PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Louise Erthal Rabelo Parente

Métodos de Extração de Característica e Aprendizado de Máquina para Avaliação da Qualidade de Cimento

Projeto de Graduação

Projeto de Graduação apresentado ao Departamento de Engenharia Mecânica da PUC-Rio

Orientador: Helon Vicente Hultmann Ayala
Coorientador: Arthur Martins Barbosa Braga

Rio de Janeiro
Junho de 2023

Agradecimentos

Agradeço ao meu orientador e ao meu coorientador.

Agradeço à PUC, às agências ANP e CNPq e ao projeto TTiLT pelo apoio concedido no desenvolvimento deste trabalho.

Resumo

Métodos de Extração de Característica e Aprendizado de Máquina para Avaliação da Qualidade de Cimento

As operações de Tamponamento e Abandono (P&A) devem ser realizadas ao final da vida útil do poço, de modo a garantir a preservação do meio ambiente. Uma etapa essencial nas operações de P&A é a avaliação da qualidade do cimento, que comumente é feita por meio de técnicas de perfilagem acústica baseadas na propagação de ondas sônicas ou ultrassônicas, onde é verificada a integridade da camada de cimento e sua capacidade de vedação. Atualmente, esse procedimento exige a retirada prévia da coluna de produção, o que aumenta os custos e a complexidade dessas operações. Assim, é de grande interesse para a indústria de óleo e gás desenvolver ferramentas e técnicas capazes de realizar a avaliação do cimento sem a retirada da tubulação. Portanto, este trabalho visa aplicar métodos de extração de características para processar sinais obtidos por meio de experimentos de perfilagem acústica em poços de petróleo com a presença da coluna de produção, e, em seguida, utilizar as características provenientes para treinar e testar redes neurais do tipo *Multi-layered Perceptron* (MLP). Resultados promissores foram obtidos para a tarefa de classificação de defeito, com acurácia de quase todos os métodos acima de 0,8 e AUC acima de 0,9.

Palavras-chave

Aprendizado de Máquina; Extração de Características; Operações de Tamponamento e Abandono (P&A); Perfilagem Acústica; Avaliação de Integridade Estrutural.

Abstract

Feature Extraction and Machine Learning Methods for Assessing Cement Qualit

Plugging and Abandonment (P&A) operations must be carried out at the end of the well's life cycle, to guarantee the preservation of the environment. An essential step in P&A operations is the evaluation of cement quality, done through acoustic logging techniques based on the propagation of sonic or ultrasonic waves, where the integrity of the cement layer and its sealing capacity are verified. Currently, this procedure requires prior removal of the production column, which increases the costs and complexity of these operations. Thus, it is of great interest to the oil and gas industry to develop tools and techniques capable of performing cement evaluation without removing the tubing. Therefore, this work aims to apply feature extraction methods to process signals obtained through acoustic logging experiments in oil wells with the presence of the production column, and then use the resulting features to train and test neural networks of the Multi-layered Perceptron (MLP) type. Promising results were obtained for the defect classification task, with an accuracy of almost all methods above 0,8 and an AUC above 0,9.

Keywords

Machine Learning; Feature Extraction; Plugging and Abandonment Operations (P&A); Acoustic Logging; Structural Integrity Assessment.

Sumário

Lista de figuras

Lista de tabelas

1 INTRODUÇÃO

As operações de tamponamento e abandono (P&A – *Plugging and Abandonment*) devem ser realizadas ao final do ciclo de vida de um poço, a fim de garantir a preservação do meio ambiente. Para o abandono permanente de um poço, é estabelecido no Sistema de Gerenciamento da Integridade de Poços (SGIP), instituído pela resolução ANP nº 46/2016, que são necessários no mínimo dois conjuntos solidários de barreira (CSB) permanentes, nos quais cimento ou outro material de desempenho similar devem ser usados como elementos de barreira. Ainda de acordo com essa resolução, o CSB permanente tem como objetivo impedir o fluxo não intencional atual e futuro de fluidos da formação, considerando todos os caminhos possíveis.

Uma importante etapa das operações de P&A é a avaliação da qualidade do cimento, que é comumente feita por meio de técnicas de perfilagem acústica. Estas técnicas baseiam-se na propagação de ondas sônicas ou ultrassônicas, permitindo verificar a integridade da ligação do cimento e sua capacidade de vedação.

Apesar do bom desempenho, as ferramentas de perfilagem acústica atuais só possuem tecnologia para serem aplicadas em poços sem a coluna de produção e com apenas uma coluna de revestimento (QI et al., 2017). Esta limitação aumenta a complexidade das operações de P&A, gerando maiores custos e prolongando o tempo necessário. Por isso, há uma demanda significativa na indústria de petróleo e gás por ferramentas e técnicas que permitam avaliar a qualidade do cimento sem a remoção da tubulação de produção.

Desse modo, esse trabalho tem como objetivo aplicar técnicas de aprendizado de máquina e de extração de características no processamento de sinais experimentais de perfilagem acústica executada com a presença da coluna de produção, avaliando se o cimento está nas condições adequadas e, caso contrário, identificando o tipo de defeito encontrado.

2 REVISÃO BIBLIOGRÁFICA

Com o objetivo de identificar o estado atual da pesquisa em relação à utilização de técnicas de perfilagem acústica para avaliar a qualidade do cimento e de métodos de aprendizado de máquina para avaliação de integridade estrutural, realizou-se uma revisão da literatura. A seguir, os estudos mais relevantes e recentes sobre esses assuntos estão sintetizados.

Viggen, Johansen e Merciu (2016) realizaram uma análise pioneira do uso da técnica de perfilagem acústica com pulso-eco ultrassônico através de múltiplos revestimentos, por meio de simulações feitas pelo método de elementos finitos. A fim de analisar inicialmente o caso mais simples possível, considerou-se uma situação composta por dois revestimentos, com a presença de água no anular A, nome dado ao espaço anular entre os revestimentos. Além disso, o transdutor emitia um pulso bem-comportado e centralizado. Para analisar essa configuração, levou-se em conta os chamados *first interface echo* (FIE) e *third interface echo* (TIE), sendo o primeiro referente às interações dos pulsos acústicos com o revestimento interno que reverberam dentro do fluido que preenche a tubulação, enquanto o segundo está associado aos pulsos acústicos que ultrapassam o revestimento interno e reverberam dentro do fluido que preenche o anular A. Assim, foram testadas as variações produzidas no TIE ao fazer alterações em três parâmetros: o material no anular B (nome dado ao espaço anular entre o segundo revestimento e a formação), a espessura do revestimento externo e a excentricidade do revestimento externo. Após o estudo, foi constatado que esses parâmetros fornecem uma variação muito pequena do TIE, tornando inviável o uso da técnica de pulso-eco ultrassônico para a determinação do material presente no anular B ou da espessura do revestimento externo. Além disso, nas simulações foram negligenciados vários mecanismos de atenuação, que, em casos reais, enfraqueceriam ainda mais os sinais recebidos. Por fim, conclui-se que essa não é a técnica mais adequada para a perfilagem acústica de poços através de tubos.

Em contraste, em outro estudo de Viggen, Johansen e Merciu (2016), através de simulações feitas com base no método de elementos finitos, foi explorado o uso da perfilagem ultrassônica pitch-catch através de múltiplos revestimentos. Considerando novamente o caso mais simples descrito anteriormente, houve a propagação de ondas

de Lamb na tubulação de produção com vazamento para o interior da tubulação e para o anular A. As frentes de onda que vazaram para o interior da tubulação foram registradas por dois receptores acústicos. Assim, foram testadas as alterações produzidas nos resultados para três casos: variando o material do anular B, variando a distância entre os revestimentos e variando a espessura do tubo externo e do material do anular B. Após as simulações serem concluídas, foi possível observar que a qualidade da ligação entre o material no anular B e a formação pode ser detectada através da amplitude do pulso secundário. Além disso, constatou-se que a variação da distância entre os revestimentos apresentou resultados coerentes com a literatura já existente. Por fim, concluiu-se que a variação da espessura do tubo externo pôde ser detectada corretamente. Dessa forma, este método se mostrou promissor para avaliar a qualidade do cimento na presença da tubulação de produção.

Já no estudo de Viggen, Merciu, Løvstakken e Måsø (2020) foi proposto o uso de aprendizado profundo para a avaliação do cimento, investigando a qualidade de sua ligação com o revestimento, assim como sua capacidade de isolamento hidráulico, por meio da interpretação de perfis acústicos. O algoritmo fez uso de redes neurais convolucionais profundas, treinadas de forma supervisionada por um conjunto de dados composto por interpretações manuais de 54 operações de perfilagem acústica. Os resultados obtidos pelo método de aprendizado profundo atingiram boas métricas de precisão, indicando ser promissor, embora os autores acreditem que possam ser melhorados aumentando o banco de dados de treinamento e diminuindo a subjetividade na avaliação dos dados utilizados.

O trabalho de Dworakowski, Dragan e Stepinski (2016) expandiu a aplicação do aprendizado de máquina para o monitoramento da integridade estrutural de aeronaves utilizando um conjunto de redes neurais artificiais (RNAs) para o processamento de sinais de transdutores piezoelétricos (PZT). Diversos tipos de RNAs com diferentes estruturas foram utilizadas, como *Multi-layered Perceptron* (MLP), *Self-organizing Map* (SOM) e *Radial Basis Function* (RBF). As redes foram avaliadas e as 10 com melhores valores de erro quadrático médio (MSE) foram incluídas no conjunto final. Analisando três casos distintos, o conjunto de RNAs apresentou excelentes resultados de MSE, sendo melhores do que aqueles produzidos ao aplicar apenas uma RNA. Assim, foi concluído que, principalmente para

casos mais complexos, utilizar um conjunto de redes neurais artificiais aumenta a confiabilidade nos resultados do monitoramento executado.

Por fim, no estudo de Melville, Alguri, Deemer e Harley (2018) foram usadas técnicas de aprendizado profundo para o monitoramento de integridade estrutural usando ondas ultrassônicas guiadas, uma vez que essas técnicas são capazes de mapear uma relação arbitrariamente complexa entre o sinal e a classificação dos danos. Para a aquisição de dados, foram considerados quatro casos, onde diferentes placas de metal foram usadas, variando o material e a espessura, além de serem simuladas variações na temperatura. Comparando os resultados obtidos, o aprendizado profundo obteve acurácia muito superior em comparação a uma *Support Vector Machine* (SVM) linear tradicional. Além disso, os autores afirmam que menos dados de treinamento são necessários no aprendizado profundo para obter previsões rápidas e precisas, fazendo com que essa técnica seja uma boa abordagem para o problema proposto.

A partir da análise dos trabalhos mencionados, é possível concluir que a aplicação conjunta de técnicas de perfilagem acústica e modelos de aprendizado de máquina tem se destacado na área de monitoramento de integridade estrutural. A evidência sugere que a integração dessas duas abordagens representa um campo de pesquisa promissor, principalmente quando aplicado à avaliação da qualidade do cimento.

Neste contexto, o propósito principal deste estudo é ampliar a compreensão desta área emergente. Para isso, esse trabalho se concentra em explorar diversas combinações de métodos de extração de características e modelos de aprendizado de máquina, usando dados de perfilagem acústica coletados experimentalmente.

3 MÉTODOS DE EXTRAÇÃO DE CARACTERÍSTICAS

Devido à grande quantidade de variáveis provenientes de uma série temporal obtida através dos experimentos de perfilagem acústica, é muito importante executar uma extração de características no conjunto de dados antes que esse seja fornecido ao modelo preditivo.

Esse procedimento garante uma redução da dimensionalidade dos dados, mantendo as informações mais relevantes, o que facilita sua visualização e sua compreensão. Além disso, com uma menor quantidade de atributos, as etapas de criação e treinamento do modelo se tornam mais rápidas e eficazes.

Os métodos de extração de características utilizados neste trabalho são apresentados a seguir.

3.1 Principal Component Analysis (PCA)

O método de extração de características conhecido como *Principal Component Analysis* (PCA) transforma as variáveis em um novo conjunto de atributos não-correlacionados, de tal forma a reter, de maneira ordenada, o máximo da variação contida nos dados originais (JOLLIFFE, 2002). A técnica é matematicamente baseada na *Singular Value Decomposition* (SVD) em que uma matriz X pode ser decomposta como (GOLUB et al., 2013):

$$\boldsymbol{X}_{n\times m} = \boldsymbol{U}_{n\times n}\boldsymbol{\Sigma}_{n\times m}\boldsymbol{V}^{T}{}_{m\times m} \tag{1}$$

Onde n é o número de amostras e m é o número de atributos iniciais que compõem o conjunto de dados de interesse, $\boldsymbol{X}$, que pode ser expandido por:

$$\boldsymbol{X}_{n\times m} = \begin{bmatrix} | & | & | \\ u_1 & \cdots & u_n \\ | & | & | \end{bmatrix}\begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & \sigma_p & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} - & v_1 & - \\ - & \vdots & - \\ - & v_m & - \end{bmatrix}^{T} \tag{2}$$

No qual $\boldsymbol{U_{n\times n}}$ e $\boldsymbol{V_{m\times m}}$ são ortogonais e Σ é uma matriz diagonal composta pelos valores singulares, de modo que $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$. Assim, a matriz $\boldsymbol{P}$ que contém os

componentes principais, *principal components* (PCs), do conjunto de dados é calculada por:

$$\boldsymbol{P} = \boldsymbol{X}.\boldsymbol{V}^T = \boldsymbol{X}_{n\times m}.\begin{bmatrix} | & | & | & | \\ v_1 & v_2 & \dots & v_p \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} PC_{1,1} & PC_{1,2} & \dots & PC_{1,q} \\ PC_{2,1} & PC_{2,2} & \dots & PC_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ PC_{n,1} & PC_{n,2} & \dots & PC_{n,q} \end{bmatrix} \quad (3)$$

A dimensão q da matriz de PCs pode ser definida com base na porcentagem de variância explicada de interesse, determinada por um limite desejado (δ), e é tipicamente menor que m, devendo satisfazer a seguinte relação:

$$\frac{\sum_{i=1}^{q} \sigma_i^2}{\sum_{i=1}^{p} \sigma_i^2} \leq \delta \quad (4)$$

3.2 Factor Analysis (FA)

Factor Analysis (FA) é um modelo de variável latente linear-gaussiana para reduzir a dimensionalidade de um conjunto de dados (BARBER, 2012). Neste modelo, as variáveis observadas $x_1, \dots, x_n$ podem ser escritas como:

$$x_i = Fh_i + c + e \quad (5)$$

Onde e é considerado um termo de ruído gaussiano distribuído ($e \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\psi})$), a constante c define a origem do sistema de coordenadas, F é a matriz de carregamento fatorial e h_i é o vetor latente. Assim, pode-se escrever este modelo na forma matricial, como segue:

$$\boldsymbol{X} = \boldsymbol{FH} + \boldsymbol{C} + \boldsymbol{E} \quad (6)$$

Além disso, ψ é dado por:

$$\boldsymbol{\psi} = diag(\psi_1, \dots, \psi_n) \quad (7)$$

A partir dessas equações, dado o vetor latente h_i, tem-se a seguinte interpretação probabilística:

$$p(x_i|h_i) = \mathcal{N}(Fh_i + c, \boldsymbol{\psi}) \tag{8}$$

Por fim, a distribuição da variável latente h é dada por:

$$p(\boldsymbol{h}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{9}$$

3.3 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) consiste na obtenção de um novo conjunto de componentes estatisticamente independentes a partir de uma transformação linear dos dados iniciais (HYVÄRINEN et al., 2000). Assim, é possível escrever cada uma das n variáveis observadas $x_1, \dots, x_n$ como uma combinação linear de n componentes independentes $s_1, \dots, s_n$, por:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \ \dots \ + \ a_{jn}s_n \tag{10}$$

Onde a_{ij} são os coeficientes de combinação. Denotando A como a matriz formada pelos elementos a_{ij}, pode-se representar o sistema na forma de matricial por:

$$\boldsymbol{x} = \boldsymbol{As} \tag{11}$$

Uma vez encontrada a matriz A, sua inversa W é calculada e os componentes independentes são obtidos por:

$$\boldsymbol{s} = \boldsymbol{Wx} \tag{12}$$

3.4 Isometric Feature Mapping (Isomap)

Uma das várias abordagens de extração de características é o *Isometric Feature Mapping* (Isomap), que é baseado em *Multidimensional Scaling* (MDS), preservando

a geometria intrínseca dos dados, uma vez que mantem a distância geodésica entre todos os pontos (TENENBAUM et. al, 2000).

O primeiro passo deste algoritmo é definir o grafo G, conectando os pontos do conjunto de dados que são vizinhos, com base nas distâncias euclidianas $d_x(x_i, x_j)$ entre dois pontos x_i e x_j. Isso pode ser feito conectando a cada ponto todos os seus K vizinhos mais próximos, definindo comprimentos de arestas iguais a $d_x(x_i, x_j)$.

Na segunda etapa, a distância geodésica $d_G(x_i, x_j)$ entre todos os pares de pontos é estimada pela menor distância entre os pontos x_i e x_j no grafo G. As distâncias calculadas são armazenadas na matriz $D_G = \{d_G(x_i, x_j)\}$. Na última etapa, o espaço de baixa dimensionalidade Y é construído e suas coordenadas y_i são definidas por:

$$y_i = \sqrt{\lambda_p \, v_p^i} \tag{13}$$

Onde λ_p é o p-ésimo autovalor em ordem decrescente da matriz $\tau(D_G)$, v_p^i é a i-ésima componente do p-ésimo autovetor, e o operador τ tem a função de converter distâncias em produtos internos.

3.5 Locally Linear Embedding (LLE)

Enquanto o Isomap tenta preservar a distância geodésica entre todos os pares de pontos em um conjunto de dados, o *Locally Linear Embedding* (LLE) recupera a estrutura não linear global de ajustes lineares localmente, eliminando a necessidade de estimar a distância entre pontos muito distantes (ROWEIS et. al, 2000).

O objetivo do algoritmo é construir estruturas lineares localmente capazes de reconstruir cada ponto a partir de seus vizinhos. Assim, na primeira etapa, devem ser definidos os K vizinhos mais próximos de cada ponto x_i, assim como feito para o Isomap. Na segunda etapa, os pesos W_{ij} que melhor reconstroem linearmente o ponto x_i a partir de seus vizinhos devem ser calculados, resolvendo o problema dos mínimos quadrados restrito:

$$min \left\| x_i - \sum_j W_{ij} x_j \right\| \quad (14)$$

Como os pesos W_{ij} determinam a contribuição do j-ésimo ponto de dados para a i-ésima reconstrução, a função está sujeita a duas restrições: i) W_{ij}= 0, se x_j não for um dos x_i vizinhos; (ii) $\sum_j W_{ij} = 1$. Na última etapa, uma vez conhecidos os valores dos pesos, o ponto x_i é mapeado para o ponto y_i, em um espaço de baixa dimensionalidade. Isso é feito escolhendo as coordenadas y_i para que a seguinte função de custo seja minimizada:

$$\Phi(y) = \sum_i \left| y_i - \sum_j W_{ij} y_j \right|^2 \quad (15)$$

3.6 Modified Locally Linear Embedding (MLLE)

Modified Locally Linear Embedding (MLLE) segue etapas semelhantes às do LLE padrão, mas em vez de calcular um único peso para cada ponto, vários vetores de peso local são calculados, o que garante uma incorporação aprimorada e mais estável (ZHANG et. al, 2006). Os pesos múltiplos $W_{ij}^{(l)}$ são obtidos resolvendo o problema dos mínimos quadrados restrito, mostrado na equação (14). Conhecendo o valor dos pesos, as coordenadas y_i em um espaço de baixa dimensionalidade devem ser escolhidas para minimizar a seguinte função de custo:

$$\Phi(y) = \sum_i \sum_l \left| y_i - \sum_j w_{ij}^{(l)} y_j \right|^2 \quad (16)$$

3.7 Características estatísticas e temporais

Por fim, na última abordagem é feita a extração das características temporais e estatísticas dos sinais acústicos. Esse procedimento pode ser realizado por meio da biblioteca *Time Series Features Extraction Library* (TSFEL), capaz de extrair mais de 60 características diferentes, nos domínios espectral, estatístico e temporal

(BARANDAS et al., 2020). Além disso, após ser feita a extração, essa biblioteca permite que atributos altamente correlacionados sejam removidos.

Ao estudar uma série temporal, é possível fazer uma análise com base nas características estatísticas do sinal. Essa análise é importante porque fornece informações sobre a estrutura de autocorrelação, sazonalidade, tendências, não linearidade dos dados, entre outras. Usando a biblioteca TSFEL e escolhendo o domínio estatístico, são extraídas dezessete características, que estão descritas na Tabela 1.

Tabela 1 – Características estatísticas extraídas usando TSFEL (BARANDAS et al., 2020).

Característica	**Descrição**	**Implementação**
ECDF	Valores de ECDF (função de distribuição cumulativa empírica) ao longo do eixo do tempo.	Implementado de acordo com (RASCHKA, 2018)
ECDF percentile	Valor percentual do ECDF.	Implementado de acordo com (RASCHKA, 2018)
ECDF percentile count	Soma cumulativa de amostras que são menores que o percentil.	Implementado de acordo com (RASCHKA, 2018)
ECDF slope	Inclinação do ECDF entre dois percentis.	Implementado de acordo com (RASCHKA, 2018)
Histogram	Histograma do sinal.	$n = \sum_{i=1}^{k} m_i$
Interquartile range	Variação interquartil do sinal.	$Q_3 - Q_1$
Kurtosis	Curtose do sinal.	Implementado de acordo com (BARANDAS *et al.*, 2020)
Max	Valor máximo do sinal.	Implementado de acordo com (KOKOSKA; ZWILLINGER, 1999)
Mean	Valor médio do sinal.	Implementado de acordo com (KOKOSKA; ZWILLINGER, 1999)
Mean absolute deviation	Desvio absoluto médio do sinal.	$\frac{\sum_{i=1}^{N} \lvert s_i^2 - média(s) \rvert}{N}$
Median	Mediana do sinal.	Implementado de acordo com (KOKOSKA; ZWILLINGER, 1999)
Median absolute deviation	Desvio absoluto mediano do sinal.	$mediana(\lvert s - mediana(s) \rvert)$
Min	Valor mínimo do sinal.	Implementado de acordo com (KOKOSKA; ZWILLINGER, 1999)
Root mean square	Raiz quadrada média do sinal.	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} s_i^2}$
Skewness	Distorção do sinal.	Implementado de acordo com (BARANDAS *et al.*, 2020)
Standard Deviation	Desvio padrão do sinal.	$\sqrt{média(\lvert s - mediana(s) \rvert)^2}$
Variance	Variação do sinal.	$média(\lvert s - mediana(s) \rvert)^2$

Outra possibilidade é estudar uma série temporal analisando suas características no domínio temporal, que explora as relações desse tipo entre os

pontos de dados e fornece uma representação destas. Usando a biblioteca TSFEL no nesse domínio, são extraídos dezoito atributos, que estão descritos na Tabela 2.

Tabela 2 - Características temporais extraídas usando TSFEL (BARANDAS et al., 2020).

Característica	**Descrição**	**Implementação**
Absolute energy	Energia absoluta do sinal.	$\sum_{i=0}^{N} s_i^2$
Area under the curve	Área sob a curva do sinal calculada com a regra do trapézio.	$\sum_{i=0}^{N} (t_i - t_{i-1}) \times \frac{s_i + s_{i-1}}{2}$
Autocorrelation	Autocorrelação do sinal.	$\sum_{n \in Z} s(n)\, \overline{s(n-l)}$
Centroid	Centroide ao longo do eixo do tempo.	$\frac{\sum_{i=0}^{N} t_i \times s_i^2}{\sum_i^N s_i^2}$
Entropy	Entropia do sinal usando a Entropia de Shannon.	$-\sum_{x \in s} P(x)\, log_2 P(x)$
Mean absolute differences	Média das diferenças absolutas do sinal.	$média(\lvert \Delta s \rvert)$
Mean differences	Média das diferenças do sinal.	$média(\Delta s)$
Median absolute differences	Mediana das diferenças absolutas do sinal.	$mediana(\lvert \Delta s \rvert)$
Median differences	Mediana das diferenças do sinal.	$mediana(\Delta s)$
Negative turning points	Número de pontos de viragem negativos do sinal.	Implementado de acordo com (BARANDAS et al., 2020)
Neighborhood peaks	Número de picos de uma vizinhança definida do sinal.	Implementado de acordo com (BARANDAS et al., 2020)
Peak to peak distance	Distância de pico a pico.	$\lvert max(s) - min(s) \rvert$
Positive turning points	Número de pontos de viragem positivos do sinal.	Implementado de acordo com (BARANDAS *et al.*, 2020)
Signal distance	Distância percorrida do sinal.	$\sum_{i=0}^{N-1} \sqrt{1 + \Delta s_i^2}$
Slope	Ajusta uma regressão linear do sinal e retorna o coeficiente m.	$y = mt + b$
Sum of absolute differences	Soma das diferenças absolutas do sinal.	$\sum_{i=0}^{N-1} \lvert \Delta s_i \rvert$
Total energy	Energia total do sinal.	$\frac{\sum_{i=0}^{N} s^2}{t_N - t_0}$
Zero crossing rate	Taxa de cruzamento zero do sinal.	Implementado de acordo com (BARANDAS et al., 2020)

4 MULTI-LAYERED PERCEPTRON (MLP)

As redes neurais artificiais (RNAs) têm como principal objetivo reproduzir o comportamento do cérebro humano, visando um alcance mais amplo para a resolução de problemas. São compostas por estruturas mais simples, denominadas neurônios, que são associados em uma rede pela conexão de sinapses com pesos definidos, com o objetivo de receber um conjunto de entradas e produzir um conjunto correspondente de saídas (NORGAARD et al., 2000). No aprendizado supervisionado, os dados de saída reais são comparados com os dados de saída desejados e são feitas alterações nos pesos, de modo que se minimize o erro de saída (BILLINGS, 2016).

Um importante exemplo de uma RNA é a *Multi-layered Perceptron* (MLP), que tipicamente consiste em uma camada de entrada, um número de camadas ocultas e uma camada de saída, como representado na Figura 1. A camada de entrada é composta por *n* entradas, que são distribuídas para a primeira camada. As saídas dos nós da primeira camada são distribuídas para a segunda camada e assim por diante (BILLINGS, 2016).

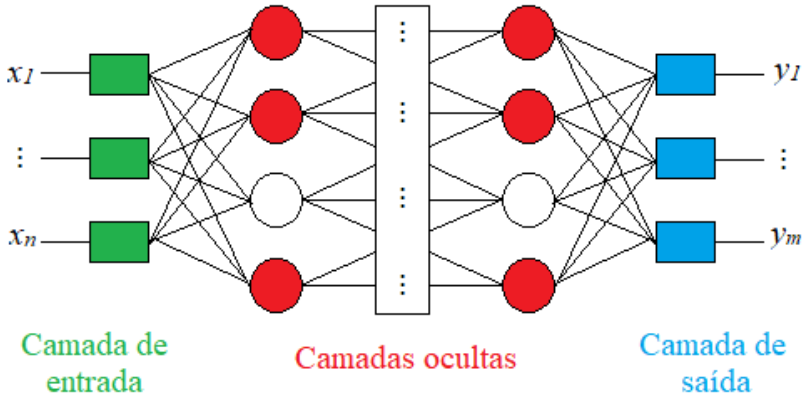


Figura 1 – Representação esquemática de uma MLP, com *n* entradas e *m* saídas (BILLINGS, 2016).

O modelo matemático da MLP é dado por:

$$\hat{p} = \phi\,[\,\textstyle\sum_k w_{ok}\phi\,[\,\sum_j w_{kj}\phi\,[\,\ldots\,\phi\,[\,\sum_i w_{li}x_i\,]]]] \tag{17}$$

Onde $\phi[\cdot]$ é a função de ativação e w são os pesos.

5 MÉTRICAS DE VALIDAÇÃO

A validação do modelo é uma etapa muito importante no aprendizado de máquina, pois é nela que será definido se o modelo é adequado para o uso pretendido. Normalmente, a validação é feita comparando-se a simulação do modelo obtido com os dados reais. Para isso, existem alguns métodos, entre os quais serão explicitados a Acurácia e AUC (*Area Under The Curve*).

5.1 Acurácia

A acurácia é uma das métricas mais usadas na etapa de validação do modelo de aprendizado de máquina, devido à sua simplicidade. Ela é capaz de determinar a proporção de classificações corretas feita pelo modelo em relação ao total de classificações e é calculada da seguinte forma (EUSEBI, 2013):

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{18}$$

Onde verdadeiros positivos (VP) e verdadeiros negativos (VN) são classificações corretas feitas pelo modelo, enquanto falsos positivos (FP) e falsos negativos (FN) são classificações incorretas.

Valores próximos de 1 indicam que o modelo é capaz de classificar os dados corretamente, enquanto valores próximos de 0 indicam que o modelo não está executando bem a tarefa de classificação.

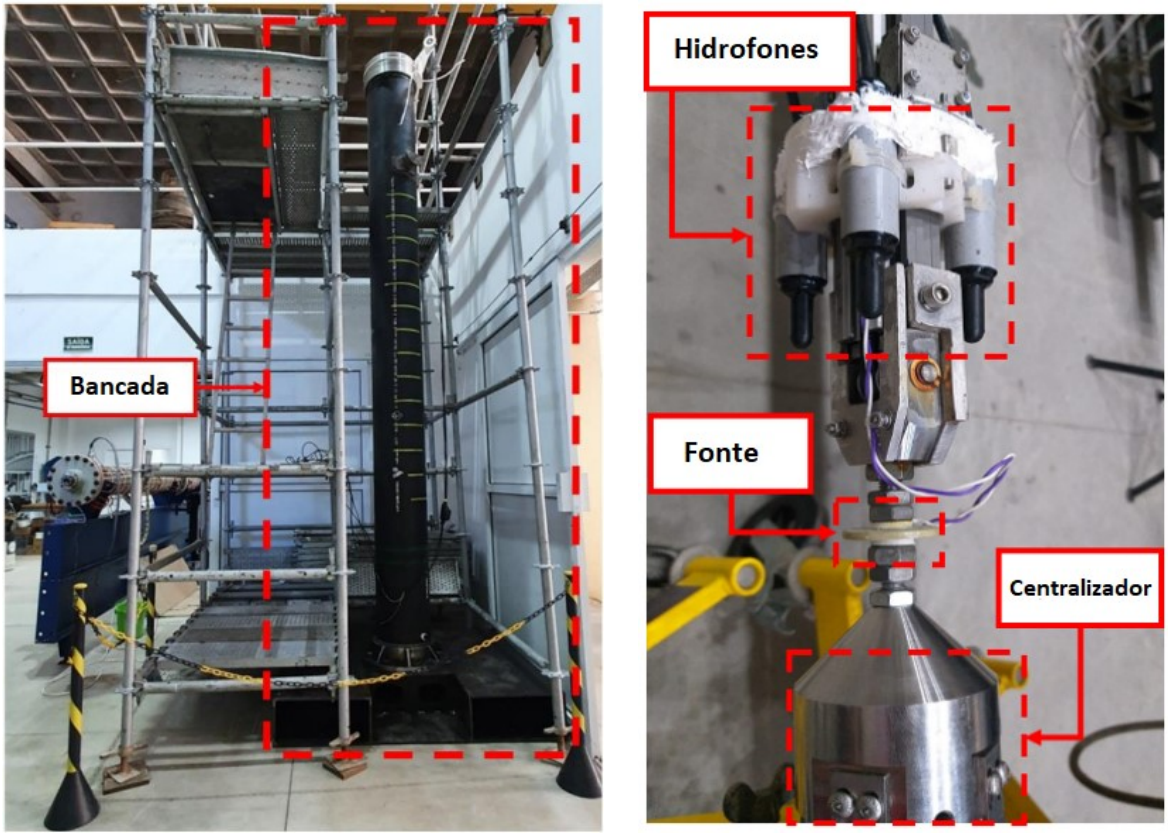
5.2 Area Under the Curve (AUC)

A métrica AUC (*Area Under the Curve*) é também uma das mais utilizadas para a validação de modelos de aprendizado de máquina, sendo exclusiva para modelos de classificação binária. Ela se refere à área sob a curva ROC (*Receiver Operating Characteristic*), que é construída relacionando a taxa de verdadeiros positivos (sensibilidade) no eixo vertical e a taxa de verdadeiros negativos (especificidade) no eixo horizonal para diferentes pontos (FAWCETT, 2006).

Assim como para a acurácia, os valores dessa métrica variam de 0 a 1, onde 1 indica um modelo de classificação perfeito, que tem 100% de sensibilidade e especificidade.

6 EXPERIMENTO

Para a obtenção dos dados de perfilagem acústica, foi montada uma bancada experimental, que pode ser vista na Figura 2a, visando reproduzir um poço de petróleo, com uma altura total de 4,2 m, sendo formada por três tubos de aço, que representam os revestimentos interno (com diâmetro de 9 5/8”) e externo, além da tubulação de produção (4 1/2").



(a) (b)

Figura 2 – Bancada (a) e ferramenta (b) montadas para o experimento (DE SOUZA, 2023).

Para os experimentos, foram considerados casos em que a tubulação e o revestimento interno foram posicionados de forma concêntrica (NE) e excêntrica, variando ainda a excentricidade entre 48,79% (E1) e 100% (E2). O posicionamento da tubulação foi feito com o auxílio do suporte apresentado na Figura 3.

Figura 3 – Suporte utilizado para a variação da excentricidade da tubulação de produção (DE SOUZA, 2023).

Além disso, também foi feita a variação do material encontrado entre as camadas, levando em conta as seguintes condições: existência apenas de água entre os revestimentos interno e externo (F), existência de uma camada de cimento sem defeito colada ao revestimento interno (C), existência de cimento com canalização colado ao revestimento interno, ou seja, em que há uma fina camada de água em seu interior (CH) e, por fim, existência de cimento de baixa qualidade colado ao revestimento interior, ao qual foram adicionadas esferas de poliestireno (CL).

Tabela 3 – Diferentes configurações encontradas no conjunto de dados produzido experimentalmente.

	Água	**Cimento sem defeito**	**Cimento com canalização**	**Cimento de baixa qualidade**
Tubulação concêntrica	F-NE	C-NE	CH-NE	CL-NE
Tubulação com 48,79% de excentricidade	F-E1	C-E1	CH-E1	CL-E1
Tubulação com 100% de excentricidade	F-E2	C-E2	CH-E2	CL-E2

Dessa forma, o conjunto de dados é formado por 12 diferentes configurações, que são obtidas através da combinação de cada um dos casos apresentados anteriormente. A Tabela 3 e a Figura 4 apresentam de forma resumida todas as configurações possíveis.

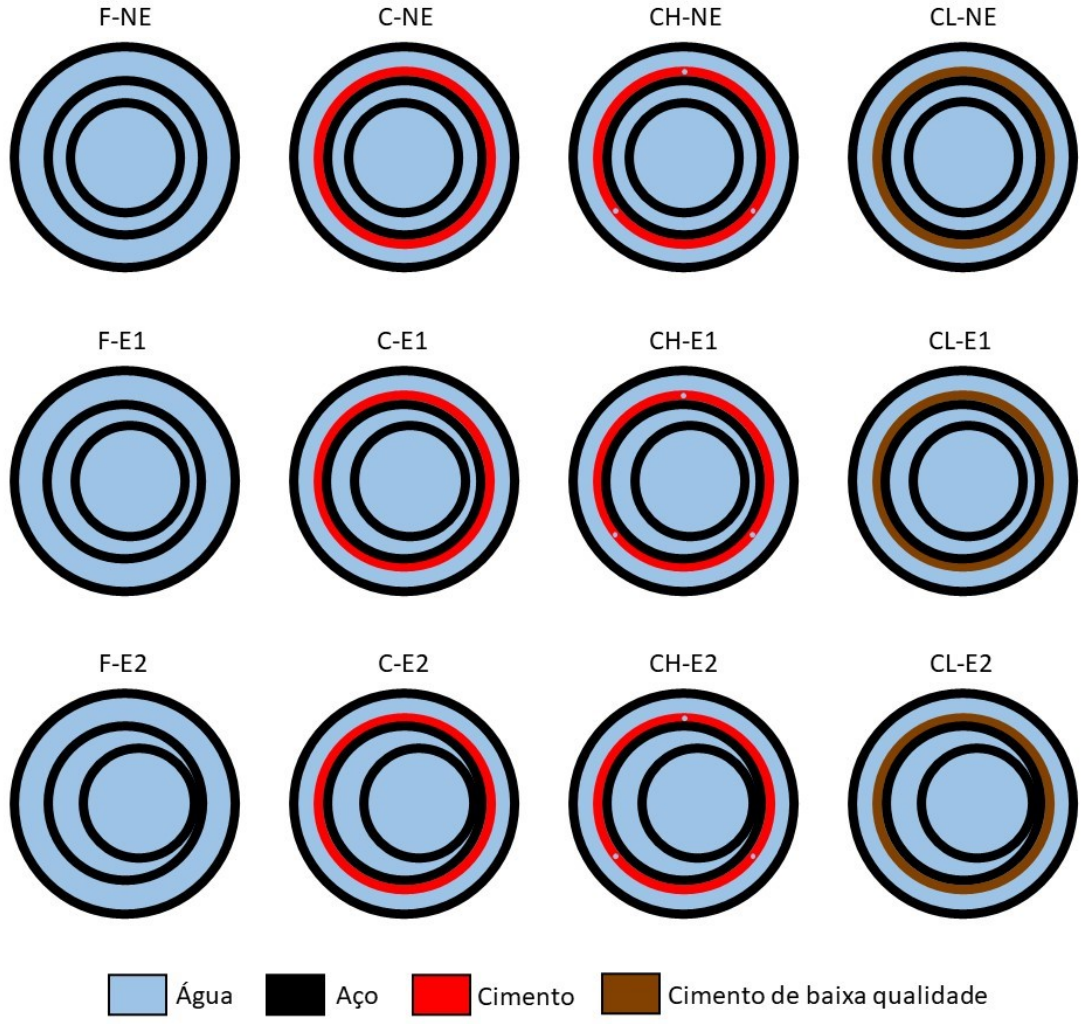


Figura 4 – Diferentes combinações de nível de excentricidade e materiais considerados ao longo do experimento (DE SOUZA, 2023).

Uma ferramenta ultrassônica foi também construída, conforme apresentado na Figura 2b, composta por um transdutor piezoelétrico APC 850, que é o emissor de sinal acústico, e por quatro hidrofones BII-7005FG da Benthowave Instrument Inc espaçados circunferencialmente a 90º um do outro, responsáveis pela recepção do sinal. A excitação do disco PZT foi feita com um gerador de forma de onda Agilent

33120A e um amplificador de potência Krohn Hite 7500, enquanto a leitura dos sinais elétricos dos hidrofones foi feita por uma unidade de aquisição digital cDAQ NI9185, NI9222 e NI9223. A posição longitudinal dos receptores é variável, sendo possível posicioná-los em mais de 6000 pontos diferentes, variando de 166,3 mm até 2171,4 mm a partir do disco PZT.

Dessa forma, a ferramenta foi posicionada dentro da tubulação de produção com o auxílio de centralizadores, conforme esquematizado na Figura 5, e um sinal *sinc* com frequência de corte de 50 kHz foi usado para excitar o transdutor e a resposta foi capturada pelos quatro hidrofones em 200 posições distribuídas ao longo do comprimento da tubulação.

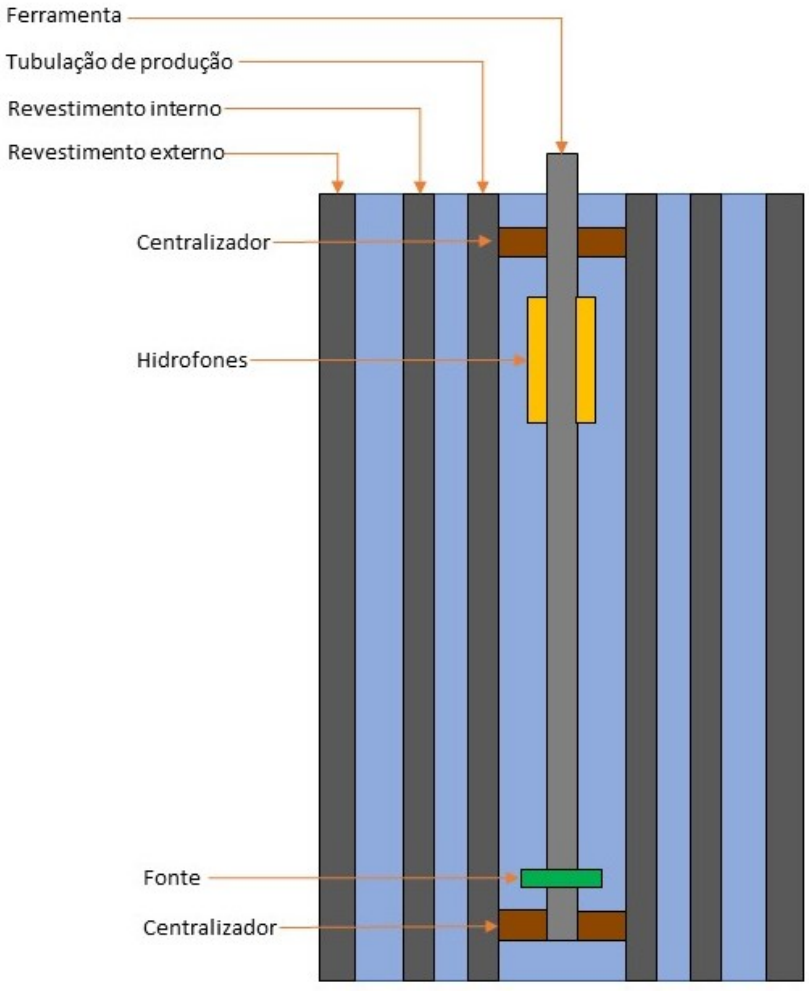


Figura 5 – Figura esquemática da bancada experimental.

Apenas 13 dos 200 pontos de medição foram utilizados, tendo sido mantidos os primeiros 6 ms dos sinais, que foram também reamostrados de 0,5 MHz para 0,25 MHz, resultando em 13 sinais com 1500 pontos cada. A fim de considerar a influência de todos os hidrofones, os dados utilizados neste trabalho correspondem à média dos sinais obtidos por cada um deles para cada instante de tempo.

O conjunto de dados é, assim, composto por 20 amostras da classe nominal (C-NE) e 10 amostras das onze classes de defeito restantes (C-E1, C-E2, F-NE, F-E1, F-E2, CH-NE, CH-E1, CH-E2, CL-NE, CL-E1, CL-E2).

A fim de determinar a melhor metodologia para detectar o tipo de defeito existente na cimentação de um poço de petróleo, inicialmente os dados experimentais foram aplicados em cada um dos métodos de extração de características mencionados na seção 3 (PCA, FA, ICA, Isomap, LLE, MLLE, TSFEL).

Em seguida, foi realizada a otimização de hiperparâmetros do modelo de aprendizado de máquina (MLP) utilizando a validação cruzada *Leave-One-Out* (LOOCV), que foi executada 50 vezes para cada método de extração de características. Nessa abordagem, o conjunto de dados de dimensão n é dividido em duas partes, uma de tamanho $n-1$, usada para treinar o modelo, e uma de tamanho 1, usada para a validação. Os hiperparâmetros considerados nessa etapa para o modelo MLP podem ser encontrados na Tabela 4.

Tabela 4 – Hiperparâmetros considerados para o modelo MLP durante o procedimento de LOOCV.

Hiperparâmetro	**Valores**
Número de neurônios	5, 10, 20, 50, 100
Número de camadas ocultas	1, 2, 3
Função de ativação	ReLU, TanH
Épocas	150, 200, 250, 300, 350
Batch size	2, 4, 6, 8, 10, 20, 30, 32

Por fim, as métricas de acurácia e AUC (*Area Under the Curve*) foram empregadas para determinar a combinação de modelo e método de extração de características com maior eficiência na avaliação da integridade da camada de cimento.

7 RESULTADOS

Após o pré-processamento dos dados coletados experimentalmente, os métodos de extração de características foram utilizados, com o objetivo de determinar o que melhor realiza a redução da dimensionalidade do conjunto de dados, mantendo o máximo de informações originais. Para auxiliar nessa análise, em um primeiro momento serão apresentadas visualizações bidimensionais e tridimensionais dos dados resultantes de cada método utilizado (PCA, FA, ICA, Isomap, LLE, MLLE, TSFEL).

Posteriormente, serão apresentadas as melhores combinações de hiperparâmetros para o modelo de aprendizado de máquina MLP, correspondentes a cada um dos métodos de extração de características empregados. A seleção dessas combinações se dará com base nos valores obtidos para as métricas AUC (Área sob a Curva ROC) e acurácia. Este passo é crucial para refinar o modelo e garantir a precisão na detecção e classificação de defeitos, contribuindo assim para a integridade e eficácia dos processos de monitoramento da cimentação de poços de petróleo.

7.1 Visualização do PCA

Para o PCA (*Principal Component Analysis)*, adotou-se como alvo o número de componentes principais (PCs) que explicassem pelo menos 95% da variância do conjunto de dados. Conforme pode ser visto na Figura 6, foram necessários 28 PCs para atingir esse objetivo.

A Figura 7 mostra as visualizações bidimensionais (todos os PCs) e tridimensionais (os três primeiros PCs). Avaliando esses gráficos, é notável que as classes C-NE, F-NE, F-E1, F-E2 e CL-E1 estão bem agrupadas, isto é, os dados pertencentes à mesma classe estão reunidos em *clusters*, distantes dos grupos formados por dados de outras classes. Contudo, os dados pertencentes às outras classes não obtiveram um desempenho satisfatório na formação de *clusters*, sugerindo uma mistura ou sobreposição dessas classes na projeção dos componentes principais.

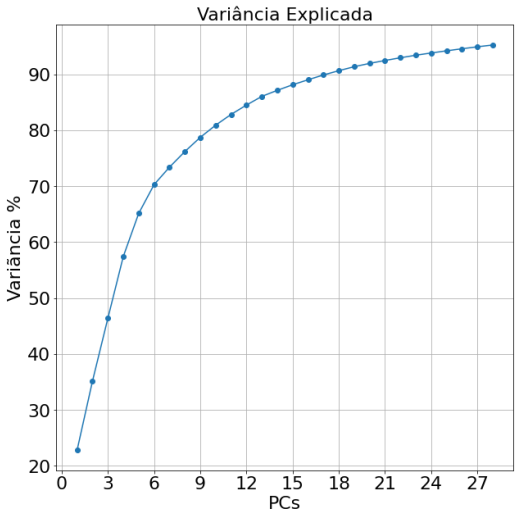


Figura 6 - Componentes principais necessários para obter pelo menos 95% da variância explicada.

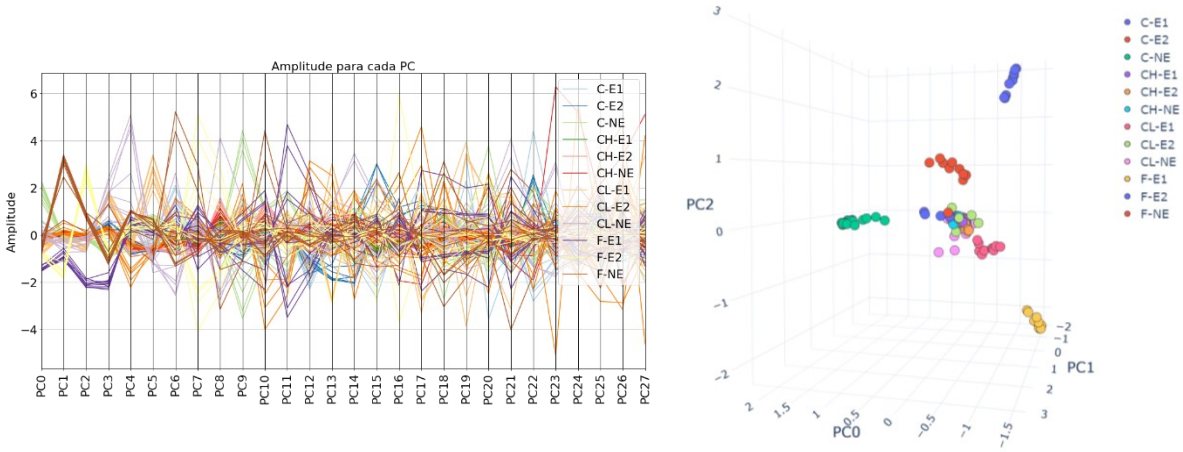


Figura 7 - Gráficos 2D e 3D para o PCA.

7.2 Visualização do FA

Para o FA (*Factor Analysis*), optou-se por um número arbitrário de três fatores. As visualizações bidimensionais e tridimensionais decorrentes desse método estão representadas na Figura 8, Observando-se esses gráficos, é possível identificar que somente quatro classes (C-NE, F-NE, F-E1 e F-E2) foram adequadamente agrupadas em *clusters.* Por outro lado, as demais classes parecem ainda mais misturadas do que na representação obtida pelo PCA, sugerindo que a Análise Fatorial pode não ter sido tão eficaz na separação dessas classes.

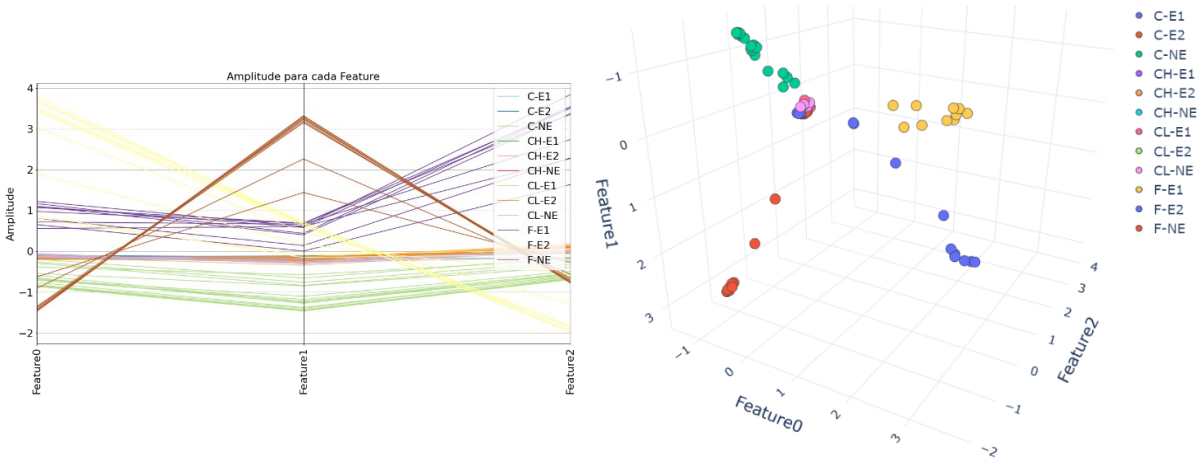


Figura 8 - Gráficos 2D e 3D para o FA.

7.3 Visualização do ICA

A Figura 9 mostra as visualizações bidimensional e tridimensional das três características extraídas usando o ICA (*Independent Component Analysis*). É relevante mencionar que a escolha de três componentes foi arbitrária. Com base nesses gráficos, pode-se observar que, similar ao resultado obtido com o PCA, cinco classes (C-NE, F-NE, F-E1, F-E2 e CL-E1) demonstraram uma boa formação de clusters. As classes de defeito restantes, porém, estão muito próximas umas das outras, indicando uma agrupamento menos eficiente quando comparado com as classes mencionadas anteriormente.

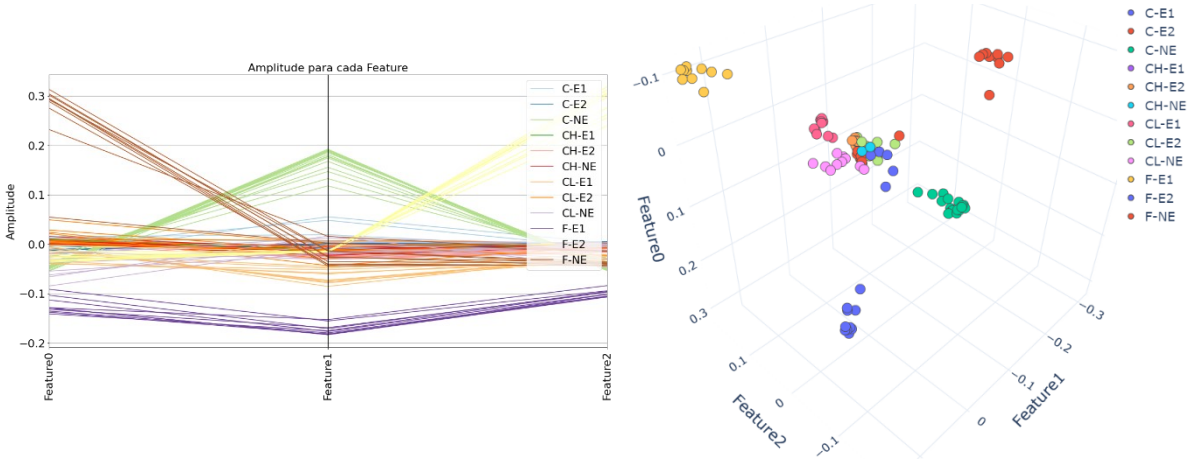


Figura 9 - Gráficos 2D e 3D para o ICA.

7.4 Visualização do Isomap

Semelhante aos casos anteriores, para o Isomap *(Isometric Feature Mapping)* optou-se por extrair somente três características, que são visualizadas na Figura 10. A partir dessa análise, nota-se uma melhoria em relação aos métodos previamente examinados, já que o Isomap é capaz de separar seis classes (C-NE, C-E1, F-NE, F-E1, F-E2, CL-E1). Entretanto, os dados dessas classes parecem mais dispersos quando comparados aos resultados dos métodos anteriores. A distinção visual das classes restantes se mostrou desafiadora, indicando um agrupamento menos eficaz.

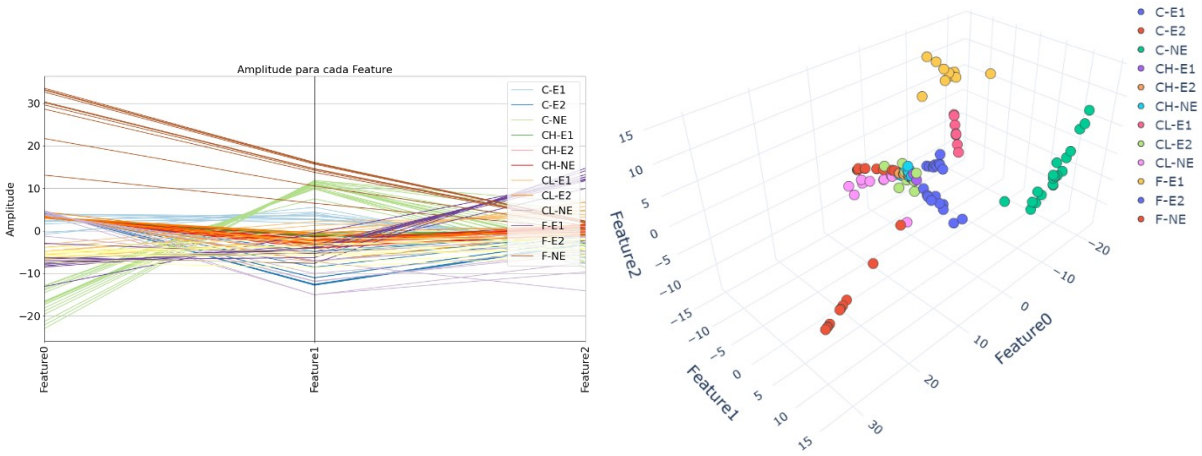


Figura 10 - Gráficos 2D e 3D para o Isomap.

7.5 Visualização do LLE

Mais uma vez, foi selecionada a extração de apenas três características, dessa vez pelo método LLE (Locally Linear Embedding), cujas visualizações bidimensionais e tridimensionais são mostradas na Figura 11. Observa-se que o LLE não apresenta uma clusterização satisfatória para nenhuma das classes de defeito, indicando que esse método pode não ser tão promissor para a tarefa em questão. A dispersão dos dados e a sobreposição entre as classes são evidentes, sugerindo que o LLE não conseguiu capturar as estruturas e relações relevantes presentes no conjunto de dados.

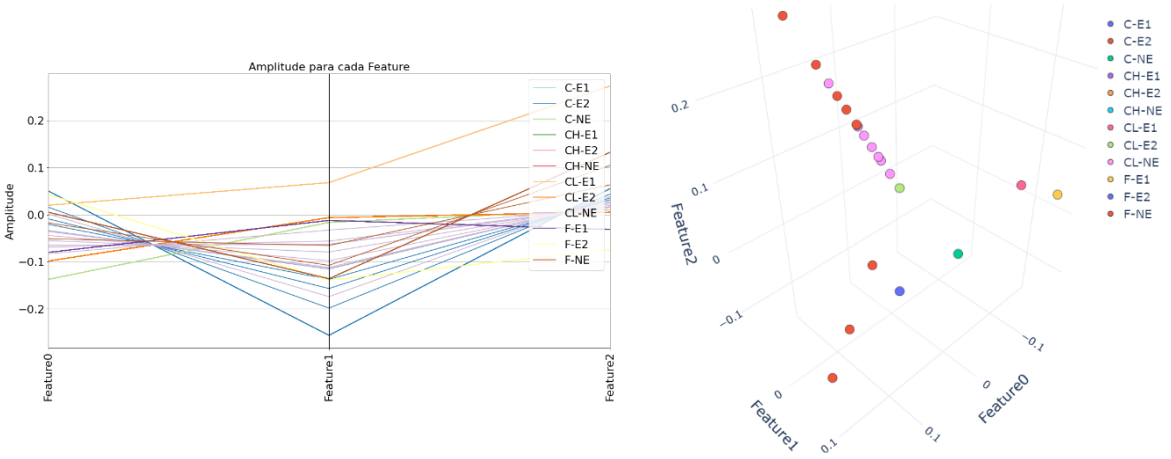


Figura 11 - Gráficos 2D e 3D para o LLE.

7.6 Visualização do MLLE

Para o método MLLE (Modified Locally Linear Embedding) também foram selecionadas três características para extração. Essas características podem ser visualizadas nos gráficos bidimensional e tridimensional na Figura 12, os quais demonstram que a clusterização foi satisfatória para quatro classes (C-NE, F-E1, F-E2 e CL-E1). No entanto, para as oito classes restantes, os resultados não foram tão bons, com uma menor separação entre os grupos e uma maior sobreposição entre as classes.

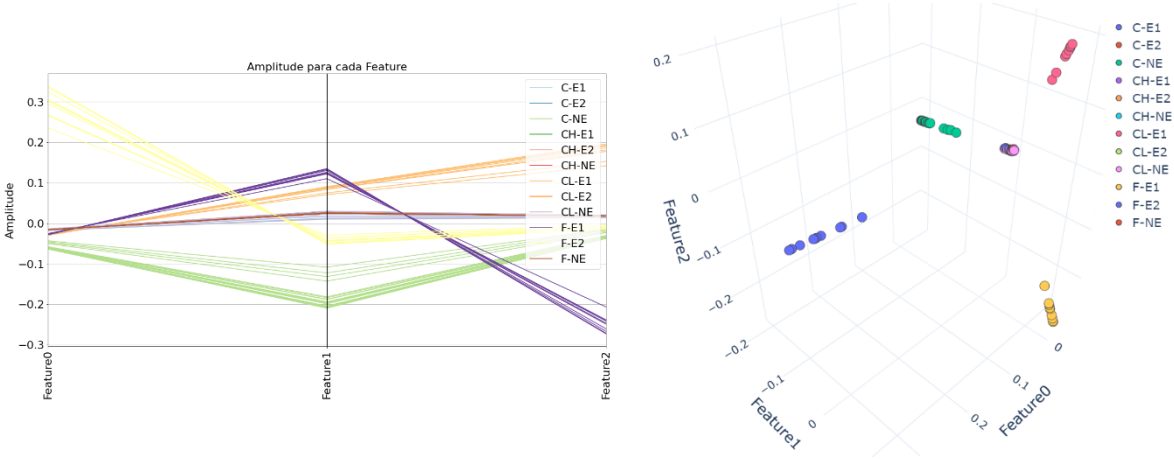


Figura 12 - Gráficos 2D e 3D para o MLLE.

7.7 Visualização do TSFEL no Domínio Estatístico

Utilizando a biblioteca TSFEL para extração de características no domínio estatístico, inicialmente foram obtidas 36 características. Após a remoção de características altamente correlacionadas e com baixa variância, restaram um total de 14 características, sendo elas ECDF *Percentile* 0, ECDF *Percentile* 1, *Histogram* 0, *Histogram* 1, *Histogram* 2, *Histogram* 3, *Histogram* 4, *Histogram* 5, *Kurtosis*, Max, *Mean*, *Median*, Min e *Skewness*.

A Figura 13 apresenta os gráficos bidimensionais com todas as características e o gráfico tridimensional com apenas três delas, selecionadas arbitrariamente. Observa-se que esse método não parece realizar uma clusterização satisfatória dos dados, uma vez que não há uma clara separação entre as classes.

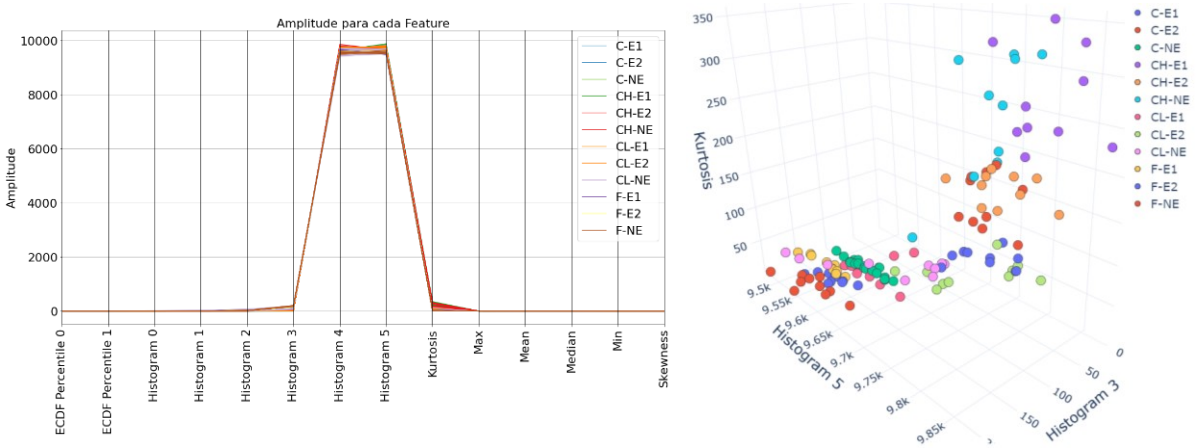


Figura 13 - Gráficos 2D e 3D para as características estatísticas.

7.8 Visualização do TSFEL no Domínio Temporal

Por fim, para as características do domínio temporal, inicialmente foram extraídas 18 características. Após a remoção das características altamente correlacionadas e com baixa variância, restaram 9 destas, denominadas *Absolute Energy, Centroid, Entropy, Mean Diff, Median Diff, Negative Turning Points, Neighbourhood Peaks, Peak to Peak Distance* e *Slope.*

A Figura 14 apresenta o gráfico bidimensional com todas as características e o gráfico tridimensional com três delas, selecionadas arbitrariamente. Novamente, esse

método não parece realizar uma clusterização adequada dos dados, uma vez que os dados de todas as classes estão muito próximos uns dos outros.

No entanto, observa-se que as características do domínio temporal parecem agrupar melhor os dados das classes C-NE, F-NE, F-E1 e F-E2 em comparação com as características estatísticas. Isso indica que as características temporais podem fornecer uma melhor separação para essas classes específicas em comparação com o outro domínio. No entanto, ainda há espaço para melhorias na clusterização dos demais grupos de defeitos.

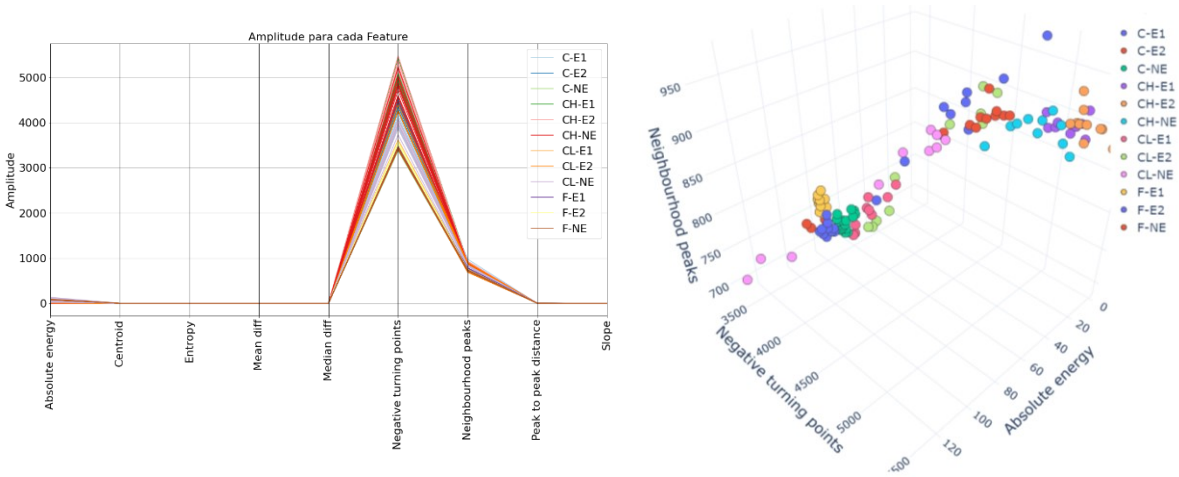


Figura 14 - Gráficos 2D e 3D para as características temporais.

7.9 Resultado do LOOCV

A fim de otimizar os hiperparâmetros do modelo de aprendizado de máquina (MLP), realizou-se a validação cruzada *Leave-One-Out* (LOOCV) 50 vezes para cada método de extração de características. A Tabela 5 apresenta o melhor resultado obtido entre as 50 iterações para cada método, com base nas métricas AUC e acurácia. Além disso, na Figura 15 é possível ver de forma gráfica os resultados, facilitando as comparações.

Analisando os resultados obtidos, pode-se notar que a maioria dos modelos apresentou altos valores de acurácia e AUC, com cinco deles tendo obtido acurácia maior que 0,8.

Levando em consideração a acurácia, o método PCA obteve o melhor desempenho (0,9231), seguido pelo TSFEL no domínio estatístico (0,8923) e pelo

Isomap (0,8769), enquanto o LLE obteve o pior desempenho (0,6615). Já em relação à métrica AUC, os métodos FA (0,9749), PCA (0,9723) e TSFEL no domínio estatístico (0,9867) tiveram um desempenho superior, enquanto o LLE novamente obteve o resultado mais baixo (0,9000).

Tabela 5 - Resultados do LOOCV.

	Acurácia	**AUC**	**Tempo de treinamento (s)**	**Parâmetros**
PCA	0.9231 ± 0.2666	0.9723	30.2028 ± 7.1667	Função de ativação: ReLU Número de neurônios: 100 Número de camadas ocultas: 2 Épocas: 350 *Batch Size*: 20
FA	0.8308 ± 0.3750	0.9749	85.0287 ± 14.5146	Função de ativação: TanH Número de neurônios: 50 Número de camadas ocultas: 3 Épocas: 300 *Batch Size*: 2
ICA	0.8231 ± 0.3816	0.9684	9.3045 ± 0.8854	Função de ativação: ReLU Número de neurônios: 50 Número de camadas ocultas: 2 Épocas: 250 *Batch Size*: 20
Isomap	0.8769 ± 0.3285	0.9643	14.8044 ± 2.0445	Função de ativação: ReLU Número de neurônios: 100 Número de camadas ocultas: 3 Épocas: 250 *Batch Size*: 10
LLE	0.6615 ± 0.4732	0.9000	22.4571 ± 1.1433	Função de ativação: TanH Número de neurônios: 100 Número de camadas ocultas: 2 Épocas: 250 *Batch Size*: 8
MLLE	0.6923 ± 0.4615	0.9530	79.6783 ± 12.2267	Função de ativação: TanH Número de neurônios: 50 Número de camadas ocultas: 3 Épocas: 300 *Batch Size*: 2
Estatísticas	0.8923 ± 0.3100	0.9867	45.1893 ± 7.3149	Função de ativação: TanH Número de neurônios: 50 Número de camadas ocultas: 3 Épocas: 350 *Batch Size*: 4
Temporais	0.7615 ± 0.4261	0.9671	33.0036 ± 2.0530	Função de ativação: TanH Número de neurônios: 10 Número de camadas ocultas: 1 Épocas: 350 *Batch Size*: 4

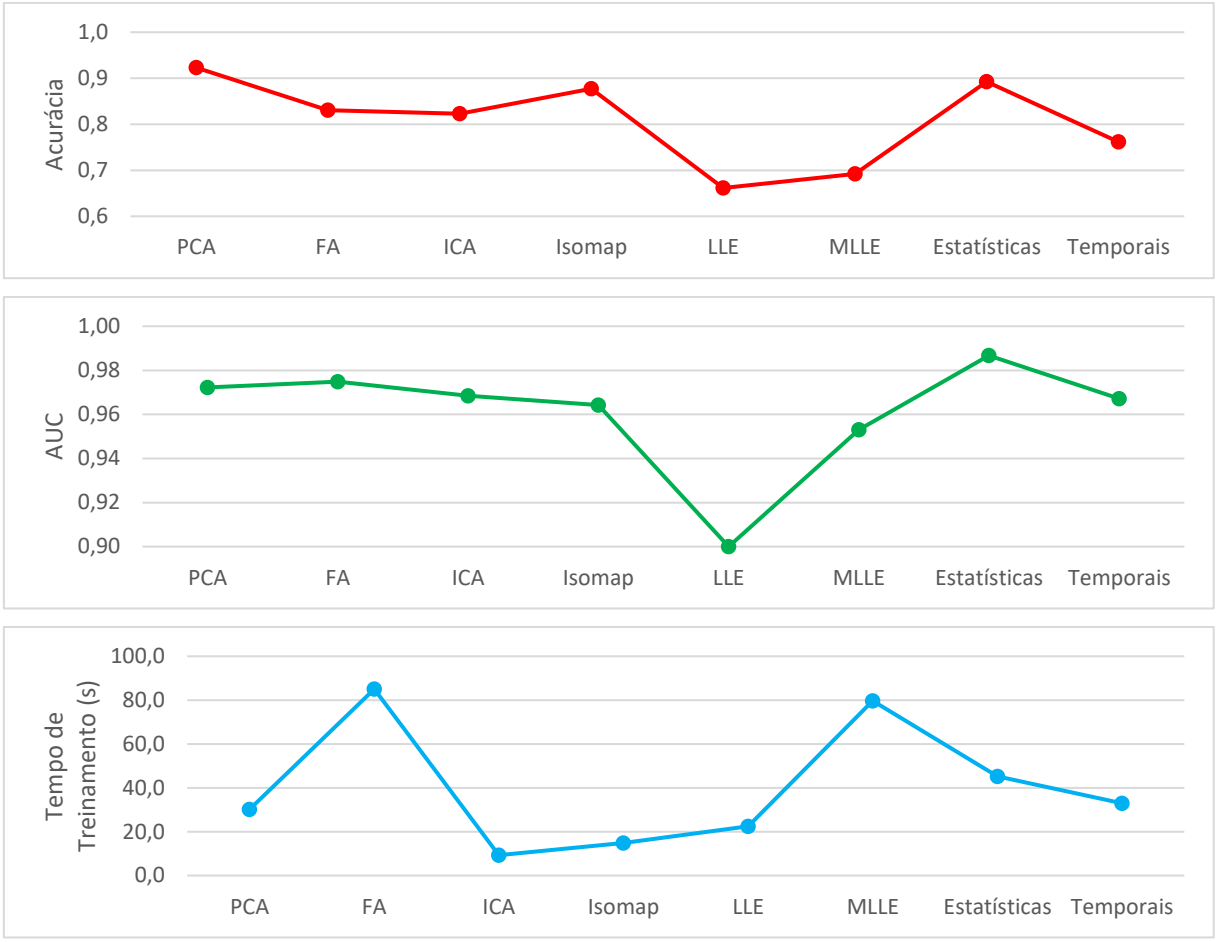


Figura 15 – Valores das métricas acurácia e AUC e do tempo de treinamento obtidos utilizando o LOOCV.

Como observado nas seções anteriores, os métodos PCA, Isomap e FA já haviam apresentado resultados promissores, sendo capazes de clusterizar visualmente quatro, seis e quatro classes, respectivamente. Por outro lado, o método TSFEL no domínio estatístico não parecia ter um desempenho satisfatório nas análises preliminares, mas acabou apresentando boas métricas para o treinamento do modelo.

No caso do LLE, ainda nas seções anteriores, analisando os gráficos bi e tridimensionais, foi constatado que esse método apresentou o pior desempenho, não sendo capaz de realizar uma clusterização adequada para nenhuma das classes. Essa conclusão é confirmada pelas métricas utilizadas.

Além das métricas de validação, outro fator importante a ser considerado na escolha de um modelo de aprendizado de máquina é o tempo de treinamento. Nesse

sentido, verificou-se que o ICA obteve o menor tempo de treinamento, apenas 9 segundos, enquanto o FA teve o pior desempenho, com 85 segundos.

Assim, levando em consideração todos esses parâmetros, pode-se concluir que a MLP com 100 neurônios, 2 camadas ocultas, 350 épocas, batch size de 20 e função de ativação ReLU, em conjunto com o método de extração de características PCA, parece ser o modelo mais promissor entre os considerados. Essa escolha é baseada na combinação de boas métricas de desempenho e um tempo de treinamento satisfatório.

8 CONCLUSÃO

Analisando os resultados obtidos neste trabalho, pode-se concluir que a maioria dos métodos de extração de características estudados em conjunto com a rede neural MLP possui bom desempenho na classificação de defeitos presentes na camada de cimento de poços de petróleo utilizando dados experimentais de perfilagem acústica com a presença da coluna de produção.

Os melhores resultados globais foram obtidos pelo método *Principal Component Analysis* (PCA), considerando as métricas acurácia (0,9231) e AUC (0,9723), assim como um baixo tempo de treinamento (30 segundos).

A melhor configuração da MLP em conjunto com o PCA consiste em uma rede com 100 neurônios e 2 camadas ocultas, treinada por 350 épocas, utilizando um *batch size* de 20 e função de ativação do tipo ReLU.

Esses resultados são promissores e indicam que as técnicas de aprendizado de máquina e de extração de características utilizadas neste estudo têm potencial para serem aplicadas em situações reais de avaliação da qualidade do cimento durante as operações de P&A. Essa abordagem pode eliminar a necessidade de remover a tubulação de produção, reduzindo significativamente o tempo e os custos envolvidos nessa etapa.

Vale ainda ressaltar que os métodos testados foram objeto de depósito de patente no INPI (RIBEIRO, 2021).

Referências bibliográficas

BARANDAS, M. et al. TSFEL: time series feature extraction library. **Softwarex**, v. 11, p. 100456, jan. 2020.

BARBER, D. **Bayesian reasoning and machine learning**. Cambridge University Press, 2012.

BILLINGS, S. A. **Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains**. Reino Unido: John Wiley & Sons, 2013.

DE SOUZA, Luis Paulo Brasil et al. Machine learning-based cement integrity evaluation with a through-tubing logging experimental setup. **Geoenergy Science and Engineering**, v. 227, p. 211882, 2023.

DWORAKOWSKI, Z.; DRAGAN, K.; STEPINSKI, T. Artificial neural network ensembles for fatigue damage detection in aircraft. **Journal Of Intelligent Material Systems And Structures**, v. 28, n. 7, p. 851-861, 28 jul. 2016.

EUSEBI, P. Diagnostic accuracy measures. **Cerebrovascular Diseases**, v. 36, n. 4, p. 267-272, 2013.

FAWCETT, T. An introduction to ROC analysis. **Pattern recognition letters**, v. 27, n. 8, p. 861-874, 2006.

GOLUB, G. H.; VAN LOAN, C. F. **Matrix computations**. JHU press, 2013.

HYVÄRINEN, A.; OJA, E. Independent component analysis: algorithms and applications. **Neural Networks**, v. 13, n. 4-5, p. 411-430, jun. 2000.

JOLLIFFE, I. T. **Principal component analysis for special types of data**. 2. ed. Springer New York, 2002.

MELVILLE, J. et al. Structural damage detection using deep learning of ultrasonic guided waves. **Aip Conference Proceedings**, 2018.

NORGAARD, M. et al. **Neural Networks for Modelling and Control of Dynamic Systems: A Practitioner's Handbook**. 1 ed. Londres: Springer-Verlag London, 2000.

TENENBAUM, J. B.; SILVA, V.; LANGFORD, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. **Science**, v. 290, n. 5500, p. 2319-2323, 22 dez. 2000.

VIGGEN, E. M.; JOHANSEN, T. F.; MERCIU, I. Analysis of outer-casing echoes in simulations of ultrasonic pulse-echo through-tubing logging. **Geophysics**, v. 81, n. 6, p. 679-685, nov. 2016.

VIGGEN, E. M.; JOHANSEN, T. F.; MERCIU, I. Simulation and modeling of ultrasonic pitch-catch through-tubing logging. **Geophysics**, v. 81, n. 4, p. 383-393, jul. 2016.

VIGGEN, E. M. et al. Automatic interpretation of cement evaluation logs from cased boreholes using supervised deep neural networks. **Journal Of Petroleum Science And Engineering**, v. 195, p. 107-124, dez. 2020.

QI, Z. B. et al. A Novel and Efficient Method for Quantitative Cement Logging using a Logging-While-Drilling Acoustic Tool. **SPE Kuwait Oil & Gas Show and Conference**, 2017.

RASCHKA, S. MLxtend: providing machine learning and data science utilities and extensions to python's scientific computing stack. **Journal Of Open Source Software**, v. 3, n. 24, p. 638, 22 abr. 2018.

RIBEIRO, M.G.C. et. al. "Método computacional de detecção e estimação de falhas de cimentação em revestimentos de poços de petróleo pela aquisição de sinais de perfilagem acústica através da coluna de produção com base no aprendizado de

máquina e em simulações de alta fidelidade”, Número do registro: BR1020210185813, 2021.

ROWEIS, S. T.; SAUL, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. **Science**, v. 290, n. 5500, p. 2323-2326, 22 dez. 2000.

ZHANG, Z.; WANG, J. MLLE: Modified locally linear embedding using multiple weights. **Advances in neural information processing systems**, v. 19, 2006.

ZWILLINGER, D.; KOKOSKA, S. **CRC standard probability and statistics tables and formulae**. Crc Press, 1999.