Detecção de anomalias em poços de petróleo surgentes com *Stacked Autoencoders*

Rodrigo Scoralick Fontoura do Nascimento [*]
Bruno Henrique Groenner Barbosa [**]
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 Exploration and Production (E&P) is responsible for most of the oil and gas production in Brazil. Due to the high level of complexity in this industry, new technologies have been proposed over the past few years. The present work aimed at developing systems for fault detection in offshore oil production wells. The public domain 3W dataset was used and stacked autoencoders were implemented for dimensionality reduction. Measurements of five process variables were used as inputs for classification with examples from a single class. Isolation Forest and Support Vector Machines of a class were the techniques used to detect anomalies in the process, such as hydrate in the production line. The results were compared with other works in the literature, and an improvement of up to eighteen percent was observed. Moreover, the designed autoencoders were effective in dimensionality reduction, helping to find more parsimonious classifiers.

Resumo: O segmento de Exploração e Produção (E&P) *offshore* é responsável pela maior parte da produção de petróleo e gás no Brasil. Devido à complexidade dessa indústria, ela vem demandando novas tecnologias ao longo dos últimos anos. O presente trabalho teve como objetivo o desenvolvimento de sistemas de detecção de falhas (anomalias) em poços de produção de petróleo *offshore* surgentes. Foram utilizados os dados de domínio público 3W *dataset* e *Autoencoders* empilhados foram empregados para redução de dimensionalidade. Medições de cinco variáveis de processos foram utilizadas como entradas para classificação com exemplos de uma única classe. Floresta de Isolamento e Máquinas de Vetores de Suporte de uma classe, são as técnicas empregadas para detectar anomalias no processo, como hidrato em linha de produção. Os resultados foram comparados com outros trabalhos da literatura, nos quais é observada uma melhora de até dezoito pontos percentuais. Ademais, foi possível observar que os *autoencoders* foram eficazes na redução de dimensão em problemas de detecção de anomalias em poços de produção de petróleo *offshore* auxiliando a obtenção de classificares mais parcimoniosos.

Keywords: Autoencoders; Fault detection; Oil well monitoring; Multivariate time series classification; one-class classification.

Palavras-chaves: Autoencoders; Detecção de falhas; Monitoramento de poços de petróleo; Classificação multivariada de séries temporais; Classificação com exemplos de uma única classe.

1. INTRODUÇÃO

O petróleo é uma grande fonte de energia do planeta, utilizado em todas as nações por meio dos seus derivados, sendo insumo para as mais variadas indústrias. Considerado como "ouro negro", o petróleo foi base de uma revolução industrial e motivo para diversas guerras ao redor do mundo (Espinola, 2013).

No Brasil a produção de petróleo *offshore* é a mais importante para a indústria petrolífera. De acordo com a ANP, os campos marítimos foram responsáveis por cerca de 96% da produção de petróleo nacional no mês de fevereiro de 2021 (ANP, 2021). Assim como cresce a produção do insumo petrolífero, cresce também a necessidade de ampliar os avanços tecnológicos em toda a cadeia de projeto e concepção de métodos e equipamentos (Espinola, 2013). Sendo

um objetivo essencial a detecção de falhas decorrentes do processo.

A ocorrência de anomalias em poços de produção de petróleo *offshore* pode gerar prejuízos de milhares de dólares para empresas produtoras. Além disso, uma complexa operação pode se suceder após a ocorrência das falhas, a fim de restabelecer a normalidade na operação dos poços (Santos et al., 2018), tornando a detecção de anomalias importante ferramenta para o negócio.

Em função da revolução da informação, dados passaram a ser gerados de forma abundante. Surgiu então o conceito de *Big Data*, o qual refere-se ao enorme volume de dados, estruturados ou não, que impacta os negócios em praticamente qualquer empresa (Machado, 2018). A informatização da indústria traz benefícios no modo de produzir, sendo a Indústria 4.0 um exemplo de como estas modificações podem desenvolver amplamente o modo de produção. Estes conceitos são fundamentais para a aplicação de técnicas que facilitam a operação na indústria de óleo e gás, mais especificamente em plataformas de produção de petróleo, onde se pretende automatizar os mais diversos processos em busca de melhorias contínuas na industria de exploração de petróleo *offshore* (Aguirre et al., 2017).

Como exemplo da aplicação de uma técnica em detecção de eventos indesejáveis na indústria de petróleo e gás, abordado por Araujo et al. (2003), um Sistema Imunológico Artificial (SIA) foi desenvolvido para detecção de falhas em poços de produção de petróleo com elevação artificial por *gas lift* . São rotuladas duas classes, que estão relacionadas às condições de operação da produção (normais e anormais) das pressões nos poços. São utilizados 80% dos dados para treinamneto dos detectores de falhas e 20% para testes do sistema. Este trabalho gerou detectores que determinam desvios no processo de produção.

Já Santos et al. (2018) propõem a detecção de eventos indesejáveis de acumulação de hidrato em linhas de injeção de água ou de produção de petróleo *offshore*. Anomalias essas que resultam em diminuição de vazão de óleo ou até perda total da produção do poço. Detectam-se as falhas em poços de produção surgentes. Com o objetivo de se retirar alarmes correlacionados, esse trabalho foi capaz de diferenciar o comportamento normal e o defeituoso com 77% de acurácia. As 12 falhas por hidrato analisadas foram detectadas no trabalho, sendo que 85% delas de forma antecipada.

Nascimento et al. (2020) implementam classificadores para detectar falhas em poços de petróleo *offshore* não surgentes com elevação artificial por *gas lift*. São desenvolvidos quatro classificadores, tendo como resultado métricas *F1 score* acima de 0,9.

Este trabalho apresenta uma nova abordagem para a construção de detectores de anomalias em poços de petróleo *offshore* surgentes por meio da implementação de *Autoencoders*. Os resultados obtidos são comparados com os obtidos em outros trabalhos como Vargas (2019) e Junior et al. (2020), onde são apresentados métodos distintos para construção de detectores de anomalias utilizando os dados 3W *dataset* (Vargas et al., 2019).

Este trabalho está estruturado da seguinte forma. A próxima seção descreve o processo de elevação e anomalias em poços de petróleo, em conjunto com 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 O Processo de Elevação e Escoamento de Petróleo Offshore

Elevação é o termo utilizado na indústria petrolífera para caracterizar o processo de ascensão do fluido contido em um reservatório até a superfície (Thomas, 2004). Os poços de petróleo em águas profundas são classificados entre surgentes e não surgentes (de Morais et al., 2019). Poços não surgentes necessitam de métodos para auxiliar o escoamento dos fluidos (água, óleo, gás e sedimentos). Já os poços 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).

2.2 Falhas em Poços de Produção de Petróleo Offshore

Falhas ou anomalias podem ocasionar desde pequenas instabilidades nas linhas de produção a total parada de fluxo do reservatório para a plataforma de produção. A identificação automática dessas falhas pode auxiliar na operação de modo a minimizar perdas em poços de produção de petróleo. Por conseguinte, diminuindo custos de manutenção, encurtando tempo de atuação no problema, evitando gastos adicionais e operações complexas no retorno à operação normal dos sistemas de produção. Os tipos de falhas presentes no 3W *dataset* e utilizados neste trabalho são listados a seguir:

- Aumento Abrupto de *BSW*;
- Fechamento Espúrio de *DHSV*;
- Intermitência Severa;
- Instabilidade de Fluxo;
- Perda Rápida de Produtividade;
- Restrição Rápida na Válvula *Choke* de Produção;
- Incrustação na Válvula *Choke* de Produção;
- Hidrato em Linha de Produção.

2.3 Autoencoder

Shin et al. (2013) explicam que o *autoencoder* é um tipo de Rede Neural Artificial (RNA) formada por três camadas, sendo o *encoder* constituído pelos neurônios das duas primeiras camadas e os neurônios das duas últimas configurando o *decoder* como apresentado na Figura 1. Corroborando, Yu and Zhang (2020) argumentam que 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 à 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.

Métodos típicos de análise de dados como o de análise dos componentes principais PCA (do inglês, *Principal Component Analysis*) e análise dos componentes independentes ICA (do inglês, *Independent Component Analysis*) se distinguem dos *autoencoders* por não se desempenharem bem em problemas não-lineares. Embora esses métodos baseados em projeção de dados capturem a formação global dos dados, a estrutura local geralmente é ignorada (Yu and Zhang, 2020). Além disso, esses métodos não são eficazes para processos não-lineares, porque os recursos extraídos não são eficientes para descrever a distribuição dos sinais de processos complexos.

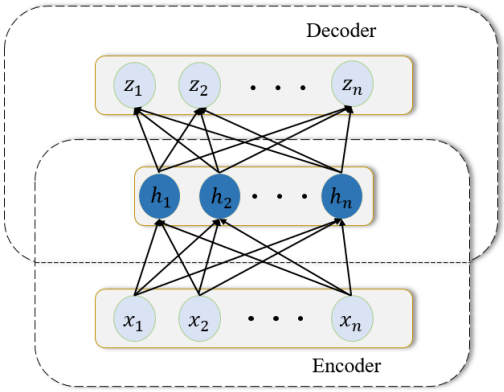

Figura 1. Estrutura de um *autoencoder*.

A Figura 1 apresenta a estrutura de um *autoencoder*, cujo vetor de dados de entrada é $\mathbf{x} = [x_1\ x_2\ ...\ x_n]$, já o vetor de dados de saída do é $\mathbf{z} = [z_1\ z_2\ ...\ z_n]$, sendo n o número de neurônios tanto na camada de entrada quanto na de saída. O vetor $\mathbf{h} = [h_1\ h_2\ ...\ h_m]$ é a representação da entrada $\mathbf{x}$ na camada oculta após a utilização de uma função de ativação sigmoide (sf) e m é a quantidade de neurônios na camada escondida. As equações que regem esse tipo de modelo são descritas por:

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

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

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

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

em que $\mathbf{W}^{(2)}$ é a matriz de pesos associados aos neurônios de saída e $\mathbf{b}^{(2)}$ 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 $\theta = \left\{\mathbf{W}^1, \mathbf{b}^1, \mathbf{W}^2, \mathbf{b}^2\right\}$ 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 quadrático médio (do inglês *Mean Squared Error*) de um *autoencoder* (Wen et al., 2019):

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

sendo N o número de amostras disponíveis.

Dada uma amostra de entrada $\mathbf{x}$, na qual ρ_j $(j = 1, ..., m)$ é a ativação média da unidade oculta j, a segunda parcela da função de otimização é definida por (Wen et al., 2019; Lu et al., 2016):

$$J_{Sparse}(\theta) = \beta\sum_{j=1}^{m_2} KL(\rho, \hat{\rho}_j), \tag{6}$$

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, m_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 *overfitting* 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 diminuir a magnitude dos pesos e S_l denota o número de neurônios totais na camada l.

2.4 Reconhecimento de Padrões

Neste trabalho são empregadas técnicas comumente utilizadas em reconhecimento de padrões, que é o ato de coletar dados brutos e tomar uma ação baseada na "categoria" do padrão. Suas etapas de desenvolvimento podem ser da

seguinte ordem: aquisição dos dados, pré-processamento, extração de características e classificação (Duda et al., 2001).

Os métodos de classificação abordados neste trabalho são: Máquina de Vetores de Suporte e Floresta de Isolamento.

2.5 Máquina de Vetores de Suporte

As máquinas de vetor de suporte (SVM, do inglês: *Support Vector Machine*) são modelos de aprendizado supervisionado que podem ser usados para detecção de falhas. Elas são classificadores de margem, que encontram um hiperplano que tem o objetivo de definir a classe para um novo ponto do conjunto de dados (Schlag et al., 2019).

Já o *one-class* SVM (OCSVM) é um algoritmo de classificação de apenas uma classe, sendo uma adaptação proposta por Schölkopf et al. (2001). O conceito do OCSVM consiste em encontrar uma hiperesfera em que a maioria das amostras de treinamento estão incluídas em um volume mínimo (Guerbai et al., 2014).

O algoritmo OCSVM primeiro mapeia os dados de entrada em um espaço de recursos de alta dimensão por meio de uma função kernel e, em seguida, encontra iterativamente o hiperplano de margem máxima, que separa melhor os dados de treinamento da origem. Assim, o hiperplano (ou limite de decisão linear) corresponde à função de classificação (Maglaras and Jiang, 2014).

2.6 Floresta de Isolamento

A Floresta de Isolamento é um algoritmo de aprendizagem não supervisionado para detecção de anomalias que funciona com base no princípio de isolar anomalias (Liu et al., 2008), em vez das técnicas mais comuns de criação de perfil de pontos normais (Chandola et al., 2009).

Trata-se de um algoritmo baseado em árvore de decisão tipo *ensemble*, que constitui uma floresta. Uma técnica é chamada de *ensemble* quando um conjunto de classificadores são treinados separadamente e as decisões são tomadas de forma combinada (Barbosa et al., 2019). Métodos *ensemble* tendem a reduzir *overtting* (Aggarwal and Sathe, 2017; Barbosa et al., 2011).

O princípio da Floresta de Isolamento é obter uma estrutura de árvores aleatórias para isolar um tipo de classe do seu conjunto de dados. As anomalias tem maior suscetibilidade ao isolamento e ficam mais perto das raízes das árvores, enquanto os pontos normais são mais difíceis de isolar e geralmente estão no extremo mais profundo das árvores (Junior et al., 2020).

O conjunto de treinamento deve ser composto por dados de apenas um tipo de classe, pois classes diferentes no treinamento tendem a abaixar a qualidade da função de decisão. Este método é, portanto, adequado para classificação *one-class* (Krawczyk et al., 2017).

3. METODOLOGIA

3.1 Banco de Dados

Vargas et al. (2019) disponibilizam dados de um conjunto de poços produtores de petróleo *offshore* chamado 3W *dataset*. Esses dados são advindos do Projeto MAE (Monitoramento de Alarmes Especialistas), da empresa Petróleo Brasileiro SA, concebido na Unidade de Negócios da Petrobras localizada no estado do Espírito Santo (UN-ES). Os dados estão no formato de Séries Temporais Multivariáveis (STM), do termo em inglês (*Multivariate Time Series*, MTS), que consistem em uma sequência de observações ordenadas em função do tempo e que apresentam intervalos de tempo iguais entre cada par de observações. O conjunto de dados é formado por instâncias, sendo que uma instância é composta por um conjunto de n séries temporais univariáveis (também referenciadas como variáveis). Todas as instâncias têm um número fixo de variáveis n, mas cada instância pode ser composta por qualquer quantidade de observações.

As observações pertencente ao conjunto de dados são variantes no tempo, nos quais as características físico-químicas dos reservatórios tendem a se alterar ao longo dos dias, meses e anos. Comumente supõe-se que existe correlação entre os dados passados e futuros (Morettin and Toloi, 2006). O conjunto de dados é dividido em oito tipos de eventos que podem causar perdas de produção em conjunto com a operação normal dos poços, conforme apresentado na Seção 2.

Para este trabalho, não foram utilizadas as variáveis diretamente relacionadas ao sistema de *gas lift*, pois a intenção é o desenvolvimento de detectores de anomalias em poços surgentes. As variáveis integrantes do processo surgente que compõe o banco de dados são:

- Pressão PDG;
- Pressão TPT;
- Temperatura TPT;
- Pressão a montante da válvula *choke* de produção;
- Temperatura a jusante válvula *choke* de produção.

3.2 Desafio Trabalhado

Vargas et al. (2019) propõem dois desafios a partir do banco de dados concebido em artigo publicado. Esses *benchmarks* são planejados apenas para detecção *online* (Vargas, 2019). Todas as observações de períodos transientes de anomalia e estados estáveis de anomalia devem ser novamente rotulados como positivos e todas as observações de períodos normais como negativas. Nessa operação, observações não rotuladas devem ser mantidas como estão. Neste trabalho, apenas um dos desafios propostos (desafio número 2) foi implementado.

3.3 Construção dos Modelos de Poços Surgentes

De acordo com as regras propostas por Vargas et al. (2019), os autores definiram uma sequência de passos na construção dos classificadores (Kadhim, 2019):

- Pré-processamento dos dados;
- Redução da dimensionalidade;
- Aplicação de técnicas de detecção;
- Análise e avaliação dos resultados obtidos.

Pré-processamento dos Dados: Observações com algum dado NaN (do inglês, *Not a Number*) foram interpoladas. As observações com transiente de anomalia foram re-rotuladas como anomalia, como apresentado na Figura 2.

Os dados foram normalizados, com média zero e variância unitária, com o objetivo de evitar que a diferença entre as escalas interferisse nos resultados (Vargas et al., 2019). Atrasos foram inseridos em cada observação, prefazendo um total 100 atrasos, ou seja, $(k-1),(k-2)...(k-100)$, conforme Figura 3. Após este procedimento, são descartadas as primeiras cem observações para treinamento. Após a inclusão dos atrasos, a quantidade de atributos de entrada passou de 5 para 500, ou seja, multiplicadas por 100. Este procedimento aumentou significativamente o volume de dados para o treinamento e teste.

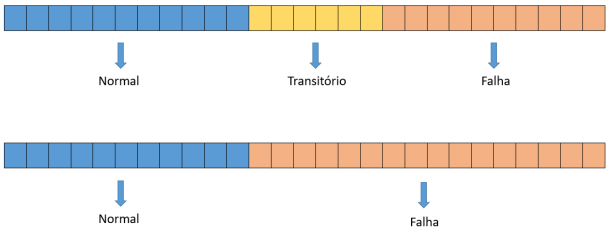

Figura 2. Dados de transitório re-rotulados como falha.

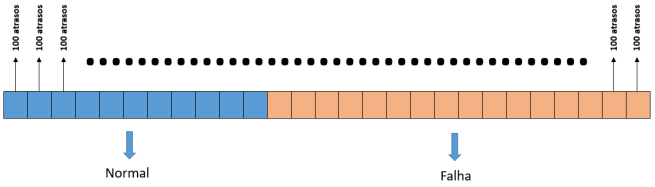

Figura 3. Atrasos nas observações.

Em períodos normais, as primeiras observações foram utilizadas para treinamento (60%) e as últimas para teste (40%). Os 40% restantes de amostras normais são misturados de forma aleatória com as amostras de anomalia, como observado na Figura 4. Esta partição dos dados normais foi baseada nos trabalhos de Vargas et al. (2019) e Junior et al. (2020).

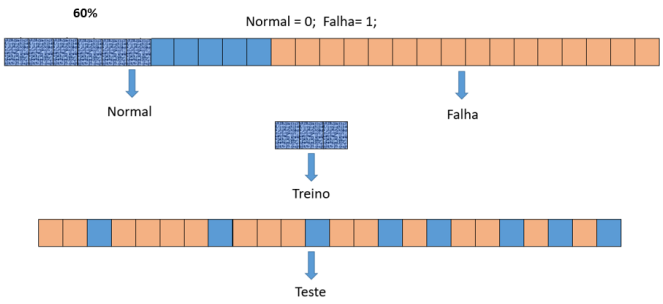

Figura 4. Observações de treino e teste.

Redução de Dimensionalidade: O processo de redução de dimensionalidade é realizado por meio de *autoencoders* em cascata. A Figura 5 mostra o processo de construção do modelo, utilizando 500 variáveis de entrada, onde $v = \{v_1, v_2, ...v_{500}\}$ são as variáveis de entrada do modelo, $\hat{v} = \{\hat{v_1}, \hat{v_2}, ...\hat{v_{500}}\}$ são as variáveis de entrada estimadas na saída dos *autoencoders*, $h^1 = \{h_1^1, h_2^1, ...h_{300}^1\}$ são os valores de saída da camada oculta do primeiro *autoencoder* e $\hat{h}^1 = \left\{\hat{h}_1^1, \hat{h}_2^1, ...\hat{h}_{300}^1\right\}$ a saída estimada de h^1 no segundo *autoencoder*. Os valores da camada escondida do segundo são $h^2 = \left\{h_1^2, h_2^2\right\}$ e também servem como parâmetros de entrada dos detectores. O número de atributos foi reduzido de 500 para 2, 10, 50 e 100, de forma a analisar o desempenho do *autoencoder*. Além disso, a técnica PCA também foi empregada para essa análise.

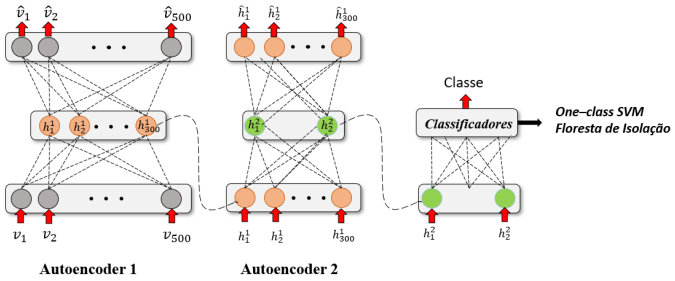

Figura 5. Redução de dimensionalidade e classificação.

Tanto no trabalho de Junior et al. (2020) como em Vargas (2019) é utilizada a estratégia específica de amostragem com janela deslizante. Também foram realizadas amostragens das instâncias com janela deslizante com a geração de até 15 amostras com 180 observações cada. Os autores extraíram e utilizaram como características: mediana, média, desvio padrão, variância, máximo e mínimo.

Aplicação de Técnicas de Detecção: Foram utilizadas duas técnicas para detecção de anomalias, OCSVM e Floresta de Isolamento, com código disponibilizado por Liu et al. (2008). Nos trabalhos de Vargas et al. (2019) e Junior et al. (2020) também são aplicadas essas técnicas e, por esse motivo, são escolhidas para criação dos modelos. Os hiperparâmetros dos detectores desenvolvidos neste trabalho não passaram por processo de otimização, eles foram obtidos do trabalho de Junior et al. (2020), onde já foram otimizados, que torna a comparação com este trabalho mais metodologicamente correta. Os hiperparâmetros dos detectores desenvolvidos podem ser consultados na Tabela 1.

Tabela 1. Hiperparâmetros dos modelos desenvolvidos para poços surgentes.

Modelos	Hiperparâmetros
Floresta de Isolamento	n estimators: 150 max samples: 1,0 max featuress: 1,0 bootstrap: False contamination: 0
OCSVM	Função kernel: rbf gama: 0,001 nu: 0,1

4. RESULTADOS E DISCUSSÕES

Foram realizadas 31 execuções de treinamento e validação para cada modelo, totalizando 496 avaliações conforme *benchmark* proposto. Os valores obtidos de *recall*, *precision* e *F1 score* (Goutte and Gaussier, 2005) dos modelos de Floresta de Isolamento e OCSVM foram calculados para cada modelo. A partir dos resultados obtidos, são usados os resultados apresentados por Vargas (2019) e Junior et al. (2020) para avaliação e comparação dos seus desempenhos.

Tabela 2. Resultados dos modelos desenvolvidos.

Modelo	OCSVM	Floresta de Isolamento
duas dimensões com *autoencoders*	0,6814 ± 0,2170	0,7997 ± 0,1532
dez dimensões com *autoencoders*	0,6926 ± 0,2068	0,8067 ± 0,1456
cinquenta dimensões com *autoencoders*	0,7033 ± 0,2069	0,8160 ± 0,1396
cem dimensões com *autoencoders*	0,7201 ± 0,1972	0,8277 ± 0,1360
cem dimensões com PCA	-	0,7956 ± 0,1428
dez dimensões com PCA	-	0,7328 ± 0,1507
sem inclusão de atrasos	-	0,6766 ± 0,1395
500 atrasos e sem redução	-	0,8345 ± 0,1329
Vargas et al. (2019)	0,5320 ± 0,0750	0,7430 ± 0,1790
Junior et al. (2020)	0,5670 ± 0,1620	0,7270 ± 0,1820

Vargas (2019) e Junior et al. (2020) apresentam os resultados de seus trabalhos por meio de tabelas que contêm os valores da média *F1* e o respectivo desvio padrão. Dessa forma, a Tabela 2 apresenta esses mesmos valores para todos os modelos desenvolvidos neste trabalho. Para os modelos desenvolvidos com PCA, sem inclusão de atrasos e sem redução são apresentados apenas detectores de Floresta de Isolamento, que obtiveram melhores resultados.

A partir da comparação dos resultados obtidos nota-se que a inserção de atrasos, promovendo uma expansão da quantidade de atributos ou características por observação, em conjunto com a aplicação de *autoencoders* na redução de dimensionalidade, conseguiu resultados satisfatórios de *F1 score* principalmente com a Floresta de Isolamento, que obteve resultados até onze pontos percentuais superiores ao OCSVM.

Observa-se também que o desempenho dos modelos obtidos neste trabalho é superior àqueles apresentados em Vargas et al. (2019) e Junior et al. (2020). O uso de uma janela de 500 atrasos é de suma importância para o desempenho dos classificadores propostos. Além disso, o modelo sem a inclusão de atrasos teve um desempenho inferior quando comparado aos com atrasos e redução de dimensionalidade. Um diferencial do trabalho proposto aos pares que são comparados neste artigo é a forma como as características são extraídas. O artigo aborda *autoencoders* para a extração de características já os demais seguem a linha de *handcrafted features*, como também explicado em Rahman et al. (2016) e Roy et al. (2018).

Com o intuito de analisar o desempenho dos *autoencoders* como técnicas de redução de dimensão, ao observar a redução de 500 para 10 atributos, é possível inferir que os *autoencoders* conseguem uma melhor representação do banco de dados nessa menor dimensão do que a PCA, sendo possível obter classificadores com melhores desempenhos, 81% contra 73%. No entanto, um fator que deve ser levado em consideração é o tempo de execução de cada abordagem, que são bem distintos. Para um poço que têm em torno de 100.000 observações, a técnica PCA aplicada para redução de 500 atributos para 100, foi aproximadamente 30 vezes mais rápida do que o *autoencoder*.

Quanto mais elevada a quantidade dimensional dos dados, melhor é o desempenho dos detectores. Isto se deve a um maior conjunto de características disponíveis na construção dos modelos. Mas, quanto maior a quantidade de entradas, crescerá a complexidade no momento do treinamento dos modelos, demandando um tempo maior neste processo. A partir destas premissas a redução de dimensionalidade contribui para a construção de modelos mais parcimoniosos para detecção de anomalias.

Quando não houve a redução de dimensionalidade, ou seja utilizando como entradas os 500 atributos para a implementação dos modelos, foram obtidos os melhores resultados. No entanto, a diferença em relação à redução para 100 atrasos foi baixa, cerca de 0,7%. Isto revela que o *autoencoder* não auxilia na melhora do desempenho do modelo, mas reduz o custo computacional de avaliação dos classificadores.

A partir dos resultados separados por classe de anomalia, conforme apresentado na Tabela 3, onde são apresentadas as estimativas de tamanhos de janelas temporais (tempo de falha) normalmente utilizados para confirmar ocorrências de anomalias, conforme apresentado em Vargas (2019). Pode ser observado que na falha 7 ocorre o pior resultado da média *F1*. Isso pode indicar que para a detecção de anomalia com uma janela de detecção (tamanho de janela) maior, esta metodologia proposta pode não ser adequada e um número maior de atrasos deveria ser escolhido. Já para a Falha 2 que tem a menor janela temporal dos casos aplicados, o detector funcionou corretamente.

Tabela 3. Resultados de acordo com a classe de anomalia e o tempo de falha, do melhor modelo proposto.

Anomalia	Média *F1*	Tamanho de janela
Falha 1	0,8505	12 h
Falha 2	0,9939	5 min -20 min
Falha 5	0,8023	12 h
Falha 6	0,8368	15 min
Falha 7	0,7330	72 h
Falha 8	0,7931	30 min - 5 h

5. CONCLUSÕES

Os detectores de anomalias desenvolvidos neste trabalho, a partir dos dados do 3W *dataset*, que são são atribuídos a poços surgentes, apresentaram resultados superiores quando comparados com outros trabalhos da literatura, com uma diferença de até dezoito pontos percentuais para os modelos OCSVM e dez pontos percentuais para os modelos de Floresta de Isolamento. Isso demonstra que o uso de um número significativo de atrasos nas variáveis em conjunto com a técnica de *autoencoders* em cascata para a redução de dimensionalidade implica em uma melhor extração de características relevantes para este conjunto de dados.

A inclusão de atrasos faz com que o modelo se torne mais eficaz, obtendo mais informações sobre o processo a ser estudado. A partir dos resultados obtidos, foi observado que quanto maior a dimensão maior foi a média *F1 score*. O contraponto do aumento no número de atributos faz com que o conjunto de dados se expanda, tornando o desenvolvimento dos modelos mais vagaroso. A inclusão de *autoencoders* auxilia nesta tarefa dando mais agilidade ao treinamento.

No momento da concepção do projeto dos detectores também deve-se levar em consideração o tempo de execução no desenvolvimento dos modelos a serem implementados. As características dos processos que se deseja detectar anomalias, são fatores determinantes para definir qual técnica de redução de dimensionalidade será empregada no processo de construção dos modelos. A estratégia dos *stacked autoencoders* conseguiu reduzir a dimensão do espaço latente e por isso o tempo de inferência foi reduzido.

AGRADECIMENTOS

Os autores agradecem 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. doi:10.1109/CISTEM.2018.8613538.

Aggarwal, C. and Sathe, S. (2017). *An Introduction to Outlier Ensembles*, 1–34. doi:10.1007/978-3-319-54765-7_1.

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. doi:https://doi.org/10.1016/j.conengprac.2017.06.002. URL `http://www.sciencedirect.com/science/article/pii/S0967066117301284`.

ANP (2021). *Boletim da Produção de Petróleo e Gás Natural*. Agência Nacial do Petroleo, Gás Natural e Biocombustíveis, Brasília - DF, 126 edition.

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. doi:10.1109/IJCNN.2003.1223658.

Barbosa, B.H.G., Aguirre, L.A., and Braga, A.P. (2019). Piecewise affine identification of a hydraulic pumping system using evolutionary computation. *IET Control Theory Applications*, 13(9), 1394–1403.

Barbosa, B.H.G., Bui, L.T., Abbass, H.A., Aguirre, L.A., and Braga, A.P. (2011). The use of coevolution and the artificial immune system for ensemble learning. *Soft Computing*, 15(9), 1735–1747.

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

de Morais, G.A., Barbosa, B.H.G., Ferreira, D.D., and Paiva, L.S. (2019). Soft sensors design in a petrochemical process using an evolutionary algorithm. *Measurement*, 148, 106920.

Duda, R.O., Hart, P.E., Stork, D.G., Duda, C.R.O., Hart, P.E., and Stork, D.G. (2001). Pattern classification, 2nd ed.

Espinola, A. (2013). *Ouro negro: petróleo no Brasil de Lobato DNPM-163 a TUPI RJS-646*. Interciência. URL `https://books.google.com.br/books?id=sY12oAEACAAJ`.

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. doi:10.1007/978-3-540-31865-1_25.

Guerbai, Y., Chibani, Y., and Hadjadji, B. (2014). The effective use of the one-class svm classifier for reduced training samples and its application to handwritten signature verification. In *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, 362–366. doi:10.1109/ICMCS.2014.6911221.

Junior, W.F., Vargas, R.E.V., Komati, K.S., and de Souza Gazolli, K.A. (2020). Detecção de anomalias em poços produtores de petróleo usando aprendizado de máquina. In *Congresso Brasileiro de Automática (CBA)*, volume 2.

Kadhim, A. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52. doi:10.1007/s10462-018-09677-1.

Krawczyk, B., Minku, L.L., Gama, J., Stefanowski, J., and Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37, 132 – 156. doi:https://doi.org/10.1016/j.inffus.2017.02.004. URL `http://www.sciencedirect.com/science/article/pii/S1566253516302329`.

Liu, F.T., Ting, K.M., and Zhou, Z. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, 413–422. doi:10.1109/ICDM.2008.17.

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. doi:10.1016/j.sigpro.2016.07.028.

Machado, F. (2018). *Big Data O Futuro dos Dados e Aplicações*. Editora Saraiva. URL `https://books.google.com.br/books?id=2LdiDwAAQBAJ`.

Maglaras, L. and Jiang, J. (2014). Intrusion detection in scada systems using machine learning techniques. doi:10.1109/SAI.2014.6918252.

Morettin, P.A. and Toloi, C.M.C. (2006). *Análise de Séries Temporais*. Blucher.

Nascimento, R.S.F., Vargas, R.E.V., Groenner, B.H., and dos Santos, I.H.F. (2020). Detecção de falhas com stacked autoencoders e técnicas de reconhecimento de padrões em poços de petróleo operados por gas lift. In *Congresso Brasileiro de Automática (CBA)*, volume 2.

Rahman, A., Smith, D., Hills, J., Bishop-Hurley, G., Henry, D., and Rawnsley, R. (2016). A comparison of autoencoder and statistical features for cattle behaviour classification. In *2016 International Joint Conference on Neural Networks (IJCNN)*, 2954–2960. doi:10.1109/

IJCNN.2016.7727573.

Roy, M., Bose, S.K., Kar, B., Gopalakrishnan, P.K., and Basu, A. (2018). A stacked autoencoder neural network based automated feature extraction method for anomaly detection in on-line condition monitoring. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1501–1507. doi:10.1109/SSCI.2018.8628810.

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.

Schlag, S., Schmitt, M., and Schulz, C. (2019). *Faster Support Vector Machines*, 199–210. doi:10.1137/1.9781611975499.16.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., and Williamson, R.C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7), 1443–1471. doi:10.1162/089976601750264965. URL `https://doi.org/10.1162/089976601750264965`.

Shin, H., Orton, M.R., Collins, D.J., Doran, S.J., and Leach, M.O. (2013). Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1930–1943.

Thomas, J. (2004). *Fundamentos de engenharia de petróleo*. Interciência. URL `https://books.google.com.br/books?id=yKyqPgAACAAJ`.

Vargas, R.E.V. (2019). *Base de Dados e Benchmarks para Prognóstico de Anomalias em Sistemas de Elevação de Petróleo*. Ph.D. thesis, Universidade Federal do Espírito Santo, Vitória - ES.

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. doi:https://doi.org/10.1016/j.petrol.2019.106223. URL `http://www.sciencedirect.com/science/article/pii/S0920410519306357`.

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.

Yu, J. and Zhang, C. (2020). Manifold regularized stacked autoencoders-based feature learning for fault detection in industrial processes. *Journal of Process Control*, 92, 119 – 136. doi:https://doi.org/10.1016/j.jprocont.2020.06.001. URL `http://www.sciencedirect.com/science/article/pii/S0959152420302353`.