DOI: 10.48011/asba.v2i1.1462

Detecção de falhas com *Stacked Autoencoders* e técnicas de reconhecimento de padrões em poços de petróleo operados por *gas lift*

Rodrigo Scoralick Fontoura do Nascimento [*]
Bruno Henrique Groenner [**] **Ricardo Emanuel Vaz Vargas** [***]
Ismael Humberto Ferreira dos Santos [****]

[*] *Programa de Pós-Graduação em Engenharia de Sistemas e Automação, Universidade Federal de Lavras, MG, (e-mail: rodrigo.nascimento2@estudante.ufla.br)*
[**] *Departamento de Automática, Universidade Federal de Lavras, MG, (e-mail: brunohb@ufla.br)*
[***] *Petróleo Brasileiro S.A., ES, (e-mail: ricardo.vargas@petrobras.com.br)*
[****] *Petróleo Brasileiro S.A., CENPES, RJ, (e-mail: ismaelh@petrobras.com.br)*

Abstract: The offshore industry is responsible for most of the oil and gas production in Brazil. When the level of complexity in this industry is high, it has been a precursor to new technologies in recent years. The main objective of the present work is the development of a system for the detection and classification of failures in oil production wells operated with elevation by gas lift. Stacked autoencoders are used and pattern recognition techniques for fault classification, verifying performance metrics and applying cross-validation to check the generalization of the models for the available observations. After the development of the classifiers, high recall values were obtained (much higher than 0.98), which shows the great applicability of the proposed system in detecting failures in offshore production wells.

Resumo: A indústria *offshore* é responsável pela maior parte da produção de petróleo e gás no Brasil. Devido ao elevado nível de complexidade nessa indústria, ela vem sendo a precursora de novas tecnologias ao longo dos últimos anos. O presente trabalho tem como objetivo o desenvolvimento de um sistema para detecção de falhas em poços de produção de petróleo operados com elevação por *gas lift*. São utilizados *autoencoders* empilhados para redução de dimensionalidade e diferentes técnicas de reconhecimento de padrões, verificando métricas de desempenho dos modelos em observações reais de operação disponíveis. Após o desenvolvimento dos classificadores, obteve-se valores de *recall* elevado (boa parte superior a 0,98), o que mostra a grande aplicabilidade do sistema proposto em detectar falhas em poços de produção *offshore*.

Keywords: Autoencoders; Fault detection; Oil well monitoring; Multivariate time series classification; Cross validation; Pattern recognition.

Palavras-chaves: Autoencoders; Detecção de falhas; Monitoramento de poços de petróleo; Classificação multivariada de séries temporais; Validação cruzada; Reconhecimento de padrões.

1. INTRODUÇÃO

No cenário atual, a indústria do petróleo e gás tem se tornado mais exigente em todas as áreas da engenharia, dentre elas, as de de segurança e produção. Diversos aspectos devem ser levados em consideração na área de petróleo e gás, por ser uma área de atuação industrial muito complexa, englobando várias áreas de engenharia que se relacionam em busca de processos e produtos de melhor qualidade, agregando tecnologia e inovação ao longo dos anos de desenvolvimento, como o abordado por Schiavi and Hoffmann (2015).

A indústria de extração de petróleo é dividida em duas modalidades de produção: *onshore* e *offshore*. A primeira é baseada em produção em terra firme, no continente. Já a segunda modalidade, tem sua produção realizada em alto mar por meio de plataformas de extração de petróleo, normalmente distantes do continente e em águas profundas. O foco deste trabalho baseia-se em estudos de aplicação em poços de petróleo e gás de plataformas marítimas, ou seja, poços *offshore*.

Os poços de petróleo em águas profundas são classificados de duas formas: os surgentes e não surgentes. Poços não surgentes necessitam de métodos para auxiliar o escoamento dos fluídos (água, óleo, gás e sedimentos). Já os

surgentes, conseguem com sua própria pressão realizar o escoamento dos fluidos de produção, ou seja, nos poços surgentes há uma elevação natural dos fluidos (Thomas, 2004).

O processo de elevação artificial por *gas lift* consiste na gaseificação da coluna de produção. Utilizando-se gás natural com a finalidade de diminuir a densidade média do fluido que está sendo produzido no reservatório (Filho, 2011). Trata-se de um processo complexo e sujeito à diversas falhas, sejam elas nos atuadores e sensores, e ainda de acordo com as características de cada poço de produção.

A ocorrência de falhas em poços de produção de petróleo com *gas lift* pode gerar prejuízos em milhares de dólares para as empresas produtoras, além de complexa operação que se sucede após tal ocorrência, para que se reestabeleça a normalidade na operação. A indústria petroleira considera que o prognóstico de anomalias em poços produtores de petróleo pode ajudar a reduzir os custos de manutenção, bem como evitar perdas de produção e acidentes ambientais e de vidas humanas (Vargas et al., 2019).

Uma forma de prever a ocorrência dessas falhas é a implementação de sistemas de reconhecimento de padrões baseados em técnicas de Inteligência Computacional. A detecção de falhas busca expor possíveis desvios apresentados em um processo, a partir de suas variáveis monitoradas. Com o advento de sistemas de medição, os valores das variáveis de processos puderam ser obtidos e armazenados em grandes quantidades e com maior precisão, permitindo-se assim um monitoramento mais eficiente (Xuewu Dai et al., 2008).

Como exemplo, Araujo et al. (2003) implementaram um sistema de detecção de falhas em poços com *gas lift* baseado em Sistemas Imunológicos Artificiais (SIA), divididos em dois padrões de operação, normal e anormal. Já Santos et al. (2018) adotaram técnicas como o PCA (do inglês, *Principal Component Analysis*) no tratamento de dados de poços de petróleo e a técnica de classificação *Random Forests* para detecção de falhas de acumulação de hidrato nas linhas de produção ou injeção de poços surgentes, sendo essas falhas catalogadas por especialistas da área de Engenharia de Petróleo. Esses trabalhos assemelham-se a este artigo, pois todos objetivam encontrar distúrbios que causam uma anormalidade na operação em poços reais de produção de petróleo.

Diversas outras técnicas de Inteligência Computacional podem ser implementadas para detecção de falhas em processos industriais. Como o apresentado por Abdellatif et al. (2018) que utiliza a saída de dois *autoencoders* como entradas de uma rede perceptron multicamadas (MLP) para detecção de barras quebradas em motores trifásicos de indução. No trabalho apresentado por Wen et al. (2018) foram aplicados *autoencoders* em conjunto como uma rede neural MLP na detecção de falhas que causam desbalanceamento em turbinas de corrente marítima.

A finalidade do presente trabalho é a implementação de um sistema de reconhecimento de padrões baseado em técnicas de Inteligência Computacional, tornando o processo mais analítico e menos operacional, sendo este uns dos objetivos principais da Industria 4.0, por meio de tecnologias disruptivas Nilchiani et al. (2019). A ideia central do artigo é a identificação de falhas em poços de produção com elevação artificial por *gas lift*, falhas essas de origens desconhecidas, determinadas por meio de inferência de operadores do processo, devido a ocorrência de perdas de produção, que sendo distintas da operação normal e estável de um poço, são diagnosticadas por intermédio de dados coletados de uma plataforma de extração de petróleo *offshore*, com o auxílio de operadores de produção.

O presente trabalho tem sua estrutura apresentada da seguinte forma: Na próxima seção tem-se o processo de extração de petróleo por *gas lift* e a fundamentação teórica das ferramentas computacionais empregadas no estudo. Na Seção 3 é apresentada a metodologia de desenvolvimento do trabalho. Na Seção 4 são apresentados os resultados e discussões. Por fim, na Seção 5, são dispostas as conclusões.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Processo

A Figura 1 mostra um esquema simplificado do diagrama de tubulação e instrumentação P&ID (do inglês, *Piping and Instrument Diagram*) de um poço de produção que utiliza elevação artificial por *gas lift*. Na Tabela 1 são apresentadas as variáveis presentes e suas unidades de grandeza. Sendo o gás de alta pressão proveniente do *header* de gás na plataforma (instrumentos marcados por 4) o qual é injetado através do anel entre a tubulação e a cadeia de revestimento até atingir uma válvula de orifício localizada a jusante na parte inferior da tubulação.

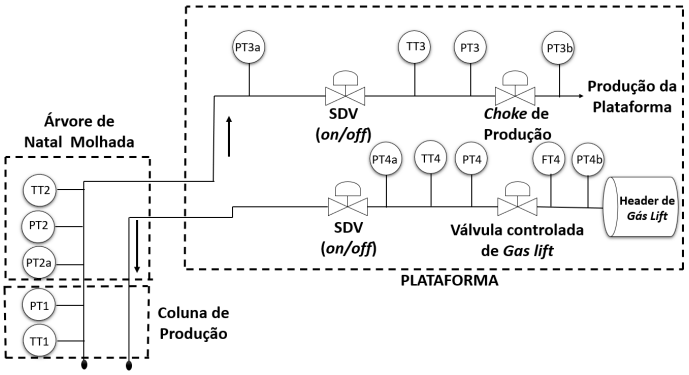

Figura 1. Diagrama P&ID simplificado de um poço de produção operando com elevação artificial por *gas lift* .

A densidade do fluido é reduzida, o que causa redução do gradiente de pressão médio ao longo da coluna e reduz a energia (pressão) necessária para que os fluidos do reservatório cheguem à plataforma. No leito do oceano, um conjunto de válvulas conhecido como árvore de natal molhada (ANM) permite a passagem dos fluidos para a plataforma e atua como barreira de segurança. Já na plataforma, uma válvula de desligamento SDV (*shutdown valve*) é disponível para interromper a produção durante uma situação de emergência e a válvula *choke*, de estrangulamento, que regula a taxa de fluxo de produção. Diferentes dinâmicas de fluxo são obtidas a depender dos valores das pressões de elevação de gás (PT4 e PT4a) e de fundo de poço (PT1).

Tabela 1. Variáveis com dados coletados, sendo P (Pressão) e T (Temperatura)

Tag	Descrição	Unidades
PT1	P PDG	kgf/cm^2
TT1	T PDG	°C
PT2	P na árvore de Natal Molhada	kgf/cm^2
TT2	T na árvore de Natal Molhada	°C
PT2a	P anular	kgf/cm^2
PT3a	P a montante da SVD de Produção	kgf/cm^2
PT3	P a montante da válvula *Choke* de Produção	kgf/cm^2
TT3	T a montante da válvula *Choke* de Produção	°C
PT4a	P a montante da SVD de *gas lift*	kgf/cm^2
TT4	T a montante da da SVD de *gas lift*	°C
FT4	Fluxo de *gas lift* instantâneo	m^3/h
FV4	Posição da válvula de *gas lift*	%
PT4	P a jusante da válvula de *gas lift*	kgf/cm^2
PT4	P do *header* de *gas lift*	kgf/cm^2
SDVL	Atuação da SDV de *gas lift*	*on/off*
SDVP	Atuação da SDV de produção	*on/off*

A árvore de natal molhada é um conjunto submarino composto com várias válvulas operadas remotamente por meio de comandos hidráulicos, onde estão contidos por exemplo os sensores TPT (Transmissor de Pressão e Temperatura) e o sensor que mede a pressão anular na válvula (PT2a) de *gas lift*. As pressões e temperaturas também são aferidas de acordo com a necessidade, bem como a montante e jusante das SDV, válvula *choke* e válvula de injeção de *gas lift*.

A válvula *choke* é um instrumento de controle de fluxo à jusante. Já a válvula de *gas lift* é um dispositivo destinado a auxiliar no controle da vazão de gás do anular para a coluna do poço. As altas pressões de *gas lift* vêm dos compressores da própria instalação. Tais equipamentos citados neste paragrafo estão instalados na plataforma de produção *offshore*. Uma descrição mais detalhada sobre este processo é feita por (Aguirre et al., 2017).

2.2 Autoencoder

O *autoencoder* é um tipo de Rede Neural Artificial (RNA) que é formada por três camadas, sendo o *encoder* constituído pelas duas primeiras camadas e as duas últimas configurando o *decoder*, como apresentado na Figura 2. O *autoencoder* tem a função de mapear o mais próximo possível a entrada em sua camada de saída. Geralmente os *autoencoders* têm em sua camada oculta um número inferior de neurônios comparado aos das suas camadas de entrada e saída. Isso é benéfico em relação a diminuição da dimensionalidade dos dados, que faz com que o *autoencoder* utilize apenas as principais características dos dados de entrada, com o intuito de eliminar descritores de pouca relevância para os modelos. Além de reduzir a dimensão o *autoencoder* também transforma os dados não linearmente, propiciando a maximização das diferenças entre as classes.

Na Figura 2 apresenta a estrutura de um *autoencoder*, onde os dados de entrada são $x = (x_1, x_2, ...x_n)$, os valores de saída do *autoencoder* são $z = (z_1, z_2, ...z_n)$, sendo n o número de neurônios tanto na camada de entrada quanto na de saída. O vetor $h = (h_1, h_2, ...h_m)$ é a representação da entrada x na camada oculta após a utilização de uma função de ativação sigmóide (sf) e m é a quantidade de neurônios na camada escondida. As equações que regem esse tipo de modelo são descritas por:

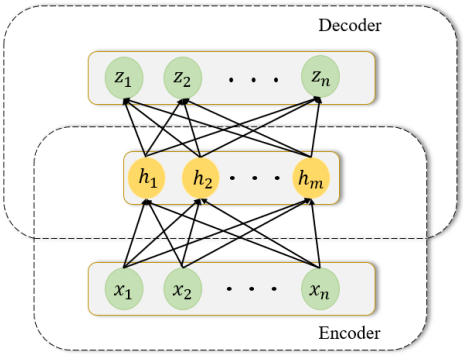

Figura 2. Estrutura de um *autoencoder*.

$$h = sf\left(W^{(1)}x + b^{(1)}\right) \tag{1}$$

$$sf\,(t) = 1/\left(1 + e^{-t}\right) \tag{2}$$

sendo $W^{(1)}$ a matriz de pesos associados aos neurônios de entrada e $b^{(1)}$ o vetor de bias da camada de entrada. Após etapa do *encoder* é necessária a reconstrução dos dados para se encontrar o vetor de saída z:

$$z = sf\left(W^{(2)}h + b^{(2)}\right) \tag{3}$$

em que $W^{(2)}$ é a matriz de pesos associados aos neurônios de saída e $b^{(2)}$ o vetor de *bias*.

As funções de otimização dos *autoencoders* são apresentadas em (Lu et al., 2016; Abdellatif et al., 2018). Elas são aplicadas para otimizar os parâmetros $\theta = \left\{W^1, b^1, W^2, b^2\right\}$ na construção do *autoencoder*. A função de custo a ser minimizada, $E(\theta)$, durante a otimização dos parâmetros da rede, é formada por três parcelas:

$$E(\theta) = J_{MSE}(\theta) + J_{Sparse}(\theta) + J_{weight}(\theta). \tag{4}$$

A primeira parcela é definida pelo erro médio quadrático MSE (do inglês, *Mean Square Error*) de um *autoencoder*, apresentada por (Wen et al., 2019).

$$J_{MSE}(\theta) = \frac{1}{n}\sum_{i=1}^{n} L_{MSE}(x_i, z_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\left\|(x_i - z_i)\right\|^2\right) \tag{5}$$

Dada uma amostra de entrada x, onde ρ_j $(j = 1, ..., s)$ é a ativação média da unidade oculta, compondo a segunda parcela da função de otimização, que pode ser definida por (Wen et al., 2019; Lu et al., 2016):

$$J_{Sparse}(\theta) = \beta\sum_{j=1}^{s_2} KL(\rho, \hat{\rho_j}), \tag{6}$$

sendo,

$$KL(\rho, \hat{\rho_j}) = \rho\,\log\frac{\rho}{\hat{\rho_j}} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho_j}}, \tag{7}$$

e

$$\hat{\rho_j} = \frac{1}{n}\sum_{i=1}^{n}\left[h_j(x_i)\right], \tag{8}$$

em que β é o parâmetro de ajuste de peso, que determina a proporção de dispersividade empregada no processo de representação esparsa, s_2 é o número de neurônios na segunda camada, $\hat{\rho_j}$ é o valor médio de ativação para a j-ésima unidade de camada escondida, ρ é o parâmetro de dispersividade e n refere-se ao número de entradas. Observa-se que um termo a mais foi adicionado na divergência de Kullback–Leibler (KL) que penaliza $\hat{\rho_j}$ ao se desviar significativamente de ρ, conforme formulado em (Lu et al., 2016).

Por fim, para evitar o *overfitting* também há um termo de decaimento que é somado aos demais termos para se encontrar a função de erro de um *autoencoder*:

$$J_{weight}(\theta) = \frac{\lambda}{2}\sum_{l=1}^{2}\sum_{i=1}^{S_l}\sum_{j=1}^{S_l+1}\left(w_{ij}^{(l)}\right)^2 \tag{9}$$

em que λ é um termo de regularização para ajudar a evitar o *overfitting* , diminuindo a magnitude dos pesos e S_l denota o número de neurônios totais na camada l.

3. MATERIAIS E MÉTODOS

3.1 Dados Experimentais de um Poço

A aquisição dos dados foi realizada em uma plataforma de petróleo *offshore*, extraídos do *plant information system* PI *System* da OSIsoft, amplamente utilizado na indústria petrolífera. O PI *System* consiste em um sistema que armazena informações da planta de processo. Por meio deste sistema, os dados foram coletados em aproximadamente 90 dias corridos, perfazendo uma janela de observações de um poço de produção de petróleo. A extração dos dados ocorreu entre os dias 06/12/2018 a 06/03/2019, totalizando 129592 observações, com intervalo de amostragem de 1 minuto.

As variáveis utilizadas na concepção dos classificadores estão na Tabela 1, no total 16 variáveis. Os sensores e atuadores na planta de processo são divididos em variáveis de topo e variáveis de fundo no supervisório. Sendo que, para as variáveis de topo, os seus sensores estão localizados na plataforma de produção de petróleo, já para as variáveis de fundo, os instrumentos são instalados no leito do mar e na coluna de produção, podendo ser mais suscetíveis à ruídos e falhas.

O diagnóstico de falha pode ser identificado a partir de certas variáveis inerentes ao processo. Essas variáveis, que se alteram na ocorrência de falhas de acordo com as características intrínsecas dos poços, assim como as condições físicas e químicas podem variar de poço a poço. Para o poço estudado neste trabalho, foi observada uma Falha Suave não definida, ou seja, não se sabe a sua origem, entre os dias 07/02/2019 e 19/02/2019, a qual ocorre a diminuição do fluxo de fluidos, reduzindo lentamente a produção até cessamento total, conforme pode ser observado na Figura 3.

As características de pressão PDG, temperatura na árvore de natal molhada e temperatura a montante da válvula *choke*, de modo a exemplificar este tipo de anomalia com esses três sensores, são apresentadas na Figura 4, onde indicam uma anormalidade no comportamento do poço descrito conforme abaixo:

- Pressão PDG (PT1): tende a ser elevada a partir da ocorrência de uma falha, devido a uma obstrução ao longo do caminho dos escoamento dos fluidos, elevando a pressão na coluna de produção;
- Temperatura na árvore de natal molhada (TT2): tende a se equilibrar com a temperatura do leito do mar, no caso de falhas de escoamento de fluidos, em água profundas estas temperaturas variam normalmente entre 2°C e 6°C;
- Temperatura a montante da válvula *choke* (TT3): tende a se igualar à temperatura na superfície do mar;

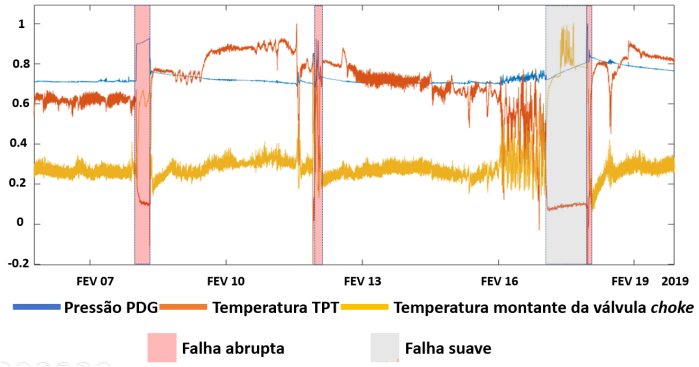

Figura 3. Falhas ocorridas entre os dias 07/02/2019 a 19/02/2019

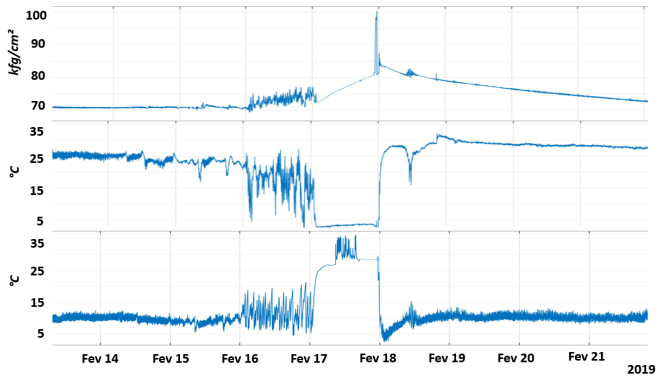

Figura 4. Comportamento da Pressão PT1, Temperatura TT2 e Temperatura TT3, respectivamente entre os dias 14/02/2019 a 21/02/2019.

A janela de 129592 observações foi dividida em duas partes de tamanhos iguais, cada uma com 64796 observações, indicada na Figura 5, com as 16 variáveis do processo exibidas. Os valores dos sensores são normalizados nas figuras para facilitar a visualização dos gráficos e as classes foram definidas como Falha e Não Falha.

A rotulagem foi proposta pelos autores, de acordo com informações repassadas por operadores da produção de poços de petróleo *offshore*, tanto operadores de campo, quanto operadores de sala de controle (supervisório). Por meio de experiências ao longo de suas carreiras na área

de óleo e gás. As falhas foram constatadas no supervisório de monitoramento da produção. Na classe Falha há dois padrões de ocorrência, Falha Abrupta e Falha Suave. A quantidade de observações para a classe Não Falha é de 126577 e 3015 para a classe Falha, sendo que Falha Abrupta e Falha Suave contem 2143 e 872 observações respectivamente. Na Tabela 2 é apresentada a divisão das classes nas duas janelas de observações. As informações sobre as classes são descritas conforme:

- Não Falha: nesta classe não ocorre variação que cause danos ao processo ou produção, a operação dos poços se comporta de maneira normal, dentro de uma estabilidade conhecida;
- Falha Abrupta: este tipo de falha decorre de um evento rápido, podendo ser a atuação de um elemento final, como válvulas ou atuadores do processo. Este tipo de falha pode gerar um alerta no supervisório da plataforma, ela é observada de maneira ágil pelos operadores do processo, agindo de forma abrupta na variável PT1, pressão PDG, por exemplo;
- Falha Suave: a ocorrência deste tipo de falha não é perceptível pelos operadores do supervisório do processo de forma ágil, a elevação da pressão PT1 acontece de forma lenta até ocorrer uma Falha Abrupta.

Tabela 2. Divisão das classes nas janela de dados

Dados	Falha	Não Falha
Primeira Janela	498	64298
Segunda Janela	2517	62279

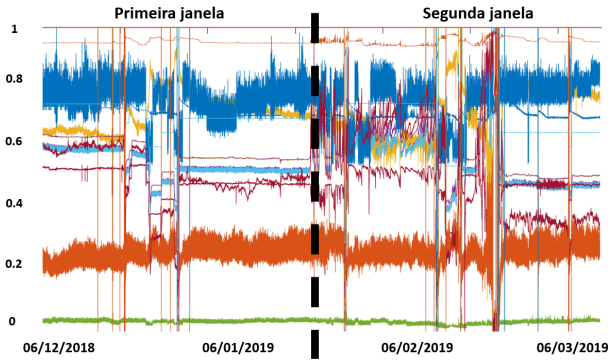

Figura 5. Conjunto de dados divididos em dois grupos

A divisão entre as classes é exibida na Figura 3, onde se visualiza os valores dos sensores PT1, TT2 e TT3, no intervalo entre os dias 07/02/2019 a 19/02/2019, vale salientar que não existe a ocorrência de Falha Suave na primeira janela de dados. A classificação das falhas para treinamento da camada de saída foi realizada aplicando os valores [0,1], para que o alvo a ser alcançado na saída do classificador seja:

- se saída = 0 : Falha;
- se saída = 1 : Não Falha.

3.2 Construção dos Detectores

Na construção dos detectores (ou modelos) de falha são empregados dois *autoencoders* empilhados, o primeiro tem a camada escondida com 9 neurônios e o segundo com 5 neurônios. O treinamento do primeiro *autoencoder* é realizado com as 16 variáveis contidas na Tabela 1 como entrada e o treino do segundo *autoencoder* utiliza a saída da camada oculta do primeiro *autoencoder* como entrada. A saída da camada oculta do segundo *autoencoder* é empregada como entrada para o treinamento dos detectores, ou seja, os *autoencoders* são utilizados como pré-processamento dos dados e redução da dimensionalidade de 16 para 5 variáveis de entrada. Foi definida previamente pelos autores uma redução de características de aproximadamente 70% da quantidade inicial de variáveis de entrada, com a intenção de reduzir a complexidade computacional do treinamento dos detectores.

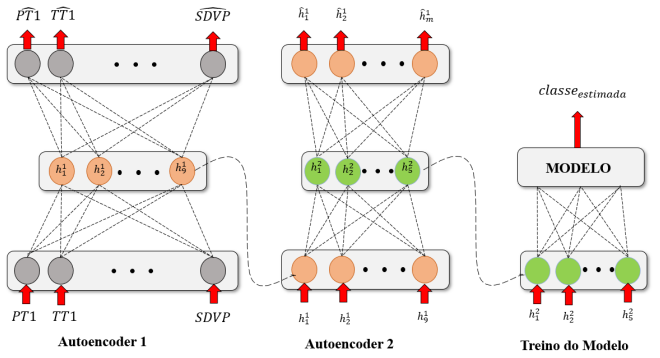

Figura 6. Processo de Construção de um modelo *stacked autoencoders* para classificação de falhas em poços de produção com elevação por *gas lift*.

A Figura 6 apresenta o processo de construção do modelo proposto, onde $v = \{PT1, TT1, ..., SDVP\}$ representa as variáveis de entrada do modelo, variáveis essas contidas na Tabela 1 do processo de extração de petróleo, $\hat{v} = \{\hat{v_1}, \hat{v_2}, ...\hat{v_{11}}\}$ são as variáveis de entrada estimadas na saída dos *autoencoders*, $h^1 = \{h_1^1, h_2^1, ...h_9^1\}$ são os valores do vetor h^1 na camada oculta do primeiro *autoencoder* e $\hat{h}_1 = \{\hat{h}_1^1, \hat{h}_2^1, ...\hat{h}_9^1\}$ a saída estimada de h^1 no segundo *autoencoder*. Os valores da camada escondida do segundo *autoencoder* são $h^2 = \{h_1^2, h_2^2, ...h_5^2\}$ que são os parâmetros de entrada dos detectores testados conforme a Figura 7, que tem como saída a classe estimada. Também são criados modelos sem a redução de dimensionalidade, utilizando as 16 características de entrada disponíveis, a fim de comparação do custo computacional e acurácia da técnica proposta.

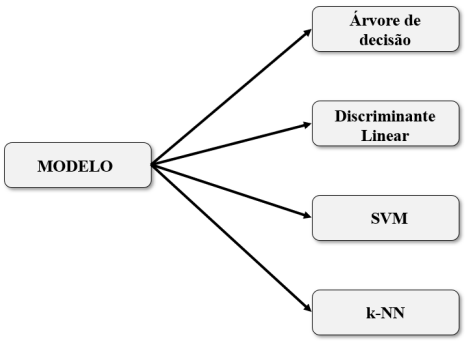

Figura 7. Modelos testados na classificação de falhas em poços de produção com elevação por *gas lift*.

São desenvolvidos nesse trabalho quatro modelos comumente empregados em reconhecimento de padrões: Árvore de Decisão (Breiman et al., 1984), Análise de Discriminante Linear (McLachlan, 1992), Máquina de Vetores de Suporte (Cortes and Vapnik, 1995) e K vizinhos mais próximos (Altman, 1992).

Para a análise do desempenho dos classificadores é utilizada as métricas *recall*, *precission* e *f1-score*. *Recall* é definido como o número de verdadeiros positivos sobre o número de verdadeiros positivos mais o número de falsos negativos. *Precission* é definido como o número de verdadeiros positivos sobre o número de verdadeiros positivos mais o número de falsos positivos. Já *F1-score* é definida como a média harmônica de *precission* e *recall* (Goutte and Gaussier, 2005).

Para o desenvolvimento dos modelos propostos são empregados classificadores disponíveis na ferramenta *Classification Learner*, aplicativo do *software* Matlab®. Para a criação dos modelos, os hiperparâmetros foram ajustados na própria ferramenta, sendo exibidos na Tabela 3.

Tabela 3. Hiperparâmetros dos modelos desenvolvidos

Modelos	Hiperparâmetros
Árvore de Decisão	Número máximo de divisões: 25 Critério de impureza: Índice Gini
Discriminante	Estrutura de covariância: Total
SVM	Função kernel: Gaussiana Escala Kernel: 0,56 Método Multiclasses: *One-vs-One*
k-NN	Número de Vizinhos: 7 Métrica de distância: Euclidiana Pesos das distâncias: Igual

A técnica de validação cruzada é aplicada para avaliar a capacidade de generalização dos modelos, a partir de um conjunto de dados (Kohavi, 1995). Dentre as técnicas de validação cruzada, as mais utilizadas são: *holdout*, *k-fold* e *leave-one-out*. Neste trabalho é aplicada a técnica *k-fold*.

No método *k-fold* os dados são divididos em parcelas (*folds*) de tamanhos iguais e exclusivas entre si, sendo k o número de divisões. Após a divisão dos dados em cada subconjunto, uma partição é utilizada para validação e as demais $(k-1)$ são utilizadas para estimação dos parâmetros, sendo este processo repetido k vezes a fim de se obter o desempenho do modelo.

Nesse trabalho, são empregados 2 *folds*. Nas observações utilizadas para treino cada classe foi particionada em dois conjuntos de tamanhos iguais, utilizados para treino, teste e validação. O segundo conjunto de observações é aplicado apenas para teste dos modelos constituídos, esse processo é repetido uma vez.

4. RESULTADOS E DISCUSSÕES

Os modelos são implementados com as classes tendo quantidades diferentes de observações, como apresentado na Seção 3. Apesar dessa discrepância, a maioria dos modelos classificaram de forma satisfatória o conjunto de observações. As métricas *recall*, *precision* e *f1-score* dos modelos desenvolvidos, em treino e teste, são apresentados, para análise do desempenho dos classificadores. Nessas tabelas são apresentados os resultados dos modelos SVM e k-NN com todas as 16 entradas, ou seja, sem implementar os *autoencoders*, pois esses classificadores obtiveram um desempenho superior na métrica *f1-score*, com a extração de características.

4.1 Resultados primeira janela de dados

Os resultados da primeira janela de dados são apresentados, sendo representados em tabelas de treino e teste nas Tabelas 4 e 5 respectivamente.

Tabela 4. Valores das métricas *recall*, *precission* e *f1-score* para os dados de treino da primeira janela de dados

Modelos	*recall*	*precission*	*f1-score*
Árvore de Decisão	0.9317	0.9299	0.9308
Discriminante	0.7871	0.9584	0.8644
SVM	0.8916	0.9933	0.9397
k-NN	0.9900	0.9801	0.9850
SVM sem redução	0.7972	1	0.8872
k-NN sem redução	0.9839	0.9899	0.9869

Tabela 5. Valores das métricas *recall*, *precission* e *f1-score* para os dados de teste da primeira janela de dados

Modelos	*recall*	*precission*	*f1-score*
Árvore de Decisão	0.8418	0.8012	0.8210
Discriminante	0.9062	0.7952	0.8471
SVM	0.7878	0.8574	0.8212
k-NN	0.8846	0.8313	0.8571
SVM sem redução	0.8692	0.8421	0.8554
k-NN sem redução	0.9195	0.8996	0.9095

4.2 Resultados segunda janela de dados

Os resultados da segunda janela de dados são apresentados, sendo representados em tabelas de treino e teste nas Tabelas 6 e 7 respectivamente.

Tabela 6. Valores das métricas *recall*, *precission* e *f1-score* para os dados de treino da segunda janela de dados

Modelos	*recall*	*precission*	*f1-score*
Árvore de Decisão	0.9050	0.9519	0.9279
Discriminante	0.8931	0.9301	0.9112
SVM	0.9750	0.9883	0.9816
k-NN	0.9825	0.9932	0.9878
SVM sem redução	0.9428	1	0.9706
k-NN sem redução	0.9944	0.9928	0.9936

Tabela 7. Valores das métricas *recall*, *precission* e *f1-score* para os dados de teste da segunda janela de dados

Modelos	*recall*	*precission*	*f1-score*
Árvore de Decisão	0.7799	0.3758	0.5072
Discriminante	0.9329	0.4529	0.6098
SVM	0.4000	0.0064	0.0125
k-NN	0.9943	0.4819	0.6492
SVM sem redução	0	0	0
k-NN sem redução	0.9956	0.5447	0.7042

4.3 Discussões

Ao utilizar o primeiro grupo de dados para treino e teste, obteve-se um resultado inferior de *f1-score*, em relação ao resultado obtido quando utilizada a segunda janela de observações para treino e teste. Esse fato ocorre devido ao número reduzido de observações da classe Falha no primeiro conjunto dados e também por este conjunto não possuir observações de Falha Suave, que são mais difíceis de serem detectadas. Os resultados dos dados de treinos são importantes para análise mais profunda dos classificadores, pois se consegue distinguir quais classificadores conseguem melhor desempenho nos dois conjuntos.

Os modelos k-NN obtiveram, de forma geral, desempenhos superiores às outras técnicas nas métricas calculadas, essa superioridade é observada especialmente no teste do modelo treinado na primeira janela de dados e testado na segunda. Apesar do modelo k-NN sem aplicação dos *autoencoders* ter alcançado melhores métricas em relação à sua versão com redução de dimensionalidade, observa-se que no quesito *recall* esta diferença no desempenho é ainda menor. Para o problema deste trabalho, essa técnica é mais importante pois a presença de falso negativo é mais problemática. O custo de um falso negativo em geral é maior, sendo esses custos diferentes, um evento anormal classificado como normal é mais prejudicial do que um evento normal classificado como anormal.

Ou seja, além do fato do uso de *autoencoders* não afetar sobremaneira os índices de desempenho, a redução de dimensionalidade propicia uma diminuição do tempo de treinamento dos modelos. Quando utilizada a técnica k-NN, esse fator está na ordem de aproximadamente oito vezes. A taxa de treinamento é de 12.000 observações por segundo para o modelo com redução e 1.600 observações por segundo sem redução.

Já os modelos treinados com a técnica SVM, quando comparados entre si, são bem próximos os seus resultados. Utilizando a segunda janela para treinamento e a primeira para teste, o desempenho é nulo no modelo com todas as variáveis de entrada, apesar do bom desempenho em dados de treino, toda a classe Falha foi classificada como Não Falha, ou seja, todos falsos negativos. Mesmo com os resultados da primeira janela sendo elevados, o desempenho deste modelo é muito inferior aos demais, com sua aplicabilidade nula para este problema em específico. Já no outro experimento, os modelos SVM obtiveram desempenho satisfatório e próximos, tanto em dados de treino, quanto em dados de teste.

Os classificadores baseados em árvore de decisão e discriminante conseguiram bons resultados. Como todos o classificadores propostos, o desempenho dessas duas técnicas é inferior quando os modelos são testados na segunda janela de observações, o modelo discriminante conseguiu atingir um elevado valor de *recall* nessa janela.

5. CONCLUSÕES

A maioria dos modelos de detecção de falhas apresentados nesse trabalho classificaram satisfatoriamente as observações, mesmo não havendo homogeneidade na quantidade de observações rotuladas de cada classe. Em especial, o modelo k-NN foi o que alcançou os melhores resultados, principalmente na métrica recall, considerada mais importante para este tipo de problema.

A rede de *autoencoders* em cascata utilizada em conjunto de outras técnicas para a classificação de falhas em poços de petróleo com elevação artificial por *gas lift*, apresenta uma grande aplicabilidade na diminuição da dimensionalidade dos dados. Os *autoencoders* possibilitaram a redução do tempo de treinamento, mantendo próximos o desempenho dos modelos ao comparar com modelos sem seu uso. Isso mostra que a utilização desta técnica pode ser viável nos mais variados tipos de indústria, onde são necessárias respostas rápidas às anormalidades nos sistemas de monitoramento do processo de produção. Esse tratamento mais ágil tende a reduzir a complexidade de ações diretas para retorno da normalidade nas plantas de processo.

Para trabalhos futuros, recomenda-se a aplicação de outros modelos de detecção de falhas, como outras técnicas de redução de características dos dados, verificando o seu desempenho em comparação às desenvolvidas. A utilização dos modelos em outros poços de produção com elevação por *gas lift* permitirá a verificação da generalização e compreensão dos resultados obtidos, além de simulações e estudos da aplicabilidade em sistemas em tempo real.

AGRADECIMENTOS

Os autores agradecem à Petróleo Brasileiro S.A. (Petrobras) pela disponibilidade dos dados para o trabalho e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo suporte financeiro.

REFERÊNCIAS

Abdellatif, S., Aissa, C., Hamou, A.A., Chawki, S., and Oussama, B.S. (2018). A deep learning based on sparse auto-encoder with mcsa for broken rotor bar fault detection and diagnosis. In *2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)*, 1–6.

Aguirre, L.A., Teixeira, B.O., Barbosa, B.H., Teixeira, A.F., Campos, M.C., and Mendes, E.M. (2017). Development of soft sensors for permanent downhole gauges in deepwater oil wells. *Control Engineering Practice*, 65, 83 – 99.

Altman, N. (1992). An introduction to kernel and nearest-neighbor nonparametric regression.

Araujo, M., Aguilar, J., and Aponte, H. (2003). Fault detection system in gas lift well based on artificial immune system. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, 1673–1677 vol.3.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and regression trees. belmont, ca: Wadsworth. *International Group*, 432, 151–166.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. In *Machine Learning*, 273–297.

Filho, H.S.R. (2011). *A Otimização de Gás Lift na Produção de Petróleo: Avaliação da Curva De Performance do Poço.* Master's thesis, Universidade Federal do Rio de Janeiro, Rio de Janeiro - RJ.

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Proceedings of the 27th European Conference on Advances in Information Retrieval Research*, ECIR'05, 345–359. Springer-Verlag, Berlin, Heidelberg.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. volume 14.

Lu, C., Wang, Z.Y., Qin, W.L., and Ma, J. (2016). Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification. *Signal Processing*, 130.

McLachlan, G. (1992). *Discriminant analysis and statistical pattern recognition*. Wiley Series in Probability and Statistics. Wiley.

Nilchiani, R., Edwards, C.M., and Ganguly, A. (2019). Introducing a tipping point measure in explaining disruptive technology. In *2019 International Symposium on Systems Engineering (ISSE)*, 1–5.

Santos, I.H., Lisboa, H.F., de S. Feital, T., Câmara, M.M., Soares, R.M., Marins, M.A., Barros, B.D., de M. Prego, T., de Lima, A.A., and Netto, S.L. (2018). Hydrate failure detection in production and injection lines using model and data-driven approaches. In *Rio Oil Gas Expo and Conference 2018*. Rio de Janeiro - RJ.

Schiavi, M.T. and Hoffmann, W.A.M. (2015). Cenário petrolífero: sua evolução, principais produtores e tecnologias. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, 13(2), 259–278.

Thomas, J. (2004). *Fundamentos de engenharia de petróleo*. Interciência.

Vargas, R.E.V., Munaro, C.J., Ciarelli, P.M., Medeiros, A.G., do Amaral, B.G., Barrionuevo, D.C., de Araújo, J.C.D., Ribeiro, J.L., and Magalhães, L.P. (2019). A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering*, 181, 106223.

Wen, L., Gao, L., and Li, X. (2019). A new deep transfer learning based on sparse auto-encoder for fault diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(1), 136–144.

Wen, P., Wang, T., Xin, B., Tang, T., and Wang, Y. (2018). Blade imbalanced fault diagnosis for marine current turbine based on sparse autoencoder and softmax regression. In *2018 33rd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, 246–251.

Xuewu Dai, Guangyuan Liu, and Zhengji Long (2008). Discrete-time robust fault detection observer design: A genetic algorithm approach. In *2008 7th World Congress on Intelligent Control and Automation*, 2843–2848.