

Metodologia para Integração de Dados Heterogêneos via IoT e Bases Públicas para o Monitoramento Preditivo de Cianobactérias em Mananciais

Fabiano N. Mendes¹, Karen P. Mendes², Diana F. Adamatti¹

¹Programa de Pós-Graduação em Modelagem Computacional – Universidade Federal de Rio Grande (FURG)
Caixa Postal 474 – 96.201-900 – Rio Grande – RS – Brasil

²Instituto Federal Sul-rio-grandense – Campus Pelotas (IFSul)
Praça Vinte de Setembro, 455 – Pelotas – RS – Brasil.

{fabianonobre, Karperes, diana}@gmail.com

Abstract. *This paper proposes a methodology for integrating heterogeneous data to monitor and predict cyanobacteria blooms in public water supply sources. The hybrid architecture combines primary temperature and luminosity data, collected via low-cost IoT sensors (ESP32), with historical nitrogen and phosphorus data from public databases. The data flow encompasses acquisition, harmonization, and processing stages through artificial intelligence. ConvLSTM architectures will be explored to capture spatio-temporal dynamics, along with XGBoost to analyze the importance of environmental variables. This work establishes a low-cost tool for preventive water resource management, aiming to anticipate public health risks.*

Keywords: *Heterogeneous Data, Cyanobacteria, Machine Learning, IoT, Water Sources.*

Resumo. Este artigo apresenta uma metodologia de integração de dados heterogêneos para o monitoramento preditivo de cianobactérias em mananciais. A arquitetura híbrida combina dados primários de temperatura e luminosidade, coletados via sensores IoT de baixo custo (ESP32), a dados históricos de nitrogênio e fósforo de bases públicas. O fluxo de dados abrange a aquisição, harmonização e o processamento por inteligência artificial. Serão exploradas as arquiteturas ConvLSTM, para dinâmicas espaço-temporais, e XGBoost, para análise de variáveis ambientais. O trabalho resultará em uma ferramenta acessível para a gestão preventiva de recursos hídricos e mitigação de riscos à saúde pública.

Palavras-chave: Dados Heterogêneos, Cianobactérias, Aprendizado de Máquina, IoT, Mananciais.

1. Introdução

A integridade dos sistemas de água doce e a confiabilidade do fornecimento hídrico em centros urbanos, enfrentam hoje o desafio das florações de cianobactérias (*HABs* - *Harmful Algal Blooms*) [He et al. 2024]. Este fenômeno é potencializado pela ação conjunta da eutrofização, fomentada pela ação humana através do descarte excessivo de nutrientes, como nitrogênio (N) e fósforo (P), bem como pelo aquecimento global, que estende os períodos de temperatura e luminosidade ideais para o crescimento desses microrganismos. O impacto além de causar uma degradação ambiental, pode atingir a saúde pública por meio da liberação de cianotoxinas, como microcistinas e saxitoxinas, cujos efeitos no organismo humano variam de reações agudas a quadros graves de neurotoxicidade e hepatotoxicidade [Wu et al. 2021].

A compreensão científica atual estabelece que a proliferação dessas algas deriva de um sistema multifatorial [Glibert et al. 2018], nos quais parâmetros físicos como a luminosidade e a temperatura da água operam como os principais indutores biológicos. Temperaturas elevadas, comumente superiores a 25°C, asseguram uma vantagem competitiva às cianobactérias frente a outros microrganismos, enquanto a luz regula processos de migração vertical e flutuabilidade [Xue et al. 2022]. Em reservatórios destinados ao consumo humano, a vigilância constante dessas variáveis é imprescindível para antecipar o aumento súbito de biomassa.

Tradicionalmente, a gestão de recursos hídricos apoia-se em monitoramento laboratorial intermitente, geralmente envolvendo alto custo operacional para amostragem e realização das análises. A quantificação de nutrientes (N e P) envolve a necessidade de infraestrutura laboratorial complexa e mão-de-obra especializada. Embora adequada e segura, a abordagem convencional é limitada para acompanhar a dinâmica temporal veloz e imprevisível das florações algais [Coffer et al. 2021].

Mesmo que a disponibilidade de nutrientes seja essencial para manter as florações, temperatura e luminosidade atuam como gatilhos para o seu surgimento. Nesse contexto, a Internet das Coisas (*Internet of Things* - *IoT*) surge como uma ruptura tecnológica que pode contribuir para uma melhor previsibilidade das florações, pois viabiliza a coleta de dados de temperatura e luminosidade em tempo real, permitindo a implementação de redes de monitoramento para empresas de saneamento de diferentes portes.

Sendo assim, este trabalho propõe a uma base de dados heterogênea, formada pela integração de bases de dados públicos com dados coletados de sensores de IoT. O núcleo do desafio científico aqui apresentado consiste em unir os dados locais de luminosidade e temperatura (via sensores IoT), com dados de fósforo e nitrogênio obtido em fontes públicas, para realização de uma modelagem preditiva.

A viabilidade da modelagem preditiva é sustentada pela evolução recente dos algoritmos de Aprendizagem de Máquina (*Machine Learning*) no campo das ciências ambientais [Camps-Valls et al. 2025]. Com base nas lacunas identificadas na Revisão Sistemática de Literatura (RSL) recentemente realizada [Mendes et al. 2026], esta pesquisa propõe o estudo de arquiteturas como Random Forest, XGBoost e redes neurais profundas do tipo ConvLSTM (Convolutional Long Short-Term Memory). O objetivo é converter grandes volumes de dados brutos em inteligência operacional, capacitando os gestores de Estações de Tratamento de Água (ETAs) a efetuarem ajustes preventivos no processo de tratamento, antecipando as florações de cianobactérias.

Diferente do levantamento exploratório realizado anteriormente com a laboração da RSL, este artigo detalha e fundamenta a aplicação prática de um sistema de monitoramento preditivo através do uso de séries temporais e dispositivos de baixo custo, buscando responder a seguinte pergunta: *"De que modo a variação na luminosidade e na temperatura, monitoradas por sensores IoT, combinada com dados de fósforo e nitrogênio provenientes de bases públicas, influencia a ocorrência de florações de cianobactérias em mananciais de abastecimento público?"*.⁷

2. Trabalhos relacionados

O monitoramento de florações de cianobactérias em mananciais de abastecimento público tem passado por uma significativa expansão tecnológica. Embora as análises laboratoriais permaneçam como o padrão de referência para a validação da qualidade da água, os estudos apontam para o surgimento de sistemas inteligentes onde os sensores baseados em Internet das Coisas (IoT) e Inteligência Artificial (IA) tornam-se alternativas complementares e de baixo custo para o monitoramento. Estas ferramentas permitem o monitoramento contínuo e em tempo real, mitigando a latência das coletas manuais e fornecendo uma camada preditiva essencial para a gestão dos mananciais. Na sequência são abordadas as principais frentes que fundamentam esta pesquisa.

2.1 Infraestrutura de IoT para monitoramento de mananciais

A viabilidade de modelos preditivos em tempo real depende intrinsecamente de uma infraestrutura de coleta de dados robusta íntegra e distribuída. O estudo de de Al-Khafaji et al. (2025), aponta que a transição de coletas pontuais para o monitoramento contínuo pode ser viabilizada pela Internet das Coisas (IoT), pois ela pode trabalhar como agente coletor de dados físico-químicos.

De acordo com a revisão de de Al-Khafaji et al. (2025), a arquitetura IoT para a gestão da água é geralmente dividida em três camadas: percepção (sensores de campo), rede (transmissão) e aplicação (análise de dados). O uso de IoT permite a diminuição de dados faltantes nos datasets, antes sujeitos a erros logísticos ou humanos. Pode também capturar variações súbitas em parâmetros como luminosidade e temperatura da água, que são gatilhos críticos para florações, conforme discutido por Brenckman et al. (2025). Além disso, a capacidade de operar em locais remotos com autonomia energética (painéis solares), torna o IoT uma alternativa possível para o monitoramento de grandes reservatórios de abastecimento público.

2.2. Monitoramento por Sensores Virtuais e Aprendizado de Máquina

Um dos grandes desafios no monitoramento de mananciais é o custo elevado de sensores químicos para nutrientes como Nitrogênio (N) e Fósforo (P). Paepae et al. (2022) introduzem o conceito de "sensores virtuais", demonstrando que é possível estimar concentrações de nutrientes utilizando variáveis de fácil medição (pH, turbidez, temperatura) processadas por algoritmos como Random Forest e XGBoost. Essa abordagem valida a viabilidade técnica de sistemas de baixo custo para preencher lacunas de dados físicos em tempo real. Esse conceito foi explorado em investigações anteriores, onde foram aplicados algoritmos de aprendizado de máquina e modelos de séries temporais (SARIMA) para prever a variabilidade da turbidez e do nível de reservatórios, demonstrando que padrões históricos podem subsidiar a gestão operacional da clarificação de água [Mendes et al. 2024].

2.3. Séries Temporais e Redes Neurais

A dinâmica das cianobactérias é não linear e influenciada por padrões sazonais. Caballero et al. (2025) destacam a superioridade das redes neurais do tipo *Long Short-Term Memory* (LSTM) para capturar essas variações temporais. Nossa RSL aponta que a integração de dados locais com modelos de *Deep Learning* permite uma antecipação de eventos de floração que modelos estatísticos tradicionais não conseguem alcançar com a mesma precisão.

2.3. Harmonização de Dados e Grandes Datasets

A robustez dos modelos de IA depende da escala, organização e integridade dos dados. Spaulding et al. (2024) apresentaram um esforço crítico de harmonização de dados de clorofila-a (2005-2022), criando o maior dataset do gênero. Este trabalho é essencial para este estudo, pois fornece a base metodológica para a calibração de modelos preditivos e sensores de campo, permitindo que a IA aprenda com uma vasta gama de variações ambientais.

2.4. Gestão Inteligente e Impacto na Saúde Pública

A transição para o monitoramento em tempo real (*Smart Water Management*) é discutida por Al-Khafaji et al. (2025), que reforçam como a IA pode otimizar o tratamento em estações de abastecimento (ETAs). A necessidade dessas tecnologias é ratificada pelos achados epidemiológicos de Wu et al. (2021), que correlacionam diretamente as florações com o aumento de atendimentos hospitalares. Finalmente, Brenckman et al. (2025) oferecem a visão de controle, enfatizando que o monitoramento contínuo via IoT é uma alternativa promissora para validar a eficácia de intervenções nos reservatórios.

A Tabela 1 apresenta uma breve descrição dos principais elementos envolvidos nos artigos base que direcionaram a proposta aqui discutida.

Tabela 1. Principais características dos estudos que norteiam esse trabalho

Autor (Ano)	Foco Principal	Tecnologia/ Algoritmo	Contribuição
Wu et al. (2021)	Saúde Pública	Satélite Epidemiologia	Justificativa do impacto social e alertas precoce
Paepae et al. (2022)	Sensores Virtuais	RF, XGBoost	Estimativa de nutrientes com sensores de baixo custo
Spaulding et al. (2024)	Big Data/ Clorofila	Harmonização de dados	Metodologia para calibração e treino de modelos
Caballero et al. (2025)	Predição de HABs	Redes LSTM	Base para modelagem temporal de séries complexas
Al-Khafaji et al. (2025)	Gestão da Água	IoT e Deep Learning	Arquitetura de sistemas inteligentes em tempo real
Brenckman et al. (2025)	Prevenção e Causas	Revisão Multidisciplinar	Contextualização das causas e métodos de controle

3. Metodologia para Integração de Dados Heterogêneos

A metodologia proposta baseia-se na criação de uma base de dados híbrida, que combina valores do monitoramento local através de sensores de baixo custo (Temperatura e Luminosidade), com a recuperação de dados históricos de bases públicas, culminando em uma arquitetura de inteligência preditiva.

O fluxo de dados é estruturado em **quatro camadas** interdependentes, conforme detalhado na Figura 1 e nas próximas seções.

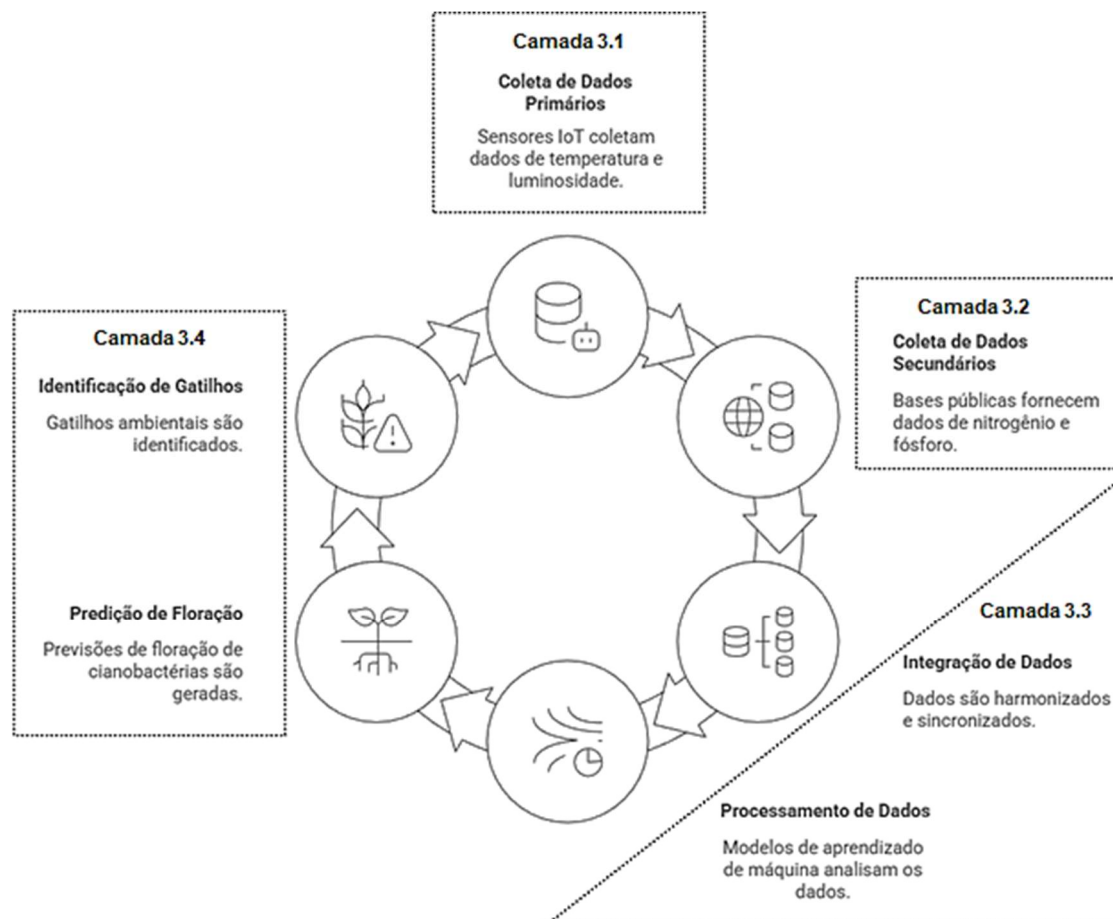


Figura 1. Etapas da Proposta Metodológica para o Ciclo de Monitoramento e Predição de Florações de Cianobactérias

3.1. Camada de Aquisição (IoT)

Esta camada é responsável pela geração de dados primários em tempo real. Utiliza-se uma arquitetura baseada em microcontroladores de baixo custo (família Arduino/ESP32), integrados a sensores de temperatura da água e luminosidade (radiação solar).

Justificativa: A escolha do ESP32 deve-se ao suporte nativo para comunicação sem fio e ao baixo consumo energético, permitindo a operação em campo. *Espressif Systems* (2023). Esta camada resolve a lacuna da amostragem discreta, fornecendo o fluxo contínuo necessário para identificar gatilhos imediatos de floração.

3.2. Camada de Resgate de dados (Bases Públicas)

Complementando os sensores físicos, esta camada executa o resgate automatizado de variáveis químicas complexas (N e P) a partir de bases de dados governamentais e repositórios públicos (como os dados harmonizados sugeridos por Spaulding et al., 2024).

Função: Como sensores de N e P são de alto custo, o resgate de bases históricas permite correlacionar os picos de nutrientes com os dados de temperatura e luminosidade coletados pela camada IoT, servindo como dados de referência para o treinamento dos modelos.

3.3. Processamento e Harmonização de Dados

Devido à periodicidade heterogênea das fontes (dados de sensores a cada hora vs. dados de bases públicas mensais), esta etapa aplica técnicas de harmonização e normalização:

- *Resampling*: Reamostragem dos dados para uma escala temporal comum.
- Tratamento de *Outliers*: Filtros estatísticos para remover ruídos dos sensores *IoT* das bases públicas.
- Imputação de Dados: Uso de técnicas de interpolação para preencher os "dados faltantes" nas bases públicas, garantindo a continuidade da série temporal.

3.4. Camada de Inteligência (*Machine Learning*)

Com os dados harmonizados, aplica-se a lógica preditiva baseada nos achados da RSL. O foco principal é a implementação de arquiteturas de ConvLSTM (*Convolutional Long Short-Term Memory*). O diferencial desta abordagem é que, diferentemente de uma rede LSTM comum, a ConvLSTM é capaz de processar as relações temporais das variáveis ambientais enquanto captura padrões, sendo ideal para prever a evolução da floração de cianobactérias no reservatório com base na dinâmica não linear (Caballero et al., 2025).

Como alternativa de alta eficiência e para fins de comparação de desempenho, a metodologia contempla o uso do algoritmo XGBoost (*Extreme Gradient Boosting*). Esta escolha justifica-se pela sua robustez em lidar com dados tabulares e sua capacidade de identificar a importância de cada característica. Enquanto a ConvLSTM foca na complexidade espaço-temporal, o XGBoost permite validar quais variáveis (Fósforo, Nitrogênio, Luminosidade ou Temperatura) exercem maior peso estatístico na predição imediata das florações, oferecendo um modelo de decisão mais interpretável e computacionalmente mais leve para processamento ou servidores com recursos limitados.

4. Conclusões e Trabalhos Futuros

Este estudo apresentou uma metodologia de integração de dados heterogêneos para o monitoramento e predição de florações de cianobactérias em mananciais de abastecimento público. Sua principal contribuição reside na junção de dados físico-químicos captados através de sensores de baixo custo com dados provenientes de bases públicas, para a criação de um banco de dados heterogêneo a ser utilizado na predição de florações de cianobactérias em mananciais de abastecimento público.

A harmonização dos dados por meio de técnicas de *resampling* e imputação irá permitir a criação de um dataset robusto, capaz de alimentar modelos de inteligência artificial de alta complexidade. A utilização da arquitetura ConvLS TM, a partir do

trabalho de Caballero et al., (2025), provou ser eficaz na captura das dinâmicas não lineares e espaço-temporais inerentes aos processos biológicos aquáticos, enquanto a aplicação do XGBoost irá oferecer uma camada essencial de interpretabilidade ao identificar o peso das variáveis ambientais como maiores gatilhos para as florações (Paepae et al., 2022).

O trabalho visa demonstrar que a integração de múltiplas tecnologias, como: IoT, *Machine Learning* e Series temporais, quando aliada a uma arquitetura de dados heterogênea, oferece uma ferramenta poderosa para a gestão preventiva de recursos hídricos, permitindo a antecipação de eventos críticos com menor investimento em infraestrutura física.

Como desdobramentos desta pesquisa, pretende-se realizar a validação experimental do modelo em um cenário real, aplicando a arquitetura de dados heterogênea em um manancial de abastecimento público para avaliar a capacidade de previsão de florações de cianobactérias. Após a implementação física dos sensores IoT e a validação da metodologia proposta, o enfoque será a transposição da arquitetura de dados híbrida idealizada para um ambiente de operação real, aplicando o monitoramento em mananciais específicos para avaliar o desempenho dos modelos ConvLSTM e XGBoost frente a dados de campo coletados em tempo contínuo. Pretende-se realizar a calibração técnica dos sensores de baixo custo e o ajuste fino dos algoritmos de harmonização, garantindo que a integração entre as leituras do ESP32 e as bases públicas apresente a acurácia necessária para obtenção de uma ferramenta que dê apoio a decisão aos gestores dos mananciais.

Adicionalmente, a etapa de implementação contemplará a análise da eficiência computacional dos modelos em cenários de uso prolongado, visando consolidar a viabilidade técnica da arquitetura para órgãos de fiscalização e empresas de saneamento. A validação empírica servirá para confirmar se as variáveis ambientais selecionadas atuam, de fato, como gatilhos preditivos confiáveis, permitindo que a metodologia evolua de uma fundamentação teórica para uma ferramenta de suporte à decisão operacional aplicada ao controle de florações de cianobactérias em mananciais de abastecimento público.

Referências

- Al-Khafaji, M. S. et al. (2025) “Revolutionizing Water Quality Monitoring with Artificial Intelligence: A Systematic Review”, *Journal of Studies in Science and Engineering*, v. 5, n. 1, p. 1-20.
- Brenckman, C. M. et al. (2025) “A Review of Harmful Algal Blooms: Causes, Effects, Monitoring, and Prevention Methods”, *Water*, v. 17, n. 13, art. 1980, <https://doi.org/10.3390/w17131980>.
- Caballero, C. B. et al. (2025) “The need for advancing algal bloom forecasting using remote sensing and modeling: Progress and future directions”, *Ecological Indicators*, v. 172, art. 113244.
- Camps-Valls, G. et al. (2025) “Artificial intelligence for modeling and understanding extreme weather and climate events”, *Nature Communications Publishing*, Cham, <https://doi.org/10.1038/s41467-025-56573-8>.

- Coffer, M. et al. (2021) “Assessing cyanobacterial frequency and abundance at surface waters near drinking water intakes across the United States”, *Water Research (ScienceDirect)*, <https://doi.org/10.1016/j.watres.2021.117377>.
- Espressif Systems (2023) “ESP32 series datasheet (versão 3.5)”, Shanghai, https://www.espressif.com/sites/default/files/documentation/esp32_datasheet_en.pdf.
- Glibert, P. M. et al. (eds.) (2018) *Global Ecology and Oceanography of Harmful Algal Blooms*, Springer International Publishing, Cham, <https://doi.org/10.1007/978-3-319-70069-4>.
- He, S. et al. (2024) “Summer heatwaves promote harmful algal blooms in the Chunjiang Reservoir, an important drinking water source”, *Journal of Environmental Management*, v. 359, p. 1–16, <https://doi.org/10.1016/j.jenvman.2024.121056>.
- Mendes, F. N. et al. (2024) “Application of Time Series in the Analysis of Physical-Chemical Data to support the Management of Water Clarification Efficiency in a Water Treatment Plant”, *Transfer Phenomena in Fluid and Heat Flows XV – Scientific.net*.
- Mendes, F. N. et al. (2026) “Inteligência Artificial e IoT para Monitoramento inteligente de cianobactérias em mananciais de abastecimento público: Uma revisão sistemática de literatura”, *Revista Políticas Públicas & Cidades*, v. 15, n. 2, p. e3157.
- Paepae, T., Bokoro, P. N. and Kyamakyia, K. (2022) “A Virtual Sensing Concept for Nitrogen and Phosphorus Monitoring Using Machine Learning Techniques”, *Sensors* (ou periódico correspondente), v. 22, n. 19, art. 7338.
- Page, M. J. et al. (2021) “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews”, *BMJ*, v. 372, n. 71, <https://doi.org/10.1136/bmj.n71>.
- Spaulding, S. A. et al. (2024) “Chlorophyll a in lakes and streams of the United States (2005–2022)”, *Scientific Data*, v. 11, art. 611.
- Wu, J. et al. (2021) “Acute health effects associated with satellite-determined cyanobacterial blooms in a drinking water source in Massachusetts”, *Environmental Health*, v. 20, p. 1–12, <https://doi.org/10.1186/s12940-021-00754-5>.
- Xue, Z. et al. (2022) “Influence of wind and light on the floating and sinking process of *Microcystis*”, *Scientific Reports*, <https://doi.org/10.1038/s41598-022-08977-5>.