

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Luiz Felipe Rodrigues Gonçalves

**Desenvolvimento de um Sensor Virtual de Vazão
Utilizando Redes Neurais para Macromedição de
uma Estação de Distribuição de Água**



João Pessoa
2025

Luiz Felipe Rodrigues Gonçalves

**Desenvolvimento de um Sensor Virtual de Vazão
Utilizando Redes Neurais para Macromedição de uma
Estação de Distribuição de Água**

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal da Paraíba como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.

Universidade Federal da Paraíba

Centro de Energias Alternativas e Renováveis

Programa de Pós-Graduação em Engenharia Elétrica

Orientador: Prof. Dr. Juan Moises Maurício Villanueva

Coorientador: Prof. Dr. Euler Cássio Tavares de Macedo

João Pessoa

2025

UNIVERSIDADE FEDERAL DA PARAÍBA – UFPB
CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS – CEAR
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA – PPGE

A Comissão Examinadora, abaixo assinada, aprova a Dissertação

**DESENVOLVIMENTO DE UM SENSOR VIRTUAL DE VAZÃO UTILIZANDO
REDES NEURAIIS PARA MACROMEDIÇÃO DE UMA ESTAÇÃO DE
DISTRIBUIÇÃO DE ÁGUA**

Elaborada por

LUIZ FELIPE RODRIGUES GONÇALVES

como requisito parcial para obtenção do grau de
Mestre em Engenharia Elétrica.

COMISSÃO EXAMINADORA


PROF. DR. JUAN MOISÉS MAURICIO VILLANUEVA
Orientador – UFPB

 Documento assinado digitalmente
EULER CASSIO TAVARES DE MACEDO
Data: 16/12/2025 22:07:37-0300
Verifique em <https://validar.it.gov.br>

PROF. DR. EULER CASSIO TAVARES DE MACÊDO
Coorientador – UFPB

 Documento assinado digitalmente
FABRICIO BRAGA SOARES DE CARVALHO
Data: 18/12/2025 07:46:51-0300
Verifique em <https://validar.it.gov.br>

PROF. DR. FABRICIO BRAGA SOARES DE CARVALHO
Examinador Interno – UFPB

 Documento assinado digitalmente
SEBASTIAN YURI CAVALCANTI CATUNDA
Data: 18/12/2025 11:52:43-0300
Verifique em <https://validar.it.gov.br>

PROF. DR. SEBASTIAN YURI CAVALCANTI CATUNDA
Examinador Externo – UFRN

AGRADECIMENTOS

A Deus, à Virgem Maria Santíssima e a São José, por sua presença constante em minha vida, sustentando-me com graças abundantes e conduzindo-me com firmeza e misericórdia em todos os momentos deste percurso acadêmico.

À minha mãe, Maria da Graça, pelo sacrifício diário, amor incansável e exemplo de fortaleza; e ao meu pai, Ricardo †, pelos ensinamentos transmitidos e pela inspiração permanente que deixou em minha vida.

Ao meu orientador, professor Juan Villanueva, pela orientação dedicada, pela atenção contínua e pela confiança depositada no desenvolvimento deste trabalho.

Aos professores Euler Macedo, Fabrício Braga, Cícero Rocha, Darlan Fernandes e Waslon Lopes, pelo apoio, incentivo e contribuições ao longo do mestrado, que enriqueceram minha formação acadêmica e profissional.

Aos demais professores que contribuíram para minha formação como engenheiro e pesquisador.

Aos meus colegas Fernanda e Lucas, cujo empenho, dedicação e parceria tornaram este trabalho possível. A colaboração de ambos, em cada etapa do desenvolvimento, representa uma parte fundamental desta pesquisa, que é tão deles quanto minha.

À CAGEPA, pela disponibilização dos dados utilizados e pelo apoio técnico durante a construção do modelo, especialmente nas pessoas de Altamar e Petrônio, que ofereceram orientações valiosas e acompanhamento nas visitas à estação de distribuição.

Ao CNPq, pelo apoio institucional e incentivo à pesquisa, fundamentais para a realização deste trabalho.

À AWS, pela concessão dos créditos computacionais utilizados no desenvolvimento, teste e implantação do sensor virtual em ambiente de nuvem.

A todos aqueles que, de alguma maneira, contribuíram, incentivaram ou rezaram por mim ao longo desta jornada, deixo meu profundo e sincero agradecimento.

A humildade é o princípio do aprendizado.

- Hugo de São Vitor

RESUMO

A Lei nº 14.026/2020, que instituiu o novo marco legal do saneamento, trouxe mudanças significativas para as empresas de distribuição de água e esgoto, estabelecendo novas exigências regulatórias e maior competitividade no setor. Entre suas diretrizes, destaca-se a necessidade de reduzir as perdas de água e ampliar os índices de macromedição. Atualmente, segundo o Sistema Nacional de Informações sobre Saneamento (SNIS), as perdas médias no Brasil chegam a 40%, e apenas 70% da água distribuída é devidamente macromedida. Apesar disso, a instalação de sensores, especialmente macromedidores de vazão, ainda é limitada devido aos custos elevados, o que leva muitas companhias a concentrarem o monitoramento apenas em regiões metropolitanas. Como alternativa, métodos de medição indireta permitem estimar a vazão com base em variáveis como pressão da rede de abastecimento e nível dos reservatórios, reduzindo a necessidade de sensores físicos em larga escala. Neste cenário, o uso de inteligência artificial (IA) tem contribuído para tornar os sistemas de abastecimento de água mais eficientes, permitindo a reconstrução de dados ausentes, a detecção de falhas e a previsão de variáveis operacionais. Sensores virtuais baseados em redes neurais possibilitam o acompanhamento contínuo do sistema e a identificação de anomalias, agregando valor à operação e auxiliando na redução de desperdícios. Este trabalho apresenta o desenvolvimento e a implementação de um sensor virtual baseado em redes neurais do tipo *Long Short-Term Memory* (LSTM) para estimar a vazão em um sistema de abastecimento de água. O modelo foi treinado com dados de pressão e nível de reservatório adquiridos pela Companhia de Água e Esgotos da Paraíba (CAGEPA) no município de Salgado de São Félix, utilizando um *broker* MQTT e monitoramento em nuvem. A validação do modelo foi realizada por meio da comparação entre os valores estimados e as medições reais obtidas com sensores físicos, resultando em um erro médio absoluto percentual menor que 5%, evidenciando a confiabilidade do modelo desenvolvido.

Palavras-chave: Sensores Virtuais, Redes Neurais, LSTM, Medição Indireta, Sistemas de Abastecimento de Água.

ABSTRACT

Law No. 14,026/2020, which established the new legal framework for sanitation in Brazil, introduced significant changes for water and sewage distribution companies by imposing stricter regulations and increasing competition in the sector. One of its key objectives is to reduce water losses and improve macromasurement rates. According to the National Sanitation Information System (SNIS), the average water loss in Brazil is around 40%, while only 70% of the distributed water is properly measured. However, the high cost of sensors, especially flow macrometers, limits their widespread deployment, leading many companies to focus monitoring efforts on metropolitan areas. To address this issue, indirect measurement methods allow flow estimation based on network pressure and reservoir level data, reducing the need for physical sensors across the entire distribution system. In this context, artificial intelligence (AI) has been increasingly applied to improve the efficiency of water supply systems, enabling the reconstruction of missing data, fault detection, and forecasting of operational variables. Soft sensors based on neural networks allow continuous system monitoring and anomaly detection, enhancing operational decision-making and minimizing resource waste. This study presents the development and implementation of a soft sensor based on Long Short-Term Memory (LSTM) neural networks to estimate flow rates in a water supply system. The model was trained using flow, pressure, and reservoir level data collected by the Companhia de Água e Esgotos da Paraíba (CAGEPA) in the municipality of Salgado de São Félix, through an MQTT broker and cloud-based monitoring. The model validation was performed by comparing the estimated values with real measurements from physical sensors, resulting in a mean absolute percentage error lower than 5%, demonstrating the reliability of the proposed approach.

Keywords: Soft Sensors, Neural Networks, LSTM, indirect Measurement, Water Supply Systems.

LISTA DE ILUSTRAÇÕES

Figura 1 – Mapa das regiões do Brasil por índice de atendimento de água (SINISA 2025).	14
Figura 2 – Mapa das regiões do Brasil por índice de perdas totais de água na distribuição (SINISA 2025).	15
Figura 3 – Estrutura típica de um sistema de abastecimento de água (SALVINO, 2012).	24
Figura 4 – Estação de tratamento de água.	25
Figura 5 – Reservatório elevado de água.	25
Figura 6 – Esquema ilustrativo de uma RNA (KUBAT, 1999).	26
Figura 7 – Exemplo da topologia de uma MLP ((LAZCANO; JARAMILLO-MORÁN; SANDUBETE, 2024)).	27
Figura 8 – Exemplo de uma CNN ((DING et al., 2024)).	28
Figura 9 – Arquitetura básica de uma RNN (adaptado de (SCHUSTER; PALIWAL, 1997)).	30
Figura 10 – Estrutura de uma célula LSTM (adaptado de (NASSER; RASHAD; HUSSEIN, 2020)).	31
Figura 11 – Diagrama do processo de desenvolvimento de um sensor virtual (adaptado de (FORTUNA et al., 2007))	33
Figura 12 – Conjunto de dados com <i>outliers</i>	35
Figura 13 – Sistema de abastecimento.	41
Figura 14 – Casa de bombas.	43
Figura 15 – Localização dos sensores na casa de bombas.	43
Figura 16 – Transdutor ultrassônico de medição de vazão.	44
Figura 17 – Estrutura da base de dados.	45
Figura 18 – Vazão de saída do reservatório ao longo de um dia.	45
Figura 19 – Topologia de rede proposta.	46
Figura 20 – Etapas de desenvolvimento do Sensor Virtual neste trabalho.	47
Figura 21 – Matriz de correlação de Pearson entre vazão, pressão e nível.	48
Figura 22 – Matriz de correlação de Spearman entre vazão, pressão e nível.	49
Figura 23 – Modelo do sensor virtual de vazão.	52
Figura 24 – Fluxo de dados do sistema.	54
Figura 25 – Visualização da pressão do sistema no Grafana (autoria própria).	55
Figura 26 – Visualização do nível do reservatório no Grafana (autoria própria).	55
Figura 27 – Visualização da vazão de distribuição no Grafana (autoria própria).	55
Figura 28 – Outlier nos dados dos sensores.	57

Figura 29 – Interface IHM em Grafana - Plataforma em nuvem AWS.	62
Figura 30 – Visualização da vazão estimada pelo LSTM.	63
Figura 31 – Visualização da vazão real (verde) e estimada (vermelho).	63

LISTA DE TABELAS

Tabela 1	–	Dados das 20 melhores <i>versus</i> 20 piores cidades (SINISA 2025)	16
Tabela 2	–	Especificações dos Transdutores de Pressão	41
Tabela 3	–	Especificações dos Macromedidores de Vazão	42
Tabela 4	–	Especificações das Sondas Hidrostáticas	42
Tabela 5	–	Especificações do Conjunto Motobomba	42
Tabela 6	–	Métricas de avaliação para diferentes modelos e configurações de neurônios.	59
Tabela 7	–	Métricas de avaliação para a arquitetura LSTM com duas camadas.	60
Tabela 8	–	Resultados da variação dos parâmetros do Monte Carlo Dropout.	61

LISTA DE ABREVIATURAS E SIGLAS

AWS	<i>Amazon Web Service</i>
CAGEPA	Companhia de Água e Esgotos da Paraíba
CLP	Computador Lógico Programável
CNN	<i>Convolutional Neural Network</i>
ETA	Estação de Tratamento de Água
GPU	<i>Graphics Processing Unit</i>
IHM	Interface Homem-Máquina
LENHS	Laboratório de Eficiência Energética e Hidráulica em Saneamento
LSTM	<i>Long Short Term Memory</i>
MAE	<i>Mean Absolute Error</i>
MAPE	<i>Mean Absolute Percentage Error</i>
MLP	<i>Multilayer Perceptron</i>
MQTT	<i>Message Qyeyubg Tekenetry Transport</i>
MSE	<i>Mean Squared Error</i>
RMSE	<i>Root Mean Squared Error</i>
RNA	Rede Neural Artificial
RNN	<i>Recurrent Neural Network</i>
SCADA	Sistemas de Supervisão, Controle e Aquisição de Dados
SINISA	Sistema Nacional de Informações em Saneamento Básico
TEDA	<i>Typicality and Eccentricity Data Analysis</i>
TPU	<i>Tensor Processing Unit</i>
UFPB	Universidade Federal da Paraíba

SUMÁRIO

	Lista de tabelas	9
1	INTRODUÇÃO	13
1.1	Objetivos	17
1.1.1	Objetivo Geral	17
1.1.2	Objetivos Específicos	17
1.2	Contribuições	18
1.3	Estrutura e organização do trabalho	18
2	ESTADO DA ARTE	20
3	FUNDAMENTAÇÃO TEÓRICA	23
3.1	Sistema de Abastecimento	23
3.2	Redes Neurais Artificiais	26
3.2.1	Redes Multilayer Perceptron	27
3.2.2	Redes Neurais Convolucionais	28
3.2.3	Redes Neurais Recorrentes	29
3.3	Memória de Curto e Longo Prazo	30
3.3.1	Porta de esquecimento	30
3.3.2	Porta de entrada e memória candidata	31
3.3.3	Porta de saída	32
3.4	Sensores Virtuais	32
3.5	Deteção de Outliers baseado no Algoritmo TEDA	34
3.6	Análise de Incerteza baseado no método Dropout Monte Carlo	37
3.7	Conclusões do Capítulo	38
4	MATERIAIS E MÉTODOS	40
4.1	Materiais	40
4.1.1	Aquisição de dados	44
4.2	Métodos	47
4.2.1	Seleção de variáveis	48
4.2.2	Correção de outliers	50
4.2.3	Modelagem do sensor virtual	50
4.2.4	Métricas para validação do sensor virtual	52
5	RESULTADOS	54

5.1	Aquisição de dados utilizando o broker MQTT	54
5.2	Correção de dados usando o algoritmo TEDA	56
5.3	Sensor virtual de vazão usando a rede neural LSTM	58
5.3.1	Variação na quantidade de camadas e análise de incertezas	59
5.3.2	Variação dos parâmetros do Monte Carlo Dropout	60
5.4	Interface usando a plataforma em nuvem	61
6	CONCLUSÃO	64
	REFERÊNCIAS	65

1 INTRODUÇÃO

A água é um recurso natural essencial para a manutenção da vida humana e o desenvolvimento das sociedades, constituindo-se como um bem de valor inestimável que deve ser preservado e utilizado de forma racional. Os sistemas de abastecimento de água desempenham um papel fundamental ao garantir o fornecimento deste recurso às comunidades em quantidade suficiente, qualidade adequada e pressão apropriada para atender às demandas domésticas, industriais e de serviços públicos. Diante do crescente processo de urbanização e da crescente escassez dos recursos hídricos, torna-se imperativo implementar práticas que minimizem as perdas e otimizem a eficiência dos sistemas de distribuição, uma vez que o desperdício não se limita apenas ao uso doméstico inadequado, mas estende-se significativamente às perdas operacionais em tubulações, reservatórios e demais componentes da infraestrutura de abastecimento.

Nesse contexto, destaca-se a importância da macromedição como instrumento fundamental de gestão operacional dos sistemas de abastecimento de água. Segundo Gomes (2009), macromedição é o conjunto de medições realizado nos sistemas de abastecimento, desde a captação de água bruta até as extremidades da rede de distribuição. Esse conceito, entretanto, não se restringe apenas à instalação de medidores físicos, mas abrange todo o processo de medição e estimação dos parâmetros hidráulicos relevantes para o monitoramento e a operação do sistema, com ênfase na etapa de distribuição. Entre os principais parâmetros de interesse no âmbito da macromedição, destacam-se a vazão, a pressão e o nível, variáveis diretamente associadas ao desempenho hidráulico, ao balanço hídrico e à identificação de perdas.

A Lei nº 14.026/2020, que instituiu o novo marco legal do saneamento, trouxe mudanças significativas para o setor de distribuição de água e esgoto, estabelecendo um novo contexto regulatório e competitivo. Entre suas diretrizes, destacam-se a redução das perdas de água e o aumento dos índices de macromedição, fatores essenciais para a sustentabilidade do setor (FLORES et al., 2020). Atualmente, segundo o Instituto Trata Brasil, com dados obtidos do Sistema Nacional de Informações em Saneamento Básico (SINISA), o índice médio de perdas nas companhias de saneamento chega a aproximadamente 40%, enquanto apenas 70% da água distribuída é devidamente macromedida, evidenciando a necessidade de soluções inovadoras que ampliem o monitoramento sem elevar substancialmente os custos operacionais (SINISA, 2025)

No contexto brasileiro, o cenário do saneamento básico apresenta disparidades regionais significativas tanto no atendimento à população quanto na eficiência operacional dos sistemas. As Figuras 1 e 2 ilustram essas diferenças através de mapas temáticos

que destacam, respectivamente, os índices de atendimento da população com rede de abastecimento de água e os percentuais de perdas na distribuição por região geográfica. A representação cartográfica evidencia claramente as desigualdades existentes entre as diferentes regiões do país, com gradações de cores que permitem visualizar de forma imediata as áreas com melhor e pior desempenho.

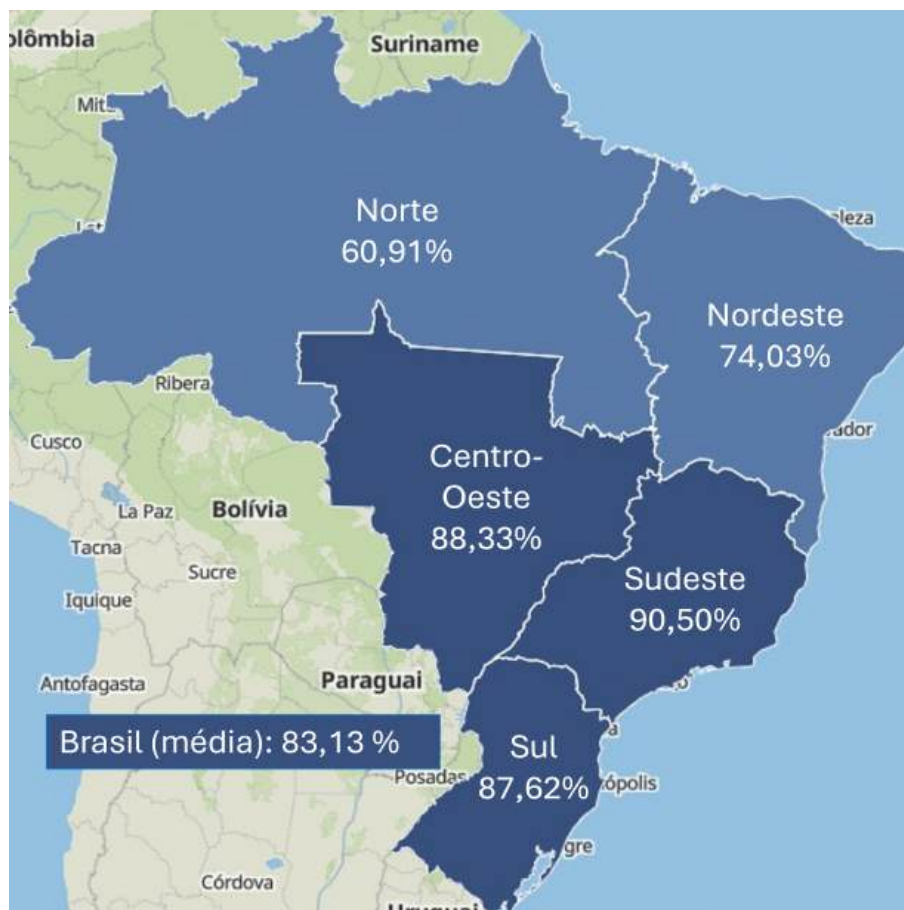


Figura 1 – Mapa das regiões do Brasil por índice de atendimento de água (SINISA 2025).

Conforme apresentado na Figura 1, os dados do SINISA revelam que o índice nacional de atendimento com rede de abastecimento de água alcança 83,13% da população. No entanto, essa média nacional mascara importantes variações regionais: enquanto a região Sudeste apresenta o melhor índice (90,50%), as regiões Norte e Nordeste apresentam os menores percentuais de atendimento, com 60,91% e 74,03%, respectivamente. Essa disparidade reflete não apenas questões geográficas e demográficas, mas também diferenças socioeconômicas e de investimento em infraestrutura.

Por outro lado, a Figura 2 apresenta um cenário preocupante em relação às perdas de água na distribuição, com uma média nacional de 40,31%. Observa-se uma correlação inversa entre atendimento e perdas: as regiões com menor cobertura de atendimento tendem a apresentar os maiores índices de perdas. A região Norte lidera negativamente com 49,78% de perdas, seguida pelo Nordeste (46,25%), enquanto as regiões Centro-Oeste

quando comparados ao grupo de pior desempenho. Esses dados reforçam que as elevadas perdas observadas nas regiões Norte (49,78%) e Nordeste (46,25%) estão diretamente associadas aos menores níveis de investimento e à necessidade urgente de implementação de soluções tecnológicas eficientes e economicamente viáveis (ITB, 2025).

Tabela 1 – Dados das 20 melhores *versus* 20 piores cidades (SINISA 2025)

Indicador	20 melhores	20 piores
População total (IBGE 2022)	22.760.067	13.999.929
Investimento total de 2019 a 2023 (R\$ MM)	20.073,33	5.488,13
Investimento médio por habitante (R\$/hab./ano)	176,39	78,40
Indicador de atendimento total de água (%)	98,85	81,50
Indicador de perdas na distribuição (%)	25,53	45,51

Para o combate às perdas de água, diversas metodologias têm sido aplicadas no setor de saneamento, baseadas em campanhas de medição, uso de inteligência artificial e gestão integrada de dados. Entre as principais abordagens destacam-se:

- macromedição por meio de sensores físicos em pontos estratégicos;
- setorização da rede para controle localizado;
- gestão de pressão através de válvulas redutoras;
- modelagem hidráulica e sistemas de monitoramento inteligente;
- detecção ativa de vazamentos com equipamentos acústicos especializados.

A instalação de sensores físicos, como macromedidores de vazão, embora ofereça medições diretas e precisas, enfrenta limitações econômicas e operacionais, restringindo sua aplicação a regiões metropolitanas e inviabilizando sua implementação em larga escala.

Nesse contexto, a medição indireta surge como uma alternativa viável, permitindo a estimativa da vazão a partir de variáveis correlacionadas, como pressão na rede de abastecimento, nível dos reservatórios e frequência de operação dos inversores. Essa abordagem reduz significativamente os custos de instrumentação e manutenção, tornando possível um monitoramento mais abrangente (FLORES et al., 2020);(LIMA et al., 2022a).

A inteligência artificial (IA) tem se mostrado uma ferramenta promissora para otimizar a gestão de sistemas de abastecimento de água, especialmente por meio de técnicas de aprendizado de máquina (*machine learning*). O desenvolvimento de sensores virtuais (*soft sensors*) permite não apenas a estimativa indireta de variáveis críticas, mas também a detecção de falhas, a reconstrução de dados ausentes e a previsão de comportamentos futuros do sistema (ALENCAR et al., 2024). Além disso, esses modelos

podem ser integrados a plataformas de monitoramento em nuvem, proporcionando uma visão em tempo real do desempenho da rede e facilitando a tomada de decisão pelos operadores.

Neste trabalho, tem-se como objetivo o desenvolvimento de um sensor virtual de vazão baseado em redes neurais artificiais, com foco na arquitetura *Long Short-Term Memory* (LSTM), conhecida por sua capacidade de modelar dependências temporais e capturar dinâmicas complexas em séries temporais. O modelo será treinado e validado com dados reais provenientes da Companhia de Água e Esgotos da Paraíba (CAGEPA), coletados na cidade de Salgado de São Félix.

Além disso, este trabalho incorpora técnicas de detecção e correção de *outliers* por meio do algoritmo TEDA (*Typicality and Eccentricity Data Analysis*), garantindo a qualidade dos dados utilizados no treinamento das redes neurais. Esse algoritmo baseia-se nos conceitos de tipicidade e excentricidade, permitindo a identificação adaptativa de valores anômalos em fluxos contínuos de dados. A integração com plataformas em nuvem não apenas centraliza a aquisição das informações, mas também possibilita o processamento e a análise em tempo real, aspectos fundamentais para a implementação de sistemas inteligentes em ambientes operacionais críticos, como as estações de distribuição de água.

Assim, o desenvolvimento deste sensor virtual representa um avanço significativo tanto do ponto de vista tecnológico quanto operacional, promovendo a integração de técnicas de IA com as necessidades práticas do setor hídrico. Ao combinar métodos de medição indireta com a robustez das redes neurais LSTM, este trabalho propõe uma solução escalável e economicamente viável, contribuindo para a modernização dos sistemas de abastecimento de água e para a gestão sustentável dos recursos hídricos.

1.1 OBJETIVOS

1.1.1 OBJETIVO GERAL

Neste trabalho tem-se como objetivo geral o desenvolvimento de um modelo de sensor virtual de vazão baseado em redes neurais artificiais com a finalidade de realizar as estimações da vazão de forma indireta e que relacione as medições diretas, tais como pressões, nível e vazões.

1.1.2 OBJETIVOS ESPECÍFICOS

Dos objetivos específicos, os principais deste trabalho são os seguintes:

- Aplicação de técnicas de IA para o desenvolvimento de *Soft sensors* de vazão aplicados na macromedição como ferramenta de combate as perdas.
- Validar o desenvolvimento do sensor virtual de vazão usando redes neurais artificiais.
- Adequar o sistema de aquisição de dados em nuvem de uma estação de abastecimento de água por meio de uso de plataformas em nuvem.
- Verificação da qualidade de dados adquiridos, *outliers*, e suas respectivas correções usando o algoritmo TEDA.

1.2 CONTRIBUIÇÕES

Como contribuição, este trabalho oferece:

- O desenvolvimento de um sensor virtual de vazão baseado em redes neurais artificiais permite estimar a vazão de forma indireta a partir de variáveis acessíveis (pressões, níveis e frequência do inversor). Isso representa um avanço em relação aos métodos tradicionais de medição direta, oferecendo uma solução inovadora e potencialmente mais econômica.
- A adequação do sistema de aquisição de dados por meio de plataformas em nuvem contribui para a centralização e o monitoramento em tempo real das informações operacionais. Essa integração melhora a gestão da estação de abastecimento de água e facilita a análise e o armazenamento dos dados coletados.
- A aplicação do algoritmo TEDA para verificação, detecção de *outliers* e correção dos dados assegura que as entradas utilizadas para o treinamento e a validação das redes neurais sejam de alta qualidade. Isso aumenta a confiabilidade do modelo e a precisão das estimativas da vazão.
- Ao proporcionar uma estimativa indireta e confiável da vazão, o sensor virtual pode apoiar decisões operacionais mais eficientes na estação de abastecimento, contribuindo para a otimização do uso da água e redução de perdas no sistema.

1.3 ESTRUTURA E ORGANIZAÇÃO DO TRABALHO

Além deste capítulo introdutório, o trabalho está composto por mais 5 capítulos, descritos a seguir:

- **Capítulo 2:** Neste capítulo, é realizada uma revisão abrangente do estado da arte relacionado ao tema deste trabalho. São exploradas as principais teorias, metodologias

e tecnologias empregadas na área, bem como as contribuições e limitações dos estudos existentes.

- **Capítulo 3:** Este capítulo apresenta a fundamentação teórica da pesquisa, oferecendo uma análise detalhada dos conceitos, teorias e modelos que embasam o desenvolvimento do trabalho. O objetivo é estabelecer uma base conceitual sólida, que não só esclareça os princípios essenciais relacionados ao tema, mas também justifique a abordagem metodológica adotada.
- **Capítulo 4:** Neste capítulo, são apresentados em detalhes os métodos e procedimentos adotados para a concepção e desenvolvimento desta pesquisa. Serão descritas as etapas metodológicas, desde a coleta e pré-processamento dos dados até o desenvolvimento, treinamento e validação dos modelos propostos. Além disso, este capítulo detalha os instrumentos, ferramentas e técnicas empregadas, justificando as escolhas metodológicas e demonstrando como elas contribuem para a evolução e robustez dos resultados obtidos.
- **Capítulo 5:** Este capítulo apresenta em detalhes os resultados obtidos.
- **Capítulo 6:** Finalmente, neste capítulo, as conclusões deste estudo são delineadas, sintetizando os resultados alcançados..

2 ESTADO DA ARTE

Neste capítulo de estado da arte serão apresentados os principais avanços relacionados ao uso de sensores virtuais em sistemas de abastecimento de água, com ênfase nas arquiteturas mais utilizadas na literatura.

O desenvolvimento de sensores virtuais, ou *soft sensors*, para a medição de vazão em sistemas de abastecimento de água tem ganhado destaque como alternativa aos métodos de medição direta, tradicionalmente associados a custos elevados e dificuldades de implementação em larga escala. A proposta de utilizar redes neurais artificiais, em particular a arquitetura *Long Short Term Memory* (LSTM), para estimar a vazão a partir de variáveis correlacionadas (como pressão, nível do reservatório e frequência do inversor) reflete um avanço significativo na área de monitoramento e gerenciamento dos recursos hídricos.

Vários estudos têm demonstrado o potencial dos *soft sensors* em substituir ou complementar os sensores físicos tradicionais. Flores et al. (2020) evidenciaram que técnicas baseadas em inteligência artificial podem estimar de forma indireta variáveis de processo, permitindo a redução dos custos de instrumentação e a ampliação do monitoramento em redes de distribuição. Nesse contexto, Lima et al. (2022b) desenvolveram um *soft sensor* para a estimação da vazão utilizando redes neurais artificiais em um ambiente de laboratório, obtendo resultados promissores com erros médios absolutos percentuais inferiores a 5%. Tais abordagens não apenas possibilitam uma melhor cobertura do sistema, mas também permitem a integração de dados operacionais em tempo real, contribuindo para a tomada de decisões mais eficientes.

Além dos métodos baseados em redes neurais tradicionais, o uso de arquiteturas LSTM tem se destacado devido à sua capacidade de modelar séries temporais e capturar dinâmicas não lineares presentes em sistemas de abastecimento de água. As LSTMs foram inicialmente propostas por Hochreiter e Schmidhuber (1997) para resolver o problema de dissipação do gradiente em redes recorrentes, e desde então têm sido aplicadas em diversas áreas, incluindo previsão de vazão e detecção de falhas em processos industriais. Essa técnica permite a consideração tanto de informações recentes (memória de curto prazo) quanto de dados históricos (memória de longo prazo), o que é fundamental para a modelagem de sistemas que apresentam variações operacionais significativas ao longo do tempo.

A literatura recente também enfatiza a importância da integração de plataformas em nuvem para a aquisição e análise dos dados. A utilização de brokers MQTT para transmissão dos dados e a implementação de sistemas SCADA em ambientes de monitoramento remoto

são práticas cada vez mais comuns, facilitando a centralização e o processamento em tempo real das informações coletadas (LIMA et al., 2022a); (ALENCAR et al., 2024). Essa integração tecnológica permite não somente a validação dos modelos de sensores virtuais, mas também a implementação de estratégias de manutenção preditiva e de otimização operacional.

Outra linha de pesquisa destacada na literatura trata da detecção e correção de outliers, visando aprimorar a qualidade dos dados empregados no treinamento dos modelos. Técnicas de pré-processamento, como o algoritmo TEDA (*Typicality and Eccentricity Data Analysis*), têm demonstrado eficácia na identificação e tratamento de valores anômalos, contribuindo para a maior confiabilidade dos *soft sensors* (ANDRADE et al., 2024; AGGARWAL, 2017).

Em suma, o estado da arte revela que a combinação de medição indireta, redes neurais LSTM e integração com plataformas de nuvem constitui uma abordagem robusta e inovadora para a estimativa de vazão em sistemas de abastecimento de água. Essa estratégia oferece vantagens significativas em termos de custo, escalabilidade e precisão, representando uma tendência promissora para o avanço da gestão de recursos hídricos e para o desenvolvimento de soluções inteligentes que atendam às exigências regulatórias e operacionais do setor.

De acordo com o estado da arte, os principais desafios e limitações para o desenvolvimento de um sensor virtual de vazão utilizando redes neurais são:

Qualidade e Representatividade dos Dados: A precisão do modelo depende fortemente da qualidade dos dados coletados. Ruídos, *outliers* e inconsistências nas medições podem comprometer a eficácia das redes neurais. Técnicas de pré-processamento, como a aplicação do algoritmo TEDA para detecção e correção de *outliers*, são essenciais, mas também representam um desafio, pois a seleção adequada de variáveis e a normalização dos dados são críticas para um treinamento robusto (AGGARWAL, 2017); (ANDRADE et al., 2024).

Estabelecimento de Relações Não Lineares: A medição indireta de vazão, baseada em variáveis correlacionadas como pressão, nível do reservatório e frequência do inversor, exige a modelagem precisa de relações não lineares. Apesar das redes neurais LSTM terem mostrado potencial para capturar dependências temporais e dinâmicas complexas, a definição de uma relação confiável entre as variáveis medidas e a vazão real pode ser bastante desafiadora (LIMA; VILLANUEVA; CATUNDA, 2022); (FLORES et al., 2020).

Generalização do Modelo: Um dos desafios críticos é a capacidade do sensor virtual de generalizar seu desempenho para diferentes condições operacionais e contextos. Modelos treinados em dados de uma estação ou em condições específicas podem não se comportar da mesma forma em outras situações, o que requer estratégias avançadas de transferência de

aprendizado ou adaptação de domínio para garantir robustez e escalabilidade (ALENCAR et al., 2024).

Integração com Plataformas em Nuvem e Sistemas SCADA: Embora a integração com plataformas em nuvem e sistemas de monitoramento remoto (via MQTT, por exemplo) ofereça vantagens na centralização e no processamento em tempo real, ela também traz desafios em termos de segurança, latência e gerenciamento de grandes volumes de dados. Garantir a comunicação estável e segura entre os sensores virtuais e os sistemas de controle é um aspecto crítico para a viabilidade prática da solução.

Limitações Computacionais e Complexidade do Modelo: O treinamento de redes neurais, especialmente arquiteturas LSTM, pode ser computacionalmente intensivo. Isso exige uma infraestrutura adequada e, muitas vezes, a utilização de aceleradores como GPUs ou TPUs para lidar com grandes volumes de dados e realizar ajustes finos nos parâmetros do modelo, o que pode aumentar os custos e a complexidade da implementação. Esses desafios refletem as dificuldades inerentes à aplicação de técnicas de aprendizado de máquina para a modelagem de sistemas complexos, mas também apontam para áreas de oportunidade onde novas pesquisas podem contribuir para superar essas limitações e aprimorar a eficiência dos sistemas de macromedição em estações de distribuição de água.

3 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os conceitos fundamentais que sustentam o desenvolvimento deste trabalho, fornecendo o embasamento teórico necessário para a compreensão da pesquisa.

Inicialmente, serão discutidos na seção 3.1 os sistemas de abastecimento de água, abordando sua estrutura e desafios operacionais, especialmente no que se refere à medição e controle de variáveis. Na sequência, será feita uma revisão sobre redes neurais artificiais na seção 3.2, com ênfase nas principais arquiteturas utilizadas em aplicações de modelagem e previsão: as redes *Multilayer Perceptron* (3.2.1), as redes neurais convolucionais (3.2.2) e as redes neurais recorrentes (3.2.3). Ademais, na seção 3.3 será dado um foco especial às redes do tipo LSTM, destacando sua capacidade de lidar com sequências temporais e sua aplicabilidade na modelagem de sistemas dinâmicos. Além disso, na seção 3.4 será abordado o conceito de sensores virtuais, sua implementação e vantagens na análise e monitoramento de processos industriais e infraestruturais. Por fim, será discutida na seção 3.5 a detecção de outliers utilizando o algoritmo TEDA, ferramenta que permite identificar anomalias em séries temporais e reforçar a confiabilidade das estimativas geradas pelo modelo.

3.1 SISTEMA DE ABASTECIMENTO

Os Sistemas de Abastecimento de Água (SAA) são estruturas formadas para garantir o fornecimento adequado de água potável a uma determinada população. Esses sistemas são compostos por um conjunto integrado de equipamentos, obras e serviços, com a finalidade de captar, tratar, armazenar e distribuir a água de maneira eficiente e segura (GOMES, 2009).

Um sistema de abastecimento é geralmente dividido em diferentes subsistemas, conforme ilustrado na Figura 3, cada um desempenhando um papel na disponibilidade de água potável para o consumidor. Os principais componentes incluem (GOMES, 2021):

- **Captação:** Pode ser realizada em mananciais superficiais, como rios e lagos, ou subterrâneos, como aquíferos. A escolha da fonte depende da disponibilidade e qualidade da água.
- **Estação Elevatória:** Instaladas quando há necessidade de impulsionar a água para regiões mais elevadas, utilizando bombas e motores para vencer desníveis topográficos.

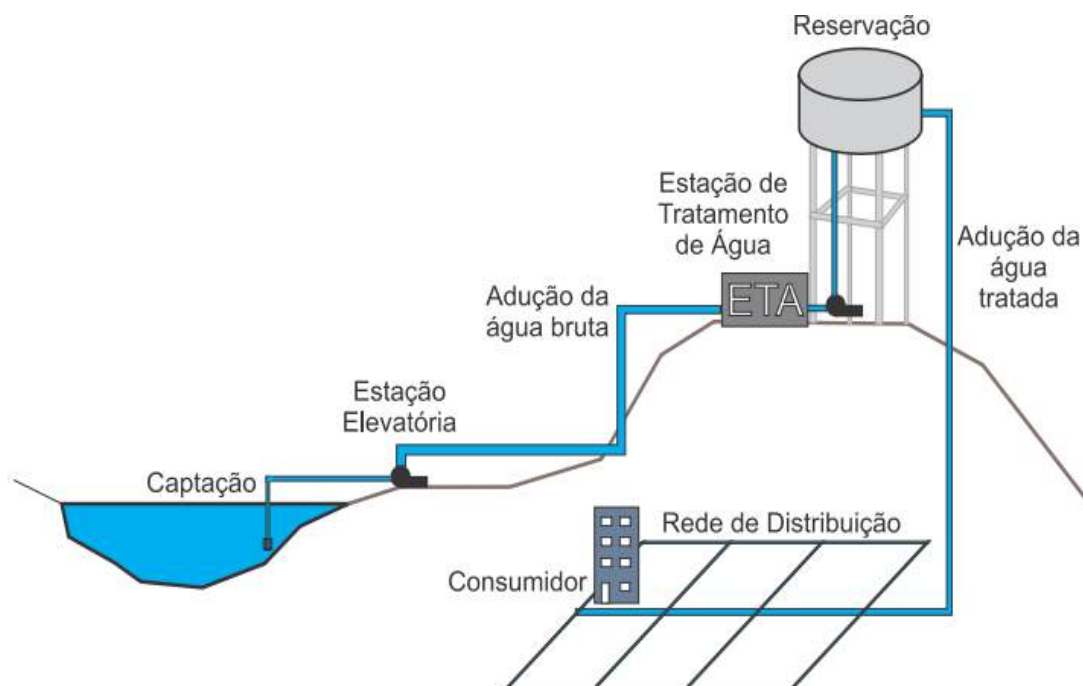


Figura 3 – Estrutura típica de um sistema de abastecimento de água (SALVINO, 2012).

- Estação de Tratamento de Água (ETA): Responsável pelo processo de purificação da água bruta, removendo impurezas físicas, químicas e biológicas para torná-la potável. Um exemplo de ETA pode ser visto na Figura 4.
- Reservatórios de distribuição e regularização: Armazenam a água tratada, regulando a pressão e garantindo o suprimento contínuo mesmo em períodos de alta demanda. Um reservatório de água elevado é ilustrado na Figura 5.
- Adutoras: Sistemas de tubulação que transportam a água tratada da ETA para os reservatórios de distribuição. Podem operar por gravidade ou exigir bombeamento.

A distribuição de água ocorre por ação da gravidade e, principalmente, através de sistemas de bombeamento, sobretudo em regiões de planície (GOMES, 2009).

Os SAAs requerem controle e automação para monitoramento de variáveis como vazão, pressão e temperatura (QUEVEDO et al., 2014). A automação promove eficiência energética e hidráulica (SALVINO et al., 2018), mas envolve altos custos devido à necessidade de sensores, bombas, controladores e instalações elétricas. A aquisição de instrumentos de medição pode elevar ainda mais os custos, tornando essencial o desenvolvimento de técnicas que reduzam essa dependência sem comprometer a precisão (LIMA et al., 2022b). Além disso, o gerenciamento dos dados coletados pelos sensores é realizado por sistemas SCADA (Supervisory Control and Data Acquisition), que, devido ao seu alto custo, levam



Figura 4 – Estação de tratamento de água.



Figura 5 – Reservatório elevado de água.

muitas companhias a buscar alternativas mais acessíveis, como softwares de baixo custo ou soluções gratuitas.

Uma alternativa ao monitoramento de grandezas hidráulicas de forma direta é a medição indireta, que estima a variável de interesse com base em outra variável correlacionada (LIMA et al., 2022a). Esse método reduz a necessidade de sensores para todas as grandezas físicas, desde que seja possível estabelecer uma relação matemática entre elas, como a correlação para dependências lineares ou a correntropia para relações não lineares (FLORES et al., 2020).

3.2 REDES NEURAIS ARTIFICIAIS

Segundo Haykin (2001), uma Rede Neural Artificial consiste em um processador maciçamente distribuído em paralelo, formado por unidades de processamento simples, capaz de armazenar conhecimento adquirido pela experiência e disponibilizá-lo para uso.

É ilustrado na Figura 6 um esquema simples de uma RNA de uma única camada oculta. De modo geral, uma Rede Neural Artificial (RNA) é composta por três tipos de camadas: a camada de entrada, onde cada neurônio recebe um sinal com informações a serem processadas; as camadas ocultas, responsáveis pelos cálculos matemáticos por meio de pesos e vieses; e a camada de saída, onde o resultado final do processamento é gerado.

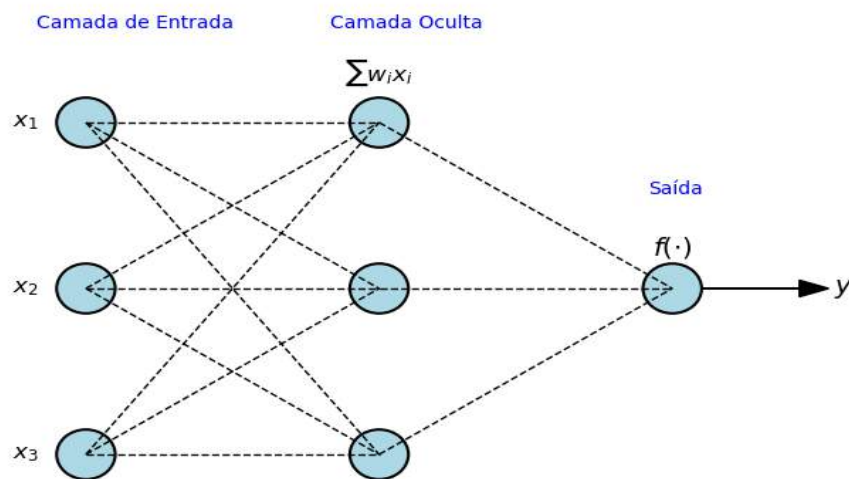


Figura 6 – Esquema ilustrativo de uma RNA (KUBAT, 1999).

O treinamento é o aspecto central de uma RNA, pois determina sua capacidade de identificar padrões e generalizar conceitos. Esse processo computacional pode variar em duração, desde alguns segundos até vários dias, dependendo da complexidade da rede e do volume de dados. Haykin et al. (2009) elenca as três principais formas de treinar uma rede neural. No treinamento supervisionado, a RNA recebe tanto os sinais de entrada quanto as saídas esperadas, permitindo que aprenda a associar corretamente os dados. Já no treinamento não supervisionado, a rede não possui conhecimento prévio das saídas desejadas, ajustando seus pesos de forma autônoma para identificar padrões nos dados. Por fim, o treinamento por reforço ocorre por tentativa e erro, onde a rede avalia suas respostas e ajusta seus parâmetros de acordo com um critério de desempenho, reforçando comportamentos que levam a melhores resultados.

A busca por otimização do treinamento é, portanto, um tema fundamental na pesquisa acadêmica, visando reduzir o tempo de processamento sem comprometer a precisão

do modelo. O custo computacional envolvido nesse processo está diretamente relacionado a diversos fatores, como o número de variáveis de entrada, o tamanho da base de dados, a quantidade de neurônios e camadas ocultas, além do algoritmo de otimização utilizado. Estratégias como o uso de técnicas avançadas de ajuste de pesos, escolha eficiente da função de ativação e implementação de métodos de regularização são essenciais para melhorar o desempenho da RNA e garantir sua capacidade de generalização.

As redes neurais artificiais têm sido amplamente utilizadas para a previsão de séries temporais devido à sua capacidade de reconhecer padrões nos dados (PALMITESSA et al., 2021). Neste contexto, é usual que o conceito seja apresentado em associação às distintas arquiteturas existentes, entre as quais se destacam a *Multilayer Perceptron* (MLP), as Redes Neurais Convolucionais (CNN) e as Redes Neurais Recorrentes (RNN). Cada uma dessas arquiteturas possui características estruturais e funcionais específicas, as quais serão detalhadas nas seções seguintes.

3.2.1 REDES MULTILAYER PERCEPTRON

As redes *Multilayer Perceptron* (MLP) são um tipo de rede neural artificial caracterizado por sua arquitetura *feedforward*, composta por três camadas principais: entrada, ocultas e saída, conforme ilustra a Figura 7. As camadas ocultas possuem neurônios interconectados por pesos ajustáveis, permitindo que a rede aprenda padrões complexos nos dados. O número de camadas ocultas e de neurônios pode variar conforme a aplicação, sendo as MLP amplamente utilizadas em tarefas como otimização, identificação de sistemas, controle de processos, robótica e reconhecimento de padrões (SALVINO et al., 2018).

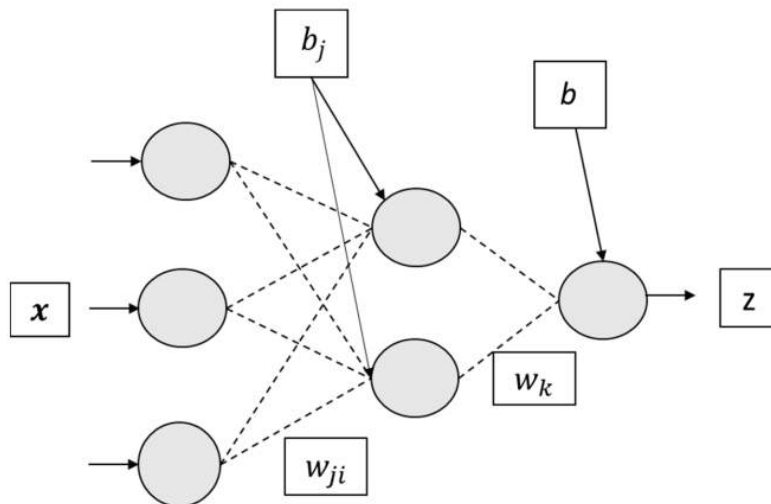


Figura 7 – Exemplo da topologia de uma MLP ((LAZCANO; JARAMILLO-MORÁN; SANDUBETE, 2024)).

O aprendizado em redes MLP ocorre por meio do algoritmo de *backpropagation*,

um processo iterativo que ajusta os pesos sinápticos para minimizar o erro do modelo. Esse treinamento ocorre em duas fases: na fase direta, a entrada é propagada pela rede até a saída, mantendo os pesos inalterados; na fase reversa, o erro entre a saída prevista e a real é calculado e propagado de volta pela rede, ajustando os pesos de acordo com um critério de minimização, como o método dos mínimos quadrados. Além disso, as redes MLP utilizam funções de ativação não lineares em suas camadas ocultas, garantindo maior capacidade de modelagem de relações complexas e não lineares entre os dados.

3.2.2 REDES NEURAIS CONVOLUCIONAIS

As Redes Neurais Convolucionais (Convolutional Neural Networks – CNNs) representam uma classe especializada de Redes Neurais Artificiais projetada para processar dados com uma estrutura de grade, como imagens e séries temporais. Inicialmente desenvolvidas para tarefas de visão computacional, as CNNs logo ganharam destaque por sua capacidade de extrair conhecimento útil e aprender representações internas de dados de alta dimensionalidade, sem a necessidade de realizar a extração manual de características ou padrões (ALENCAR et al., 2024). Um exemplo de rede convolucional se ilustra na Figura 8.

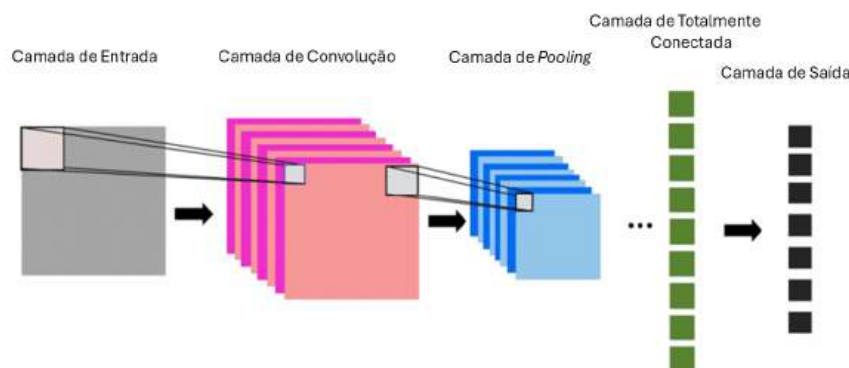


Figura 8 – Exemplo de uma CNN ((DING et al., 2024)).

Diferentemente das redes Multilayer Perceptron (MLP), as CNNs utilizam operações matemáticas chamadas convoluções para capturar automaticamente padrões relevantes nos dados de entrada. Sua arquitetura é composta por três tipos principais de camadas:

- Camadas Convolucionais: responsáveis pela extração de características, aplicam filtros sobre os dados de entrada, operando como pequenas janelas deslizantes que detectam padrões espaciais ou temporais relevantes.

- Camadas de *Pooling*: reduzem a dimensionalidade dos mapas de características por meio de operações de subamostragem, como o *max pooling*, que extrai os valores máximos de pequenas regiões da camada convolucional, mantendo as informações mais significativas e reduzindo a sensibilidade a pequenas variações nos dados.
- Camadas Totalmente Conectadas: recebem os features extraídos pelas camadas anteriores e os transformam em uma saída estruturada, garantindo que a rede neural possa realizar previsões de maneira eficiente.

O treinamento das CNNs segue o mesmo princípio das MLPs, baseado no algoritmo de *backpropagation*, onde os pesos dos filtros são ajustados para minimizar o erro da previsão. Entretanto, devido ao alto número de parâmetros e cálculos envolvidos, o treinamento eficiente das CNNs geralmente exige grande capacidade computacional, sendo realizado com aceleradores como GPUs e TPUs.

3.2.3 REDES NEURAIIS RECORRENTES

As redes neurais recorrentes (RNNs) são uma extensão das redes neurais convencionais, projetadas para processar dados sequenciais. Diferente das redes *feedforward*, que tratam cada entrada de forma independente, as RNNs possuem estados ocultos recorrentes que armazenam informações de etapas anteriores, permitindo a modelagem de dependências temporais. Embora, em teoria, possam capturar longos contextos, na prática, sua memória é limitada, restringindo a influência de sequências muito extensas (MIGLIATO, 2021).

A Figura 9 ilustra o funcionamento de uma RNN, onde a entrada X_t é processada pelo bloco de estado oculto H_t , que armazena informações dos dados ao longo do tempo. Diferente das redes neurais tradicionais, a RNN inclui uma conexão recorrente, representada pelo caminho de realimentação na figura, que permite que o estado oculto de um instante anterior H_{t-1} influencie o cálculo do estado atual. Esse mecanismo é formalizado pelo operador Z^{-1} , que indica o atraso de uma unidade de tempo. A saída Y_t é então gerada com base no estado oculto atualizado.

Entretanto, conforme mais dados passados são utilizados o gradiente do erro, devido à sua natureza multiplicativa, tende a se aproximar de zero ao longo das iterações caso seja menor que 1, caracterizando o problema de dissipação do gradiente. Por outro lado, se o valor do gradiente for maior que 1, ele pode crescer indefinidamente, resultando no problema de estouro do gradiente. No primeiro caso, a dissipação do gradiente dificulta o aprendizado, pois a rede neural perde a capacidade de ajustar adequadamente seus pesos. Já no segundo caso, o estouro do gradiente provoca a divergência dos valores da rede em iterações mais longas, comprometendo a estabilidade do modelo.

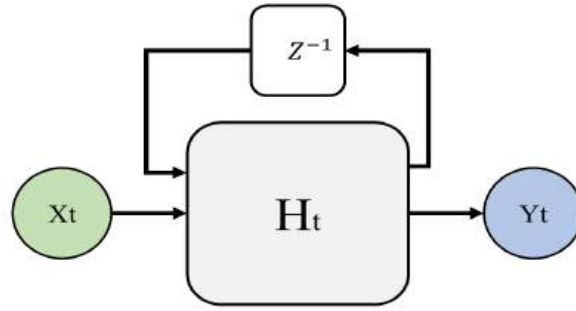


Figura 9 – Arquitetura básica de uma RNN (adaptado de (SCHUSTER; PALIWAL, 1997)).

3.3 MEMÓRIA DE CURTO E LONGO PRAZO

Hochreiter e Schmidhuber (1997), buscando resolver o problema da dissipação do gradiente, desenvolveram, em 1997, o modelo LSTM, uma variante das redes neurais recorrentes que permite armazenar informações importantes por períodos mais longos.

As redes LSTM se inspiram no processo de esquecimento da mente humana para otimizar a utilização dos dados na previsão de novos valores. Desse modo, os dados recentes (memória de curto prazo) são avaliados em conjunto com informações previamente acumuladas (memória de longo prazo), possibilitando que características pouco relevantes ou desnecessárias sejam descartadas (HE et al., 2023).

A arquitetura do LSTM é composta por unidades de memória, chamadas células LSTM, conforme ilustrado na Figura 10. Cada célula inclui três portas com papéis distintos: a porta de esquecimento descarta informações irrelevantes, a porta de entrada incorpora novas informações ao estado da célula, e a porta de saída regula quais informações armazenadas serão usadas para gerar a saída atual (NASSER; RASHAD; HUSSEIN, 2020).

3.3.1 PORTA DE ESQUECIMENTO

A finalidade da porta de esquecimento é determinar que informações do estado anterior S_{t-1} devem ser esquecidas e quais serão mantidas através da Equação 3.1, na medida que mantêm o gradiente do erro sob controle.

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (3.1)$$

Na Equação f_t é o valor da porta de esquecimento, onde a entrada atual x_t , o estado oculto da iteração anterior h_{t-1} e um *bias* b_f (em português "viés") são concatenados na

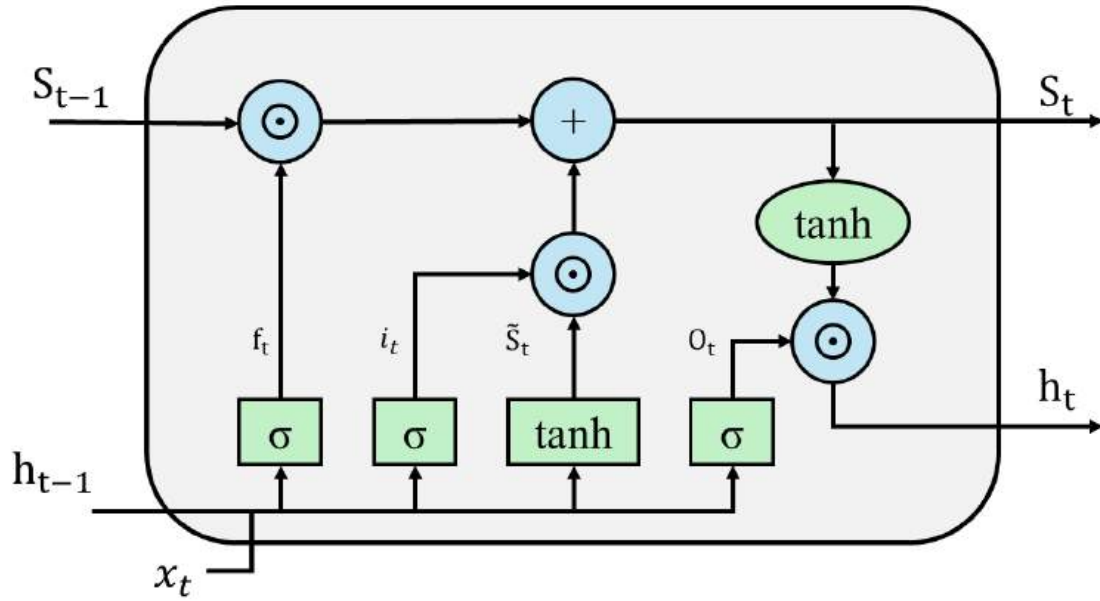


Figura 10 – Estrutura de uma célula LSTM (adaptado de (NASSER; RASHAD; HUSSEIN, 2020)).

função sigmoide σ , dada pela Equação 3.2. W_x e W_h são as matrizes de pesos associadas à x_t e h_{t-1} , respectivamente.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (3.2)$$

O vetor resultante da Equação 3.1 é multiplicado elemento a elemento com S_{t-1} , de modo que, para cada posição no vetor, um valor de 0 em f_t esquece a informação do estado anterior, enquanto um valor de 1 o deixa inalterado.

3.3.2 PORTA DE ENTRADA E MEMÓRIA CANDIDATA

A porta de entrada i_t insere novas informações que serão armazenadas no estado atual da célula. De forma similar à porta de esquecimento, o vetor de entrada x_t e o de estado h_{t-1} são concatenados e passam por uma função sigmoide, como é definido na Equação 3.3.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (3.3)$$

O segundo elemento desta etapa é chamado memória candidata $\tilde{S}_t$ (ou estado candidato (HE et al., 2023)) cuja fórmula é expressa na Equação 3.4.

$$\tilde{S}_t = \tanh(W_{xs}x_t + W_{hs}h_{t-1} + b_s) \quad (3.4)$$

A memória candidata também recebe x_t e h_{t-1} , mas difere das portas de entrada e esquecimento por utilizar a função tangente hiperbólica. Diferente da função sigmoide, a função $\tanh$ varia de -1 a 1.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3.5)$$

A porta de entrada e a memória candidata são multiplicadas elemento a elemento e, posteriormente, somadas ao resultado da etapa de esquecimento, conforme a Equação 3.6.

$$S_t = f_t \odot S_{t-1} + i_t \odot \tilde{S}_t \quad (3.6)$$

em que: S_t é o vetor de estado atual que será utilizado na próxima iteração e para calcular o vetor de estado oculto atual h_t .

3.3.3 PORTA DE SAÍDA

A função da porta de saída é gerar o estado oculto h_t . Ele é calculado a partir do estado atual S_t e mais uma vez pelo vetor concatenado da entrada e do estado oculto anterior, que por sua vez é dado pela Equação 3.7. O estado S_t passa por uma função $\tanh$ e então é multiplicado elemento a elemento com a porta de saída o_t , conforme a Equação 3.8. A função $\tanh$ é utilizada para auxiliar na estabilidade do sistema.

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (3.7)$$

$$h_t = o_t \odot \tanh(S_t) \quad (3.8)$$

3.4 SENSORES VIRTUAIS

Um sensor virtual, comumente chamado pelo seu nome em inglês *soft sensor*, é um modelo computacional que simula, em um ambiente virtual, o funcionamento de um sensor real (HAYKIN et al., 2009). O termo *soft sensor* surge da junção das palavras "software" e "sensor", pois esses modelos são desenvolvidos em programas computacionais e fornecem informações comparáveis às dos sensores físicos (LUTTMANN et al., 2012).

Esses modelos utilizam dados de variáveis secundárias para calcular a variável de interesse, podendo atuar tanto como substitutos quanto como verificadores do desempenho de sensores tradicionais (LOTUFO; GARCIA, 2008). Sua implementação é motivada por

fatores como erros de medição, atrasos na aquisição de dados, inviabilidade do uso de sensores físicos em determinados ambientes e altos custos de instrumentação (LOTUFO; GARCIA, 2008). Dentre as diversas abordagens para o desenvolvimento de sensores virtuais, uma das mais eficazes é o uso de redes neurais artificiais, capazes de modelar relações complexas entre variáveis e melhorar a precisão das estimativas (LIMA et al., 2022b).

O design de um sensor virtual é feito, segundo Fortuna et al. (2007) e conforme ilustrado na Figura 11, a partir das seguintes etapas: seleção de dados, correção de *outliers*, estruturação do modelo, estimação do modelo e validação do modelo.

Segundo Lima, Villanueva e Catunda (2022), é comum que o desenvolvimento de sensores virtuais dependa exclusivamente de dados históricos de uma planta industrial, já que a realização de testes nem sempre é viável. Por isso, torna-se essencial avaliar as informações disponíveis sobre a operação do processo a fim de verificar sua adequação ao projeto. Desse modo, a primeira etapa consiste na seleção das variáveis adequadas ao objetivo.

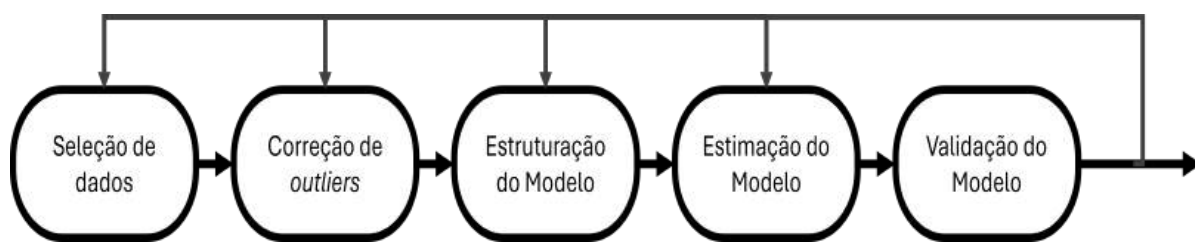


Figura 11 – Diagrama do processo de desenvolvimento de um sensor virtual (adaptado de (FORTUNA et al., 2007))

Jiang et al. (2021) e Kadlec, Gabrys e Strandt (2009) destacam diferentes formas de pré-processamento, tais como filtragem de ruído, normalização dos dados, verificação da integridade dos dados e correção de *outliers*. Jiang et al. (2021) explica que a correção de *outliers* consiste em aplicar métricas para detectar e eliminar valores espúrios advindos das medições dos sensores ou fatores externos à planta. A correção desses dados é de suma importância, pois os sensores virtuais são fortemente influenciados pela correlação entre as grandezas físicas. Ao continuar o desenvolvimento do modelo de sensor virtual, deve-se definir a representação matemática e a estrutura apropriada para o sistema em questão. Esta etapa envolve a escolha entre diversas arquiteturas de modelagem, como redes neurais artificiais, modelos estatísticos ou baseados em primeiros princípios (LOTUFO; GARCIA, 2008). A estimação do modelo busca otimizar sua capacidade de representar com precisão o fenômeno físico. Para isso, diferentes métodos de otimização podem ser empregados, dependendo da complexidade do sistema e da disponibilidade de dados. Entre as abordagens mais comuns, destacam-se os métodos de mínimos quadrados, otimização baseada em gradiente e algoritmos evolutivos, como redes neurais artificiais e algoritmos

genéticos.

Após a estimação, a etapa seguinte é a validação, onde o desempenho do modelo é testado em condições variadas para verificar sua precisão e capacidade de generalização. Caso os resultados não atendam aos critérios desejados, ajustes nos parâmetros ou até na estrutura do modelo podem ser necessários.

Uma vez que essas cinco etapas tenham sido concluídas e a estimação realizada esteja dentro dos limites desejados, pode-se fazer a implementação do sensor virtual. Entre as principais estratégias, destacam-se a integração do modelo em controladores programáveis (CLPs), o desenvolvimento de aplicações dedicadas em linguagens como Python ou MATLAB, e a incorporação em sistemas embarcados para processamento em tempo real (KADLEC; GABRYS; STRANDT, 2009).

3.5 DETECÇÃO DE OUTLIERS BASEADO NO ALGORITMO TEDA

A análise estatística de dados é uma importante etapa no desenvolvimento de um sensor virtual, através dela decide-se quais variáveis são mais propensas a resultados confiáveis. Dentro da análise de dados se destaca a detecção de *outliers*. Um *outlier* pode ser definido como um valor que, quando comparado aos outros dados amostrados, diverge da normalidade apresentada levando à suspeita de ter sido gerado por um outro mecanismo (HAWKINS, 1980). De forma prática, os dados são gerados por uma ou mais entidades, quer sejam elementos naturais ou artificiais. Quando uma dessas entidades não opera conforme o esperado, os *outliers* podem surgir (AGGARWAL, 2017).

Os *outliers* podem ser classificados como pontuais, contextuais e coletivos. Um outlier pontual é uma ocorrência de dados que se mostra incoerente em relação ao restante das informações disponíveis. Esse tipo de outlier é considerado um dos mais simples de identificar e representa o foco de grande parte das pesquisas na área de detecção de *outliers* (BANSAL; GAUR; SINGH, 2016). Por sua vez, *outliers* contextuais são aqueles que se desviam do comportamento esperado em relação a um contexto específico, considerando atributos que definem esse enquadramento. Eles são identificados com base em dois conjuntos principais de atributos: os atributos contextuais, que determinam o ambiente ou situação da análise (como latitude e longitude em dados espaciais), e os atributos comportamentais, que descrevem características observadas ou medidas em relação ao contexto (como a quantidade de chuva em uma região específica). A identificação de *outliers* contextuais depende da relação entre essas duas categorias, destacando ocorrências que não condizem com o padrão esperado dentro do contexto analisado (BANSAL; GAUR; SINGH, 2016). Já os *outliers* coletivos são conjuntos de dados que, quando considerados

individualmente, podem não ser anômalos, mas tornam-se anômalos quando analisados em conjunto. Um exemplo de *outlier* pode ser observado na Figura 12, na área delimitada pelo círculo azul, onde um conjunto de dados provenientes de um sensor de pressão apresenta um comportamento constante, enquanto os demais dados exibem alta variabilidade.

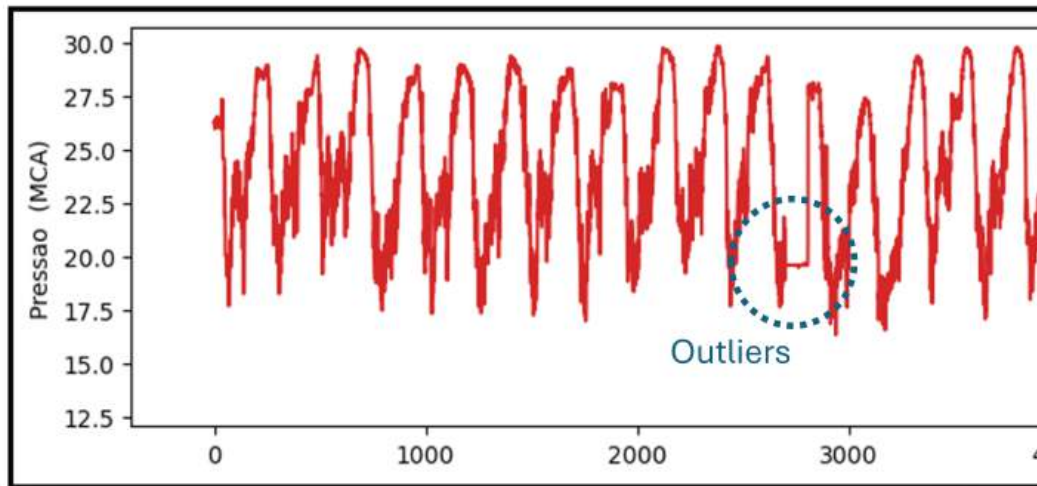


Figura 12 – Conjunto de dados com *outliers*.

Existem diversas maneiras de detectar *outliers* em um conjunto de dados. O método mais simples consiste em definir um intervalo de confiança com base no desvio padrão, considerando como *outliers* aqueles que estão fora do intervalo $\mu \pm 3\sigma$. Isso ocorre porque, em uma distribuição normal, 99,7% dos dados geralmente se encontram dentro desse intervalo. Outros métodos, como a clusterização utilizando o algoritmo K-means, podem também ser aplicados para identificar *outliers*. No entanto, métodos convencionais enfrentam dificuldades na detecção de *outliers* contextuais, que não necessariamente se desviam do intervalo de confiança, mas apresentam padrões anômalos dependendo do contexto em que ocorrem.

Uma forma de sanar esse problema é utilizar os conceitos de typicalidade e excentricidade, que ajudam a identificar *outliers* contextuais. A typicalidade refere-se à medida em que um dado se encaixa dentro de um padrão ou comportamento esperado, considerando as características de dados semelhantes. Já a excentricidade avalia a discrepância ou a singularidade de um dado em relação aos outros, focando na sua "distância" em relação aos dados mais próximos em um determinado contexto (ANGELOV, 2014b). Esse conceito é utilizado no algoritmo TEDA (*Typicality and Eccentricity Data Analysis*), que oferece uma abordagem matemática eficiente para identificar padrões e anomalias em fluxos de dados em tempo real. O algoritmo desenvolvido para tratar de novos dados dentro do sensor virtual será abordado em detalhes no capítulo de materiais e métodos.

No contexto do TEDA, a identificação de padrões e anomalias é realizada por meio

da proximidade acumulada (π), uma métrica derivada da soma das distâncias entre um ponto x e os demais pontos do conjunto de dados até o instante k . Em um espaço de dados n -dimensional $X \in \mathbb{R}^n$, as amostras de dados são organizadas em uma sequência ordenada $X = [x_1, x_2, x_3, \dots, x_k, \dots]$, onde cada x_k representa o sistema em um instante k . A distância entre dois pontos, x e x_i , pode ser calculada de diferentes formas, como a distância Euclidiana, Mahalanobis, etc (ANGELOV, 2014a). A proximidade acumulada de um ponto x , denotada por $\pi_k(x)$, é dada por

$$\pi_n(x) = \sum_{i=1}^n d(x, x_i), \quad n \geq 2 \quad (3.9)$$

Dessa forma, pode-se calcular a excentricidade ξ de uma amostra x no instante k através da Equação 3.10, sendo ela a razão entre o dobro da proximidade acumulada de x e a soma das proximidades acumuladas de todas as amostras até aquele instante.

$$\xi_k(x) = \frac{2\pi_k(x)}{\sum_{i=1}^k \pi_k(x_i)}, \quad k \geq 2 \quad \text{e} \quad \sum_{i=1}^k \pi_k(x_i) > 0 \quad (3.10)$$

Ou ainda, considerando a média da base de dados μ_k e a variância σ_k^2 , temos a excentricidade recursiva (ANDRADE et al., 2024):

$$\xi_k(x) = \frac{1}{k} + \frac{(\mu_k - x_k)^T(\mu_k - x_k)}{k\sigma_k^2} \quad (3.11)$$

A média μ_k e a variância σ_k^2 são calculadas de forma incremental, permitindo que novos dados sejam incorporados sem necessidade de armazenar todas as amostras anteriores. Essas definições recursivas são fundamentais para a eficiência computacional do TEDA, pois reduzem o custo de armazenamento e processamento dos dados ao longo do tempo.

$$\mu_k(x) = \frac{k-1}{k} \mu_{k-1} + \frac{1}{k} x_k, \quad \mu_1 = x_1 \quad (3.12)$$

$$\sigma_k^2(x) = \frac{k-1}{k} \sigma_{k-1}^2 + \frac{1}{k-1} |x_k - \mu_k|^2, \sigma_1^2 = 0 \quad (3.13)$$

Por sua vez, a tipicidade τ é definida matematicamente pela Equação 3.14. Enquanto a excentricidade mede o grau de desvio, a tipicidade é definida como seu complemento, ou seja, valores altos de τ indicam que a amostra está alinhada ao comportamento predominante dos dados.

$$\tau_k(x) = 1 - \xi_k(x) \quad (3.14)$$

E de modo análogo à Equação 3.11, temos a tipicidade recursiva:

$$\tau_k(x) = \frac{k-1}{k} - \frac{(\mu_k - x_k)^T(\mu_k - x_k)}{k\sigma_k^2} \quad (3.15)$$

Sempre que a excentricidade ultrapassa um limiar definido, o algoritmo identifica o ponto como *outlier*, sinalizando que ele se desvia de maneira significativa do restante dos dados. Para a detecção dos outliers, empregou-se a Equação 3.16, que define a medida de excentricidade, considerando-se um limiar de $m = 2,5$ (ANDRADE et al., 2024).

$$\zeta_k(x) > \frac{m^2 + 1}{2k}, \quad (3.16)$$

Em que $\zeta_k(x)$ é a excentricidade normalizada dos dados e calculada pela seguinte Equação (3.17):

$$\zeta_k(x) = \frac{\xi_k(x)}{2}, \quad \sum_{i=1}^k \zeta_i(x) = 1, \quad k \geq 2 \quad (3.17)$$

Sendo m um parâmetro que define a sensibilidade do limiar, um valor maior de n reduz a sensibilidade do método, resultando na detecção de uma menor quantidade de *outliers*.

3.6 ANÁLISE DE INCERTEZA BASEADO NO MÉTODO DROPOUT MONTE CARLO

A avaliação da incerteza em modelos de aprendizado profundo é um aspecto fundamental em aplicações que demandam alta confiabilidade, como é o caso dos sensores virtuais utilizados em sistemas de abastecimento de água. Entre os métodos disponíveis para estimar incertezas de natureza epistemológica, destaca-se o *Monte Carlo Dropout*, proposto por Gal e Ghahramani (2016a). Essa técnica interpreta o *dropout*, originalmente concebido como método de regularização, como uma aproximação variacional bayesiana, permitindo quantificar a incerteza das predições sem alterar significativamente a arquitetura da rede neural.

O *dropout* convencional, introduzido por Srivastava et al. (2014), consiste em desativar aleatoriamente unidades da rede durante o treinamento, reduzindo o sobreajuste e promovendo modelos mais generalizáveis. Gal et al. (2016) demonstrou que esse procedimento pode ser visto como uma aproximação variacional ao modelo bayesiano, onde a distribuição posterior exata dos pesos $p(W|D)$ é aproximada por uma distribuição estocástica induzida pelo *dropout*. Assim, manter o *dropout* ativo na fase de inferência equivale a realizar amostragens estocásticas dos pesos da rede.

No método *Monte Carlo Dropout*, uma mesma entrada x é propagada pela rede T vezes, cada vez com um subconjunto aleatório de neurônios desativados. Isso resulta em um conjunto de predições estocásticas

$$\hat{y}^{(1)}, \hat{y}^{(2)}, \dots, \hat{y}^{(T)},$$

a partir das quais se calcula a predição média:

$$\mathbb{E}[\hat{y}] = \frac{1}{T} \sum_{t=1}^T \hat{y}^{(t)},$$

e a variância associada, que representa a incerteza epistemológica do modelo:

$$\text{Var}(\hat{y}) \approx \tau^{-1} + \left(\frac{1}{T} \sum_{t=1}^T \hat{y}^{(t)} \hat{y}^{(t)T} \right) - \mathbb{E}[\hat{y}] \mathbb{E}[\hat{y}]^T.$$

O termo τ^{-1} corresponde ao nível de ruído modelado e é calculado segundo

$$\tau = \frac{l^2(1 - \rho)}{2N\lambda},$$

em que l é o *length scale*, ρ é a taxa de *dropout*, N é o número de amostras de treinamento e λ representa o peso da regularização L2. Dessa forma, o *Monte Carlo Dropout* fornece um intervalo de confiança associado às predições, aspecto essencial para aplicações industriais sensíveis a riscos operacionais.

Embora o método tenha sido inicialmente formulado para redes neurais *feed-forward*, Gal e Ghahramani (2016b) estenderam sua aplicação para redes recorrentes, como as LSTM. Nessa formulação, utiliza-se um padrão fixo de desativação ao longo do tempo, conhecido como dropout variacional, garantindo estabilidade nas conexões recorrentes. Esse procedimento torna o método particularmente adequado para sensores virtuais baseados em LSTM, que lidam com dados temporais e variabilidade intrínseca dos processos hidráulicos.

Em aplicações de sensores virtuais, o *Monte Carlo Dropout* apresenta diversas vantagens: permite estimar a confiabilidade de cada predição, auxilia na detecção de anomalias, contribui para a interpretação da resposta do modelo e possibilita a integração com técnicas de detecção de inconsistências, como o algoritmo TEDA.

Em síntese, o *Monte Carlo Dropout* combina simplicidade computacional, fundamentação bayesiana e aplicabilidade prática, tornando-se uma das estratégias mais eficientes e acessíveis para quantificação de incertezas em modelos de aprendizado profundo empregados no desenvolvimento de sensores virtuais.

3.7 CONCLUSÕES DO CAPÍTULO

Neste capítulo, foram apresentados os fundamentos teóricos que embasam o desenvolvimento do sensor virtual de vazão, proporcionando uma base conceitual sólida

para o trabalho. Iniciamos com a contextualização dos sistemas de abastecimento de água, destacando seus componentes críticos e os desafios operacionais que justificam a adoção de abordagens inovadoras para a medição indireta. Em seguida, exploramos detalhadamente as principais arquiteturas de redes neurais artificiais, incluindo as redes Multilayer Perceptron, Convolucionais e Recorrentes, com ênfase especial nas LSTM, cuja capacidade de lidar com sequências temporais complexas se mostra essencial para modelar as dinâmicas dos sistemas hídricos.

Adicionalmente, abordamos o conceito de sensores virtuais (*soft sensors*) e a importância do pré-processamento dos dados, destacando técnicas de correção de outliers, como o algoritmo TEDA, que garantem a qualidade dos dados para o treinamento dos modelos. Também foi introduzido o método de Monte Carlo *Dropout*, empregado para estimar a incerteza nas previsões do sensor virtual. A aplicação dessa técnica permite realizar múltiplas inferências estocásticas durante a fase de predição, fornecendo não apenas a estimativa da vazão, mas também intervalos de confiança associados, o que aumenta a confiabilidade do modelo em cenários operacionais reais. Essa combinação de técnicas de inteligência artificial com uma compreensão aprofundada dos sistemas de abastecimento de água estabelece um alicerce robusto para a implementação do sensor virtual proposto.

Em resumo, o capítulo não apenas esclarece os princípios teóricos que fundamentam o projeto, mas também evidencia as inter-relações entre os componentes do sistema, preparando o terreno para a aplicação prática e o desenvolvimento do modelo nos capítulos subsequentes. Essa integração teórica é fundamental para a compreensão dos desafios e das soluções adotadas, consolidando a relevância da abordagem proposta para a melhoria da macromedição e do gerenciamento de recursos hídricos.

4 MATERIAIS E MÉTODOS

Neste capítulo, são descritos os materiais e métodos empregados na condução da pesquisa, incluindo os recursos necessários para alcançar os objetivos propostos. A base de dados utilizada foi obtida a partir de sensores e atuadores instalados em uma estação elevatória de água localizada em Salgado de São Félix, Paraíba, com a transmissão dos dados realizada via *broker* MQTT. A modelagem do sensor virtual foi desenvolvida em Python e implementada em uma máquina virtual da Amazon Web Services (AWS), com os resultados visualizados por meio da plataforma Grafana.

4.1 MATERIAIS

A Figura 13 apresenta um sistema de abastecimento de água típico, composto por diferentes dispositivos de monitoramento e controle distribuídos ao longo do processo. Inicialmente, observa-se a captação de água, que é conduzida ao sistema por meio de adutoras equipadas com válvulas borboleta, responsáveis pela regulação do fluxo. Em pontos estratégicos, encontram-se sondas hidrostáticas, utilizadas para medir o nível da água em reservatórios, bem como um transmissor de pressão, que fornece informações sobre as condições de bombeamento e distribuição.

O sistema também conta com medidores ultrassônicos de vazão, instalados em diferentes trechos: tanto na saída do reservatório elevado quanto na linha de distribuição para a comunidade. Esses medidores permitem quantificar o volume de água transportado, sem contato direto com o fluido, utilizando ondas ultrassônicas para estimar a velocidade média do escoamento.

A etapa final é composta pela rede de distribuição, equipada com hidrômetros residenciais, que contabilizam o consumo individual dos usuários. Todo esse conjunto de equipamentos gera dados fundamentais para o balanço hídrico, controle operacional e combate a perdas no sistema.

No contexto da pesquisa proposta, destaca-se o medidor ultrassônico de vazão (FT1) como elemento central para a elaboração de um sensor virtual de vazão. Embora os medidores físicos forneçam medições diretas, sua instalação e manutenção apresentam custos elevados, especialmente em redes extensas. Dessa forma, o desenvolvimento de um sensor virtual permitirá estimar a vazão em tempo real a partir de variáveis já monitoradas no sistema (pressão, nível e operação das bombas), aproveitando os dados provenientes dos sistemas SCADA e de automação. Essa abordagem possibilitará maior cobertura de monitoramento, redução de custos e aumento da confiabilidade no acompanhamento do

desempenho hidráulico do sistema de abastecimento. O sensor virtual foi modelado a partir de dados coletados na estação elevatória ilustrada na Figura 13 na qual são apresentados os elementos físicos presentes na medição da pressão, vazão e nível.

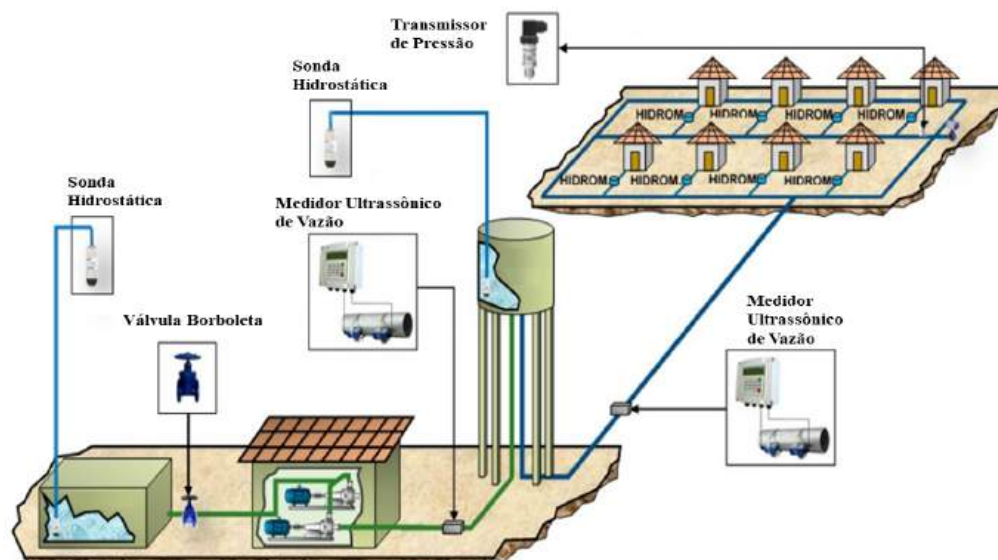


Figura 13 – Sistema de abastecimento.

Na Figura 13, ilustra-se de forma detalhada a instrumentação empregada na estação de distribuição de água, fundamental para a coleta de dados que alimentam o sensor virtual de vazão. Nela, podem ser identificados os seguintes componentes:

Transdutores de Pressão: Estes sensores medem a pressão em pontos estratégicos da rede de abastecimento. Os dados obtidos permitem inferir variações no fluxo e identificar possíveis anomalias que possam afetar a operação do sistema.

Tabela 2 – Especificações dos Transdutores de Pressão

Parâmetro	Descrição
Tipo	Transdutor piezorresistivo
Faixa de medição	0 - 16 bar
Precisão	$\pm 0,25\%$ FS
Saída	4-20 mA
Aplicação	Monitoramento da pressão na rede de distribuição

Macromedidores de Vazão: Utilizados para a medição direta do fluxo de água, estes dispositivos servem como referência para a validação do sensor virtual. Mesmo sendo de custo elevado, sua instalação em pontos-chave possibilita a calibração e comparação dos dados estimados pelo modelo de redes neurais.

Tabela 3 – Especificações dos Macromedidores de Vazão

Parâmetro	Descrição
Tipo	Ultrassônico
Faixa de medição	10 - 1000 m ³ /h
Precisão	±1% da leitura
Saída	Pulso ou 4-20 mA
Aplicação	Medição direta da vazão para calibração do sensor virtual

Sondas Hidrostáticas: Instaladas nos reservatórios, as sondas medem o nível da água, o que, quando correlacionado com os dados de pressão, contribui para uma estimativa mais robusta da vazão.

Tabela 4 – Especificações das Sondas Hidrostáticas

Parâmetro	Descrição
Tipo	Sensor piezorresistivo submersível
Faixa de medição	0 - 10 m de coluna d'água
Precisão	±0,1% FS
Saída	4-20 mA
Aplicação	Monitoramento do nível dos reservatórios

Atuadores (Conjunto Motobomba): Responsáveis pela movimentação da água dentro do sistema, os atuadores garantem que a rede opere dentro dos parâmetros ideais, influenciando diretamente as medições realizadas pelos sensores.

Tabela 5 – Especificações do Conjunto Motobomba

Parâmetro	Descrição
Tipo	Motobomba centrífuga
Potência	30 - 100 CV
Vazão nominal	50 - 500 m ³ /h
Pressão de operação	2 - 10 bar
Aplicação	Movimentação da água na rede de abastecimento

Sistema de Aquisição de Dados e Comunicação: Os sinais provenientes dos transdutores, macromedidores e sondas são coletados por um sistema de aquisição de dados, que realiza o pré-processamento e a correção de eventuais *outliers*. Posteriormente, esses dados são enviados em tempo real para uma plataforma em nuvem por meio do protocolo MQTT, possibilitando a integração com o SCADA e a análise contínua do desempenho do sistema.

Essa instrumentação, cuidadosamente distribuída pela rede, forma a base para o desenvolvimento do sensor virtual, permitindo uma monitorização abrangente e precisa das variáveis críticas. Ao garantir a qualidade dos dados coletados, o sistema de instrumentação

é essencial para o treinamento, validação e eventual implementação do modelo virtual de vazão.

Na Figura 14 é ilustrada a casa de bombas da estação de abastecimento de água, enquanto na Figura 15 é ilustrado de forma mais próxima a localização dos sensores de vazão e pressão, a Figura 16 mostra em detalhe o transdutor ultrassônico de vazão que será utilizado para desenvolver o sensor virtual.



Figura 14 – Casa de bombas.



Figura 15 – Localização dos sensores na casa de bombas.



Figura 16 – Transdutor ultrassônico de medição de vazão.

4.1.1 AQUISIÇÃO DE DADOS

O DataFrame utilizado na análise é composto por cinco variáveis, das quais quatro são variáveis secundárias e a vazão (FIT-DMC-01) é a variável de interesse. Cada uma dessas variáveis possui 14320 amostras. Após consultas e estudos com profissionais da área, optou-se por manter todas as variáveis disponíveis no processo, pois cada uma delas contribui de forma significativa para a compreensão do comportamento do sistema. As variáveis utilizadas foram:

- LIT-REL-01 – Nível do reservatório de distribuição (m);
- PIT-DMC-01 – Pressão da rede de distribuição (mca);
- FIT-DMC-01 – Vazão de saída do reservatório de distribuição (l/s);
- FIT-REL-01 – Vazão de recalque (l/s);
- Pressão recalque – Pressão no recalque (mca).

Os dados originais foram recebidos por colaboradores da CAGEPA em um arquivo CSV e estão ilustrados na Figura 17.

Os dados foram coletados ao longo de 2 meses, com um intervalo de amostragem de 6 minutos. A Figura 18 apresenta o comportamento da vazão durante um dia típico. Observa-se um aumento significativo na demanda de água no período da manhã, seguido por uma redução durante a tarde e um novo pico de consumo por volta das 18 horas.

A arquitetura para comunicação de dados foi baseada no protocolo MQTT (*Message Queuing Telemetry Transport*) cuja configuração foi utilizada como interface intermediária

```

1 Tempo,vazao,pressao,nivel
2 2023-01-01 02:00:27,6.54869022222222,26.271505472222223,2.07240825
3 2023-01-01 02:06:27,6.67920436111111,26.277494222222217,2.042771305555556
4 2023-01-01 02:12:27,6.939879444444446,26.334971277777777,2.015526805555555
5 2023-01-01 02:18:27,6.621289166666668,26.223744,1.992755694444445
6 2023-01-01 02:24:27,6.120632666666665,26.410189777777777,1.9658470833333328
7 2023-01-01 02:30:27,6.862923749999998,25.979011833333333,1.940286361111111
8 2023-01-01 02:36:27,6.396704916666667,26.085559027777778,1.9152345000000004
9 2023-01-01 02:42:27,6.327152527777777,26.368661666666664,1.8897918611111104
10 2023-01-01 02:48:27,6.597680444444445,26.19194813888889,1.86201375
11 2023-01-01 02:54:27,5.956724055555556,26.30490575,1.840677805555555
12 2023-01-01 03:00:27,5.946236083333332,26.295211833333337,1.820743083333333
13 2023-01-01 03:06:27,5.624145972222224,26.365831666666666,1.799114583333333
14 2023-01-01 03:12:27,5.177469166666667,26.383251833333333,1.7770251944444442
15 2023-01-01 03:18:27,5.785042388888889,26.134415111111111,1.750353666666666
16 2023-01-01 03:24:27,5.472062805555556,26.241115027777774,1.731167472222222
17 2023-01-01 03:30:27,4.982986722222221,26.561713194444447,1.709531138888889
18 2023-01-01 03:36:27,5.535073305555556,26.352084611111111,1.691997083333333
19 2023-01-01 03:42:27,5.599440777777778,26.11112372222222,1.676008083333333
20 2023-01-01 03:48:27,5.526220444444444,26.137325777777778,1.6556863611111112

```

Figura 17 – Estrutura da base de dados.



Figura 18 – Vazão de saída do reservatório ao longo de um dia.

entre a CAGEPA (Companhia de Água e Esgotos da Paraíba) e a UFPB (Universidade Federal da Paraíba). O MQTT é um protocolo de comunicação leve e eficiente, ideal para a transmissão de dados em tempo real, com baixa latência e mínimo consumo de largura de banda. O *broker* MQTT é uma entidade centralizada responsável por gerenciar e distribuir as mensagens entre os dispositivos conectados. Os sensores na estação de distribuição de água, instalados pela CAGEPA, publicam dados operacionais, como pressão e nível da água, após a coleta. O *broker* MQTT recebe essas mensagens e as encaminha para os sistemas de assinatura da UFPB, permitindo o recebimento e processamento dos dados pelos pesquisadores que atuam na universidade. A utilização do *broker* MQTT permite a integração entre a infraestrutura de TI (Tecnologia da Informação) da CAGEPA e o sistema de monitoramento da UFPB, assegurando uma comunicação eficiente e segura.

Além disso, garante uma escalabilidade no sistema, permitindo a adição de novos sensores e dispositivos sem impactar a estrutura da rede. Esse modelo de comunicação viabiliza uma gestão em tempo real dos recursos hídricos, contribuindo para a tomada de decisões mais precisas e otimizando o controle e a eficiência do fornecimento de água.

O processo de aquisição de dados do sensor virtual é realizado através de um cliente MQTT baseado na biblioteca *paho-mqtt*, que recebe os dados de pressão e nível dos sensores físicos. O código implementa um sistema de *callback* que processa as mensagens recebidas, verificando se há duplicação de leituras para evitar redundâncias. Após validar os dados, as leituras são encaminhadas para o modelo de sensor virtual, que utiliza redes neurais profundas, incluindo LSTM, MLP e CNN, para estimar a vazão. Os valores resultantes são armazenados em um banco de dados MySQL para posterior análise e enviados de volta ao *broker* MQTT para integração com o sistema SCADA da CAGEPA-PB. A estrutura do código também inclui mecanismos de reconexão automática em caso de falhas de conexão e tratamento de exceções para garantir robustez na transmissão e armazenamento dos dados.

Na Figura 19 ilustra-se, de maneira simplificada, o fluxo de dados entre a CAGEPA e a UFPB por meio de um *broker* MQTT, evidenciando como as informações coletadas na estação de distribuição são transmitidas, processadas e disponibilizadas para análises em tempo real.

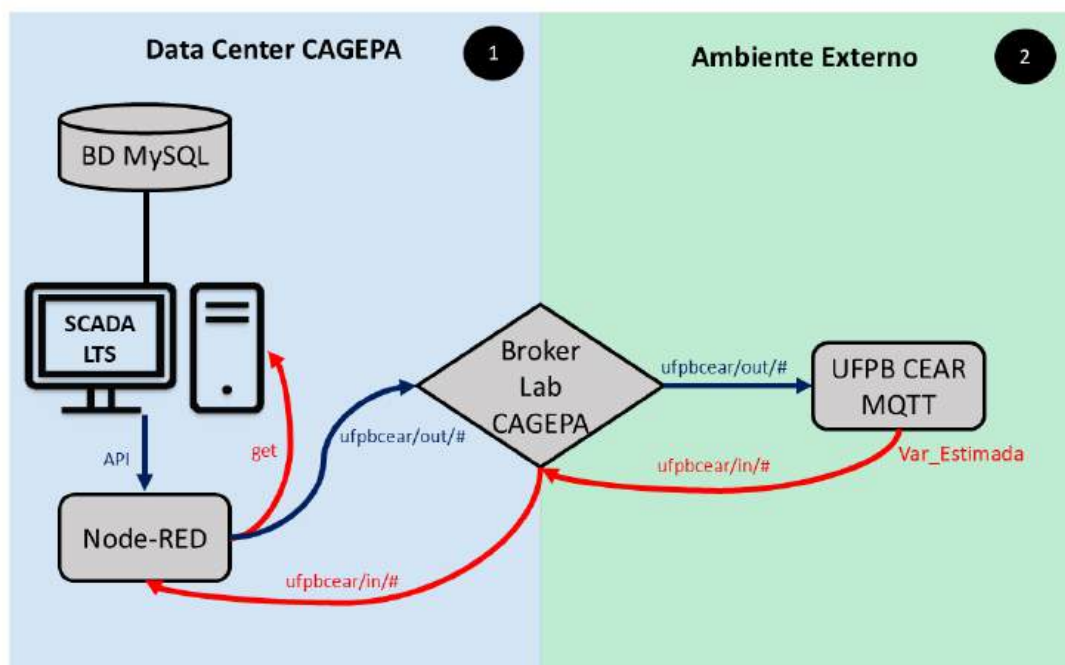


Figura 19 – Topologia de rede proposta.

No lado da CAGEPA (Região 1), um banco de dados MySQL e o sistema SCADA

LTS armazenam e exibem as variáveis operacionais provenientes da estação de distribuição. O Node-RED atua como camada intermediária, obtendo dados via API e publicando-os no *broker* MQTT interno, utilizando tópicos do padrão *ufpbcear/out/#*. O Broker Lab CAGEPA, localizado na infraestrutura da companhia, gerencia as mensagens enviadas e recebidas através do modelo *publish/subscribe*, encaminhando-as para os clientes conectados quando há interesse nos tópicos correspondentes.

Na UFPB (Região 2), o laboratório se conecta ao *broker* e recebe os dados operacionais em tempo real, podendo realizar estimativas (VAR_ESTIMADA) ou análises avançadas de forma remota. Dessa forma, a arquitetura MQTT possibilita a troca de informações de maneira eficiente e escalável, garantindo que as variáveis coletadas sejam compartilhadas de forma segura e em tempo real entre a CAGEPA e a UFPB, atendendo às necessidades de monitoramento, controle e pesquisa no setor de abastecimento de água.

4.2 MÉTODOS

Tomando como base as etapas clássicas de desenvolvimento de sensores virtuais descritas por Fortuna et al. (2007) (Figura 11), o presente trabalho estrutura sua metodologia em um fluxo sequencial que abrange desde a aquisição inicial dos dados até a aplicação prática do modelo em ambiente operacional. Na Figura 20 ilustra-se esse processo, evidenciando as fases de preparação, modelagem e validação que compõem o ciclo de construção do sensor virtual.

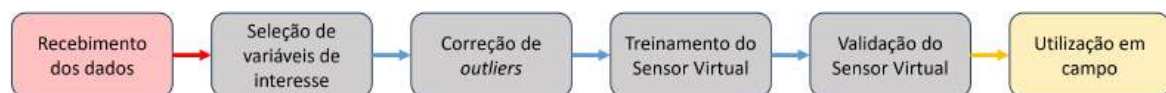


Figura 20 – Etapas de desenvolvimento do Sensor Virtual neste trabalho.

Inicialmente, os dados brutos provenientes do sistema supervisório são recebidos e organizados, conforme foi descrito na seção 4.1.1; em seguida, procede-se à seleção das variáveis de interesse com base em critérios físicos, estatísticos e procedurais. Após essa filtragem, é realizada a correção de *outliers* para assegurar a consistência das séries temporais utilizadas no treinamento, essa etapa será feita através do algoritmo TEDA, conforme se verá adiante. Na etapa subsequente, o modelo baseado em redes neurais LSTM é desenvolvido, ajustado e validado por meio de métricas de desempenho adequadas. Por fim, o sensor virtual resultante é implantado no ambiente de campo, onde sua capacidade de estimar a variável de vazão em tempo real pode ser avaliada em operação contínua,

conforme se verá no capítulo de resultados. A seguir, cada uma dessas etapas intermediárias é detalhada.

4.2.1 SELEÇÃO DE VARIÁVEIS

A seleção das variáveis de entrada do sensor virtual é uma etapa fundamental, pois influencia diretamente a qualidade das previsões do modelo. Para isso, foi realizada uma análise de correlação entre as variáveis disponíveis na base de dados, considerando os coeficientes de Pearson e Spearman.

O coeficiente de correlação de Pearson (ρ) mede a relação linear entre duas variáveis X e Y , sendo definido por:

$$\rho_{X,Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}, \quad (4.1)$$

em que: $\bar{X}$ e $\bar{Y}$ são as médias das variáveis X e Y , respectivamente. A Figura 21 apresenta a matriz de correlação entre as três principais grandezas medidas.

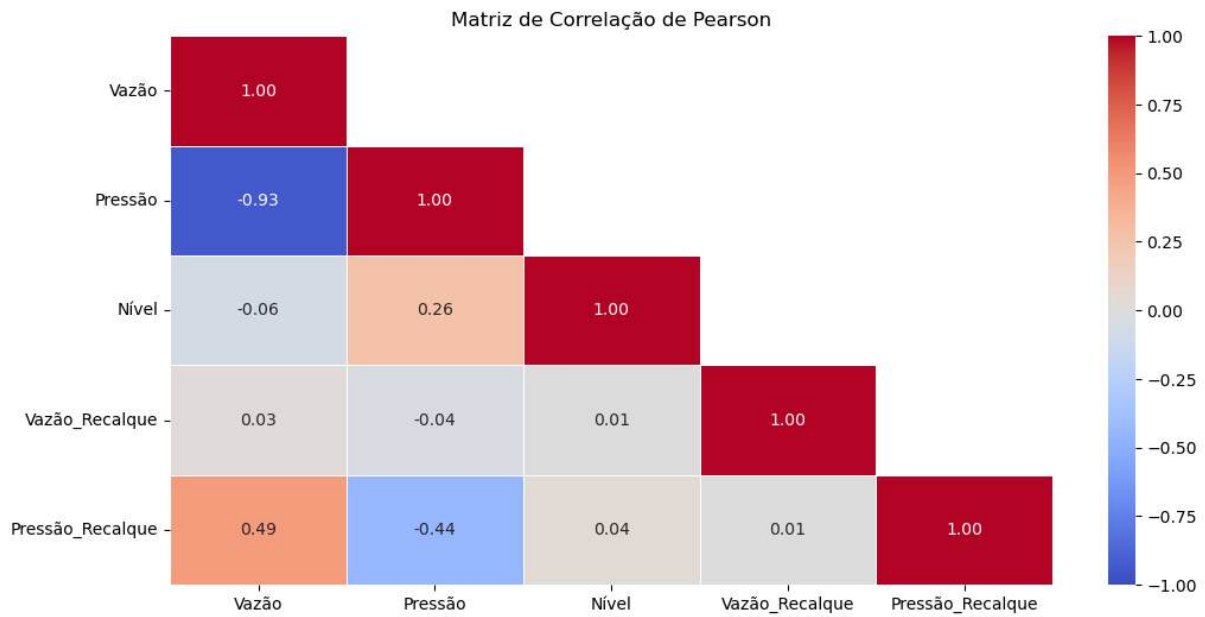


Figura 21 – Matriz de correlação de Pearson entre vazão, pressão e nível.

Como pode ser observado nesta análise, a vazão e a pressão possuem uma correlação linear muito próxima de -1, o que evidencia uma relação inversa entre elas.

Por outro lado, o coeficiente de Spearman (ρ_s) avalia a relação monotônica entre as variáveis, baseado nos postos de cada observação. Ele é dado por:

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (4.2)$$

em que: d_i é a diferença entre os postos das observações e n é o número total de observações. A Figura 22 apresenta a matriz de correlação de Spearman, de forma semelhante à figura anterior.

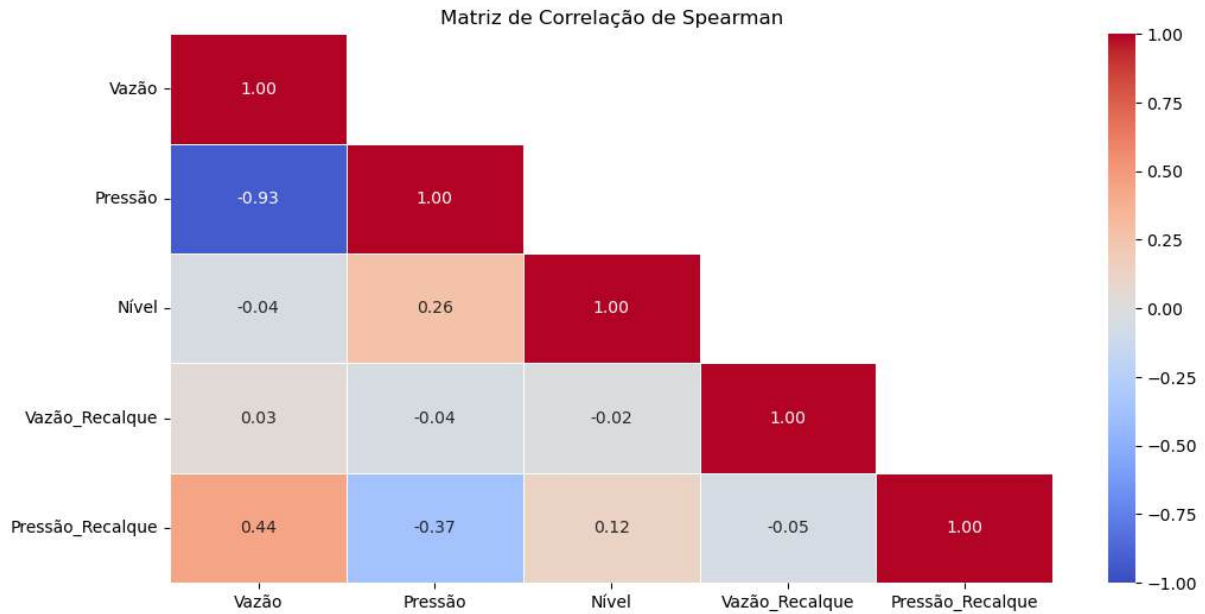


Figura 22 – Matriz de correlação de Spearman entre vazão, pressão e nível.

A análise estatística inicial mostrou que apenas a vazão e a pressão apresentam entre si uma correlação linear elevada, com coeficiente próximo de 1 em módulo. As demais variáveis exibiram correlações lineares próximas de zero, indicando fraca associação sob uma perspectiva estritamente linear. Contudo, deve-se ressaltar que a ausência de correlação linear não implica inexistência de relação funcional: dependências não lineares podem estar presentes e, para identificá-las, são necessárias ferramentas matemáticas mais adequadas, como a correntropia.

Além disso, considerando as condições operacionais dos pontos mais isolados do sistema, nos quais, em muitos casos, apenas o nível é medido, e reconhecendo que essa variável apresenta baixa correlação linear com a vazão e a pressão, decidiu-se, em conjunto com a equipe técnica da CAGEPA, empregar não apenas vazão e pressão na construção do sensor virtual, mas também as demais variáveis disponíveis. Essa escolha visa ampliar a capacidade do modelo de capturar relações potenciais, inclusive não lineares, que possam contribuir para a estimação da vazão.

4.2.2 CORREÇÃO DE OUTLIERS

Nos sistemas de macromedição, é comum surgirem valores anômalos devido a interferências operacionais e limitações dos instrumentos. Esses desvios podem resultar de instabilidades na comunicação, ruídos do ambiente elétrico, surtos momentâneos provocados por manobras hidráulicas ou falhas temporárias dos sensores. Em geral, manifestam-se como picos isolados, sequências de valores artificiais ou alterações abruptas que não condizem com o comportamento físico esperado.

Diante dessas limitações, o método TEDA mostra-se especialmente adequado para o tratamento de anomalias, pois opera de forma incremental e acompanha a evolução natural dos dados sem exigir janelas fixas ou processamento em blocos. Sua capacidade de atualizar métricas estatísticas a cada nova amostra o torna compatível com fluxos contínuos, como os provenientes de sistemas de monitoramento hidráulico. Além disso, a análise baseada em tipicidade e excentricidade permite identificar desvios sutis que métodos tradicionais tendem a ignorar, oferecendo maior robustez ao processo de depuração dos dados antes do treinamento do sensor virtual.

Uma etapa importante no desenvolvimento desta pesquisa consiste no pré-processamento dos dados, que envolve a análise das variáveis quanto à sua relevância e a correção de valores anômalos, conhecidos como *outliers*, presentes na base de dados. Esses *outliers* podem comprometer o desempenho dos algoritmos de aprendizado, levando-os a interpretar valores incorretos como padrões válidos durante a fase de validação, o que afeta diretamente a precisão do modelo.

O algoritmo TEDA, apresentado no Algoritmo 1 a seguir, é projetado para identificar *outliers* em medições contínuas de grandezas como vazão, pressão e nível, utilizadas em sistemas de sensores virtuais. Ele atualiza iterativamente a média e a variância do conjunto de dados, utilizando essas métricas para calcular a excentricidade de cada novo ponto. A excentricidade normalizada é então comparada a um limiar dinâmico, que se ajusta conforme o número de pontos processados, para determinar se o ponto em questão é um *outlier*. O algoritmo pode ser executado em modo *offline*, processando dados em batch, ou em modo *online*, onde os dados são analisados individualmente em tempo real. Essa abordagem permite que o sistema de medição seja capaz de detectar anomalias de forma adaptativa e dinâmica, ajustando-se à evolução dos dados ao longo do tempo.

4.2.3 MODELAGEM DO SENSOR VIRTUAL

O modelo de sensor virtual de vazão ilustrado na Figura 23 tem como objetivo prever o valor futuro da vazão com base em um conjunto de variáveis operacionais relacionadas ao sistema de abastecimento de água. A abordagem adota uma rede neural LSTM, que é particularmente eficaz para lidar com dependências temporais e não lineares em séries

Algorithm 1 TEDA: Identificação de Outlier

```

1: Inicializar  $k \leftarrow 1$ ,  $\mu_k \leftarrow 0$ ,  $\sigma_k^2 \leftarrow 0$ 
2: while  $X$  estiver ativo do
3:   Ler a amostra  $x_k \in X$ 
4:   if  $k = 1$  then
5:      $\mu_k \leftarrow x_k$ 
6:      $\sigma_k^2 \leftarrow 0$ 
7:   else
8:     Atualizar  $\mu_k$  usando a equação 3.12:
       $\mu_k \leftarrow \frac{k-1}{k}\mu_{k-1} + \frac{1}{k}x_k$ 
9:     Atualizar  $\sigma_k^2$  usando a equação 3.13:
       $\sigma_k^2 \leftarrow \frac{k-1}{k}\sigma_{k-1}^2 + \frac{\|x_k - \mu_{k-1}\|^2}{k}$ 
10:    Calcular  $\xi_k$  usando a equação 3.11:
       $\xi_k \leftarrow \frac{1}{k} + \frac{(x_k - \mu_k)^T (x_k - \mu_k)}{k\sigma_k^2}$ 
11:    Calcular  $\zeta_k$  usando a equação 3.17:
       $\zeta_k \leftarrow \frac{\xi_k}{2}$ 
12:    Definir novo limiar para outlier usando a equação 3.16:
       $\tau_k \leftarrow \frac{m^2+1}{2k}$ 
13:    if  $\zeta_k > \tau_k$  then
14:       $outlier \leftarrow verdadeiro$ 
15:    else
16:       $outlier \leftarrow falso$ 
17:    end if
18:  end if
19:  Atualizar  $k \leftarrow k + 1$ 
20: end while

```

históricas. Na sequência, são explicadas as etapas do modelo do sensor virtual:

Variáveis de entrada (Inputs): Foram selecionadas cinco variáveis relevantes para estimar o comportamento da vazão ao longo do tempo, pois refletem tanto as condições de operação quanto a dinâmica de entrada e saída de água no sistema. As variáveis de entrada são: Vazão (saída do reservatório); Vazão de Recalque; Nível do Reservatório; Pressão (na rede de distribuição); e Pressão de Recalque.

Janela de Observação (Window-size): O modelo considera 10 passos de tempo anteriores (equivalentes a 1 hora, caso cada passo corresponda a 6 minutos). Isso significa que, para cada previsão, o modelo analisa o histórico das cinco variáveis ao longo desses 10 passos anteriores, capturando tendências e relações de curto e médio prazo.

Rede LSTM (Long Short-Term Memory): A rede LSTM processa as sequências de dados geradas pela janela de observação. Devido ao seu mecanismo de memória, a LSTM é capaz de reter informações relevantes sobre a evolução das variáveis ao longo do tempo, ignorando o que for irrelevante (porta de esquecimento) e agregando novas informações (porta de entrada).

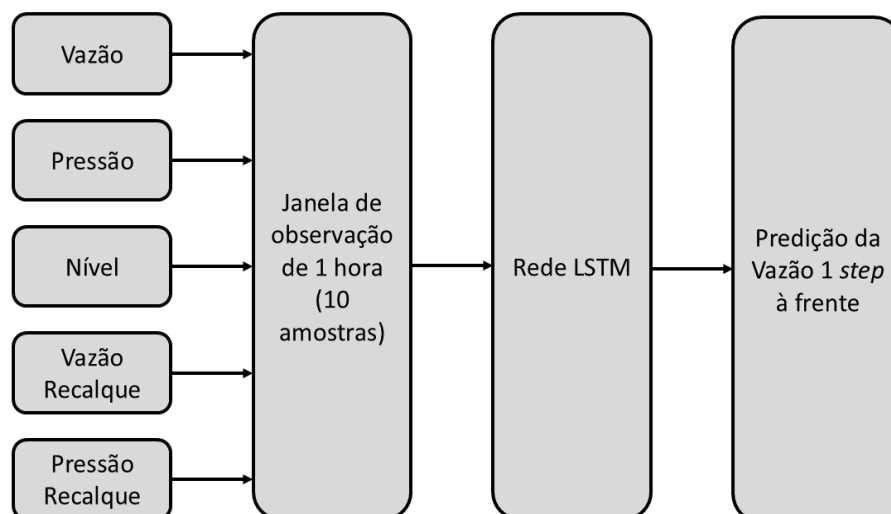


Figura 23 – Modelo do sensor virtual de vazão.

Saída (Previsão de 1 passo à frente): O modelo gera uma estimativa da vazão no próximo intervalo de 6 minutos. Assim, a rede neural LSTM antecipa possíveis variações na vazão, fornecendo informações para o controle e a tomada de decisão na operação do sistema de distribuição de água.

Portanto, o sensor virtual utiliza dados históricos de múltiplas variáveis para estimar o valor futuro da vazão em intervalos regulares. Esse tipo de abordagem é especialmente útil em situações onde a medição direta da vazão pode ser custosa ou inviável, ou quando se deseja um mecanismo de redundância para validar ou complementar sensores físicos já existentes.

Os dados do sensor virtual foram divididos da seguinte forma: 70% para treinamento, 10% para teste e 20% para validação. Estes são valores usualmente utilizados na literatura para a modelagem de redes neurais.

4.2.4 MÉTRICAS PARA VALIDAÇÃO DO SENSOR VIRTUAL

A última etapa corresponde à validação do sensor virtual. Para isso foram utilizadas métricas estatísticas de avaliação do erro.

As métricas de avaliação utilizadas foram o Erro Médio Absoluto (MAE), o Erro Médio Absoluto Percentual, o Erro Médio Quadrático (MSE), a Raiz do Erro Médio Quadrático (RMSE) e o Coeficiente de Determinação (R^2).

O MAE mede a média dos erros absolutos entre os valores reais e os valores preditos. Ele expressa o erro em unidades da variável de interesse, sendo útil para avaliar o desempenho geral do modelo. A equação 4.3 apresenta a fórmula matemática dessa

métrica, onde y_i é o valor real atual da amostra, $\hat{y}_i$ é o valor estimado pelo sensor virtual e n é o número de amostras.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (4.3)$$

O MAPE é similar ao MAE, mas expressa o erro como uma porcentagem dos valores reais, facilitando a interpretação do desempenho do modelo em diferentes escalas.

$$MAPE(\%) = 100 \cdot \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \quad (4.4)$$

O MSE calcula a média dos erros quadráticos, penalizando erros maiores de forma mais intensa do que o MAE.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (4.5)$$

O RMSE é simplesmente a raiz quadrada do MSE. Ele mantém a mesma unidade da variável de interesse e, assim como o MSE, penaliza mais os erros grandes.

$$RMSE = \sqrt{MSE}. \quad (4.6)$$

O Coeficiente de Determinação (R^2), definido como o quadrado do coeficiente de correlação entre os valores reais e estimados. Essa métrica indica quanto da variação da variável de saída é explicada pelas variáveis utilizadas como entrada no modelo, ou seja, o quanto mudanças nas variáveis preditoras produzem mudanças na variável estimada. Valores próximos de 1 indicam uma forte relação entre os valores reais e preditos. A formulação matemática é apresentada na Equação 4.7, em que y_i representa o valor real, $\hat{y}_i$ o valor estimado e $\bar{y}$ a média dos valores reais.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.7)$$

5 RESULTADOS

Neste capítulo, serão apresentados e analisados os resultados alcançados nesta pesquisa, focando na aquisição dos dados utilizando o *broker* MQTT, a correção de dados usando o algoritmo TEDA, e a estimativa da vazão utilizando o sensor virtual baseado na rede neural do tipo LSTM.

5.1 AQUISIÇÃO DE DADOS UTILIZANDO O BROKER MQTT

A etapa inicial deste projeto contou com a disponibilização, pela CAGEPA, do acesso ao seu *broker* MQTT pré-existente, desenvolvido com Node-RED. Esse *broker* é responsável por coletar, a cada minuto, os dados mais recentes armazenados no banco de dados da CAGEPA, garantindo a atualização contínua das informações. Os sensores instalados na Estação de Salgado de São Félix enviam suas medições por meio de um Controlador Lógico Programável (CLP) para o banco de dados da CAGEPA. A partir desse repositório, o Node-RED realiza a leitura automatizada das últimas leituras de cada sensor e encaminha os dados, organizados em tópicos específicos, para o *broker* MQTT. Essa abordagem padronizada e modular facilita a integração dos dados e a sua distribuição para as etapas subsequentes. A Figura 24 ilustra a estrutura idealizada para o fluxo de dados e a integração entre os componentes do sistema.

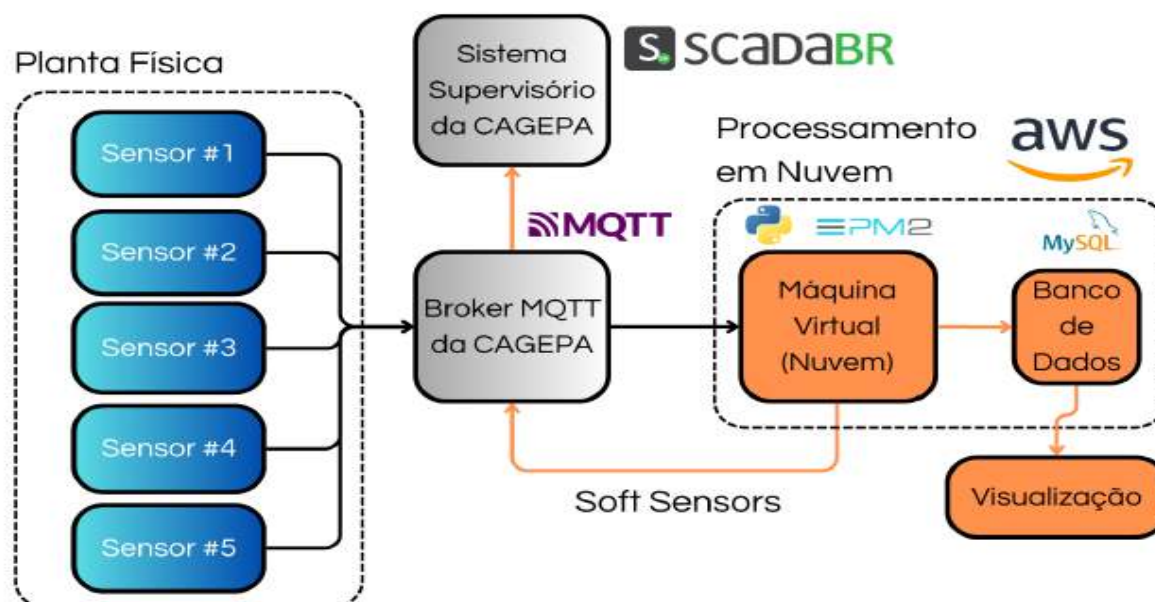


Figura 24 – Fluxo de dados do sistema.

As Figuras de 25 a 27 ilustram, respectivamente, os gráficos de pressão, vazão e

nível, Esses gráficos representam as medições coletadas ao longo de uma semana no mês de janeiro de 2025, proporcionando uma visão detalhada do comportamento das variáveis ao longo do tempo. Os dados exibidos nesses gráficos permitem a análise das flutuações dessas variáveis, facilitando a compreensão das tendências operacionais no período em questão.

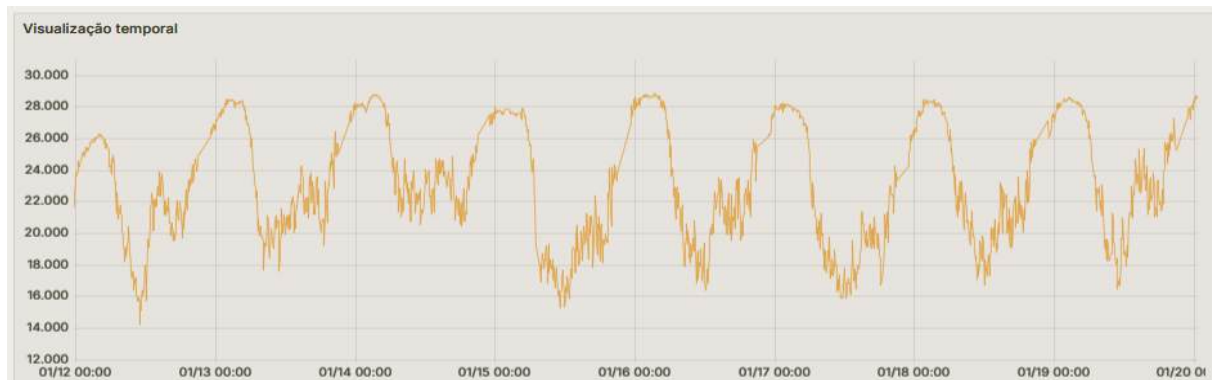


Figura 25 – Visualização da pressão do sistema no Grafana (autoria própria).

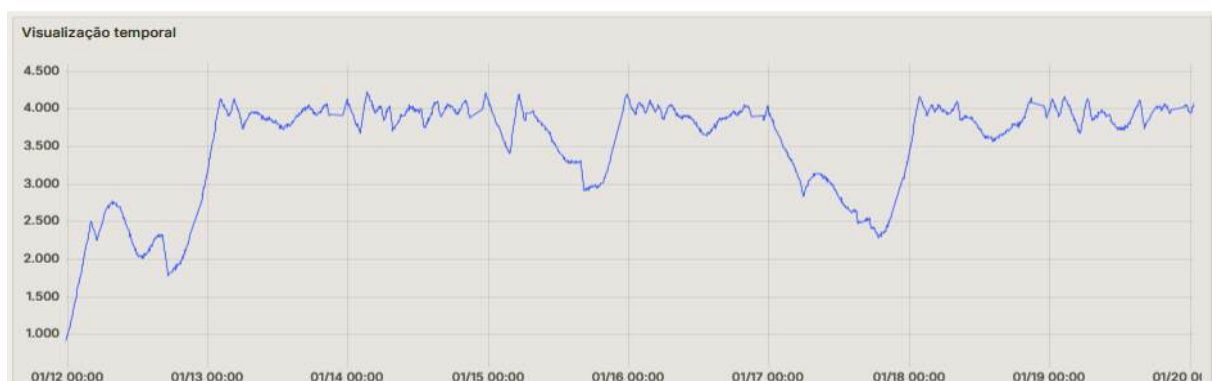


Figura 26 – Visualização do nível do reservatório no Grafana (autoria própria).

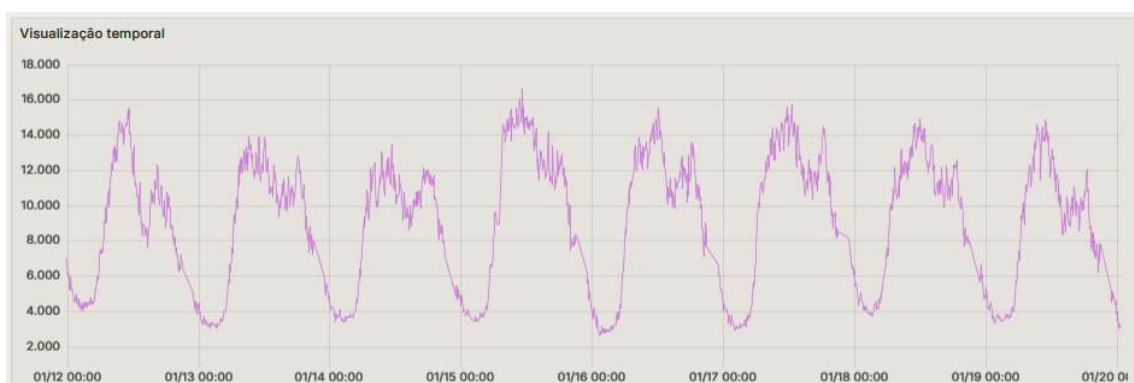


Figura 27 – Visualização da vazão de distribuição no Grafana (autoria própria).

A dinâmica conjunta entre nível, pressão e vazão revela um comportamento cíclico característico do funcionamento hidráulico da estação de distribuição. Observa-se que, nos períodos de menor consumo, tipicamente durante a madrugada, a vazão demandada pela rede diminui, permitindo que o reservatório se recupere e o nível se eleve gradualmente. Esse aumento do nível, associado à menor perda de carga no sistema, resulta em pressões mais altas na linha, evidenciadas pelos picos observados no período. Em contraste, durante os horários de maior demanda, o reservatório entra em fase de esvaziamento, levando à redução do nível e ao consequente alívio da pressão na tubulação, fenômeno corroborado pelas quedas simultâneas nas curvas de pressão e nível. Esse ciclo diário de enchimento e esvaziamento descreve o modo normal de funcionamento da estação de distribuição e explica como nível, pressão e vazão variam ao longo do dia. Essas variáveis não se relacionam de forma simples ou proporcional, mas apresentam dependências não lineares que se manifestam no tempo. Compreender esse comportamento é essencial para representar corretamente o sistema e permite que o sensor virtual identifique e reproduza esses padrões temporais de forma adequada.

5.2 CORREÇÃO DE DADOS USANDO O ALGORITMO TEDA

Na Figura 28 são ilustrados os *outliers* presentes nos dados das grandezas hidráulicas adquiridas, como pressão, vazão e nível dos reservatórios. Esses valores atípicos caracterizam-se por se afastarem significativamente do comportamento esperado do sistema em condições normais de operação. Tais ocorrências podem estar associadas a falhas nos sensores, erros de medição, mudanças abruptas nas condições operacionais ou ainda a fatores externos imprevistos, como intervenções na rede de distribuição ou eventos climáticos extremos.

Quando não corrigidos, os *outliers* podem comprometer a qualidade dos dados, afetando a precisão das análises e modelos preditivos, como será o caso do sensor virtual de vazão que utiliza redes neurais. A detecção e correção eficaz de *outliers* é fundamental para garantir a confiabilidade dos dados utilizados no treinamento e validação de modelos, como os de sensores virtuais. Entre as técnicas de correção, o algoritmo TEDA se destaca por sua eficácia na identificação e remoção ou ajuste de valores atípicos. O TEDA se baseia na análise da tipicidade e excêntricidade dos dados em relação ao comportamento global da série temporal. Ele é capaz de distinguir os dados normais dos *outliers*, considerando os padrões de variabilidade típicos do sistema e as anomalias que se afastam desses padrões.

No contexto das grandezas hidráulicas, o TEDA adapta os dados para que as flutuações inesperadas, causadas por dados errôneos ou imprecisos, sejam corrigidas adequadamente, permitindo que as análises do sensor virtual e os sistemas de controle operacional mantenham um nível de precisão elevado. Uma vez aplicados esses algoritmos, as séries temporais de dados tornam-se mais consistentes, melhor representando o compor-

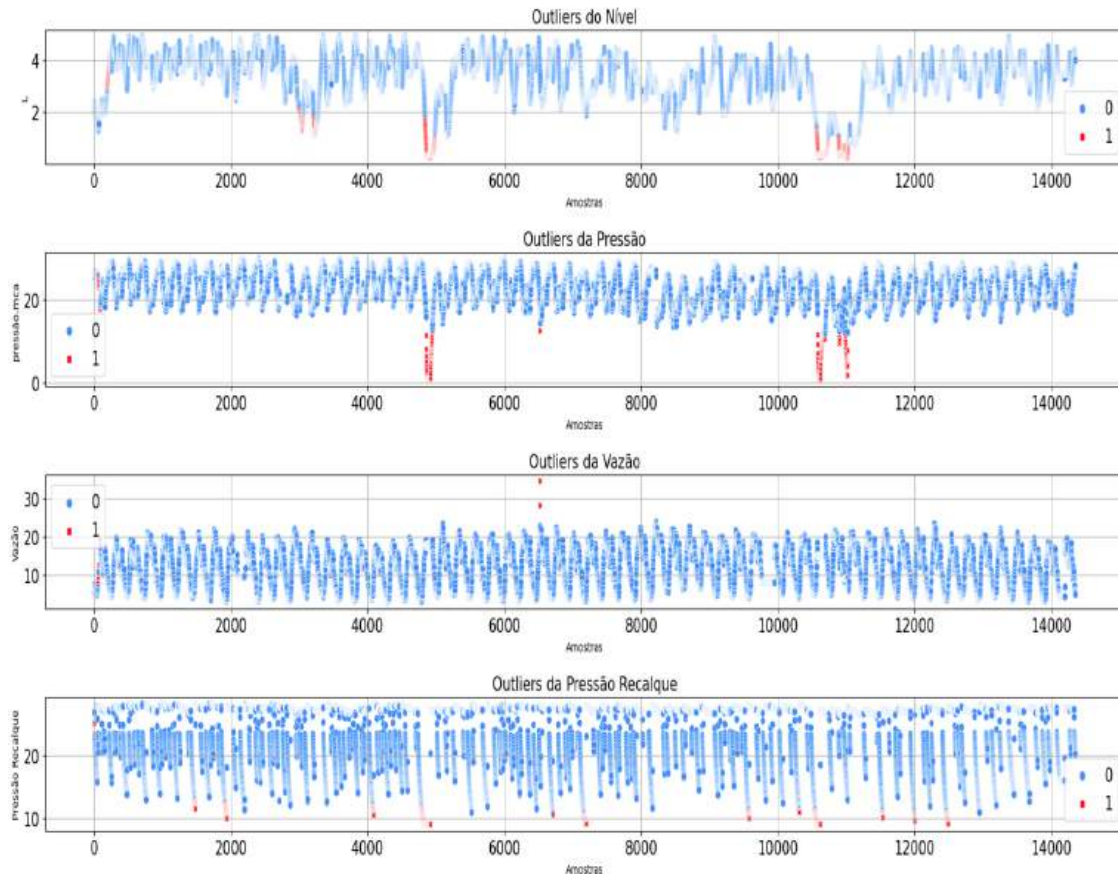


Figura 28 – Outlier nos dados dos sensores.

tamento real do sistema, o que, por sua vez, contribui para melhorias nos diagnósticos e decisões operacionais, além de otimizar a eficiência do modelo de *soft sensor*. Portanto, o uso de TEDA para a correção de *outliers* torna-se uma ferramenta computacional efetiva para a obtenção de dados confiáveis que possam ser usados com precisão tanto para o treinamento de redes neurais quanto para a análise e operação em tempo real dos sistemas hidráulicos.

Na Figura 28 é apresentada a ocorrência de pontos discrepantes (*outliers*) nas séries temporais das variáveis hidráulicas analisadas. Esses *outliers* podem surgir por diversos motivos, como falhas temporárias nos sensores, interferências no processo de transmissão ou condições operacionais atípicas. Quando não tratados, podem comprometer a qualidade dos dados e, conseqüentemente, a eficácia de modelos de previsão ou controle que se baseiem nesses valores.

Portanto, o TEDA detecta anomalias ao calcular continuamente duas métricas: a excentricidade, que avalia o quão distante cada ponto está em relação ao padrão predominante da amostra, e a tipicidade, que mede o quanto esse ponto se assemelha ao comportamento esperado. Sempre que a excentricidade ultrapassa um limiar definido,

o algoritmo identifica o ponto como *outlier*, sinalizando que ele se desvia de maneira significativa do restante dos dados. Para a detecção dos *outliers*, foi utilizada a seguinte equação para a excentricidade (com limiar de $m=2,5$):

$$\xi_k(x) > \frac{m^2 + 1}{2k}. \quad (5.1)$$

Uma vez detectados, os *outliers* podem ser tratados de diferentes formas, como a remoção pontual dos valores anômalos ou a substituição por estimativas plausíveis baseadas em dados históricos ou modelos de previsão. Dessa forma, o algoritmo TEDA contribui para melhorar a consistência dos dados, tornando-os mais representativos das condições reais de operação. A eliminação ou correção desses valores discrepantes garante maior confiabilidade no treinamento de redes neurais e na aplicação de técnicas de controle ou monitoramento avançadas, elevando a robustez de todo o processo de gestão dos recursos hídricos. Neste trabalho, para a correção dos *outliers* foi realizado o procedimento de substituição pela média geral naquela hora.

5.3 SENSOR VIRTUAL DE VAZÃO USANDO A REDE NEURAL LSTM

Com a finalidade de implementar o modelo de sensor virtual mais adequado, foi feita uma comparação entre as diferentes estruturas de redes neurais artificiais. A Tabela 6 apresenta as métricas de avaliação para as arquiteturas MLP, CNN e LSTM, com as devidas alterações na quantidade de neurônios por camada, conforme descrito na seção 4.2.3.

A partir da Tabela 6, observa-se que a rede LSTM apresentou o melhor desempenho entre as arquiteturas analisadas. O modelo 12, com 20 neurônios, alcançou um MAPE de 3,67% e um RMSE de 0,35, estabelecendo-se como a configuração mais precisa na estimativa da vazão.

O segundo melhor resultado foi obtido pela rede MLP, cujo modelo 5 com 50 neurônios atingiu um MAPE de 4,09% e um RMSE de 0,47. Embora a MLP seja estruturalmente mais simples que as outras arquiteturas testadas, seu desempenho próximo ao da LSTM indica que, com um número adequado de neurônios, ela pode oferecer uma boa capacidade preditiva.

O segundo melhor desempenho foi obtido pela rede MLP, cujo modelo 5, com 50 neurônios, apresentou um MAPE de 4,09% e um RMSE de 0,47. Embora a MLP possua uma arquitetura estruturalmente mais simples quando comparada às redes recorrentes, como a LSTM, esse resultado indica que, com um número adequado de neurônios na

Tabela 6 – Métricas de avaliação para diferentes modelos e configurações de neurônios.

Nº	Modelo	Neurônios	MAE	MAPE	MSE	RMSE	R ²
1	MLP	10	0.58	7.82%	0.45	0.67	0.78
2	MLP	20	0.52	6.91%	0.39	0.62	0.81
3	MLP	30	0.48	5.74%	0.34	0.58	0.83
4	MLP	40	0.41	4.98%	0.28	0.53	0.86
5	MLP	50	0.36	4.09%	0.22	0.47	0.89
6	CNN	10	0.56	7.35%	0.42	0.65	0.79
7	CNN	20	0.50	6.57%	0.37	0.61	0.82
8	CNN	30	0.45	5.89%	0.31	0.56	0.85
9	CNN	40	0.40	5.23%	0.26	0.51	0.88
10	CNN	50	0.35	4.61%	0.21	0.46	0.90
11	LSTM	10	0.50	6.74%	0.35	0.59	0.84
12	LSTM	20	0.29	3.67%	0.12	0.35	0.95
13	LSTM	30	0.32	4.21%	0.14	0.37	0.93
14	LSTM	40	0.38	4.84%	0.19	0.44	0.89
15	LSTM	50	0.43	5.46%	0.24	0.49	0.87

camada oculta, a MLP é capaz de aproximar satisfatoriamente as relações não lineares entre as variáveis hidráulicas do sistema. Tal comportamento sugere que, apesar de não explorar explicitamente dependências temporais, a MLP consegue capturar padrões médios e relações estáticas relevantes presentes nos dados, o que justifica seu desempenho competitivo em relação a arquiteturas mais complexas.

Já a arquitetura CNN apresentou resultados intermediários em relação ao MLP e ao LSTM. Seu melhor desempenho foi registrado no modelo 10 com 50 neurônios, com um MAPE de 4,61% e um RMSE de 0,46, valores que, embora inferiores ao LSTM, ainda demonstram boa capacidade de modelagem do problema.

Com base nesses testes, cada uma das três arquiteturas foi implementada na máquina virtual da AWS, permitindo que o sensor virtual continue recebendo dados em tempo real do sistema de abastecimento. Assim, o modelo continua gerando previsões contínuas da vazão com base nos valores de pressão e nível do reservatório, garantindo a atualização constante da estimativa.

5.3.1 VARIAÇÃO NA QUANTIDADE DE CAMADAS E ANÁLISE DE INCERTEZAS

Após a definição da arquitetura base do sensor virtual, investigou-se a influência da profundidade da rede na capacidade de modelagem temporal. Para isso, foram realizados experimentos considerando exclusivamente arquiteturas LSTM compostas por duas camadas recorrentes, variando-se o número de neurônios em cada camada (10, 20, 30, 40 e 50 unidades). O objetivo foi verificar se o acréscimo de uma segunda camada seria capaz

de capturar dinâmicas mais complexas do sistema hidráulico, proporcionando melhorias mensuráveis nas métricas de desempenho.

A Tabela 7 ilustra os resultados obtidos, demonstrando que a adição da segunda camada LSTM produziu apenas ganhos marginais, sem impacto significativo na precisão das estimativas. Em todas as configurações testadas, o desempenho permaneceu muito próximo ao observado nas arquiteturas com apenas uma camada. Mesmo nos casos em que houve pequena melhora em algum indicador isolado, o comportamento geral do modelo não justificou o aumento de complexidade — especialmente considerando o maior custo computacional associado à utilização de múltiplas camadas recorrentes.

Tabela 7 – Métricas de avaliação para a arquitetura LSTM com duas camadas.

Modelo	Neurônios	MSE	RMSE	MAE	MAPE	R ²
16	10	0.1860	0.4313	0.3577	4.810%	0.912
17	20	0.2018	0.4492	0.3831	3.973%	0.905
18	30	0.1431	0.3782	0.3187	4.087%	0.932
19	40	0.2444	0.4944	0.4168	5.772%	0.887
20	50	0.1382	0.3718	0.3125	4.084%	0.934

Essa tendência ficou evidente ao comparar os resultados das arquiteturas de duas camadas com aqueles obtidos pela melhor configuração de camada única (20 neurônios). Ainda que a inclusão de mais unidades recorrentes pudesse potencialmente ampliar a capacidade de representação do modelo, observou-se que, para os dados e o problema em questão, a estrutura mais profunda não trouxe benefícios práticos. Dessa forma, confirmou-se que a arquitetura com uma única camada LSTM constitui a solução mais eficiente e equilibrada, apresentando melhor desempenho global entre todas as métricas analisadas.

Com base nessas evidências, estabeleceu-se que as etapas subsequentes, incluindo os experimentos envolvendo o algoritmo Monte Carlo Dropout, seriam conduzidas utilizando exclusivamente a configuração de uma camada LSTM com 20 neurônios, uma vez que ela se mostrou a opção mais apropriada tanto em termos de desempenho quanto de simplicidade estrutural.

5.3.2 VARIAÇÃO DOS PARÂMETROS DO MONTE CARLO DROPOUT

Após a definição da arquitetura da rede, composta por uma única camada LSTM com 20 neurônios, procedeu-se à análise da incerteza do modelo por meio da técnica de Monte Carlo Dropout. Nessa etapa, variaram-se sistematicamente dois parâmetros fundamentais do método: o número de amostras de predições e o peso de regularização L2, enquanto os demais parâmetros foram mantidos constantes. O objetivo central foi identificar as combinações capazes de reduzir simultaneamente os erros de predição e a

incerteza associada ao modelo, permitindo avaliar a robustez do sensor virtual diante das variações naturais dos dados medidos.

A Tabela 8 reúne os resultados obtidos para todos os cenários testados, apresentando métricas como MSE, RMSE, MAE, MAPE, o valor estimado de incerteza e o coeficiente de determinação.

Tabela 8 – Resultados da variação dos parâmetros do Monte Carlo Dropout.

Modelo	DP	MSE	RMSE	MAE	MAPE	R ²	Incerteza	λ
21	5.0%	0.1905	0.4365	0.3758	4.6533%	0.78	2.0603	4.733×10^{-7}
22	5.0%	0.0974	0.3121	0.2450	2.5368%	0.89	1.8290	4.733×10^{-8}
23	5.0%	0.2896	0.5382	0.4786	5.3481%	0.72	1.8070	9.466×10^{-9}
24	5.0%	0.1797	0.4239	0.3592	4.7209%	0.80	1.8041	4.734×10^{-9}
25	5.0%	0.2548	0.5048	0.4403	5.1836%	0.74	1.8013	4.733×10^{-11}
26	10.0%	0.2847	0.5336	0.4528	6.1147%	0.73	2.0610	4.484×10^{-7}
27	10.0%	0.1794	0.4236	0.3522	4.2931%	0.81	1.8298	4.484×10^{-8}
28	10.0%	0.1880	0.4336	0.3518	4.8005%	0.80	1.8077	8.968×10^{-9}
29	10.0%	0.1855	0.4307	0.3610	4.5060%	0.81	1.8051	4.484×10^{-9}
30	10.0%	0.1800	0.4243	0.3487	4.6788%	0.82	1.8021	4.484×10^{-11}
31	15.0%	0.2288	0.4783	0.3901	5.2112%	0.77	2.0620	4.235×10^{-7}
32	15.0%	0.2975	0.5454	0.4759	5.8930%	0.70	1.8307	4.235×10^{-8}
33	15.0%	0.4890	0.6993	0.5743	8.5758%	0.28	1.8085	8.470×10^{-9}
34	15.0%	0.1366	0.3697	0.2918	3.4911%	0.88	1.8060	4.235×10^{-9}
35	15.0%	0.1437	0.3790	0.3062	3.5855%	0.87	1.8034	4.235×10^{-11}

De maneira geral, observa-se que a incerteza tende a permanecer próxima de 1,8 L/s, independentemente do conjunto de parâmetros, o que indica alta estabilidade do modelo sob o uso do Monte Carlo Dropout. Entre as variações testadas, destaca-se a configuração correspondente ao modelo de número 22, que apresentou simultaneamente os menores valores de MAPE e RMSE, evidenciando o melhor equilíbrio entre desempenho preditivo e confiabilidade estatística. Esse comportamento confirma a eficiência da rede LSTM em estimar a vazão mesmo sob cenários de perturbação controlada, mantendo baixa dispersão entre as predições e uma forte aderência aos valores reais.

5.4 INTERFACE USANDO A PLATAFORMA EM NUVEM

A fim de disponibilizar de forma clara e acessível as medições do sistema de abastecimento e as estimativas geradas pelo sensor virtual, foi desenvolvida uma interface visual na plataforma Grafana. Essa interface permite o monitoramento contínuo dos valores obtidos pelos sensores físicos, bem como das previsões geradas por cada modelo de rede neural implementado.

A Figura 29 ilustra o painel desenvolvido para atuar como IHM (Interface Homem-Máquina) do sistema, recebendo dados a cada 6 minutos via *broker* MQTT. Todo o

ambiente de monitoramento está hospedado na plataforma em nuvem da AWS (Amazon Web Services), garantindo alta disponibilidade, escalabilidade e segurança no gerenciamento das informações.



Figura 29 – Interface IHM em Grafana - Plataforma em nuvem AWS.

O uso do protocolo MQTT possibilita uma comunicação eficiente e leve entre os dispositivos de campo e a camada de supervisão, enquanto o Grafana oferece recursos avançados para visualização e análise dos dados. A interface inclui gráficos temporais, indicadores numéricos e painéis personalizados, facilitando a interpretação dos resultados e apoiando a tomada de decisões na operação e otimização do sistema de distribuição de água.

A Figura 30 apresenta um gráfico detalhado da vazão estimada pelo modelo LSTM, destacando seu desempenho ao longo do tempo. A escolha dessa visualização se justifica pelo fato de a LSTM ter obtido os melhores resultados nas métricas de erro, conforme discutido anteriormente. A curva gerada demonstra a capacidade do modelo de capturar padrões dinâmicos na vazão, fornecendo estimativas alinhadas às variações observadas nos sensores físicos.

Essa abordagem baseada em *dashboards* melhora significativamente a usabilidade do sistema, permitindo que operadores e engenheiros monitorem o desempenho do sensor virtual em tempo real e identifiquem eventuais discrepâncias entre os modelos. Além disso, a integração dos dados na nuvem assegura que as informações estejam sempre atualizadas, contribuindo para uma gestão mais eficiente do sistema de abastecimento.

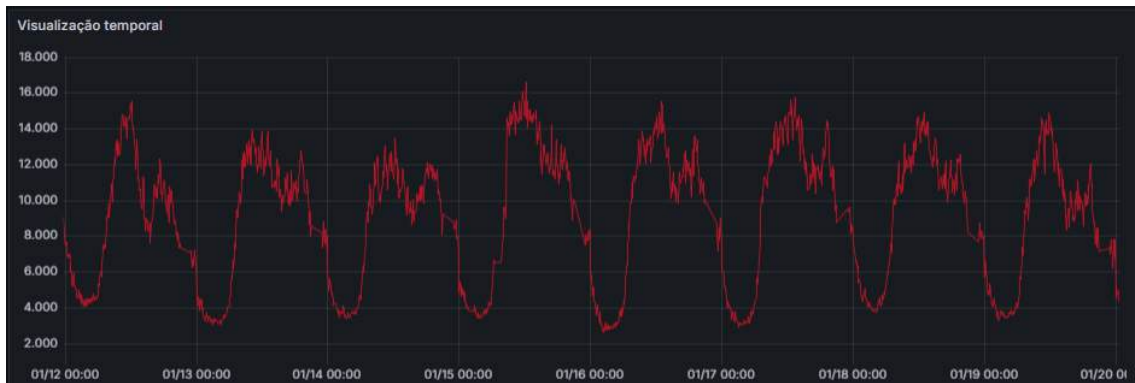


Figura 30 – Visualização da vazão estimada pelo LSTM.

A plataforma também permite a sobreposição de gráficos, facilitando a comparação entre os valores medidos e estimados. A Figura 31 exibe a sobreposição dos gráficos das Figuras 27 e 30, correspondentes, respectivamente, à medição do sensor físico (curva verde) e à estimativa do modelo LSTM (curva vermelha).



Figura 31 – Visualização da vazão real (verde) e estimada (vermelho).

Essa visualização evidencia a semelhança entre os valores obtidos, destacando a capacidade do sensor virtual de aprender e reproduzir o comportamento dinâmico do sistema. Vale ressaltar que os dados apresentados são inéditos, ou seja, resultam da generalização do modelo a novas condições operacionais, reforçando sua eficácia na previsão da vazão.

6 CONCLUSÃO

O presente trabalho demonstrou a viabilidade do desenvolvimento de um sensor virtual de vazão baseado em redes neurais do tipo LSTM, aplicado à macromedição em sistemas de abastecimento de água. A estratégia de medição indireta adotada mostrou-se eficiente ao permitir a estimação da vazão a partir de variáveis hidrodinâmicas como pressão e nível, reduzindo a dependência de sensores físicos adicionais e trazendo economia operacional para a companhia de saneamento.

A integração da solução com o broker MQTT possibilitou a transmissão contínua e confiável das medições realizadas na estação de distribuição da CAGEPA, garantindo o fluxo de dados em tempo real entre o ambiente de campo e a infraestrutura de nuvem na AWS. Essa arquitetura permitiu o monitoramento direto via Grafana, bem como a inserção do *soft sensor* em um contexto operacional realista.

A aplicação do algoritmo TEDA para detecção e correção de *outliers* configurou-se como um recurso fundamental para a melhoria da qualidade dos dados utilizados no treinamento e validação do modelo. Sua capacidade de adaptação e baixo custo computacional contribuiu significativamente para a robustez do sistema como um todo. Além disso, a adoção do método Monte Carlo Dropout permitiu estimar a incerteza associada às previsões do sensor virtual. Por meio da geração de múltiplas inferências estocásticas, foi possível quantificar a variabilidade do modelo, oferecendo uma medida adicional de confiabilidade das estimativas e ampliando a interpretabilidade do *soft sensor* em um ambiente operacional real.

Os experimentos realizados evidenciaram que a arquitetura LSTM proposta alcançou erros percentuais médios inferiores a 4%, indicando que o sensor virtual apresenta desempenho compatível com aplicações reais de apoio à tomada de decisão em sistemas de abastecimento.

Do ponto de vista computacional, a adoção de uma infraestrutura em nuvem permitiu flexibilidade, escalabilidade e potencial de replicação do método para cenários distintos, abrindo perspectivas para sua utilização ampliada em operações de saneamento. A integração de soft sensors a sistemas SCADA representa um avanço promissor para o setor, favorecendo o controle operacional, a redução de perdas e a gestão inteligente dos recursos hídricos.

REFERÊNCIAS

- AGGARWAL, C. C. An introduction to outlier analysis. In: _____. *Outlier Analysis*. Cham: Springer International Publishing, 2017. p. 1–34. ISBN 978-3-319-47578-3. Disponível em: <https://doi.org/10.1007/978-3-319-47578-3_1>. Citado 2 vezes nas páginas 21 e 34.
- ALENCAR, G. M. R. d. et al. A soft sensor for flow estimation and uncertainty analysis based on artificial intelligence: A case study of water supply systems. *Automation*, MDPI, v. 5, n. 2, p. 106–127, 2024. Citado 4 vezes nas páginas 16, 21, 22 e 28.
- ANDRADE, P. et al. Teda-rls: A tinyml incremental learning approach for outlier detection and correction. *IEEE Sensors Journal*, v. 24, n. 22, p. 38165–38173, 2024. Citado 3 vezes nas páginas 21, 36 e 37.
- ANGELOV, P. Anomaly detection based on eccentricity analysis. In: *2014 IEEE Symposium on Evolving and Autonomous Learning Systems (EALS)*. [S.l.: s.n.], 2014. p. 1–8. Citado na página 36.
- ANGELOV, P. Outside the box: an alternative data analytics framework. *Journal of Automation Mobile Robotics and Intelligent Systems*, Sieć Badawcza Łukasiewicz-Przemysłowy Instytut Automatyki i Pomiarów, v. 8, n. 2, p. 29–35, 2014. Citado na página 35.
- BANSAL, R.; GAUR, N.; SINGH, S. N. Outlier detection: Applications and techniques in data mining. In: *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*. [S.l.: s.n.], 2016. p. 373–377. Citado na página 34.
- DING, Y. et al. A cnn-lstm model for electric load forecasting. In: *2024 International Conference on Networking, Sensing and Control (ICNSC)*. [S.l.: s.n.], 2024. p. 1–6. Citado 2 vezes nas páginas 7 e 28.
- FLORES, T. K. S. et al. Indirect feedback measurement of flow in a water pumping network employing artificial intelligence. *Sensors*, MDPI, v. 21, n. 1, p. 75, 2020. Citado 5 vezes nas páginas 13, 16, 20, 21 e 25.
- FORTUNA, L. et al. *Soft sensors for monitoring and control of industrial processes*. [S.l.]: Springer, 2007. v. 22. Citado 3 vezes nas páginas 7, 33 e 47.
- GAL, Y.; GHAHRAMANI, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: PMLR. *international conference on machine learning*. [S.l.], 2016. p. 1050–1059. Citado na página 37.
- GAL, Y.; GHAHRAMANI, Z. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, v. 29, 2016. Citado na página 38.
- GAL, Y. et al. Uncertainty in deep learning. phd thesis, University of Cambridge, 2016. Citado na página 37.

- GOMES, H. P. *Sistemas de bombeamento*. [S.l.]: João Pessoa, Editora Universitária UFPB, 2009. Citado 3 vezes nas páginas 13, 23 e 24.
- GOMES, H. P. *Abastecimento de água*. [S.l.: s.n.], 2021. 2º ed., ISBN 978-65900599-0-0. Citado na página 23.
- HAWKINS, D. *Identification of outliers*. [S.l.]: Chapman and Hall, 1980. Citado na página 34.
- HAYKIN, S. *Redes neurais: princípios e prática*. [S.l.]: Bookman Editora, 2001. Citado na página 26.
- HAYKIN, S. S. et al. *Neural networks and learning machines/Simon Haykin*. [S.l.]: New York: Prentice Hall, 2009. Citado 2 vezes nas páginas 26 e 32.
- HE, Y. et al. An improved industrial process soft sensor method based on lstm. In: IEEE. *2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS)*. [S.l.], 2023. p. 1750–1755. Citado 2 vezes nas páginas 30 e 31.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 1997. Citado 2 vezes nas páginas 20 e 30.
- ITB. *Instituto Trata Brasil - Ranking do Saneamento*. 2025. Acessado em 02/09/2025. Disponível em: <<https://tratabrasil.org.br/ranking-do-saneamento-2025>>. Citado 2 vezes nas páginas 15 e 16.
- JIANG, Y. et al. A review on soft sensors for monitoring, control, and optimization of industrial processes. *IEEE Sensors Journal*, v. 21, n. 11, p. 12868–12881, 2021. Citado na página 33.
- KADLEC, P.; GABRYS, B.; STRANDT, S. Data-driven soft sensors in the process industry. *Computers & chemical engineering*, Elsevier, v. 33, n. 4, p. 795–814, 2009. Citado 2 vezes nas páginas 33 e 34.
- KUBAT, M. Neural networks: a comprehensive foundation by simon haykin, macmillan, 1994, isbn 0-02-352781-7. *The Knowledge Engineering Review*, v. 13, n. 4, p. 409–412, 1999. Citado 2 vezes nas páginas 7 e 26.
- LAZCANO, A.; JARAMILLO-MORÁN, M. A.; SANDUBETE, J. E. Back to basics: The power of the multilayer perceptron in financial time series forecasting. *Mathematics*, MDPI, v. 12, n. 12, p. 1920, 2024. Citado 2 vezes nas páginas 7 e 27.
- LIMA, J. dos S.; VILLANUEVA, J. M. M.; CATUNDA, S. Y. C. Modeling a virtual flow sensor in a sugar-energy plant using artificial neural network. In: *2022 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. [S.l.: s.n.], 2022. p. 1–6. Citado 2 vezes nas páginas 21 e 33.
- LIMA, R. P. G. et al. Desenvolvimento de um soft sensor para estimação da vazão em sistemas de abastecimento de água utilizando redes neurais artificiais. Universidade Federal da Paraíba, 2022. Citado 3 vezes nas páginas 16, 21 e 25.
- LIMA, R. P. G. et al. Development of a soft sensor for flow estimation in water supply systems using artificial neural networks. *Sensors*, MDPI, v. 22, n. 8, p. 3084, 2022. Citado 3 vezes nas páginas 20, 24 e 33.

- LOTUFO, F. A.; GARCIA, C. Sensores virtuais ou soft sensors: Uma introdução. In: *7th Brazilian Conference on Dynamics, Control and Applications, Presidente Prudente*. [S.l.: s.n.], 2008. Citado 2 vezes nas páginas 32 e 33.
- LUTTMANN, R. et al. *Soft sensors in bioprocessing: a status report and recommendations*. [S.l.]: Wiley Online Library, 2012. 1040–1048 p. Citado na página 32.
- MIGLIATO, A. L. T. *Detecção de Outliers em Dados não Vistos de Séries Temporais por meio de Erros de Predição com SARIMA e Redes Neurais Recorrentes LSTM e GRU*. Tese (Doutorado) — Universidade de São Paulo, 2021. Citado na página 29.
- NASSER, A. A.; RASHAD, M. Z.; HUSSEIN, S. E. A two-layer water demand prediction system in urban areas based on micro-services and lstm neural networks. *IEEE Access*, IEEE, v. 8, p. 147647–147661, 2020. Citado 3 vezes nas páginas 7, 30 e 31.
- PALMITESSA, R. et al. Soft sensing of water depth in combined sewers using lstm neural networks with missing observations. *Journal of Hydro-Environment Research*, Elsevier, v. 38, p. 106–116, 2021. Citado na página 27.
- QUEVEDO, J. et al. Combining learning in model space fault diagnosis with data validation/reconstruction: Application to the barcelona water network. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 30, p. 18–29, 2014. Citado na página 24.
- SALVINO, L. R. et al. Eficiência energética em sistemas pressurizados de distribuição de água com o uso de redes neurais artificiais. Universidade Federal da Paraíba, 2018. Citado 2 vezes nas páginas 24 e 27.
- SALVINO, M. M. *Modelagem Computacional Visando a Reabilitação de Redes Hidráulicas*. Tese (Doutorado em Engenharia Mecânica), UFPB, João Pessoa. 2012. Citado 2 vezes nas páginas 7 e 24.
- SCHUSTER, M.; PALIWAL, K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, v. 45, n. 11, p. 2673–2681, 1997. Citado 2 vezes nas páginas 7 e 30.
- SINISA. *Índice de Perdas*. 2025. Acessado em 02/09/2025. Disponível em: <<https://indicadores-sinisa-2025.cidades.gov.br/dashboard?modulo=agua>>. Citado na página 13.
- SRIVASTAVA, N. et al. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, JMLR. org, v. 15, n. 1, p. 1929–1958, 2014. Citado na página 37.
-