Trabalho de Conclusão de Curso

PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

ALUNO: BRUNO EDUARDO MADEIRA
ORIENTADOR: FLÁVIO ARTHUR LEAL FERREIRA

PUCRS online uol edtech

Sumário

1. RESUMO

Este trabalho tem por objetivo determinar o quão representativos são os dados, gerados por simulação, presentes no conjunto de dados 3W, fornecido pela empresa Petrobras, com o propósito de criação de detectores de anomalias. Mais precisamente, é apresentada uma avaliação de um detector de anomalias treinado utilizando-se um conjunto de séries temporais geradas por simulação e testado em um conjunto de dados de sensores reais de plataformas de petróleo. As anomalias são detectadas empregando-se um classificador com uma arquitetura definida por uma rede convolucional causal dilatada. Como resultado, chegou-se à conclusão de que o uso de classificadores desse tipo foi capaz de detectar falhas com uma acurácia de 93% e f1 igual a 0,93, considerando-se um treinamento sobre dados simulados e um conjunto de teste também formado por dados sintéticos. No entanto, quando esse mesmo classificador, treinado com dados simulados, foi aplicado sobre dados reais, observou-se que foi incapaz de detectar falhas, classificando todas as ocorrências como normais. Esse é um forte indicativo de que existem características nos dados reais que diferem dos dados simulados, dificultando o processo de classificação.

2. INTRODUÇÃO

A prevenção de falhas de operação de plataformas de petróleo é um problema importante, de alto grau de interesse para as empresas que atuam nesse setor, tendo em vista que sua ocorrência pode significar perda de vidas, perda de produtividade e aumento de custos de manutenção. Por esse motivo, visando melhorar o seu sistema de prevenção de falhas em plataformas de petróleo, a empresa Petrobras disponibilizou, em 2019, um conjunto de dados aberto, que reúne várias informações sobre diversos tipos de falhas de operação.

Tendo em vista a relevância do problema e a raridade dos dados disponíveis, são utilizados dados simulados em combinação com dados reais para se detectar e classificar intercorrências.

2.1. TECNOLOGIA USADA NA EXTRAÇÃO DE PETRÓLEO EM ALTO-MAR

Faremos agora uma caracterização resumida do sistema utilizado para extração de petróleo. Mais detalhes podem ser encontrados em Vargas (2019).

Abordamos exclusivamente um problema relacionado a sistemas de extração em alto-mar, baseados em poços de petróleo cuja pressão natural é suficiente para gerar uma produção de hidrocarbonetos em uma taxa comercial, sem a necessidade de energia adicional.

Esses sistemas são compostos por um conjunto de sensores acoplados a sistemas mecânicos, hidráulicos e pneumáticos.

Mais precisamente, existe uma plataforma em alto-mar conectada a uma estrutura submarina chamada de Árvore de Natal Molhada (ANM), que faz a extração dos hidrocarbonetos.

Para saber se o sistema está operando de forma adequada, são monitorados diversos sensores instalados em válvulas do sistema, denominadas: Permanent Downhole Gauge (PDG); Temperature and Pressure

Transducer (TPT); e Production Choke (PCK).

Esses sensores realizam medições de temperatura e pressão das válvulas. Dessa forma, são obtidas medições com uma taxa de amostragem de 1Hz dos seguintes sinais:

- Pressão no PDG.
- Pressão no TPT.
- Temperatura no TPT
- Pressão no fluido montante à válvula PCK.
- Temperatura do fluido jusante à válvula PCK.

2.2. TIPOS DE FALHA NA PRODUÇÃO DE PETRÓLEO EM ALTO-MAR

Durante a operação de um sistema de extração de petróleo em alto-mar, podem ocorrer diversos tipos de eventos indesejados, e a Petrobras procura evitá-los utilizando o monitoramento de seus sensores instalados no sistema de extração de petróleo. Dentre esses eventos indesejados, podemos citar (Vargas, 2019):

- Aumento abrupto de *Basic Sedimentar and Water*.
- Fechamento espúrio da *Downhole Safety Valve*.
- Intermitência severa.
- Instabilidade de fluxo.
- Perda rápida de produtividade.
- Restrição rápida em PCK.
- Incrustação em PCK.

A relação entre os sinais descritos na seção anterior e esses diferentes tipos de falhas constituem o conjunto de dados disponibilizados pela Petrobras, cujas características serão descritas a seguir.

3. CARACTERÍSTICAS DOS DADOS FORNECIDOS PELA PETROBRAS

O conjunto de dados fornecido pela Petrobras, batizado de 3W (Vargas et al., 2019), tem as seguintes características:

• Reúne dados reais, dados sintéticos produzidos por simulações em computador, e dados feitos manualmente por especialistas.

• Representa os cinco sinais, descritos na Seção 2, obtidos por sensores instalados no sistema. Tais sinais são associados a uma classificação do estado do sistema, que pode ser: normal; em estado transiente para falha; e em estado de falha.

• Fornece diversas séries temporais associadas a informações dos sete diferentes tipos de falhas descritos na Seção 2. Ou seja, para cada tipo de falha são fornecidas diversas séries temporais que caracterizam a dinâmica do estado de monitoramento dos sensores, bem como as respectivas classes associadas.

O principal interesse da Petrobras é conseguir, conforme explicado em Vargas *et al.* (2019, identificar as falhas ainda durante o estágio transiente, evitando o estado estabilizado.

4. TRABALHOS RELACIONADOS

O conjunto de dados disponibilizado pela empresa Petrobras foi utilizado como base para a realização de diversos trabalhos. Dentre tais publicações, podemos destacar Vargas *et al.* (2019), em que os autores explicam como os dados foram gerados e organizados. Na sequência, muitos outros estudos foram feitos usando esse mesmo conjunto de dados, dos quais podemos citar, por exemplo: Turan e Jäschke (2021); Vignoli (2021); Momm (2022); e Figueirêdo (2023).

Uma característica compartilhada pela maioria dos trabalhos anteriores é que o enfoque dado por eles foi o de propor arquiteturas de classificadores

capazes de identificar falhas e de diferenciar o tipo de falhas. Esse enfoque, conforme será explicado na seção a seguir, é diferente do enfoque do presente trabalho.

5. PROBLEMA DE INTERESSE

Neste artigo abordamos o problema de identificar qual é a real efetividade do uso de simulações em sistemas de detecção de falhas na produção de petróleo. Esse estudo é relevante tendo em vista que a maioria dos dados do conjunto 3W é formada por dados simulados.

Com o objetivo de simplificar o estudo, consideramos, exclusivamente, o processo de detecção de falhas definidas como Perda Rápida de Produtividade.

Um estudo similar poderia ser feito para todos os demais tipos de falha citados na Seção 2. Mas optou-se por estudar este em particular, tendo em vista que é o tipo de falha com a maior quantidade de dados simulados fornecida.

6. METODOLOGIA

Para fazer a avaliação da efetividade do uso de simulações, seguimos a seguinte estratégia:

1) Escolha de uma arquitetura de classificador.
2) Treinamento de um classificador A usando dados reais.
3) Teste do classificador A usando dados reais.
4) Treinamento de um classificador B usando dados simulados.
5) Teste do classificador B usando dados simulados.
6) Teste aplicando o classificador B sobre dados reais.

Em relação ao classificador, escolhemos fazer uso de redes convolucionais causais dilatadas, conforme será explicado na Seção 7 a seguir.

Os resultados dos itens 3 e 5 do passo a passo da estratégia proposta permitem dizer se o classificador, construído usando redes convolucionais

causais dilatadas, funciona bem para o propósito de detecção de anomalias tanto em dados simulados quanto em dados reais.

Em outras palavras, se os resultados dos itens 3 e 5 forem satisfatórios, conclui-se que redes convolucionais causais dilatadas são suficientes para tratar problemas de classificação, tanto com dados reais quanto simulados.

Já o resultado do item 6 indica a efetividade do uso de simulações na representação de sinais de sensores de plataformas de petróleo com o propósito de detecção de falhas. Se o resultado for bom – ou seja, de acurácia alta e f1 alto –, isso é um indicativo de que o classificador está utilizando as mesmas características nos dados simulados e reais para fazer a classificação.

7. REDES CONVOLUCIONAIS CAUSAIS DILATADAS

O modelo de redes convolucionais causais dilatadas foi introduzido em Oord *et al.* (2016) com o nome WaveNet. Sua motivação inicial foi definir uma arquitetura útil para modelos generativos, cuja saída foi utilizada para geração de áudio, podendo, por exemplo, ser empregada em sistemas de geração de fala a partir de texto, em que o áudio é representado por milhares de amostras por segundo.

No entanto, esse tipo de rede também pode ser aplicado na previsão de séries temporais, tendo em vista que as convoluções empregadas não violam a sequencialidade dos dados, como no caso de redes convolucionais tradicionais (Géron, 2021, p. 403).

Um exemplo de rede convolucional causal é ilustrado na Figura 1, em que cada seta indica os valores que são processados por cada núcleo de convolução. Nessa arquitetura, verifica-se que o valor de saída é definido exclusivamente a partir da entrada atual e de entradas do passado.

É fácil de verificar que o número de camadas no exemplo da figura cresce linearmente com o número de amostras. Dessa forma, essa arquitetura torna-se proibitiva para modelar séries temporais em que são utilizados muitos estados anteriores na previsão do estado futuro.

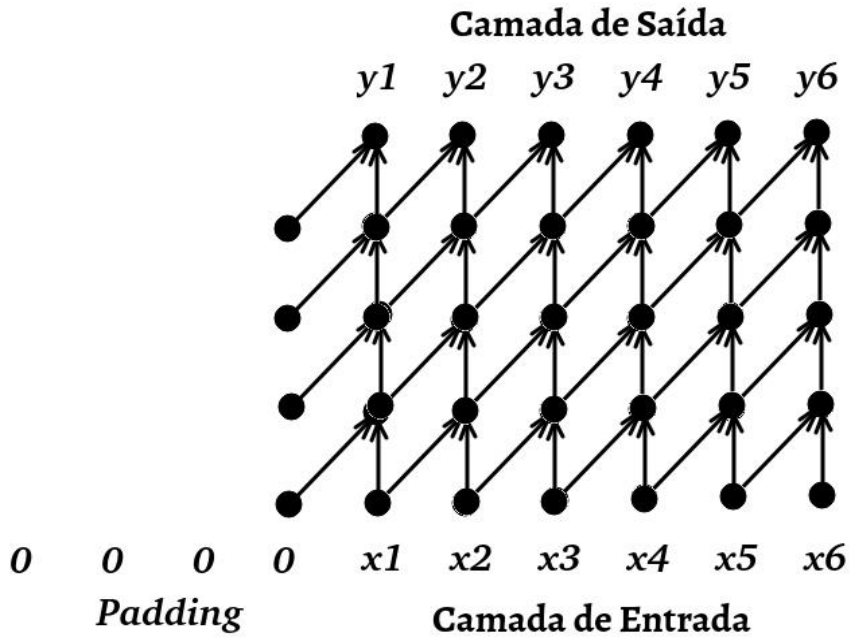

Figura 1. Rede convolucional causal não dilatada.

Uma solução para esse problema é a utilização de redes convolucionais causais dilatadas, cuja arquitetura é ilustrada na Figura 2, que mostra como o valor de saída $\boldsymbol{y}$ calculado no estado atual depende da entrada atual e de diversos outros valores da camada de entrada. Nesse caso, será utilizado *padding* 0 caso existam valores anteriores não definidos na entrada, considerando-se o uso da implementação fornecida pela biblioteca TensorFlow2.

A maior motivação para o uso de rede neurais convolucionais causais dilatadas nesse problema deve-se ao fato de elas não necessitarem de um processo de engenharia de características, podendo ser empregadas em predições de séries temporais a partir de um longo período de amostras.

Além disso, até onde vai nosso conhecimento, tal abordagem ainda não foi experimentada em nenhum dos trabalhos anteriores feitos com o mesmo conjunto de dados.

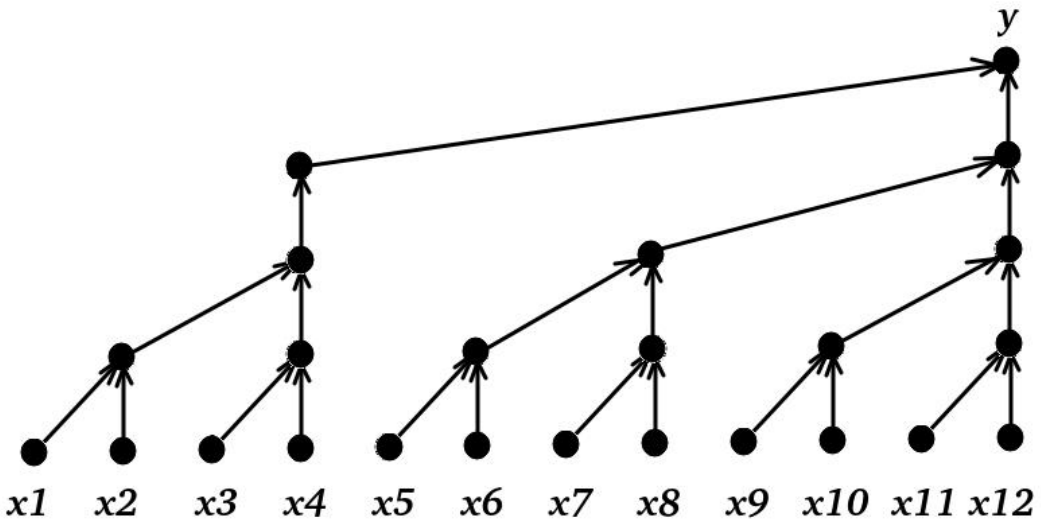

Figura 2. Relação de dependência entre o valor de uma das saídas da rede e todos os diversos valores de entrada definidas por uma arquitetura convolucional causal dilatada com taxa de dilatação que dobra a cada camada.

8. ARQUITETURA DA SOLUÇÃO

Em nossa proposta de abordagem para a solução do problema, adotamos uma arquitetura similar àquela ilustrada na Figura 3. Ou seja, uma sequência de camadas convolucionais causais dilatadas com filtros com núcleo 2, em que se dobra a taxa de dilatação conforme nos aproximamos da camada de saída. Entre cada uma dessas camadas colocou-se uma camada de *Batch Normalization*, tendo em vista que isso demonstrou gerar um aumento na taxa de convergência do treinamento.

Após as camadas convolucionais causais dilatadas, o tensor resultante foi processado por uma camada de média global ponderada. Essa camada é relevante tendo em vista que a quantidade de *padding* de cada elemento desse tensor é diferente. Assim, tal ponderação é ajustada pela referida camada.

Por fim, é aplicada uma camada de classificação logística na saída, gerando como resultado a probabilidade de que a sequência de medições de entrada contenha algum elemento com falha ou em estado transiente de falha.

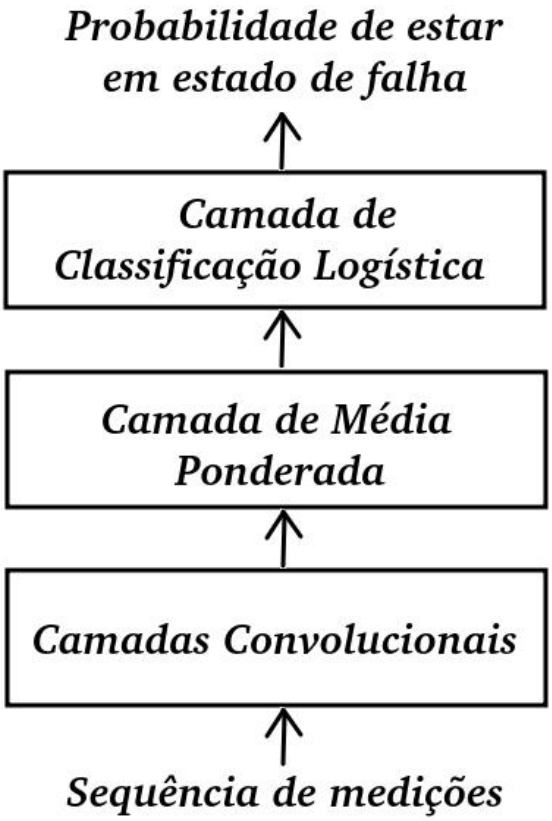

Figura 3. Arquitetura da rede.

9. PREPARAÇÃO DOS DADOS E HIPERPARÂMETROS

Foram utilizados os dados disponibilizados do conjunto de dados 3W da Petrobras na Versão 2, que se encontra em Melo (2024).

Os dados de sensores ausentes foram obtidos por meio da propagação dos valores anteriores e, em seguida, da propagação dos valores posteriores para trás. Depois, os dados foram subamostrados, usando uma taxa de uma amostra a cada 3 segundos.

Tendo em vista a ausência dos sinais de PJUS-CKGL, T-JUS-CKGL e QGL em dados simulados, as medições desses três sensores foram totalmente descartadas. Provavelmente essas medições foram inseridas em atualizações

do conjunto de dados 3W, pois não constam da descrição dos dados feita em Vargas *et al.* (2019).

Foram usadas redes convolucionais causais de taxa de dilatação 2, nas quais é aplicada uma quantidade adaptativa de filtros. Mais precisamente, foi utilizada uma quantidade de núcleos de convolução igual à taxa de dilatação em cada camada. Dessa forma, a quantidade de informação foi preservada na passagem de uma camada convolucional para a camada posterior.

No total, foram empilhadas 5 camadas convolucionais causais dilatadas antes da camada de média ponderada global. Foram consideradas sequências de 120 medições na entrada do classificador, representando janelas de 6 minutos.

O treinamento do modelo feito com dados reais (classificador A) utilizou 8 séries temporais, sobre as quais foram definidas janelas deslizantes a cada 3 segundos, totalizando cerca de 162 mil janelas.

Já o treinamento do modelo feito com dados simulados (classificador B) também foi feito com janelas deslizantes, intervaladas de 3 segundos, utilizando 20 séries temporais, totalizando cerca de 376 mil janelas.

Em ambos os casos, o conjunto de dados foi dividido na fração de 80% para treinamento e 20% para teste.

Cada instância foi definida associando-se cada janela de minutos com o valor 1 se existe a presença de falha ou de transiente para falha em seu interior, e valor 0 caso contrário.

Um ponto a se destacar é que, no caso de dados reais (classificador A), para cada janela sem presença de falha, existem cerca de 4 janelas com falha. Já no caso de dados simulados (classificador B), para cada janela sem falha, existem cerca de 192 janelas com falha. Dessa forma, para cada janela com falha, foram introduzidas múltiplas cópias de janelas sem falha nos treinamentos dos classificadores. Isso foi feito para não enviesar os classificadores durante o treinamento.

10. RESULTADOS

Obtivemos, após o treinamento com dados reais (classificador A), uma acurácia de 84% e medida f1 igual a 0,85 no conjunto de validação formado por dados reais. A matriz de confusão obtida foi a apresentada na Figura 4.

Para o caso do classificador treinado com dados simulados (classificador B) e testado sobre dados simulados, obtivemos uma acurácia de 93% e medida f1 igual a 0,93. A matriz de confusão obtida foi apresentada na Figura 5.

Por outro lado, ao aplicarmos o classificador B sobre os dados reais, obtivemos um péssimo resultado, com acurácia de 52%, conforme ilustrado na matriz de confusão da Figura 6. Além disso, verifica-se que o classificador foi incapaz de detectar falhas.

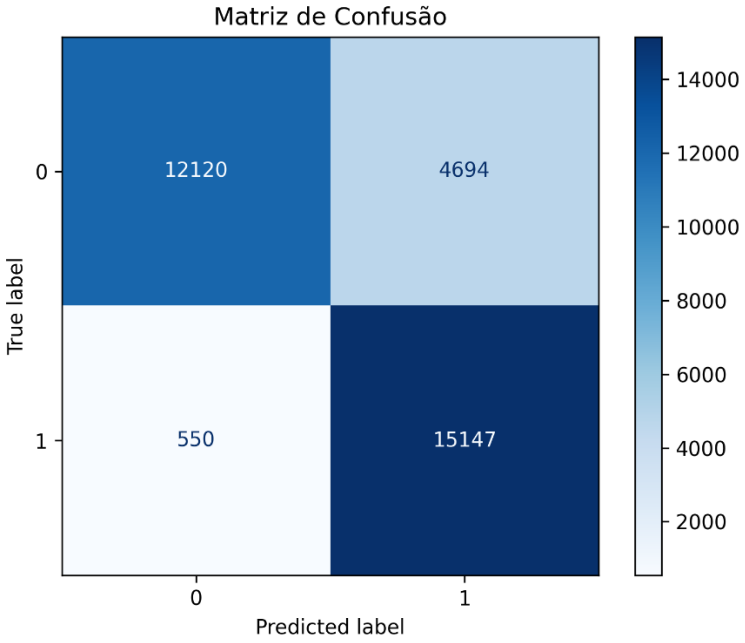

Figura 4. Matriz de confusão obtida com classificador treinado em dados reais (classificador A) e testado em dados reais. O valor 0 representa ausência de falha e o valor 1 representa ocorrência de falha ou estado transiente para falha.

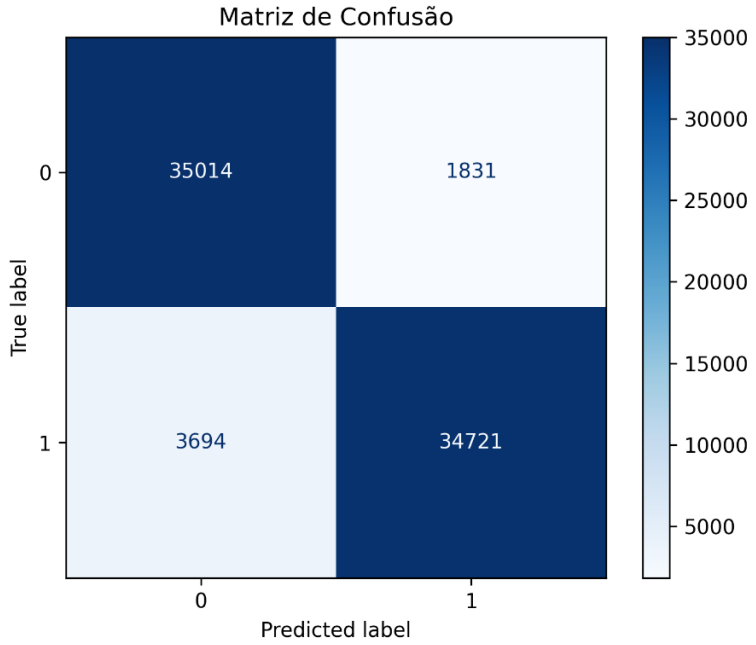

Figura 5. Matriz de confusão obtida aplicando-se o classificador B, treinado e testado em dados simulados. A interpretação dos valores 0 e 1 é a mesma da Figura 4.

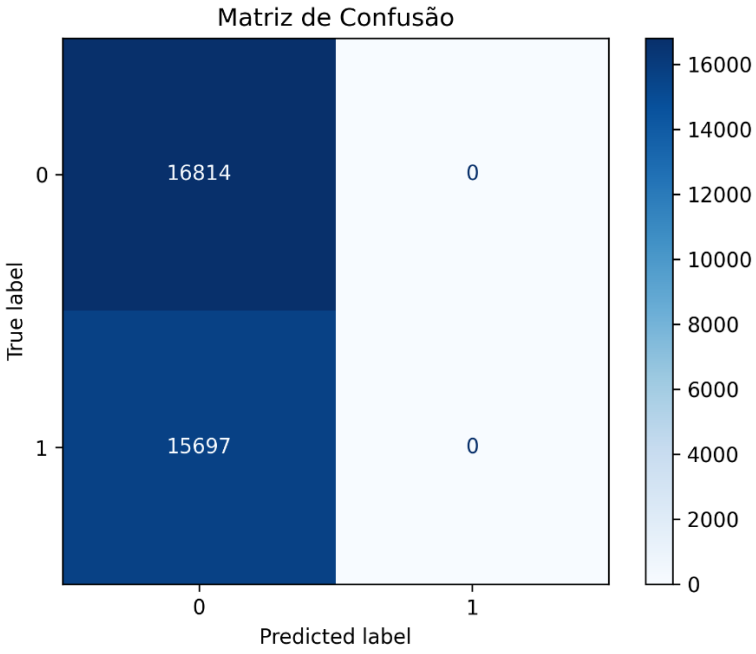

Figura 6. Matriz de confusão obtida aplicando-se o classificador B, treinado em dados simulados e testado em dados reais. A interpretação dos valores 0 e 1 é a mesma das figuras anteriores.

11. DISCUSSÃO

Os resultados indicados pelas matrizes de confusão das Figuras 4 e 5 sugerem que redes convolucionais causais dilatadas podem ser usadas em classificadores para detectar anomalias tanto em dados simulados quanto reais, tendo sido obtidos bons resultados de classificação quando os conjuntos de treino e de teste eram originários do mesmo conjunto de dados.

Por outro lado, a matriz de confusão da Figura 6 indica que, quando se realizou o treinamento com dados simulados (classificador B) e testou-se esse modelo aplicado sobre dados reais, o classificador não foi capaz de detectar falha alguma, obtendo-se uma péssima acurácia de 52%. Isso é um indicativo de que as características dos sinais dos sensores simulados que estão sendo utilizadas para diferenciar sinais normais de sinais com anomalias são diferentes para os casos de dados reais e para o caso de dados simulados.

Uma possível explicação para os resultados observados é que os modelos usados em simulações são simplificações da realidade, portanto geram sinais de sensores com diferenças em relação aos sinais reais. Consequentemente, esse fenômeno indesejado indica que o emprego de dados simulados do conjunto 3W no treinamento de classificadores deve ser feito com cautela, principalmente se for utilizado um classificador que não seja baseado em uma engenharia de características especificamente projetada para o problema de classificação em questão.

12. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho teve por objetivo analisar a representatividade dos dados simulados fornecidos pelo conjunto de dados 3W, com o propósito de detecção de falhas em plataformas de extração de petróleo a partir do processamento de sinais de sensores instalados nelas. Esse problema é relevante pois a maioria dos dados disponíveis sobre problemas desse tipo são simulados.

Optou-se por se utilizar redes convolucionais causais dilatadas para o

processamento das séries temporais, tendo em vista que essas não exigem nenhum tipo de engenharia de características.

Chegou-se à conclusão de que redes convolucionais causais dilatadas são adequadas para o propósito de classificação de falhas tanto para o caso de dados reais quanto para o caso de dados simulados. No entanto, se o treinamento é feito com dados simulados, o modelo não é capaz de detectar falha alguma em sinais reais, indicando que é temerário o uso de classificadores que não fazem uso de engenharia de características durante o treinamento com os dados simulados do conjunto 3W.

Como possível trabalho futuro, podemos considerar o uso combinado de dados reais e simulados, no momento do treinamento, com o objetivo de introduzir um viés no classificador que melhor se ajuste aos dados reais.

13. BIBLIOGRAFIA

FIGUEIRÊDO, Ilan Sousa Figueirêdo. *Uma nova abordagem de Inteligência Artificial baseada em autoaprendizagem profunda para manutenção preditiva em um ambiente de produção de petróleo e gás offshore*. 142f. Tese (Doutorado em Modelagem Computacional e Tecnologia Industrial). Centro Universitário SENAI CIMATEC, Salvador, 2023. Disponível em: http://repositoriosenaiba.fieb.org.br/handle/fieb/1730. Acesso em: 1 Dez. 2024.

GÉRON, A. *Mãos à obra: aprendizado de máquina com Scikit-Learn, Keras & TensorFlow*. Rio de Janeiro: Alta Books, 2021.

MELO, Afrânio Melo. *3W Dataset - Undesirable events in oil wells*. Kaggle, 2024 URL: https://www.kaggle.com/datasets/afrniomelo/3w-dataset. Acesso em: 2 Dez. 2024.

MOMM, Gustavo Gacia Momm. *Detecção de anomalias em sensores de poços submarinos com uso de redes neurais artificiais*. Trabalho de Conclusão de Curso (MBA) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022. Disponível em: https://bdta.abcd.usp.br/directbitstream/8ff5c042-9393-4e36-9748-1de3f339b0fa/Gustavo%20Garcia%20Momm_TCC%20Gustavo%20Momm%20Final_206491.pdf. Acesso em: 2 Dez. 2024.

OORD, Aäron van den Oord *et al*. “WaveNet: A Generative Model for Raw Audio”. In: CoRR abs/1609.03499 (2016). arXiv: 1609.03499. Disponível em: http://arxiv.org/abs/1609.03499. Acesso em: 1 Dez. 2024.

TURAN, Evren M. Turan; and JÄSCHKE, Johannes Jäschke. “Classification of undesirable events in oil well operation”. In: *2021 23rd International Conference on Process Control (PC)*. 2021, pp. 157–162. DOI: 10.1109/PC52310.2021.9447527.

VARGAS, Ricardo Emanuel Vaz *et al*. “A realistic and public dataset with rare undesirable real events in oil wells”. In: *Journal of Petroleum Science and Engineering*. V. 181, p. 106223, 2019. ISSN: 0920-4105. DOI:https://doi.org/10.1016/j.petrol.2019.106223. Disponível em: https://www.sciencedirect.com/science/article/pii/S0920410519306357. Acesso em: 1 Dez. 2024.

VARGAS, Ricardo Emanuel Vaz. *Base de dados e benchmarks para prognóstico de anomalias em sistemas de elevação de petróleo*. Tese (Doutorado em Engenharia Elétrica). Universidade Federal do Espírito Santo, Vitória, 2019. Disponível em: https://repositorio.ufes.br/items/317715bb-60c4-4735-9872-e5f106cd958a. Acesso em: 1 Dez. 2024.

VIGNOLI, Luciana Escobar Gonçalves Vignoli. “*Análise comparativa de métodos para detecção de eventos em séries temporais”*. Tese (Mestrado em Ciência da Computação). Centro Federal de Educação Tecnológica Celso Sukow da Fonseca, Rio de Janeiro, 2021. Disponível em: https://eic.cefet-rj.br/ppcic/wp-content/uploads/2020/11/30-Luciana-Escobar-Goncalves-Vignoli.pdf. Acesso em: 1 Dez. 2024.