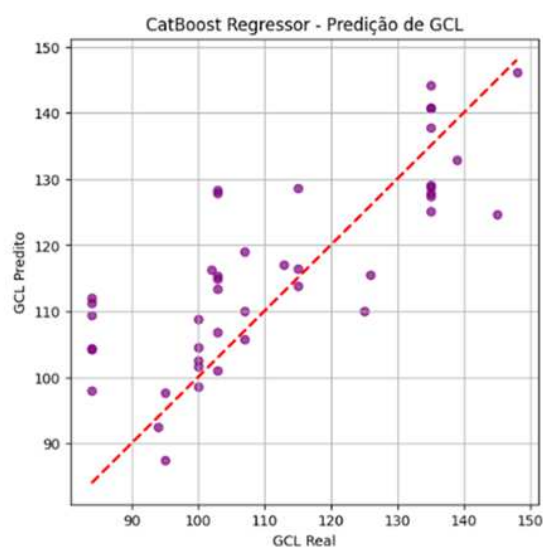
Fonte: Autor.

Figura 18 – Glicose Real x Predita para o modelo CatBoost para janelas com qualidade acima de 0.7.



Fonte: Autor.

Figura 19 – Glicose Real x Predita para o modelo CatBoost para janelas com qualidade acima de 0.9.

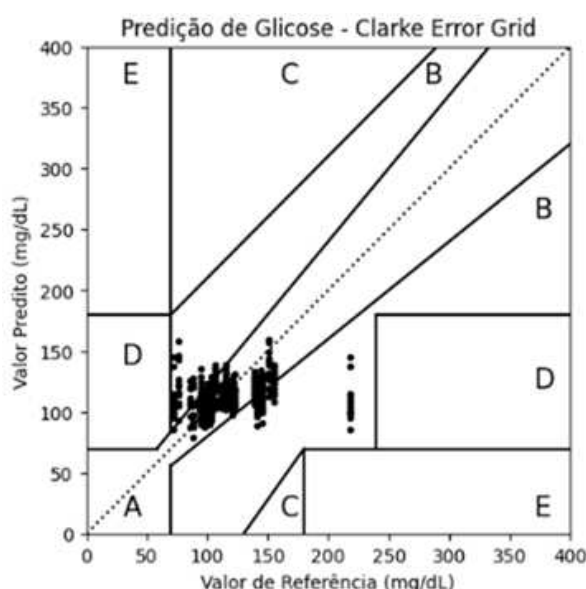


Fonte: Autor.

Nos gráficos das figuras 16, 17, 18 e 19, usando como referência um dos modelos utilizados, o *CatBoost*, observamos que ao aumentar o valor da qualidade das janelas, o modelo tende a aproximar os valores preditos da reta de referência em vermelho, indicando um comportamento que conforme a glicose real aumenta, o modelo tende a prever e acompanhar, prevendo valores maiores de glicose.

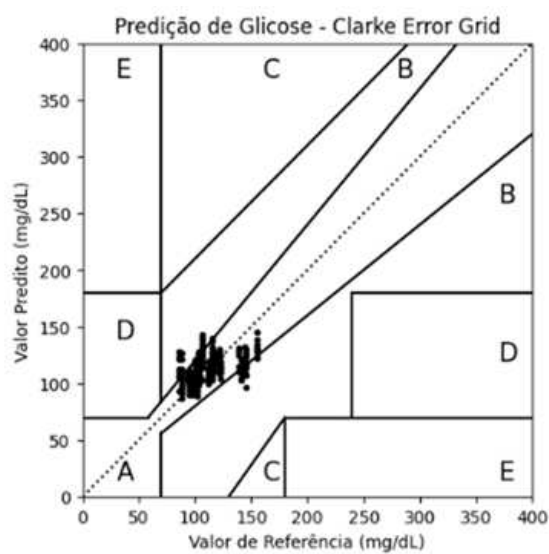
Nas figuras 20, 21, 22 e 23, temos o gráfico de Clarke, fazendo a comparação para o caso de todas as janelas e para as janelas filtradas com qualidade acima de 0.2, 0.7 e 0.9. Na figura 20, com todas as janelas presentes, temos apenas 72% dos valores preditos na Zona A, e podemos observar um conjunto de valores preditos se aproximando da Zona E, uma região indesejada, pois valores de glicose altos estão sendo previstos como baixos. Na Figura 21 para valores de qualidade acima de 0.2 temos ainda 72% dos valores preditos na Zona A, mas não temos mais o *outlier* como na Figura 20, os valores apesar de fora da Zona A, estão bem próximos. Já para valores de qualidade acima de 0.7, como mostram a figura 22, temos um avanço, com 81% dos valores preditos na Zona A. Por fim para valores de qualidade acima de 0.9 temos 83,7% dos valores dentro da Zona A.

Figura 20 – Gráfico de Clarke para todas as janelas.



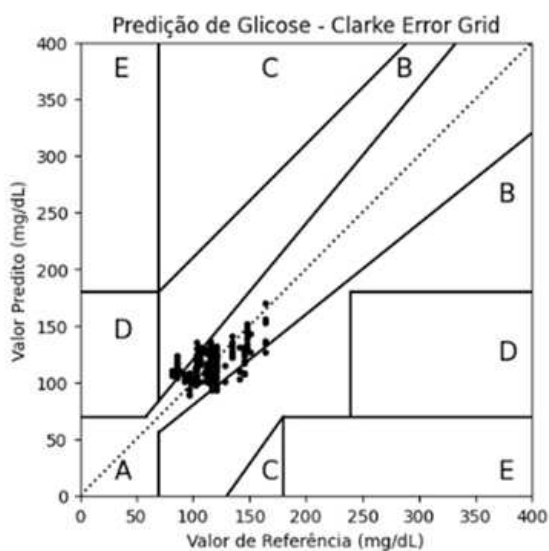
Fonte: Autor.

Figura 21 – Gráfico de Clarke para janelas de qualidade acima de 0.2.



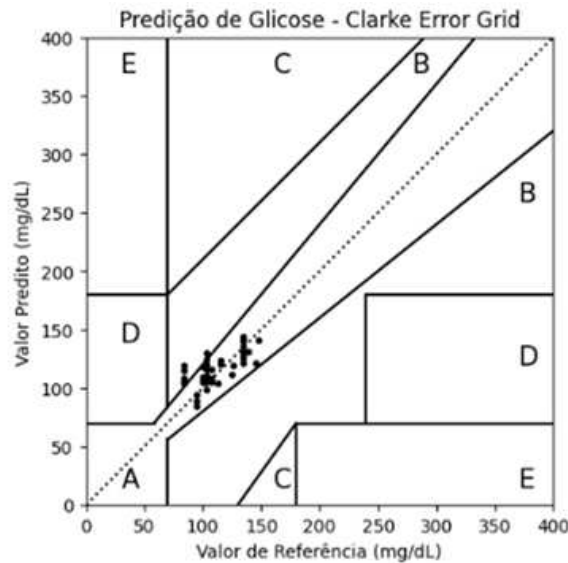
Fonte: Autor.

Figura 22 – Gráfico de Clarke para janelas de qualidade acima de 0.7.



Fonte: Autor.

Figura 23 – Gráfico de Clarke para janelas de qualidade acima de 0.9.



Fonte: Autor.

Portanto com essas análises fica nítida a importância da seleção e análise da qualidade do sinal, sendo para treinar o modelo ou para verificar a qualidade do modelo com o conjunto de teste.

A biblioteca *NeuroKit2*, por padrão, utiliza o método de análise de qualidade de sinal *AverageQRS* para avaliar a qualidade do sinal. Esse método calcula a similaridade entre cada segmento QRS e o QRS médio obtido a partir dos dados, de modo que o índice gerado é relativo: valores próximos de 1 indicam batimentos mais semelhantes ao QRS médio, enquanto valores próximos de 0 representam batimentos mais distantes desse padrão.

É importante destacar que um valor de qualidade igual a 1 não implica necessariamente boa qualidade de sinal. Caso a maioria dos batimentos esteja ruidosa, o QRS médio também será afetado, e estar próximo dele não garante que o batimento seja realmente limpo. No entanto, após uma inspeção visual dos sinais, observou-se que, apesar do nível elevado de ruído, a maior parte dos batimentos permanece legível, com apenas alguns trechos comprometidos por distorções mais fortes ou picos de ruído. Assim, considera-se que valores de qualidade mais próximos de 1 refletem, na prática, batimentos de melhor qualidade que aqueles associados a valores mais baixos.

Também foram testados outros métodos de avaliação de qualidade, como *zhao2018* e *templatematch*, porém nenhum deles apresentou desempenho superior ao método adotado.

4.10 OTIMIZAÇÃO DE HIPERPARÂMETROS

Nesta etapa, buscou-se otimizar os hiperparâmetros dos modelos com o objetivo de melhorar os resultados previamente obtidos. Inicialmente, foi aplicada a técnica de

GridSearch, que realiza uma busca exaustiva testando todas as combinações possíveis dos hiperparâmetros definidos. Contudo, essa abordagem não proporcionou melhorias significativas no desempenho dos modelos.

Em seguida, foi utilizado o método de otimização de hiperparâmetros *Optuna*, que apresentou resultados superiores. Diferentemente do *GridSearch*, o *Optuna* utiliza algoritmos inteligentes, para explorar o espaço de hiperparâmetros de forma otimizada. A técnica aprende com os resultados anteriores e decide dinamicamente quais combinações testar, priorizando as mais promissoras.

Dessa forma, o *Optuna* permite ampliar o espaço de busca, tornando possível avaliar um número maior de combinações com menor custo computacional. Assim, mostrou-se mais eficiente e eficaz para encontrar configurações de hiperparâmetros que elevassem o desempenho dos modelos. Os hiperparâmetros otimizados são mostrados no Quadro 1, e as descrições de cada um mostrados no Apêndice (A).

Quadro 1 – Hiperparâmetros otimizados.

Algoritmo	Hiperparâmetros
SVM	C, epsilon, gamma, kernel
Random Forest	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features
Gradient Boosting	n_estimators, max_depth, min_samples_split, min_samples_leaf, learning_rate, subsample, max_features
KNN	n_neighbors, weights, p
CatBoost	iterations, learning_rate, depth, l2_leaf_reg, bagging_temperature, border_count
XGBoost	n_estimators, learning_rate, max_depth, subsample, colsample_bytree, gamma, min_child_weight, lambda, alpha, booster, rate_drop, max_delta_step

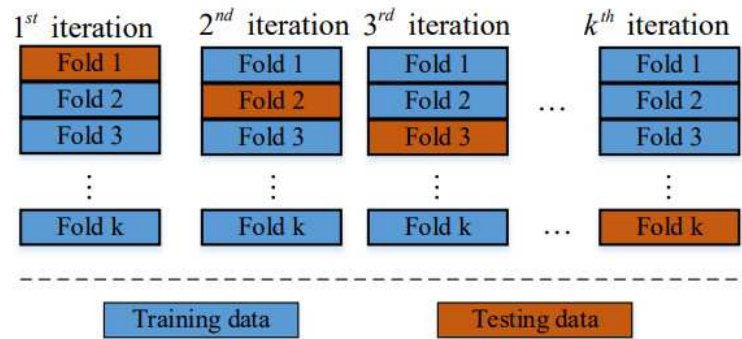
4.10.1 K-Fold Cross-Validation

Juntamente com o *optuna*, foi utilizado o método de validação *K-Fold Cross-Validation*, adotando-se k igual a 5. Nesse procedimento, o conjunto de treinamento é dividido em cinco partes iguais, chamadas *folds*. Em cada iteração, quatro *folds* são utilizados para treinar o modelo, enquanto o *fold* restante é empregado para validação. Esse processo se repete cinco vezes, de forma que, a cada rodada, um *fold* diferente atua como conjunto de validação.

Ao final das cinco iterações, calcula-se a média da métrica utilizada, no caso, o RMSE. Assim, durante o processo de otimização, a combinação de hiperparâmetros selecionada é aquela que apresenta a menor média dos 5 valores de RMSE entre todos os testes com as possíveis combinações de hiperparâmetros.

Essa abordagem reduz o risco de *Overfitting*, uma vez que o modelo é avaliado em múltiplas partições dos dados, e contribui para a escolha de hiperparâmetros com maior capacidade de generalização.

Figura 24 – K-Fold Cross-Validation.



Fonte: Adaptado de (NTI; NYARKO-BOATENG; ANING, 2021).

4.11 DIVISÃO DOS DADOS

A divisão do conjunto de dados é uma etapa fundamental em modelos de aprendizado de máquina, e sua escolha deve ser cuidadosamente analisada conforme os objetivos do estudo. Neste projeto, adotou-se o método intraparticipante, no qual amostras do mesmo indivíduo podem aparecer tanto no conjunto de treino quanto no de teste. Essa abordagem foi escolhida visando uma aplicação personalizada para o paciente, com o modelo podendo ser refinado com seus próprios dados, ou de pessoas com características semelhantes, buscando identificar se o modelo é capaz de detectar variações morfológicas no ECG de um mesmo paciente associadas às mudanças nos níveis de glicose.

Entretanto, como o sinal foi previamente segmentado em janelas, a separação dos dados foi realizada por sinal completo, e não por janelas isoladas. Dessa forma, todas as janelas de um mesmo sinal estão presentes no conjunto de treino ou no conjunto de teste, evitando que janelas do mesmo sinal apareçam em ambos os conjuntos.

Foi definido 80% dos dados para treinamento, que consiste em 190 janelas, e 20% dos dados para teste, 43 janelas. Em relação ao conjunto de validação como já explicado,

foi utilizado apenas na otimização de hiperparâmetros, onde foi usado um valor de $k=5$ no método *K-fold*, ou seja a cada iteração é usado 20% do conjunto de treinamento para se criar o conjunto de validação.

5 RESULTADOS

Este capítulo tem como objetivo apresentar detalhadamente os resultados obtidos ao longo do desenvolvimento deste trabalho. Nele são discutidas as análises realizadas, bem como a comparação entre os diferentes modelos utilizados. Além disso, busca-se identificar quais métodos demonstraram o melhor desempenho, evidenciando suas principais características, vantagens e limitações.

Tabela 5 – Resultado dos modelos após otimização de hiperparâmetros

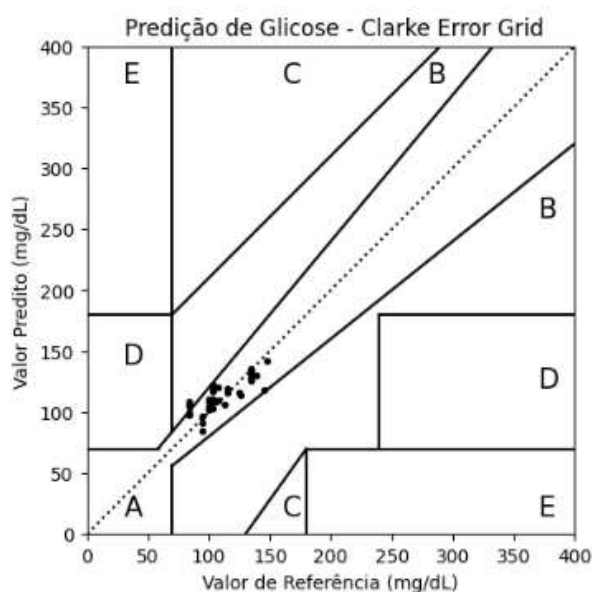
Modelo	ME	MAE	MAPE (%)	RMSE	R ²
SVM	-4.954	13.780	13.49	18.419	0.048
Regressão Linear	-8.080	12.006	11.55	14.981	0.370
Random Forest	-3.223	8.873	8.63	11.295	0.642
Gradient Boosting	-3.598	9.366	9.09	11.768	0.611
KNN	-5.845	12.051	11.67	15.742	0.304
CatBoost	-4.997	9.715	9.53	12.512	0.56
XGBoost	-5.822	11.17	10.99	14.0	0.449

A Tabela 5 apresenta os resultados obtidos pelos modelos após o processo de otimização de hiperparâmetros. A Regressão Linear não passou por essa etapa, uma vez que não possui hiperparâmetros ajustáveis relevantes. Entre os modelos avaliados, aqueles que apresentaram o melhor desempenho foram o *Random Forest*, o *Gradient Boosting* e o *CatBoost*, melhores hiperparâmetros são mostrados na Tabela 6.

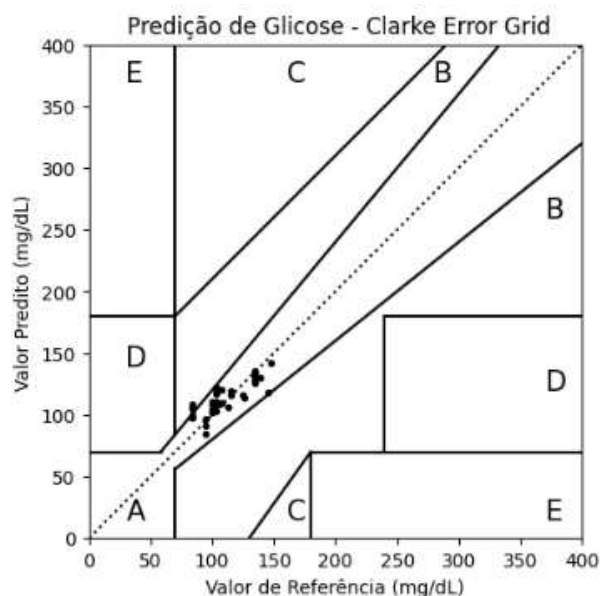
Tabela 6 – Melhores hiperparâmetros encontrados para cada modelo

Modelo	Hiperparâmetros
Random Forest	n_estimators = 285 max_depth = 46 min_samples_split = 3 min_samples_leaf = 2 max_features = None
Gradient Boosting	n_estimators = 391 learning_rate = 0.01341 max_depth = 6 min_samples_split = 8 min_samples_leaf = 3 subsample = 0.5634 max_features = None
CatBoost	iterations = 725 learning_rate = 0.02262 depth = 4 l2_leaf_reg = 6.579 random_strength = 9.964 border_count = 241

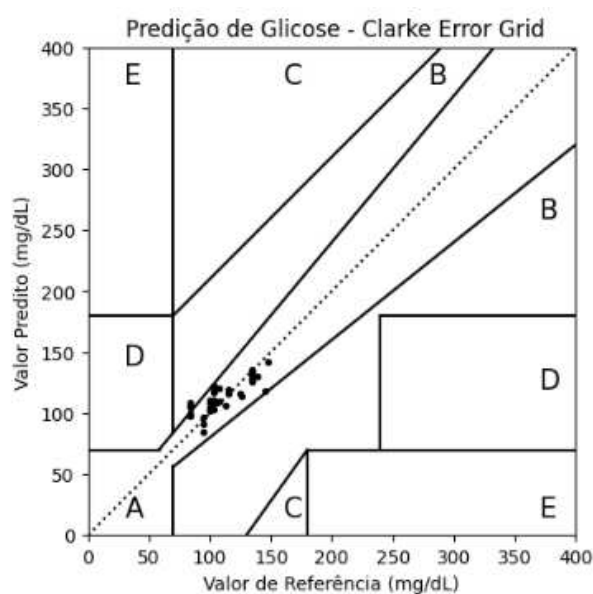
Para esses modelos também foi comparado o Gráfico de Clarke com o *Random Forest* com 93% dos pontos na Zona A, *Gradient boosting* com 84% na Zona A e o *CatBoost* com 81% dos pontos previstos na Zona A.

Figura 25 – Gráfico de Clarke para o modelo *Random Forest*.

Fonte: Autor.

Figura 26 – Gráfico de Clarke para o modelo *Gradient Boosting*.

Fonte: Autor.

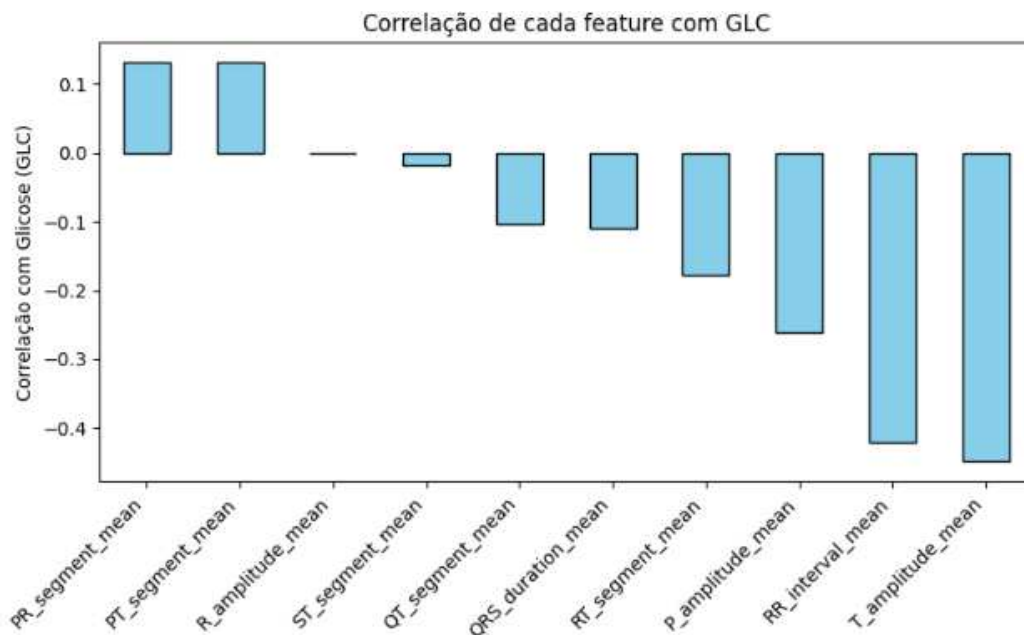
Figura 27 – Gráfico de Clarke para o modelo *CatBoost*.

Fonte: Autor.

A Figura 28 apresenta a correlação entre as *features* extraídas do sinal de ECG e a variação da glicose. Observa-se que o intervalo RR e a amplitude da onda T foram as características mais influentes, mostrando que alterações dos intervalos cardíacos e na morfologia das ondas do ECG estão diretamente relacionadas às mudanças nos níveis glicêmicos.

No estudo de Tobore et al. (2018), que investigou as modificações morfológicas do ECG durante um experimento de ingestão controlada de glicose, verificou-se que, em mais de 80% dos participantes, houve um aumento da frequência cardíaca após a ingestão. Além disso, observou-se uma redução nos intervalos QRS e QT para a maior parte dos indivíduos. Assim, os resultados deste trabalho reforçam e confirmam os resultados observados naquele estudo.

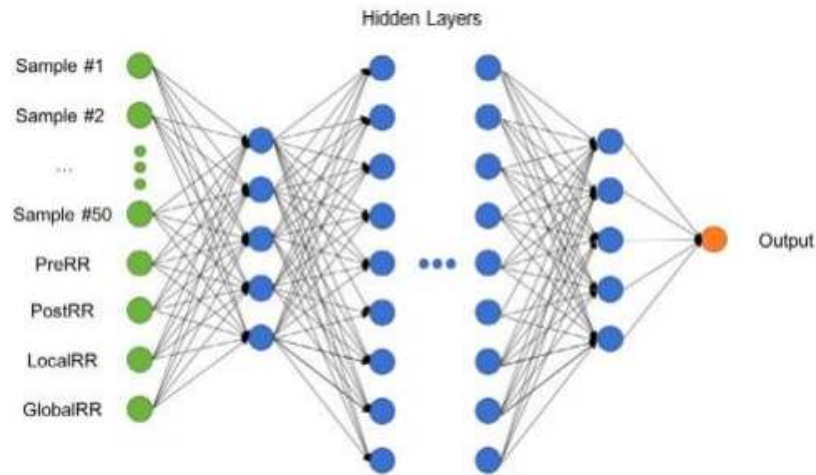
Figura 28 – Correlação das *Features* com a Glicose.



Fonte: Autor.

5.1 REDE NEURAL

Até então, devido ao número limitado de dados, foram utilizados apenas modelos clássicos, uma vez que redes neurais necessitam de grandes volumes de dados para serem treinadas. No entanto, para fins de comparação de resultados, foi adotada uma arquitetura baseada no trabalho de Sannino e De Pietro (2018), que empregou uma DNN (*Deep Neural Network*) com 7 camadas ocultas, contendo 5, 10, 30, 50, 30, 10 e 5 neurônios, respectivamente, conforme mostrado na Figura 29. Essa rede foi utilizada para a detecção de arritmias cardíacas em sinais de ECG.

Figura 29 – DNN (*Deep Neural Network*).

Fonte: Sannino e De Pietro (2018).

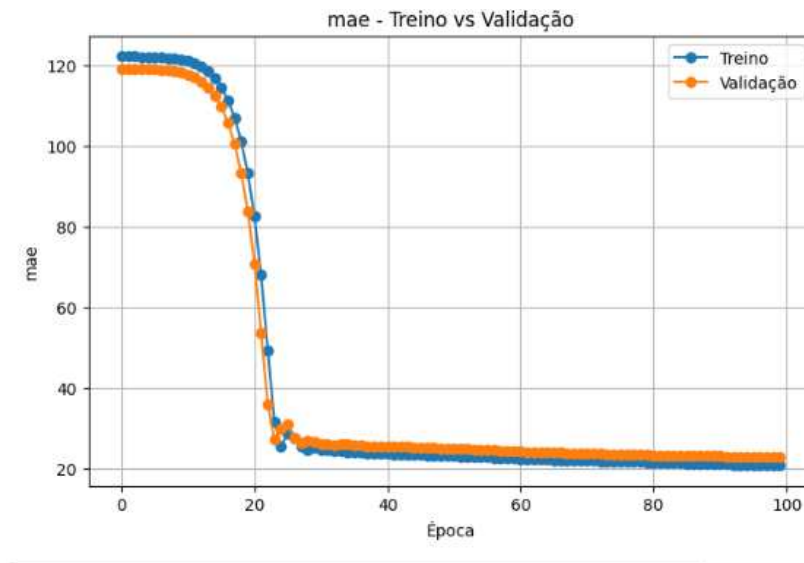
Foi criado um conjunto de validação correspondente a 20% do conjunto de treinamento, e foram realizados três experimentos distintos: o primeiro utilizando apenas as features extraídas, o segundo empregando os sinais completos como entrada para a rede neural, e o terceiro seguindo a abordagem proposta por Sannino e De Pietro (2018), em que cada sinal de ECG foi interpolado para 50 pontos e combinado com as features extraídas. Os resultados desses testes são apresentados na Tabela 7.

Tabela 7 – Resultados dos diferentes conjuntos de entrada na rede neural

Entrada	ME	MAE	MAPE (%)	RMSE	R ²
Features extraídas	-9.808	18.513	18.34	22.325	-0.399
Sinal completo	-13.921	29.377	28.86	38.505	-3.162
50 pontos + Features	-10.179	19.948	19.50	25.075	-0.765

A figura 30 mostra o Erro Absoluto Médio do conjunto de treino e validação ao longos das épocas para o treinamendo apenas com as *Features*, que obteve melhor desempenho.

Figura 30 – MAE treino x validação.

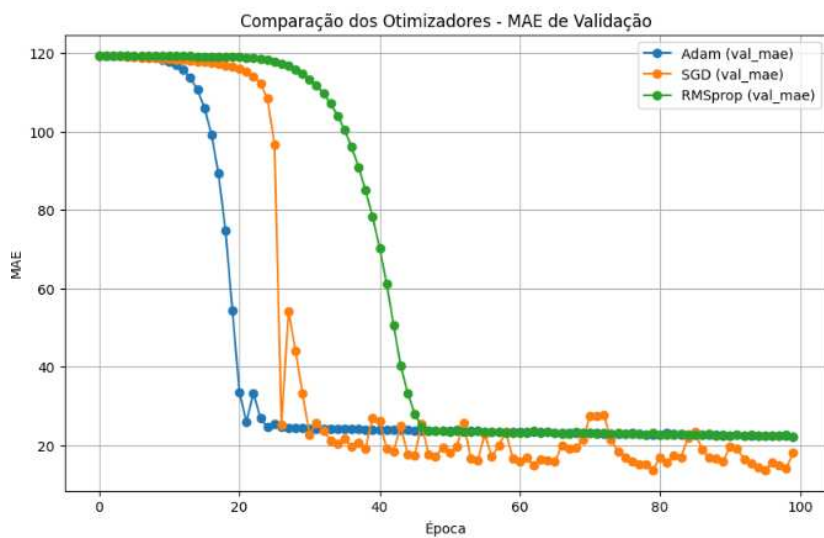


Fonte: Autor.

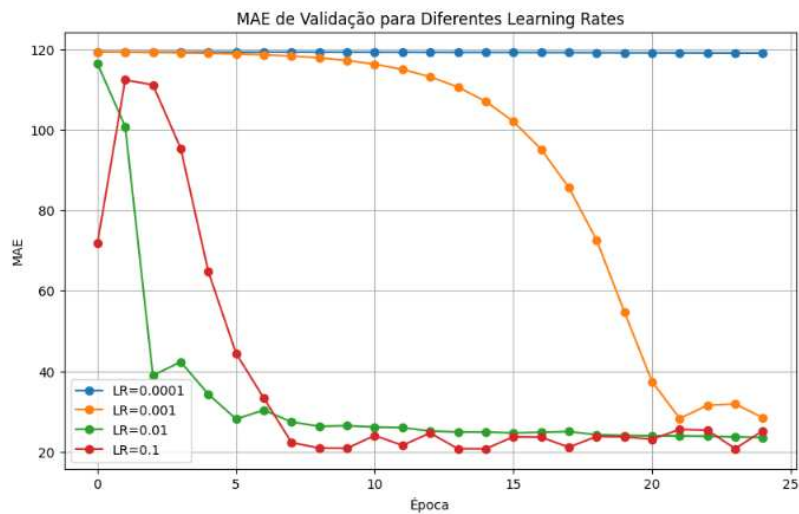
O melhor resultado obtido, conforme observado anteriormente, foi com o uso apenas das features extraídas, uma vez que elas ressaltam as características mais importantes do sinal. Para otimizar ainda mais os resultados, foram realizados testes variando diferentes parâmetros de treinamento, incluindo otimizadores, taxas de aprendizado (*Learning rates*) e *Batch Size*.

Os otimizadores testados foram *Adam*, *SGD* e *RMSprop*, as taxas de aprendizado testadas foram 0.0001, 0.001, 0.01 e 0.1, e os tamanhos de *Batch Size* avaliados foram 8, 16, 32, 64, 128 e 512. As Figuras 31, 32 e 33 apresentam os resultados obtidos nesses treinamentos.

Figura 31 – MAE de validação para diferentes otimizadores

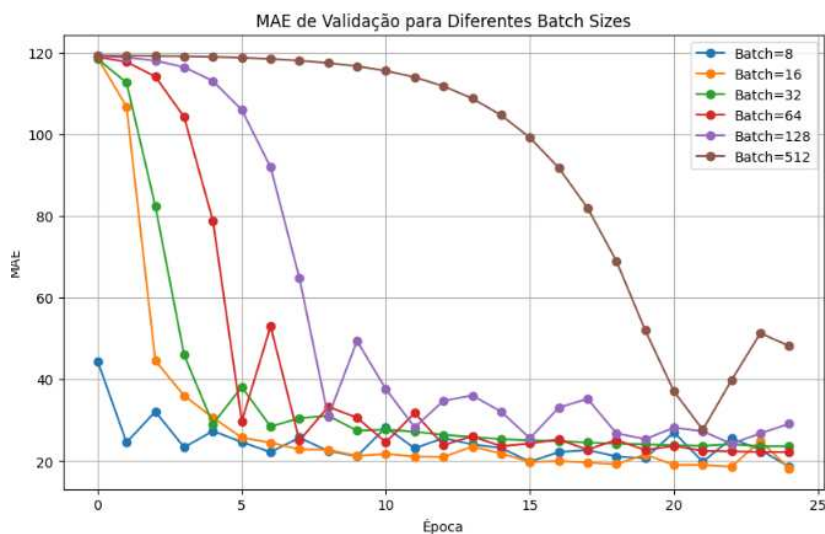


Fonte: Autor.

Figura 32 – MAE de validação para diferentes *Learning rates*

Fonte: Autor.

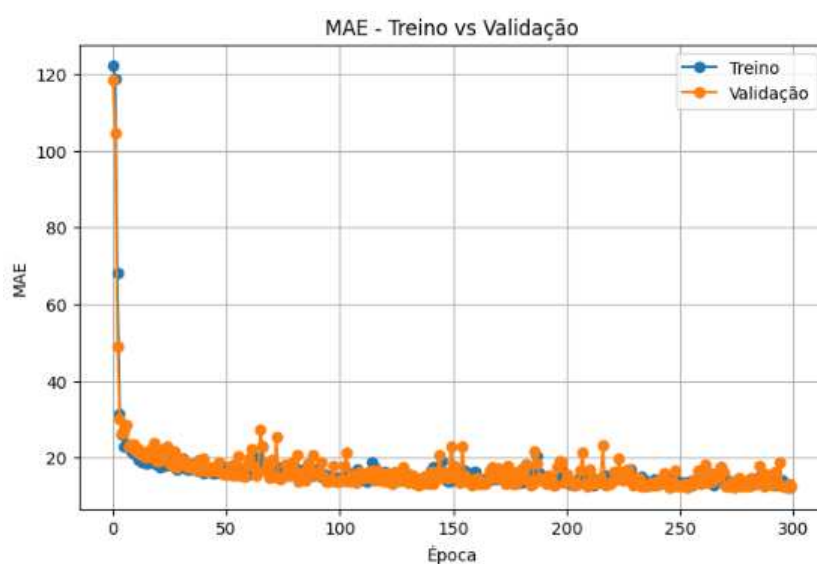
Figura 33 – MAE de validação para diferentes Batch Sizes



Fonte: Autor.

Analisando os resultados e comparando desempenho e estabilidade, foram escolhidos os seguintes parâmetros: otimizador *Adam*, *Learning rate* 0.01 e *Batch size* 16. Os resultados do treinamento final, utilizando esses melhores parâmetros, são apresentados na Figura 34, Tabela 8 e no Gráfico de Clarke a seguir.

Figura 34 – MAE treino x validação

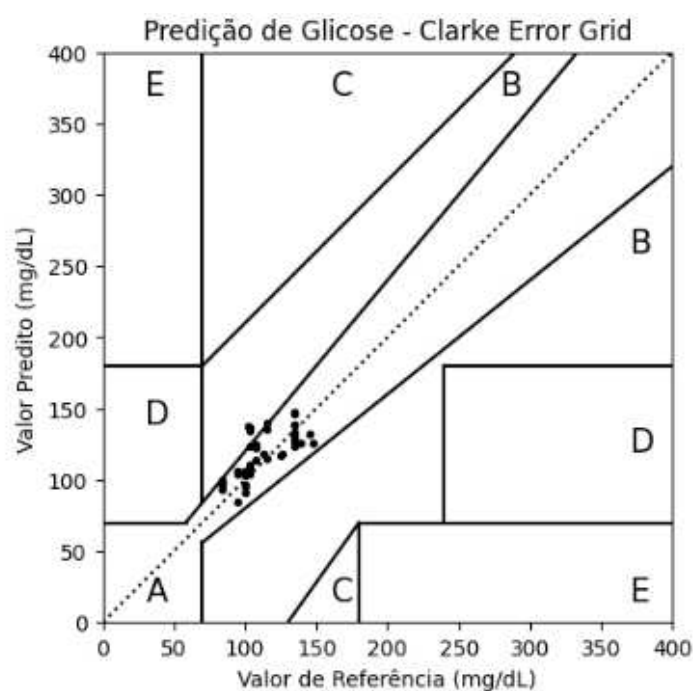


Fonte: Autor.

Tabela 8 – Métricas do treinamento final com os melhores parâmetros

Métrica	Valor
Mean Error (ME)	-6.461
Mean Absolute Error (MAE)	12.169
Mean Absolute Percentage Error (MAPE)	11.43%
Root Mean Squared Error (RMSE)	15.067
R ² (Coeficiente de Determinação)	0.363

Figura 35 – Gráfico de Clarke para a Rede Neural



Fonte: Autor.

Portanto, os resultados finais das métricas são apresentados na Tabela 11, e o modelo obteve 86% das estimativas na Zona A de Clarke.

6 DISCUSSÃO

Neste capítulo, serão discutidos os resultados obtidos, bem como possíveis melhorias, trabalhos futuros, dificuldades encontradas no projeto e a interpretação dos resultados.

A principal dificuldade enfrentada foi o elevado nível de ruído presente nos sinais, o que exigiu grande ênfase em um bom pré-processamento, filtragem e análise de qualidade. Um dos fatores que contribuíram para esse ruído foi o método de aquisição, os sinais foram coletados por meio de um dispositivo vestível, o relógio Galaxy Watch 6. Em comparação com o método clássico de aquisição de ECG, que utiliza eletrodos e um eletrocardiógrafo, o sinal coletado pelo relógio não possui a mesma qualidade. Entretanto, o objetivo do projeto era justamente investigar a viabilidade da coleta de ECG por um dispositivo prático, que o participante pudesse utilizar ao longo do dia sem dificuldades. Com o avanço das pesquisas na área, espera-se que novos dispositivos vestíveis ofereçam melhor qualidade na aquisição de ECG.

Outro ponto de destaque foi a análise da qualidade dos sinais. Durante o processamento dos dados, observou-se que diversos registros apresentavam baixa qualidade em todos os seus trechos, o que impossibilitou sua utilização e levou à exclusão desses sinais por completo. Essa análise abre espaço para uma aplicação prática importante, a implementação de sistemas automáticos de detecção de qualidade da coleta nos próprios dispositivos vestíveis. Dessa forma, o dispositivo poderia alertar o usuário imediatamente quando a medição estivesse inadequada e solicitar uma nova coleta, garantindo dados mais confiáveis e de melhor qualidade.

Outra limitação do projeto foi a dificuldade em obter valores de glicemia realmente baixos, por exemplo, abaixo de 70 mg/dL, faixa considerada hipoglicêmica. O conjunto de dados continha valores entre 72 mg/dL e 218 mg/dL. Assim, trabalhos futuros que incluam uma faixa mais ampla de níveis glicêmicos podem alcançar resultados mais robustos.

A tentativa de utilização de redes neurais também foi um ponto discutido no projeto. Embora essas técnicas sejam amplamente reconhecidas por seu alto desempenho em diversos problemas complexos, foi observado que, neste caso específico, elas não apresentaram resultados superiores aos métodos clássicos de aprendizado de máquina. Isso reforça a ideia de que redes neurais não são necessariamente a melhor solução para todos os problemas, especialmente quando a base de dados é limitada e apresenta ruído. Neste projeto, a combinação de uma boa extração de features e a aplicação de modelos clássicos resultou em um desempenho superior, demonstrando a importância de compreender profundamente o problema.

Por fim, como este banco de dados foi criado especificamente para o projeto, não há estudos anteriores que permitam uma comparação direta dos resultados. Ainda assim, mesmo com a existência de erro nos modelos, a capacidade de um dispositivo identificar se o participante está em hiperglicemia ou hipoglicemia já representa um avanço significativo.

Vale destacar que até mesmo o glicosímetro e sensores utilizado como referência possui uma tolerância de erro.

De acordo com a Sociedade Brasileira de Diabetes, o desempenho dos glicosímetros varia conforme o tipo de monitoramento. Para a monitorização de glicose no sangue capilar, os resultados devem apresentar acurácia de ± 15 mg/dL para valores laboratoriais inferiores a 100 mg/dL e $\pm 15\%$ para valores iguais ou superiores a 100 mg/dL, sendo que ao menos 99% das medições devem estar nas zonas A e B da análise de grade de erro, o que indica baixo impacto de medições equivocadas. Já para a monitorização contínua de glicose, realizada no fluido intersticial, os sensores devem garantir que mais de 87% das medições estejam dentro de $\pm 20\%$ dos valores de referência, assegurando confiabilidade adequada para acompanhamento das tendências glicêmicas e suporte à decisão clínica.

7 CONCLUSÃO

Este projeto apresentou um estudo baseado em sinais de eletrocardiograma (ECG) para a estimativa de níveis de glicose utilizando um dispositivo vestível não invasivo. Os resultados reforçam estudos recentes, que afirmam que variações nos níveis de glicose podem provocar alterações na morfologia do sinal de ECG, permitindo que tais variações sejam utilizadas como indicadores do estado glicêmico do indivíduo.

O conjunto de dados utilizado foi desenvolvido especificamente para este projeto e consistiu em 165 sinais de ECG coletados de 9 participantes, com níveis de glicose variando entre 72 mg/dL e 218 mg/dL. A partir desses registros, foi realizado um processo completo que incluiu pré-processamento, filtragem, normalização, análise visual, extração de características e avaliação da qualidade dos sinais. Entre as características morfológicas analisadas, as que apresentaram maior correlação com mudanças nos níveis de glicose foram a amplitude da onda T e a distância entre picos R (intervalo RR).

Diferentes modelos de aprendizado de máquina foram treinados e avaliados. Os melhores desempenhos foram obtidos pelos modelos *Random Forest* e *Gradient Boosting*. O modelo *Random Forest* apresentou erro RMSE de 11.295, MAPE de 8.63% e 93% das estimativas classificadas na Zona A do Gráfico de Clarke, indicando alta precisão clínica. O *Gradient Boosting* apresentou RMSE de 11.768, MAPE de 9,09% e 84% das estimativas na Zona A. O modelo *CatBoost* também obteve desempenho competitivo, com RMSE de 12.512, MAPE de 9,53% e 81% das estimativas na Zona A. Esses resultados evidenciam que, mesmo utilizando um dispositivo vestível com limitações de qualidade de sinal, foi possível alcançar modelos com desempenho adequado para aplicações práticas de monitoramento.

Além de contribuir com o desenvolvimento de metodologias para estimar glicemia de forma não invasiva, este estudo contribui com a literatura ao demonstrar como sinais fisiológicos obtidos no cotidiano podem auxiliar no monitoramento contínuo da saúde. A pesquisa reforça o potencial dos dispositivos vestíveis em promover bem-estar, conforto e autonomia ao usuário, reduzindo a necessidade de medições invasivas e oferecendo suporte adicional ao acompanhamento clínico.

REFERÊNCIAS

- ACHARYA, Deepshikha et al. Application of adaptive Savitzky–Golay filter for EEG signal processing. **Perspectives in Science**, v. 8, p. 677–679, 2016. Recent Trends in Engineering and Material Sciences. ISSN 2213-0209.
- ALEXAKIS, C. et al. Feature Extraction and Classification of Electrocardiogram (ECG) Signals Related to Hypoglycaemia. In: **COMPUTERS in Cardiology**. Greece: [s.n.], 2003. p. 537–540.
- ANDELLINI, Martina et al. Are the variations in ECG morphology associated to different blood glucose levels? Implications for non-invasive glucose monitoring for T1D paediatric patients. **Diabetes Research and Clinical Practice**, v. 212, p. 111708, 2024. ISSN 0168-8227.
- APPATHURAI, Ahilan. A Study on ECG Signal Characterization and Practical Implementation of Some ECG Characterization Techniques. **Measurement**, Elsevier BV, 2019.
- ARBI, K. Fellah et al. Blood glucose estimation based on ECG signal. **Physical and Engineering Sciences in Medicine**, v. 46, n. 1, p. 255–264, 2023. DOI: 10.1007/s13246-022-01214-3.
- BENT, B.; CHO, P. J.; HENRIQUEZ, M. et al. Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches. **NPJ Digital Medicine**, v. 4, n. 1, p. 89, 2021. DOI: 10.1038/s41746-021-00465-w.
- CORONADO-REYES, O.I. et al. Non-invasive Hyperglycemia Detection via Electrocardiogram Using Discrete Wavelet Transform and Machine Learning. **Cureus**, v. 17, n. 3, e80548, mar. 2025. DOI: 10.7759/cureus.80548.
- DEUS, Luis Felipe de. **Estimativa dos níveis de glicose através do processamento de sinais de fotopletismografia (PPG) e modelos de inteligência artificial**. 2023. Dissertação de Mestrado – Centro Tecnológico, Programa de Pós-Graduação em Engenharia Elétrica, Florianópolis.
- DOMBI, József; DINEVA, Adrienn. Adaptive Savitzky-Golay filtering and its applications. **International Journal of Advanced Intelligence Paradigms**, v. 16, p. 145, 2020. DOI: 10.1504/IJAIP.2020.107011.
- FELLAH ARBI, Khadidja; SOULIMANE, Sofiane; SAFFIH, Faycal. Non-invasive method for blood glucose monitoring using ECG signal. **Polish Journal of Medical Physics And Engineering**, v. 29, n. 1, p. 1, jan. 2023. DOI: 10.2478/pjmpe-2023-0001.
- FRIEDMAN, Jerome. Stochastic Gradient Boosting. **Computational Statistics & Data Analysis**, v. 38, p. 367–378, 2002. DOI: 10.1016/S0167-9473(01)00065-2.

GALLAGHER, Neal B. **Savitzky–Golay Smoothing and Differentiation Filter**. 196 Hyacinth Road, Manson, WA 98831, jan. 2020. DOI: 10.13140/RG.2.2.20339.50725. Disponível em: <https://eigenvector.com/wp-content/uploads/2020/01/SavitzkyGolay.pdf>.

GONG, Youdi et al. A Survey on Dataset Quality in Machine Learning. **Information and Software Technology**, v. 162, p. 107268, 2023. ISSN 0950-5849. DOI: 10.1016/j.infsof.2023.107268. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950584923001222>.

GUO, G. et al. KNN Model-Based Approach in Classification. In: MEERSMAN, R.; TARI, Z.; SCHMIDT, D. C. (ed.). **On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE**. Berlin, Heidelberg: Springer, 2003. v. 2888. (Lecture Notes in Computer Science). DOI: 10.1007/978-3-540-39964-3_62. Disponível em: https://doi.org/10.1007/978-3-540-39964-3_62.

HANCOCK, Jay T.; KHOSHGOFTAAR, Taghi M. CatBoost for big data: an interdisciplinary review. **Journal of Big Data**, v. 7, n. 1, p. 94, 2020. DOI: 10.1186/s40537-020-00369-8.

HESAR, Hamed Danandeh; HESAR, Amin Danandeh. Efficient Bayesian ECG denoising using adaptive covariance estimation and nonlinear Kalman Filtering. **Computers and Electrical Engineering**, v. 120, p. 109869, 2024. ISSN 0045-7906.

HOSSEINZADEHKETILATEH, MohammadReza; ADAMI, Banafsheh; KARIMIAN, Nima. Advancements in Continuous Glucose Monitoring: Integrating Deep Learning and ECG Signal. Morgantown, USA, 2024. arXiv:2403.07296 [eess.SP]. Disponível em: <https://arxiv.org/abs/2403.07296>.

ISLAM, Tanvir Tazul et al. Blood Glucose Level Regression for Smartphone PPG Signals Using Machine Learning. **Applied Sciences**, MDPI, Basel, Switzerland, v. 11, n. 2, p. 618, 2021. DOI: 10.3390/app11020618.

KAPLAN BERKAYA, Selcan et al. A survey on ECG analysis. **Biomedical Signal Processing and Control**, v. 43, p. 216–235, 2018. DOI: 10.1016/j.bspc.2018.03.003.

LIU, Y.; WANG, Y.; ZHANG, J. New Machine Learning Algorithm: Random Forest. In: LIU, B.; MA, M.; CHANG, J. (ed.). **Information Computing and Applications**. Berlin, Heidelberg: Springer, 2012. v. 7473. (Lecture Notes in Computer Science). DOI: 10.1007/978-3-642-34062-8_32. Disponível em: https://doi.org/10.1007/978-3-642-34062-8_32.

MEJÍA-MEJÍA, E. et al. Heart Rate Variability (HRV) and Pulse Rate Variability (PRV) for the Assessment of Autonomic Responses. **Frontiers in Physiology**, v. 11, p. 779, 2020. DOI: 10.3389/fphys.2020.00779.

MEJÍA-MEJÍA, Elisa et al. Photoplethysmography signal processing and synthesis. In: **PHOTOPLETHYSMOGRAPHY: Technology, Signal Analysis, and Applications**. [S.l.]:

Elsevier, 2022. p. 69–146. ISBN 9780128233740. DOI: 10.1016/B978-0-12-823374-0.00015-3.

NGUYEN, L.; SU, S.; NGUYEN, H. T. Effects of Hyperglycemia on Variability of RR, QT and Corrected QT Intervals in Type 1 Diabetic Patients. In: PROCEEDINGS of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). [S.l.: s.n.], 2013. p. 1819–1822. DOI: 10.1109/EMBC.2013.6609876.

NOBLE, William S. What is a Support Vector Machine? **Nature Biotechnology**, v. 24, n. 12, p. 1565–1567, 2006. DOI: 10.1038/nbt1206-1565.

NTI, Isaac; NYARKO-BOATENG, Owusu; ANING, Justice. Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation. **International Journal of Information Technology and Computer Science**, v. 6, p. 61–71, 2021. DOI: 10.5815/ijitcs.2021.06.05.

OLIVEIRA, Filipe A. C. et al. Machine Learning-Based Diabetes Detection Using Photoplethysmography Signal Features. In: ANAIS do 24^a Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS). Goiânia, GO: Sociedade Brasileira de Computação, 2024. p. 94–105. DOI: 10.5753/sbcas.2024.1889.

PALERMI, Stefano et al. Artificial intelligence and the electrocardiogram: A modern renaissance. **European Journal of Internal Medicine**, v. 140, p. 106329, 2025. ISSN 0953-6205.

PANDEY, Harshita; TIWARI, Rajinder. An Innovative Design Approach of Butterworth Filter for Noise Reduction in ECG Signal Processing based Applications. **Progress In Science and Engineering Research Journal (PISER)**, PISER Journal, v. 2, n. 2/06, p. 332–337, 2014. ISSN 2347-6680. Disponível em: <http://piserjournal.org/>.

SANAI, Farhan; SAHID, A. S.; HUVANANDANA, J. et al. Evaluation of a Continuous Blood Glucose Monitor: A Novel and Non-Invasive Wearable Using Bioimpedance Technology. **Journal of Diabetes Science and Technology**, v. 17, n. 2, p. 336–344, 2023. DOI: 10.1177/19322968211054110.

SATTER, Shama et al. EMD-Based Noninvasive Blood Glucose Estimation from PPG Signals Using Machine Learning Algorithms. **Applied Sciences**, v. 14, n. 4, 2024. DOI: 10.3390/app14041406.

SU, Xian; YAN, Xiang; TSAI, Chih-Ling. Linear Regression. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 4, n. 3, p. 275–294, 2012. DOI: 10.1002/wics.1198.

SUN, Yan; CHAN, Kap Luk; KRISHNAN, Shankar Muthu. ECG signal conditioning by morphological filtering. **Computers in Biology and Medicine**, v. 32, n. 6, p. 465–479, 2002. ISSN 0010-4825.

TOBORE, I. et al. Statistical and spectral analysis of ECG signal towards achieving non-invasive blood glucose monitoring. **BMC Medical Informatics and Decision Making**, v. 19, Suppl 6, p. 266, 2019. DOI: 10.1186/s12911-019-0959-9.

WANG, R. et al. A Systematic Review on the Use of Consumer-Based ECG Wearables on Cardiac Health Monitoring. **IEEE Journal of Biomedical and Health Informatics**, v. 28, n. 11, p. 6525–6537, 2024. DOI: 10.1109/JBHI.2024.3456028.

APÊNDICE A –

Quadro 2 – Descrição dos Hiperparâmetros otimizados.

C	Controla o equilíbrio entre margem larga (generalização) e poucos erros (complexidade).
gamma (SVM)	Controla o alcance da influência de cada ponto de treino.
kernel	Função que transforma os dados em um espaço de maior dimensão para torná-los linearmente separáveis.
n_estimators	Número de árvores que compõem a floresta.
max_depth	Profundidade máxima de cada árvore.
min_samples_split	Número mínimo de amostras necessárias para dividir um nó.
min_samples_leaf	Número mínimo de amostras exigidas em uma folha.
max_features	Número máximo de features consideradas em cada divisão de nó.
learning_rate	Taxa que controla o quanto o modelo ajusta seus parâmetros a cada iteração.
subsample	Fração de amostras usadas para treinar cada árvore.
n_neighbors	Número de vizinhos mais próximos usados para determinar a classe de um ponto.
weights	Fine como os vizinhos contribuem para a classificação.
p	Define a métrica de distância usada.
iterations	Número de árvores (ou iterações) que o modelo irá construir durante o treinamento.
depth	Profundidade máxima de cada árvore.
l2_leaf_reg	Regularização L2 aplicada às folhas das árvores para reduzir overfitting.
bagging_temperature	controla a aleatoriedade no subsampling das árvores.
border_count	Número de divisões binárias usadas para transformar features contínuas em discretas.
colsample_bytree	Fração de features escolhidas aleatoriamente para cada árvore.
gamma (XGBoost)	Define o ganho mínimo necessário para realizar uma divisão em uma árvore.
lambda	Parâmetro de regularização L2 para os pesos das folhas das árvores.
alpha	Parâmetro de regularização L1 para os pesos das folhas das árvores.
booster	Tipo de algoritmo de boosting usado.
rate_drop	Probabilidade de descartar uma árvore durante o treinamento.
max_delta_step	Limite máximo de incremento em cada árvore.

Fonte: Elaborada pelo autor (2025).