



**DETECCIÓN TEMPRANA DE  
CONTAMINANTES EN AGUAS  
SUPERFICIALES USANDO INTELIGENCIA  
ARTIFICIAL**

**MARÍA JOSÉ ERAZO GONZÁLEZ**

**Tesis para optar al título de Ingeniero Civil en Informática y  
Telecomunicaciones**

**Profesor guía: Diego Dujovne**

**FACULTAD DE INGENIERÍA Y CIENCIAS  
ESCUELA DE INFORMÁTICA Y TELECOMUNICACIONES**

**Santiago, Chile  
19 de julio de 2025**





**DETECCIÓN TEMPRANA DE  
CONTAMINANTES EN AGUAS  
SUPERFICIALES USANDO INTELIGENCIA  
ARTIFICIAL**

**MARÍA JOSÉ ERAZO GONZÁLEZ**

**Tesis para optar al título de Ingeniero Civil en Informática y  
Telecomunicaciones**

**Profesor guía: Diego Dujovne  
Comité: Karol Suchan**

**FACULTAD DE INGENIERÍA Y CIENCIAS  
ESCUELA DE INFORMÁTICA Y TELECOMUNICACIONES**

**Santiago, Chile  
19 de julio de 2025**

 María José Erazo González  
✉ maria.erazo@mail\_udp.cl

A mis padres y hermano, por enseñarme que nada está perdido hasta el final.

A mi abuela, que desde donde esté celebra este logro.

*"To the girl who reads by flashlight, who sees dragons in the clouds, who feels most alive in worlds that never were, who knows magic is real, who dreams. This is for you."*

— Meagan Spooner, Hunted

## Agradecimientos

---

Quiero tener este momento para agradecer a mis padres por siempre ayudarme a seguir intentando terminar lo que empecé, y por ayudarme a entender que nada realmente está perdido hasta el final. Su constante apoyo y fe en mis capacidades fueron fundamentales para superar los momentos más desafiantes de este proceso. A mi hermano, por darme consejos para poder ir superando las dificultades que se han presentado a lo largo de la carrera. Sus palabras de aliento y perspectiva práctica fueron invaluables durante todo este tiempo. A mi abuela, que a pesar de que ya no está con nosotros, aún siento su presencia y las ganas que tenía de que nosotros fuéramos profesionales y que cumpliéramos todo lo que nos propusieramos. Su legado de perseverancia sigue siendo una fuente de inspiración. A mis grandes amigos, que son increíbles personas que no pensé que iba a encontrar. Su apoyo y la forma en que congeniamos a pesar de ser tan diferentes ha sido uno de los regalos más valiosos de esta etapa universitaria. A mi profesor guía Diego Dujovne, por la oportunidad de tener esta idea que me impulsó a crecer y investigar un mundo nuevo a través de este proyecto. Su orientación técnica y académica fue esencial para transformar una idea inicial en una investigación sólida y contribuir al conocimiento en el área de monitoreo ambiental automatizado.

## Resumen

---

Esta tesis propone un sistema para la detección temprana de contaminantes en aguas superficiales mediante el uso combinado de espectroscopía de reflectancia UV-Vis e inteligencia artificial. Se desarrolló una metodología integral que incluye preprocesamiento espectral, clasificación binaria y multiclasificación, validación cruzada temporal y evaluación robusta de modelos. Los algoritmos utilizados (SVM, XGBoost y LSTM) fueron aplicados a 29 contaminantes, logrando detecciones confiables en 8 casos, lo que representa una tasa de éxito del 27,6 %. Los resultados demuestran que, incluso ante datos espectrales ruidosos o parcialmente estructurados, es posible construir modelos capaces de detectar contaminantes específicos con una precisión aceptable. Este trabajo sienta las bases para futuras implementaciones de sistemas de monitoreo ambiental autónomos, escalables e interpretables.



## **Abstract**

---

This thesis presents a system for the early detection of contaminants in surface water sources using UV-Vis reflectance spectroscopy combined with machine learning techniques. The proposed methodology includes spectral preprocessing, binary and multiclass classification, temporal cross-validation, and robust model evaluation using SVM, XGBoost, and LSTM algorithms. Applied to 29 contaminants, the system achieved reliable detection in 8 cases, corresponding to a 27.6 % success rate. The results demonstrate that even with noisy and partially structured spectral data, it is possible to develop models capable of detecting specific contaminants with acceptable accuracy. This work lays the foundation for future implementations of autonomous, scalable, and interpretable environmental monitoring systems.



# Contenido

---

|   |     |
|---|-----|
| <b>Resumen</b>  | i   |
| <b>Abstract</b>   | iii |
| <b>Lista de tablas</b>  | vii |
| <b>Lista de figuras</b>   | ix  |
| <b>Capítulo 1. Introducción</b>   | 1   |
| 1.1. Motivación y Contexto . . . . .  | 1   |
| 1.2. Planteamiento del Problema . . . . .   | 2   |
| 1.3. Objetivos . . . . .  | 3   |
| 1.3.1. Objetivo General . . . . .   | 3   |
| 1.3.2. Objetivos Específicos . . . . .  | 3   |
| 1.4. Hipótesis . . . . .  | 3   |
| 1.5. Metodología General . . . . .  | 4   |
| 1.6. Justificación Científica y Social . . . . .                                      | 4   |
| 1.7. Alcance y Limitaciones . . . . .   | 5   |
| <b>Capítulo 2. Marco Teórico</b>  | 7   |
| 2.1. Espectroscopía de Reflectancia UV-Vis . . . . .                                  | 7   |
| 2.2. Métodos Tradicionales de Monitoreo . . . . .                                     | 8   |
| 2.3. Aplicación de inteligencia artificial . . . . .                                  | 10  |
| 2.4. Comparación entre Métodos . . . . .  | 11  |
| 2.5. Justificación del Enfoque Propuesto . . . . .                                    | 11  |
| 2.6. Limitaciones y Desafíos . . . . .  | 12  |
| <b>Capítulo 3. Estado del Arte</b>  | 13  |
| 3.1. Monitoreo de la Calidad del Agua . . . . .                                       | 13  |
| 3.2. Trabajos Previos sobre el conjunto de datos Lechevallier et al. (2024) . . . . . | 14  |
| 3.3. Espectroscopía UV-Vis vs. Imágenes Hiperespectrales . . . . .                    | 15  |
| 3.4. Modelos de inteligencia artificial en Clasificación Espectral . . . . .          | 16  |
| 3.5. Brechas Identificadas en el Estado Actual . . . . .                              | 17  |

|  |           |
|--|-----------|
| <b>Capítulo 4. Metodología</b>                                     | <b>19</b> |
| 4.1. Enfoque Experimental y Computacional . . . . .                | 20        |
| 4.2. Metodología de Preprocesamiento Espectral . . . . .           | 22        |
| 4.3. Diseño del flujo de procesamiento de Clasificación . . . . .  | 25        |
| 4.4. Metodología de Validación y Criterios de Aceptación . . . . . | 27        |
| 4.5. Resumen de la Metodología . . . . .                           | 29        |
| <b>Capítulo 5. Implementación</b>                                  | <b>31</b> |
| 5.1. Arquitectura General del Sistema . . . . .                    | 31        |
| 5.2. Metodología de Construcción de conjuntos de datos . . . . .   | 32        |
| 5.3. Ingeniería de Características Espectrales Avanzada . . . . .  | 34        |
| 5.4. Análisis de Viabilidad Espectral . . . . .                    | 36        |
| 5.5. Estrategias Diferenciadas de Entrenamiento . . . . .          | 37        |
| 5.6. Evaluación y Validación de Modelos . . . . .                  | 38        |
| <b>Capítulo 6. Resultados y Análisis</b>                           | <b>41</b> |
| 6.1. Evaluación Integral por Contaminante . . . . .                | 41        |
| 6.2. Análisis Comprehensivo de Resultados . . . . .                | 46        |
| 6.3. Síntesis del Desempeño Final . . . . .                        | 48        |
| <b>Capítulo 7. Conclusiones Generales</b>                          | <b>51</b> |
| 7.1. Estrategia de Evaluación y Síntesis de Resultados . . . . .   | 54        |
| 7.2. Contribuciones Metodológicas del Sistema . . . . .            | 55        |
| 7.3. Direcciones para Desarrollo Futuro . . . . .                  | 56        |
| 7.4. Reflexiones Finales y Perspectivas . . . . .                  | 57        |
| <b>Referencias bibliográficas</b>                                  | <b>59</b> |
| <b>Anexo A. Evaluación de la Calidad de los Datasets</b>           | <b>65</b> |
| <b>Anexo B. Resultados de Modelos de Machine Learning</b>          | <b>67</b> |
| <b>Anexo C. Metodología Detallada</b>                              | <b>69</b> |
| <b>Anexo D. Especificaciones Técnicas del Sistema</b>              | <b>71</b> |
| <b>Anexo E. Firmas Espectrales por Contaminante</b>                | <b>73</b> |
| <b>Anexo F. Síntesis de Resultados</b>                             | <b>81</b> |

## **Lista de tablas**

---

|      |  |    |
|------|--|----|
| 1.1. | Resumen de contaminantes por categoría analizados en este estudio . . . . .        | 6  |
| 2.1. | Comparación entre enfoques de monitoreo según literatura . .                       | 11 |
| 3.1. | Aplicaciones documentadas de modelos de IA en espectroscopía UV-Vis . . . . .      | 17 |
| 4.1. | Criterios de filtrado para validación de modelos . . . . .                         | 28 |
| 5.1. | Estrategias de procesamiento según características del conjunto de datos . . . . . | 33 |
| 5.2. | Estrategias de entrenamiento diferenciadas por categoría de contaminante . . . . . | 38 |
| 6.1. | Resultados detallados de modelos exitosos . . . . .                                | 43 |
| 6.2. | Resultados detallados de los mejores modelos con criterios de producción . . . . . | 44 |
| 6.3. | Casos de estudio específicos con estrategias diferenciadas . .                     | 45 |
| A.1. | Clasificación de calidad de datasets y recomendaciones . . . .                     | 65 |
| B.1. | Resumen de rendimiento por algoritmo . . . . .                                     | 67 |
| B.2. | Mejores resultados por contaminante . . . . .                                      | 68 |
| B.3. | Tasa de éxito por categoría química . . . . .                                      | 68 |
| C.1. | Top 5 características espectrales más importantes . . . . .                        | 70 |
| D.1. | Especificaciones clave del sistema MV.X . . . . .                                  | 71 |
| D.2. | Estadísticas descriptivas del dataset completo . . . . .                           | 72 |
| E.1. | Ánálisis detallado de firmas espectrales por contaminante . .                      | 80 |



## **Lista de figuras**

---

|   |    |
|---|----|
| 4.1. Arquitectura metodológica del sistema propuesto. Procesamiento del conjunto de datos de Lechevallier et al. (2024) mediante tres algoritmos de aprendizaje automático con clasificación jerárquica de dos etapas y validación temporal estricta. | 29 |
| E.1. Firmas espectrales diferenciales de todos los contaminantes evaluados, mostrando patrones de absorción característicos en el rango UV-Vis (400-800 nm) . . . . .   | 74 |
| E.2. Mapa de calor de diferencias espectrales por contaminante. Los colores rojos indican mayores diferencias espectrales, revelando los contaminantes más discriminables . . . . .   | 74 |
| E.3. Firma espectral comparativa de Benzotriazole, mostrando excelente separabilidad entre concentraciones altas y bajas . . . . .  | 75 |
| E.4. Firma espectral de 24-D, evidenciando las mayores diferencias espectrales observadas en el estudio . . . . .   | 75 |
| E.5. 6PPD-quinone mostrando firmas espectrales casi idénticas entre clases, explicando las dificultades en clasificación automatizada . . . . .   | 76 |
| E.6. Acesulfame con firmas superpuestas a pesar del amplio rango de concentraciones (5859-152831 ng/L) . . . . .  | 76 |
| E.7. Firma espectral de PO <sub>4</sub> (Fosfatos) . . . . .  | 77 |
| E.8. Firma espectral de Turbidez . . . . .  | 77 |
| E.9. Firma espectral de Hydrochlorothiazide . . . . .   | 78 |
| E.10. Firma espectral de Diclofenac . . . . .   | 78 |
| E.11. Firma espectral de Diuron . . . . .   | 79 |
| E.12. Firma espectral de Mecoprop . . . . .   | 79 |



# Capítulo 1

## Introducción

---

### 1.1. Motivación y Contexto

En Chile, la calidad del agua ha sido objeto de atención creciente por parte de la institucionalidad, especialmente en zonas con alta demanda hídrica y uso intensivo, como áreas agrícolas e industriales, donde la DGA ha debido intervenir con medidas de redistribución y control de extracciones. Según el informe nacional de la Dirección General de Aguas de 2023, el 61 % de las cuencas hidrográficas del país presenta al menos una estación con calidad de agua superficial en categoría no buena [1].

El mismo informe evidencia que hasta hace pocos años la cobertura y frecuencia de monitoreo eran limitadas, lo que motivó una expansión significativa de la red, alcanzando más de 1300 estaciones de monitoreo activo en 2023, con capacidad de transmisión en tiempo real [1]. Esta situación previa dificultaba la detección de eventos transitorios o de corta duración, restringiendo la capacidad de respuesta oportuna ante episodios de contaminación.

El desafío técnico principal consiste en distinguir las variaciones naturales de la calidad del agua de aquellas causadas por contaminantes específicos. Los métodos tradicionales presentan limitaciones significativas: requieren personal especializado, son intensivos en costos debido al alto mantenimiento de instalaciones de laboratorio, utilizan materiales químicos, y los resultados pueden obtenerse después de varios días sin capacidad de monitoreo en tiempo real [2]. Además, debido a la configuración compleja y el

## INTRODUCCIÓN

tiempo requerido, el contenido de las muestras puede cambiar durante el proceso, produciendo datos menos valiosos para el monitoreo [2].

### 1.2. Planteamiento del Problema

El sistema de monitoreo hídrico en Chile presenta limitaciones significativas que afectan su capacidad de detección oportuna de eventos de contaminación. La densidad actual de estaciones de monitoreo (una estación por cada 818 km<sup>2</sup>) es significativamente inferior a los estándares internacionales recomendados (una estación cada 5 km<sup>2</sup>), lo que genera importantes brechas en la cobertura territorial [3].

Adicionalmente, el sistema actual opera con frecuencias de muestreo mensual, bimestral o trimestral, y no incluye el monitoreo de contaminantes emergentes como microplásticos, fármacos o disruptores endocrinos [1]. A diferencia de otros países, Chile aún no ha implementado sistemas de alerta temprana para eventos agudos de contaminación, limitando su capacidad de respuesta ante situaciones críticas [1].

Estas limitaciones se ven agravadas por un marco normativo que requiere la superación de normas específicas para declarar contaminación, impidiendo la detección preventiva cuando dichas normas no existen [3]. Esta situación reduce significativamente la capacidad del sistema para distinguir de manera confiable entre variaciones naturales y alteraciones causadas por contaminantes, con implicancias directas en la gestión ambiental y la protección de recursos hídricos.

### Pregunta de investigación

¿Es posible desarrollar un sistema de detección temprana de contaminantes en aguas superficiales que opere en tiempo real, utilizando espectroscopía UV-Vis e inteligencia artificial, para mejorar la capacidad actual de distinguir patrones de contaminación de las variaciones naturales?

## 1.3. Objetivos

### 1.3.1. Objetivo General

Explorar la viabilidad técnica de un sistema basado en espectroscopía UV-Vis e inteligencia artificial para la detección temprana de contaminantes, como prueba de concepto para futuras aplicaciones en monitoreo de aguas superficiales, utilizando datos de aguas residuales urbanas como caso de estudio representativo.

### 1.3.2. Objetivos Específicos

- Evaluar la capacidad de la espectroscopía UV-Vis para capturar firmas espectrales distintivas de diferentes contaminantes.
- Entrenar modelos de inteligencia artificial (Support Vector Machines, XGBoost y Long Short-Term Memory) que permitan clasificar dichas firmas.
- Comparar el rendimiento de los modelos en términos de sensibilidad, precisión y generalización.
- Implementar un flujo de procesamiento de análisis automatizado capaz de integrarse en entornos de monitoreo continuo.
- Analizar el comportamiento del sistema ante datos con ruido, variabilidad ambiental y condiciones reales.

Los objetivos planteados se fundamentan en la siguiente hipótesis de trabajo.

## 1.4. Hipótesis

Diferentes tipos de contaminantes en agua generan firmas espectrales distintivas en el rango UV-Vis, como ha sido demostrado para contaminantes iónicos específicos como carbonato, cloruro, fluoruro y sulfato [4]. La espectroscopía UV-Vis puede detectar efectivamente materia orgánica, nitratos, metales pesados y cloro residual, aunque presenta limitaciones conocidas relacionadas con interferencias por turbidez y solapamiento espectral [5].

Estas firmas espectrales pueden ser reconocidas y clasificadas automáticamente mediante modelos de inteligencia artificial, como se ha demostrado

## INTRODUCCIÓN

en sistemas de detección automatizada de metales pesados usando análisis de componentes principales y métodos de regresión [6].

Esta hipótesis se evalúa mediante la metodología descrita a continuación.

### 1.5. Metodología General

Para abordar esta problemática, este trabajo propone un enfoque secuencial con retroalimentación iterativa. La estrategia se fundamenta en dos etapas principales que operan de manera complementaria:

1. **Clasificación binaria:** Permite distinguir entre estados normales del agua y potenciales eventos de contaminación.
2. **Clasificación multiclasa:** Una vez detectado un evento anómalo, identifica el tipo específico de contaminante presente.

Esta aproximación utiliza datos espectrales obtenidos de fuentes reales, procesados mediante modelos de aprendizaje automático que se ajustan a la dinámica espectral característica de cada sustancia analizada.

### 1.6. Justificación Científica y Social

El enfoque desarrollado en esta investigación representa un complemento tecnológico a los sistemas de monitoreo existentes, ya que la espectroscopía UV-Vis permite realizar detección automatizada de contaminantes específicos como materia orgánica, nitratos y metales pesados [5], con capacidad de clasificación automatizada mediante inteligencia artificial [6].

Esta aproximación tecnológica responde a necesidades identificadas en el contexto chileno, donde las autoridades han reconocido deficiencias en personal, gestión de información y cobertura del monitoreo hídrico, así como la necesidad de fortalecer las redes de medición e información hídrica a nivel nacional [7]. La automatización del proceso de detección podría contribuir a abordar estas limitaciones de recursos humanos y cobertura territorial.

El desarrollo de esta metodología, aunque validada inicialmente con datos de aguas residuales urbanas, establece principios técnicos transferibles para el monitoreo de aguas superficiales. Considerando que el 61 % de las cuencas chilenas presenta problemas de calidad de agua [1], y que el sistema actual opera con frecuencias de muestreo mensual o trimestral [1], el establecimiento

de metodologías automatizadas representa una contribución relevante para la futura implementación de tecnologías complementarias en la gestión de recursos hídricos y el cumplimiento de estándares de calidad.

## 1.7. Alcance y Limitaciones

Esta investigación se enfoca específicamente en la detección de contaminantes mediante espectroscopía UV-Vis en el rango de 400–800 nm, correspondiente al espectro visible optimizado para condiciones de monitoreo en campo urbano [8]. El estudio analiza exactamente 29 contaminantes específicos, detallados en la Tabla 1.1, que incluyen 9 parámetros fisicoquímicos convencionales y 20 compuestos orgánicos emergentes de diversas fuentes (municipal, agrícola, industrial y escorrentía vial) [9].

El trabajo utiliza exclusivamente datos del conjunto de datos público de Lechevallier et al. (2024) [9], desarrollado en EAWAG mediante espectrofotometría de absorción y reflectancia UV-Vis con imágenes hiperespectrales en formato ENVI<sup>1</sup>. Esto garantiza reproducibilidad completa pero limita la validación a condiciones específicas de aguas residuales urbanas [8].

Se reconoce que las propiedades ópticas y la dinámica de contaminantes pueden diferir significativamente entre aguas residuales urbanas y aguas superficiales naturales. Esta limitación no invalida la metodología desarrollada, sino que establece la necesidad de validación adicional para su transferencia a otras matrices acuáticas. Los resultados obtenidos sientan las bases metodológicas para futuras investigaciones que podrían expandir el enfoque a aguas superficiales y diferentes condiciones ambientales.

---

<sup>1</sup>ENVI es un formato estándar para datos espectrales hiperespectrales, compuesto por un archivo binario (.bin) y un encabezado asociado (.hdr) con metadatos.

## INTRODUCCIÓN

**Tabla 1.1.** Resumen de contaminantes por categoría analizados en este estudio

| Categoría                 | Cantidad  | Fuente principal            | Método                 |
|---------------------------|-----------|-----------------------------|------------------------|
| Parámetros fisicoquímicos | 9         | Urbana                      | Métodos convencionales |
| Fármacos y edulcorantes   | 8         | Municipal                   | LC-HRMS/MS             |
| Pesticidas                | 5         | Agrícola                    | LC-HRMS/MS             |
| Aditivos industriales     | 4         | Escoorrentía/<br>Industrial | LC-HRMS/MS             |
| Inhibidores de corrosión  | 3         | Mixta                       | LC-HRMS/MS             |
| Total                     | <b>29</b> | Diversas                    | Múltiples              |

# Capítulo 2

## Marco Teórico

---

### 2.1. Espectroscopía de Reflectancia UV-Vis

La espectroscopía UV-Vis constituye una técnica analítica no invasiva que permite caracterizar muestras basándose en su interacción específica con la luz a diferentes longitudes de onda [10]. A lo largo de las últimas décadas, esta metodología ha ganado reconocimiento en análisis medioambientales debido a su capacidad para identificar sustancias disueltas sin requerir reactivos químicos ni procesos destructivos, ofreciendo ventajas significativas sobre métodos tradicionales de laboratorio [10].

El fundamento de esta técnica reside en la medición de la luz absorbida por una muestra después de ser irradiada con una fuente de espectro amplio. Diferentes sustancias químicas presentan respuestas ópticas características, lo que puede resultar en firmas espectrales distintivas, aunque con posibles solapamientos que requieren técnicas de procesamiento avanzadas para su diferenciación [10]. Esta propiedad permite detectar efectivamente ciertos contaminantes específicos, incluyendo nutrientes como nitrato, carbono orgánico disuelto y algunos productos farmacéuticos con estructuras aromáticas, aunque presenta limitaciones conocidas para pesticidas y compuestos en concentraciones traza [10].

En el contexto de esta investigación, la aplicación de esta técnica se evalúa utilizando datos de aguas residuales urbanas como caso de estudio, estableciendo bases metodológicas transferibles para futuro monitoreo de aguas

superficiales. Los contaminantes analizados incluyen nutrientes, productos farmacéuticos y materia orgánica, representativos de los retos de calidad hídrica en diferentes matrices acuáticas [9].

La implementación de espectroscopía UV-Vis para monitoreo automatizado representa una oportunidad estratégica para complementar la vigilancia de calidad del agua, especialmente considerando sus ventajas de medición continua, bajo costo operacional y capacidad de integración en sistemas de monitoreo en tiempo real [10].

## 2.2. Métodos Tradicionales de Monitoreo

El monitoreo convencional de calidad del agua se basa en análisis de laboratorio que, aunque proporcionan alta precisión y especificidad, presentan limitaciones operacionales significativas. Estos métodos requieren personal especializado, son intensivos en costos, utilizan reactivos químicos y requieren varios días para obtener resultados, sin capacidad de monitoreo en tiempo real [2].

Los sistemas de monitoreo convencionales operan con frecuencias limitadas de muestreo [1], lo que puede limitar la detección oportuna de eventos de contaminación de corta duración. Estas limitaciones operacionales han motivado el interés en desarrollar tecnologías complementarias que permitan mayor frecuencia de monitoreo y capacidad de respuesta más rápida.

### Fundamentos Matemáticos y Pertinencia Específica

#### Support Vector Machines para Datos Espectrales

El fundamento matemático de SVM reside en la búsqueda del hiperplano óptimo que maximiza el margen entre clases en el espacio de características. Para datos espectrales, esto se formaliza como:

$$f(x) = \text{sgn} \left( \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \right)$$

donde  $K(x_i, x)$  representa la función kernel que permite proyectar los espectros a espacios de mayor dimensionalidad.

**Pertinencia para Espectroscopía UV-Vis** Los datos espectrales presentan características que hacen a SVM particularmente apropiado:

1. **Alta dimensionalidad:** Los espectros contienen 200+ bandas, creando espacios de características de alta dimensión donde SVM excela.
2. **Separabilidad no lineal:** Las firmas espetrales de diferentes contaminantes pueden no ser linealmente separables en el espacio original, pero sí en espacios transformados por kernels RBF o polinomiales.
3. **Robustez al ruido:** La formulación de margen suave de SVM maneja efectivamente el ruido instrumental típico en espectroscopía de campo.

### XGBoost para Variables Correlacionadas

XGBoost implementa gradient boosting mediante la optimización iterativa:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

**Pertinencia para Análisis Espectral** La arquitectura de árboles de decisión de XGBoost maneja naturalmente:

1. **Correlaciones espetrales:** Las bandas adyacentes están naturalmente correlacionadas; los árboles capturan estas dependencias sin requerir decorrelación previa.
2. **Importancia de características:** Proporciona rankings interpretables de bandas espetrales más discriminativas.
3. **Manejo de valores atípicos:** La segmentación por árboles es robusta frente a picos espetrales anómalos.

### LSTM para Secuencias Espectrales

Las redes LSTM procesan secuencias mediante gates que controlan el flujo de información:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

**Pertinencia para Firmas Espectrales** Aunque los espectros son mediciones instantáneas, la estructura secuencial por longitud de onda presenta ventajas:

1. **Dependencias espectrales:** Los valores de absorción en diferentes longitudes de onda no son independientes, sino que siguen patrones físicos de transiciones electrónicas.
2. **Características de forma:** LSTM puede aprender patrones como pendientes, valles y picos que son diagnósticos para contaminantes específicos.
3. **Memoria selectiva:** Los gates permiten “recordar” características espectrales relevantes mientras ignoran ruido instrumental.

### 2.3. Aplicación de inteligencia artificial

El aprendizaje automático ha demostrado potencial significativo para el análisis de datos espectrales, particularmente debido a su capacidad para reconocer patrones complejos en grandes volúmenes de información [10]. Esta investigación emplea tres algoritmos seleccionados por sus características complementarias para el análisis de datos espectrales:

**Support Vector Machines (SVM)** Algoritmos de clasificación especializados en encontrar fronteras óptimas entre clases en espacios de alta dimensionalidad. Resultan particularmente adecuados para el análisis de espectros multivariados debido a su robustez frente al ruido y su capacidad de generalización [10].

**XGBoost** Técnica avanzada de boosting fundamentada en árboles de decisión. Su aplicación a espectroscopía UV-Vis ha demostrado eficacia excepcional para identificación de fuentes de agua mediante análisis de espectros de absorción [11] y detección cuantitativa de metales pesados en aguas residuales [12]. Esta técnica destaca por su robustez frente a ruido espectral y capacidad de manejar variables altamente correlacionadas, condiciones típicas en datos espectrales de campo.

**Long Short-Term Memory (LSTM)** Arquitectura de redes neuronales recurrentes diseñada para procesar secuencias complejas. En aplicaciones de calidad de agua, estos modelos han demostrado capacidad para analizar espectros UV-Vis y predecir concentraciones de nutrientes como nitrato con alta precisión [13].

La aplicación de estos algoritmos a datos espectrales del conjunto de datos de Lechevallier et al. (2024) permite desarrollar modelos de clasificación automática para distinguir entre condiciones normales y eventos de contaminación [9]. La efectividad de este enfoque depende de la calidad de los datos espectrales de entrada y la representatividad del conjunto de entrenamiento.

## 2.4. Comparación entre Métodos

Para contextualizar las ventajas del enfoque propuesto, se presenta a continuación una comparación sistemática entre los métodos tradicionales de monitoreo y las técnicas basadas en inteligencia artificial combinada con espectroscopía:

**Tabla 2.1.** Comparación entre enfoques de monitoreo según literatura

| Criterio de Evaluación     | Métodos Tradicionales | Espectroscopía + IA |
|----------------------------|-----------------------|---------------------|
| <b>Tiempo de respuesta</b> | Varios días           | Respuesta rápida    |
| <b>Reactivos químicos</b>  | Requeridos            | No requeridos       |
| <b>Mantenimiento</b>       | Alto                  | Mínimo              |
| <b>Aplicación en campo</b> | Limitada              | Factible            |

Los métodos tradicionales de laboratorio presentan limitaciones operacionales significativas relacionadas con tiempos de análisis prolongados, dependencia de reactivos químicos y restricciones para implementación en campo, mientras que las técnicas basadas en espectroscopía UV-Vis e inteligencia artificial ofrecen ventajas en términos de respuesta rápida, operación libre de reactivos y factibilidad de monitoreo continuo. Fuentes: [1, 2, 10, 11]

## 2.5. Justificación del Enfoque Propuesto

La integración de espectroscopía UV-Vis con algoritmos de aprendizaje automático permite superar de manera efectiva las principales limitaciones identificadas en los métodos tradicionales de monitoreo. Esta combinación tecnológica ofrece la capacidad de adaptarse dinámicamente a las condiciones cambiantes del entorno, reducir significativamente los costos operacionales y escalar hacia el desarrollo de redes de monitoreo distribuidas y completamente autónomas.

La viabilidad de esta estrategia ha sido reconocida ampliamente en la literatura especializada como una solución prometedora para monitoreo ambiental continuo, especialmente en contextos caracterizados por infraestructura limitada o recursos técnicos restringidos [10, 14, 15]. Esta aproximación resulta especialmente relevante para países como Chile, donde la cobertura de monitoreo hídrico permanece insuficiente en numerosas regiones rurales y agrícolas.

Desde una perspectiva operacional, el sistema propuesto no solo promete mejorar la frecuencia y cobertura del monitoreo, sino que también puede contribuir significativamente a la detección temprana de eventos de contaminación, facilitando así una respuesta más rápida y efectiva por parte de las autoridades competentes.

## 2.6. Limitaciones y Desafíos

Como cualquier tecnología emergente, la integración de espectroscopía UV-Vis con inteligencia artificial presenta desafíos técnicos que deben considerarse en su implementación. Basándose en la literatura revisada, se identifican las siguientes áreas de atención prioritaria:

1. **Calibración y mantenimiento:** Los sistemas requieren procedimientos de calibración apropiados para aplicaciones de campo.
2. **Representatividad de datos:** La efectividad depende de la calidad y representatividad de los datos de entrenamiento.
3. **Interferencias técnicas:** Factores como ruido espectral y variaciones en la matriz pueden afectar el rendimiento.
4. **Contexto de aplicación:** Los resultados requieren interpretación específica según las características locales.

Estas consideraciones, documentadas en estudios previos sobre aplicaciones de espectroscopía e inteligencia artificial [10, 15], son abordadas sistemáticamente en la metodología propuesta.

Los conceptos y herramientas presentados en este capítulo constituyen el fundamento teórico sobre el cual se construye la metodología propuesta, cuyo diseño e implementación se describe en detalle en el Capítulo 4.

# Capítulo 3

## Estado del Arte

---

### 3.1. Monitoreo de la Calidad del Agua

Los sistemas oficiales de monitoreo en América Latina se caracterizan por presentar frecuencias de muestreo notablemente bajas, realizando en muchos casos monitoreo solo una o dos veces al año [16]. Esta problemática limita significativamente la capacidad de detectar eventos transitorios, como se ha documentado en diferentes contextos regionales [1].

Los sistemas de monitoreo regionales presentan limitaciones sistemáticas, con falta de enfoques sistemáticos y capacidad limitada para detectar episodios de contaminación o tendencias a largo plazo [16]. Esta situación es especialmente crítica para eventos de corta duración, que pueden pasar inadvertidos debido a la baja resolución temporal del monitoreo convencional.

Además, la cobertura geográfica limitada de los programas de monitoreo tradicionales deja vastas áreas sin vigilancia adecuada, especialmente en regiones rurales o de difícil acceso, donde la infraestructura de laboratorio es escasa o inexistente. Esta situación evidencia la necesidad de desarrollar soluciones tecnológicas complementarias que permitan un monitoreo continuo y la emisión de alertas tempranas ante cambios relevantes en la calidad del agua.

### **3.2. Trabajos Previos sobre el conjunto de datos Lechevallier et al. (2024)**

El conjunto de datos utilizado en esta investigación fue desarrollado y publicado por Lechevallier et al. (2024) en su estudio pionero sobre monitoreo de calidad de aguas residuales utilizando espectrofotometría UV-Vis [9]. Este trabajo representa el primer conjunto de datos abierto específicamente diseñado para aplicaciones de monitoreo continuo mediante sensores espectrales.

#### **Características del conjunto de datos Original**

El conjunto de datos comprende 5,801 imágenes hiperespectrales de aguas residuales crudas, capturadas durante una campaña experimental de 25 semanas (mayo-octubre 2023). Las principales características incluyen:

- Mediciones espectrales en el rango 400–800 nm (rango efectivo de procesamiento) con resolución de 2 nm
- Datos de laboratorio para 29 contaminantes diferentes
- Información ambiental complementaria (temperatura, pH, turbidez, flujo)
- Condiciones operacionales reales en canal abierto

#### **Enfoque de Esta Investigación**

Esta investigación utiliza el conjunto de datos de Lechevallier et al. (2024) como base para desarrollar y evaluar técnicas de aprendizaje automático aplicadas a la detección automatizada de contaminantes. El enfoque se diferencia del trabajo original al:

1. Implementar algoritmos de aprendizaje automático para clasificación automatizada
2. Desarrollar un flujo de procesamiento de dos etapas para detección y identificación
3. Evaluar la aplicabilidad para monitoreo en tiempo real
4. Establecer métricas específicas para validación operacional

Esta aplicación contribuye al campo del monitoreo automatizado mediante inteligencia artificial, aprovechando las características únicas del conjunto de datos público disponible.

### 3.3. Espectroscopía UV-Vis vs. Imágenes Hiperespectrales

La elección entre espectroscopía UV-Vis puntual e imágenes hiperespectrales representa una decisión estratégica crucial en el diseño de sistemas de monitoreo ambiental. Cada enfoque presenta ventajas y limitaciones específicas que determinan su aplicabilidad en diferentes contextos operacionales.

#### Ventajas de la Espectroscopía UV-Vis Puntual

La espectroscopía UV-Vis permite la captura de firmas espectrales puntuales mediante sensores portátiles que pueden desplegarse en terreno con costos relativamente bajos y alta velocidad de adquisición [17, 18]. Esta técnica ha ganado reconocimiento por su practicidad en entornos operacionales caracterizados por restricciones de infraestructura, mostrando particular eficacia para monitoreo continuo de calidad del agua [10].

Las principales ventajas operacionales incluyen:

- Menor complejidad instrumental y requisitos de mantenimiento
- Capacidad de operación autónoma en condiciones adversas
- Velocidades de adquisición compatibles con monitoreo en tiempo real
- Costos de implementación accesibles para redes distribuidas

#### Limitaciones de las Imágenes Hiperespectrales

Las imágenes hiperespectrales, aunque ofrecen una representación espacial detallada del medio acuático, presentan desafíos significativos para aplicaciones de monitoreo continuo. Estos sistemas requieren dispositivos considerablemente más complejos, mayor capacidad de almacenamiento y procesamiento computacional intensivo [2].

En el contexto específico de monitoreo de calidad de agua, la espectroscopía UV-Vis puntual emerge como una opción más viable para tareas de

detección continua en entornos de campo, especialmente cuando se considera la relación costo-beneficio y los requerimientos operacionales [10, 14].

### **3.4. Modelos de inteligencia artificial en Clasificación Espectral**

La aplicación de técnicas de inteligencia artificial al análisis espectral de calidad del agua ha experimentado un desarrollo acelerado durante la última década. Diversos estudios han evaluado sistemáticamente el desempeño de diferentes algoritmos aplicados a esta problemática específica.

#### **Hallazgos Principales de la Literatura**

Los resultados reportados en la literatura especializada revelan tendencias respecto a la aplicación de diferentes enfoques algorítmicos en espectroscopía:

**Espectroscopía UV-Vis combinada con aprendizaje automático** Ha demostrado ser efectiva para la detección de múltiples parámetros de calidad del agua, ofreciendo ventajas como análisis no invasivo y capacidad de monitoreo continuo [10].

**Algoritmos basados en árboles de decisión** XGBoost ha mostrado efectividad en aplicaciones de monitoreo hídrico con espectroscopía UV-Vis, demostrando robustez ante ruido espectral y capacidad para manejar variables correlacionadas [11, 12].

**Redes neuronales recurrentes** Long Short-Term Memory presenta potencial para el análisis de datos espetrales, como se ha documentado en aplicaciones de predicción de nutrientes [13].

#### **Análisis Comparativo de Algoritmos**

Basándose en aplicaciones documentadas, cada familia de algoritmos presenta características específicas:

- **XGBoost** ha demostrado efectividad en aplicaciones de identificación de fuentes de agua y detección de metales pesados mediante espectroscopía UV-Vis [11, 12]
- **Long Short-Term Memory** muestra capacidad para análisis de espectros UV-Vis en aplicaciones de predicción de nutrientes [13]

- SVM ha sido aplicado exitosamente en detección de contaminantes con espectroscopía UV-Vis [10, 14]

**Tabla 3.1.** Aplicaciones documentadas de modelos de IA en espectroscopía UV-Vis

| Modelo                        | Aplicación Documentada   | Fuente |
|-------------------------------|--|--------|
| <b>SVM</b>                    | Detección de contaminantes orgánicos en monitoreo continuo       | [14]   |
| <b>XGBoost</b>                | Identificación de fuentes de agua mediante espectros UV-Vis      | [11]   |
| <b>Long Short-Term Memory</b> | Predicción de concentraciones de nutrientes con espectros UV-Vis | [13]   |

### 3.5. Brechas Identificadas en el Estado Actual

El análisis de la literatura especializada revela áreas de oportunidad que motivan el enfoque propuesto en esta investigación:

#### Oportunidades en Arquitecturas de Modelos

Los enfoques documentados en la literatura utilizan generalmente modelos individuales para clasificación espectral [2, 10]. Existe oportunidad para explorar flujo de procesamientos secuenciales que combinen detección binaria y clasificación multiclas para optimizar eficiencia computacional y precisión.

#### Potencial de Long Short-Term Memory en Datos Espectrales

Aunque se ha documentado la aplicación de Long Short-Term Memory para análisis de espectros UV-Vis en predicción de nutrientes [13], el uso de estas arquitecturas para análisis de patrones temporales en datos espectrales puntuales representa un área con potencial de desarrollo.

#### Sistemas Adaptativos

Existe oportunidad para desarrollar sistemas que se adapten a la calidad variable de conjuntos de datos y a las características específicas de diferentes

## ESTADO DEL ARTE

tipos de contaminantes, mejorando la robustez y aplicabilidad práctica.

Estas oportunidades identificadas motivan el enfoque propuesto en esta tesis, que explora la combinación de monitoreo espectral, detección automática de eventos y clasificación multietapa usando modelos de inteligencia artificial seleccionados según las características del problema.

# Capítulo 4

## Metodología

---

Este capítulo describe la metodología integral diseñada para detectar tempranamente contaminantes en aguas superficiales mediante el uso combinado de espectroscopía de reflectancia en el rango UV-Vis y algoritmos de inteligencia artificial. El sistema se estructura como un pipeline modular compuesto por cinco etapas: (1) caracterización de contaminantes y selección del dataset, (2) preprocesamiento espectral y control de calidad, (3) extracción y selección de características relevantes, (4) entrenamiento jerárquico de modelos clasificadores y (5) validación cruzada temporal con criterios estrictos de aceptación. Cada componente del pipeline ha sido optimizado para asegurar robustez, evitar fuga de información (*data leakage*) y garantizar la adaptabilidad a distintos tipos de contaminantes. A continuación, se detallan los elementos metodológicos que componen este sistema.

### Contaminantes Evaluados

El sistema propuesto fue evaluado utilizando el conjunto completo de 29 contaminantes documentados en el conjunto de datos de Lechevallier et al. (2024) [9], distribuidos según las categorías presentadas en la Tabla 1.1. Las muestras fueron analizadas mediante espectrofotometría UV-Vis y cromatografía líquida de alta resolución acoplada a espectrometría de masas (LC-HRMS/MS).

## METODOLOGÍA

Los contaminantes evaluados incluyen:

- **Parámetros fisicoquímicos convencionales (9):** Carbono orgánico disuelto (DOC), fosfato ( $PO_4$ ), amonio ( $NH_4$ ), nitrógeno total (NTOT), nitrógeno disuelto (NSOL), sólidos suspendidos totales (TSS), turbidez (NTU), sulfato ( $SO_4$ ), y carbono orgánico total (TOC).
- **Compuestos orgánicos emergentes (20):** Incluyendo fármacos (cafeína, citalopram, diclofenaco, candesartán, hidroclorotiazida), pesticidas (diuron, carbendazim, MCPA, mecoprop, 2,4-D), biocidas (triclosán, OIT), inhibidores de corrosión (benzotriazol, 4-&5-metilbenzotriazol), y compuestos industriales emergentes (HMMM, 6PPD-quinona, 1,3-difenilguanidina, DEET, acesulfame, ciclamato).

Esta selección abarca una diversidad química suficiente para validar la efectividad del sistema en la detección de patrones espectrales complejos, estableciendo una base metodológica adaptable a escenarios locales mediante calibración posterior.

### 4.1. Enfoque Experimental y Computacional

Esta investigación adopta un enfoque metodológico que integra técnicas de espectroscopía UV-Vis con algoritmos de aprendizaje automático para abordar el problema de detección de contaminantes. El fundamento conceptual se basa en la hipótesis de que diferentes compuestos químicos presentan firmas espectrales distintivas, detectables mediante espectroscopía UV-Vis y reconocibles automáticamente por modelos de clasificación apropiadamente entrenados.

La arquitectura del sistema se estructura en dos fases complementarias que operan de manera secuencial: una primera etapa de detección binaria para identificar eventos de contaminación, seguida de una etapa de clasificación multiclasa para determinar el tipo específico de contaminante presente. Esta aproximación jerárquica permite optimizar la eficiencia computacional utilizando modelos binarios como filtro inicial, aplicando posteriormente clasificadores especializados únicamente cuando se detecta una anomalía.

## **Selección y Justificación del conjunto de datos**

Esta investigación utiliza exclusivamente el conjunto de datos público desarrollado por Lechevallier et al. (2024) [9], que representa una contribución significativa para aplicaciones de monitoreo automatizado mediante espectrofotometría UV-Vis.

Los criterios de selección incluyeron:

**Diversidad de contaminantes** 29 compuestos diferentes que abarcan desde parámetros fisicoquímicos hasta compuestos orgánicos emergentes

**Calidad espectral** Mediciones estandarizadas en el rango 400–800 nm con resolución de 2 nm

**Escala temporal** Campaña de 25 semanas con 5,801 mediciones espetrales

**Condiciones reales** Datos obtenidos en condiciones operacionales documentadas

**Reproducibilidad** Formato estándar ENVI y disponibilidad pública verificada

## **Contribución Metodológica**

Esta investigación contribuye al desarrollo de metodologías de inteligencia artificial para análisis de datos espetrales. El trabajo se diferencia del estudio original de Lechevallier et al. (2024), que se enfocó en la validación técnica del sistema de adquisición, mediante:

- Implementación de flujo de procesamiento de dos etapas (binaria + multiclase)
- Aplicación de algoritmos de aprendizaje automático (Support Vector Machines, XGBoost, Long Short-Term Memory) para clasificación automatizada
- Validación específica para aplicaciones de monitoreo continuo
- Desarrollo de criterios de evaluación para robustez operacional

## 4.2. Metodología de Preprocesamiento Espectral

El preprocesamiento de datos espectrales constituye una etapa crítica que determina la calidad de los modelos de clasificación subsiguientes. La metodología se basa en fundamentos establecidos de análisis de firmas espectrales [19] y técnicas modernas de procesamiento de imágenes adaptadas para aplicaciones de aprendizaje automático [20].

### Estrategia de Procesamiento

La estrategia de preprocesamiento se estructura en cuatro etapas principales, adaptadas a las características del conjunto de datos de Lechevallier et al. (2024), que proporciona datos espectrales ya calibrados y en formato estándar ENVI [9]:

1. **Carga y validación:** Lectura de datos en formato ENVI y verificación de integridad
2. **Filtrado de calidad:** Eliminación de mediciones con datos faltantes o inconsistentes
3. **Extracción de espectros:** Generación de firmas espectrales representativas para cada muestra
4. **Preparación para aprendizaje automático:** Normalización y estructuración según el algoritmo de destino

### Procesamiento de Datos Espectrales

Los datos del conjunto de datos de Lechevallier se proporcionan como imágenes hiperespectrales calibradas en el rango 400–800 nm. El procesamiento incluye:

- Extracción de espectros promedio por región de interés
- Filtrado de longitudes de onda con ruido excesivo
- Normalización espectral para comparabilidad entre muestras
- Sincronización temporal con mediciones de laboratorio

## Selección de Características Espectrales

El sistema implementa dos estrategias diferenciadas de preparación de datos según las características específicas de cada algoritmo de machine learning:

**Estrategia para Modelos Clásicos (SVM, XGBoost)** Para estos algoritmos se desarrolló un conjunto de 84 características espectrales interpretables, organizadas en seis categorías principales:

- **Estadísticas básicas:** Media, mediana, desviación estándar por regiones espectrales
- **Índices espectrales establecidos:** Turbidez, CDOM, índices de materia orgánica
- **Características de picos:** Detección y cuantificación de máximos locales
- **Relaciones espectrales:** Ratios entre zonas UV, visible e infrarrojo cercano
- **Derivadas espectrales:** Primera y segunda derivada para detectar pendientes
- **Patrones de forma:** Curvatura, ancho de picos, simetría espectral

La selección final de características se realizó mediante un sistema automático de ranking que combina tres métricas complementarias: mutual information (dependencia no lineal), importancia por Random Forest (relevancia predictiva) y correlación de Pearson (relación lineal). Este enfoque multi-métrica optimiza tanto interpretabilidad física como rendimiento predictivo.

**Estrategia para Redes LSTM** Para preservar las dependencias secuenciales naturales entre longitudes de onda adyacentes, se mantiene la secuencia espectral completa sin reducción dimensional. Los datos se estructuran como series temporales donde cada banda espectral representa un paso temporal secuencial, permitiendo que la red capture patrones de absorción que dependen del contexto espectral circundante.

La Tabla C.1 del Anexo C presenta las cinco características más importantes identificadas por el sistema de ranking, donde *UV\_Peak\_280nm*

## METODOLOGÍA

(importancia 0,147) y *Turbidity\_Index* (0,132) emergen como los indicadores más discriminativos para la detección de contaminantes.

### Control de Calidad

Se implementó un sistema dual de control de calidad que opera en dos niveles complementarios: evaluación de datasets individuales y clasificación automática de calidad por contaminante.

**Control de Calidad de Datos Espectrales** Para cada dataset individual se aplicaron verificaciones específicas:

- **Compleitud espectral:** Verificación de ausencia de bandas faltantes o corrompidas en el rango 400-800 nm
- **Detección de outliers:** Identificación automática de mediciones con valores espectrales anómalos usando criterios estadísticos (IQR y Z-score)
- **Validación cruzada:** Correlación entre firmas espectrales y concentraciones medidas por LC-HRMS/MS para verificar consistencia
- **Filtrado de calidad:** Eliminación de muestras con ruido excesivo o inconsistencias que comprometían la confiabilidad

**Sistema de Clasificación Automática de Calidad** Adicionalmente, se desarrolló un sistema que evalúa cada dataset de contaminante con una puntuación de 0-100 basada en cinco criterios cuantitativos:

- **Varianza espectral:** Separabilidad entre clases alta/baja concentración
- **Balance de clases:** Distribución equilibrada de muestras positivas/negativas
- **Cobertura temporal:** Representatividad a lo largo del período de estudio
- **Estabilidad instrumental:** Consistencia de mediciones en el tiempo
- **Coherencia de etiquetas:** Ausencia de inconsistencias en clasificación

Esta evaluación multidimensional permite clasificar automáticamente cada dataset en cinco categorías (Excelente, Buena, Regular, Pobre, Problemática) y ajustar dinámicamente las estrategias de preprocesamiento, regularización y selección de algoritmos según la puntuación obtenida.

Los criterios específicos de puntuación y las recomendaciones de procesamiento por categoría se detallan en el Anexo A. La implementación técnica de ambos sistemas se describe en el Capítulo 5.

## Fundamento Espectral del Sistema

El dataset de Lechevallier fue adquirido utilizando un sistema de imágenes hiperespectrales MV.X, diseñado específicamente para monitoreo de calidad del agua. Este sistema opera en el rango visible-infrarrojo cercano de 400–800 nm con resolución espectral de 2,0 nm FWHM.

La selección de este rango espectral se fundamenta en que captura las transiciones electrónicas más relevantes para contaminantes orgánicos (absorción UV 280-400 nm) y las propiedades ópticas de parámetros fisicoquímicos como turbidez y materia orgánica disuelta (región visible 400-700 nm). Las imágenes se almacenan en formato ENVI (.hdr/.bin), facilitando su procesamiento con bibliotecas especializadas.

Las especificaciones completas del sistema se detallan en el Anexo D.

### 4.3. Diseño del flujo de procesamiento de Clasificación

El diseño del sistema de clasificación se fundamenta en una arquitectura modular de dos etapas que optimiza tanto la eficiencia computacional como la precisión de detección. Esta aproximación permite abordar de manera sistemática la complejidad inherente al problema de clasificación multicon-taminante.

#### Arquitectura Jerárquica de Clasificación

La estrategia de clasificación jerárquica implementada opera mediante dos niveles secuenciales:

**Nivel 1 - Detección Binaria** Identificación de eventos anómalos versus condiciones normales del agua, funcionando como filtro de primera

## METODOLOGÍA

instancia

**Nivel 2 - Clasificación Multiclasé** Determinación del tipo específico de contaminante entre 29 categorías predefinidas, activado únicamente tras detección positiva

Esta aproximación reduce significativamente el costo computacional y minimiza falsos positivos al concentrar la capacidad de procesamiento en muestras identificadas como potencialmente contaminadas.

### Adaptación por Tipo de Algoritmo

La metodología reconoce las diferencias fundamentales entre algoritmos y adapta la estructura de datos según sus requerimientos específicos:

#### Modelos Estáticos (Support Vector Machines, XGBoost)

- Representación matricial bidimensional (muestras × características)
- Reducción dimensional basada en importancia de características
- Normalización por característica para garantizar comparabilidad

#### Modelos Secuenciales (Long Short-Term Memory)

- Estructura tensorial tridimensional preservando secuencias espectrales
- Mantenimiento del orden espectral natural (400-800 nm)
- Procesamiento secuencial por longitud de onda para capturar dependencias entre bandas adyacentes

### Estrategia de Entrenamiento y Validación

Cada modelo se entrena y evalúa de manera independiente por contaminante, permitiendo:

- Comparación objetiva de desempeño individual
- Detección temprana de sobreajuste específico por clase
- Estimación de intervalos de confianza por tipo de contaminante
- Evaluación de viabilidad operacional bajo condiciones reales

#### 4.4. Metodología de Validación y Criterios de Aceptación

La robustez del sistema propuesto requiere una metodología de validación que simule fielmente las condiciones de operación en terreno. Para ello, se implementó un protocolo de validación temporal estricta que evita la fuga de información y asegura la capacidad de generalización prospectiva.

##### **Validación Cruzada Temporal**

La validación temporal implementada preserva la estructura cronológica de los datos, simulando condiciones realistas de despliegue donde los modelos deben predecir eventos futuros basándose en información histórica. Esta aproximación es crítica para detectar sobreajuste temporal y evaluar la estabilidad del sistema ante variaciones estacionales.

##### **Métricas de Evaluación Multidimensional**

El desempeño de cada modelo se evalúa mediante un conjunto comprensivo de métricas que capturan diferentes aspectos de la calidad predictiva:

**Area Under the Curve (AUC)** Capacidad de discriminación entre clases independiente del umbral de decisión

**F1-score** Balance entre precisión y sensibilidad, crítico para aplicaciones de detección

**Brecha de Generalización** Diferencia entre rendimiento en entrenamiento y prueba, indicador de sobreajuste

**Estabilidad Temporal** Consistencia de predicciones a lo largo del tiempo

## Criterios de Filtrado Estrictos

Solo se consideran aceptables aquellos modelos que satisfacen simultáneamente todos los criterios establecidos en la Tabla 4.1. Estos umbrales fueron establecidos mediante evaluación iterativa de múltiples configuraciones en un conjunto de validación independiente, balanceando la precisión requerida para detección confiable (minimizar falsos negativos) con la robustez necesaria para evitar falsos positivos en condiciones variables de campo.

**Tabla 4.1.** Criterios de filtrado para validación de modelos

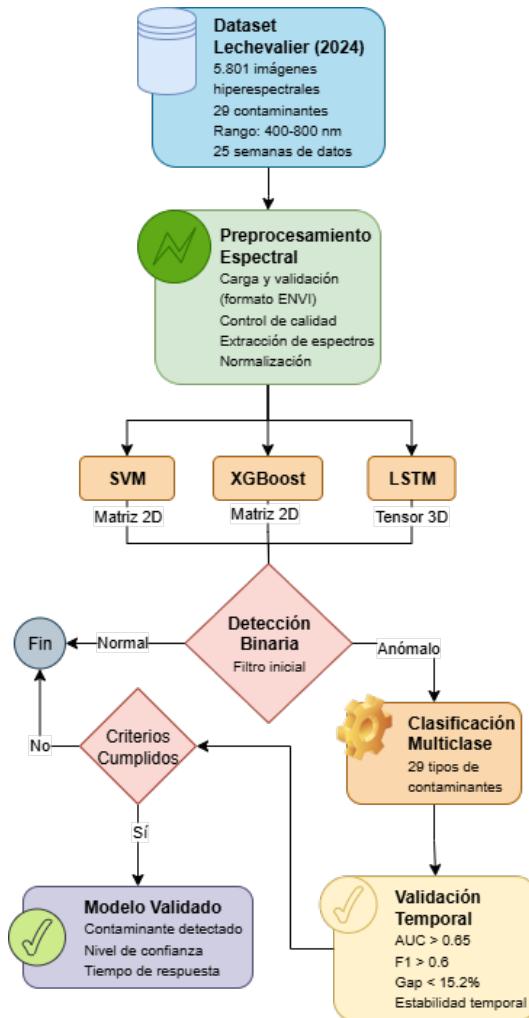
| Criterio de Evaluación               | Umbral Mínimo |
|--------------------------------------|---------------|
| <b>AUC (Área bajo la curva)</b>      | 0,65          |
| <b>F1-score</b>                      | 0,60          |
| <b>Brecha F1-score (train/test)</b>  | <15 %         |
| <b>Brecha exactitud (train/test)</b> | <20 %         |

Vale la pena señalar que estos umbrales fueron calibrados empíricamente después de evaluar múltiples configuraciones, ya que los valores inicialmente propuestos resultaron demasiado permisivos para garantizar robustez operacional.

El protocolo utilizado está basado en validación temporal estricta mediante *forward chaining*, donde se entrena con ventanas cronológicas crecientes respetando el orden temporal. Además, se introduce un *gap* de al menos 24 horas entre conjuntos de entrenamiento y prueba, con el objetivo de minimizar correlaciones falsas o artificiales. Esta estrategia garantiza una evaluación robusta y prospectiva, detallada en el Anexo C.

## 4.5. Resumen de la Metodología

En este trabajo se integra de manera sistemática técnicas de espectroscopía avanzada con algoritmos de inteligencia artificial de última generación. La Figura 4.1 ilustra el flujo metodológico completo, desde el conjunto de datos de Lechevallier et al. (2024) hasta la validación final del sistema, mostrando la arquitectura jerárquica de clasificación de dos etapas y los criterios estrictos de validación temporal implementados.



**Figura 4.1.** Arquitectura metodológica del sistema propuesto. Procesamiento del conjunto de datos de Lechevallier et al. (2024) mediante tres algoritmos de aprendizaje automático con clasificación jerárquica de dos etapas y validación temporal estricta.

El enfoque se caracteriza por tres elementos distintivos: (1) separación clara entre etapas de detección y clasificación que optimiza la eficiencia computacional.

## METODOLOGÍA

tacional, (2) estructuras de datos adaptadas específicamente a los requerimientos de cada familia de algoritmos, y (3) implementación de controles de calidad automatizados que aseguran confiabilidad práctica.

La metodología establecida proporciona el marco conceptual y técnico para la implementación práctica del sistema, cuyos aspectos específicos se detallan en el Capítulo 5, seguido del análisis comprensivo de resultados presentado en el Capítulo 6.

# Capítulo 5

## Implementación

---

El desarrollo del sistema se llevó a cabo mediante un proceso iterativo y adaptativo, incorporando mejoras progresivas conforme se identificaban limitaciones en versiones anteriores del flujo de procesamiento y se profundizaba en la comprensión del comportamiento de los datos espectrales. Esta aproximación permitió construir un sistema robusto y flexible, capaz de ajustarse automáticamente a las características del conjunto de datos disponible.

### 5.1. Arquitectura General del Sistema

El sistema desarrollado procesa datos hiperespectrales del conjunto de datos de Lechevallier et al. (2024) [9] siguiendo un flujo estructurado que abarca desde la lectura de archivos en formato ENVI hasta la generación de predicciones confiables. Los datos espectrales, calibrados en el rango 400–800 nm, requieren adaptaciones específicas antes de ser utilizados efectivamente por los modelos de aprendizaje automático.

El procesamiento se basa en principios establecidos de análisis espectral [20], adaptados específicamente para el contexto de monitoreo de calidad del agua. Las etapas principales incluyen extracción de regiones de interés, normalización adaptativa, filtrado por calidad de datos y estructuración según el algoritmo de destino.

Una característica distintiva del sistema es su capacidad para estructurar los datos en dos formatos paralelos optimizados según el tipo de algoritmo.

## IMPLEMENTACIÓN

Para modelos clásicos como SVM y XGBoost, se genera una matriz tabular donde cada fila representa una muestra y cada columna una banda espectral. Para redes Long Short-Term Memory, se preserva la estructura secuencial, manteniendo el orden natural de las longitudes de onda.

Esta dualidad estructural permite implementar eficientemente el esquema de clasificación jerárquica propuesto: detección binaria inicial para identificar eventos anómalos, seguida de clasificación multiclas para determinar el tipo específico de contaminante presente.

El sistema fue implementado en Python, utilizando bibliotecas como NumPy, Pandas, scikit-learn, XGBoost y TensorFlow. El flujo completo fue automatizado para permitir el procesamiento de múltiples contaminantes de forma secuencial. Al detectar un nuevo contaminante, el sistema genera los conjuntos de datos requeridos, aplica validaciones de calidad y estructura los archivos para su posterior uso por los modelos.

## 5.2. Metodología de Construcción de conjuntos de datos

El sistema genera automáticamente conjuntos de datos individualizados para cada contaminante, adaptando su configuración según múltiples factores que incluyen el tipo de sustancia analizada, el balance observado entre clases y la calidad de los datos espectrales disponibles.

Para cada contaminante procesado, se generan simultáneamente estructuras de datos optimizadas para diferentes familias de algoritmos. La matriz tabular, destinada a modelos SVM y XGBoost, organiza las características espectrales en formato rectangular estándar con etiquetas binarias o multiclas según corresponda. Paralelamente, se construye una representación secuencial para Long Short-Term Memory que preserva las correlaciones naturales entre bandas espectrales adyacentes.

La división de datos implementa estrategias de validación temporal para preservar la estructura cronológica del conjunto de datos y evitar fuga de información. Esta aproximación resulta especialmente importante para evaluar la capacidad de generalización prospectiva del sistema en condiciones operacionales reales.

## Sistema de Evaluación de Calidad

Cada conjunto de datos generado es sometido a una evaluación integral de calidad que considera múltiples dimensiones críticas para el éxito del entrenamiento. El sistema analiza el balance entre clases, la varianza espectral por banda, la presencia de columnas constantes o altamente redundantes, la detección automática de valores atípicos y la identificación de posibles inconsistencias en las etiquetas.

Basándose en estos análisis, se implementó un sistema de clasificación de calidad que guía automáticamente las estrategias de procesamiento según las características del conjunto de datos. Esta clasificación proporciona una evaluación objetiva de la viabilidad del conjunto y orienta decisiones sobre estrategias de regularización y selección de algoritmos específicos.

La clasificación automática de calidad implementada (detallada en el Anexo A) permitió categorizar los datasets en cinco niveles, desde ‘Exce-lente’ (85–88 puntos) hasta ‘Problemática’ (<50 puntos), con estrategias de procesamiento específicas para cada categoría. Este sistema de evaluación resultó crítico para identificar los 21 datasets válidos de los 29 analizados, descartando 4 conjuntos que no cumplían los criterios mínimos para apren-dizaje automático.

**Tabla 5.1.** Estrategias de procesamiento según características del conjunto de datos

| Características            | Tamaño de muestra | Estrategia de proce-samiento                       |
|----------------------------|-------------------|--|
| <b>Alta calidad</b>        | >100 muestras     | Optimización con hi-perparámetros complejos        |
| <b>Calidad moderada</b>    | 50–100 muestras   | Configuración balan-cieada con regulariza-ción     |
| <b>Baja calidad</b>        | 20–50 muestras    | Regularización aumen-tada y modelos simpli-ficados |
| <b>Datos insuficientes</b> | <20 muestras      | Evaluación para viabi-lidad del análisis           |

### Protocolo de Validación Temporal

Para garantizar la robustez temporal del sistema y prevenir la fuga de información, se desarrolló un protocolo de validación cronológica específicamente adaptado a las características del conjunto de datos de Lechevallier et al. (2024) [9].

El sistema preserva la integridad temporal mediante particionado cronológico estricto, donde los conjuntos de entrenamiento utilizan exclusivamente datos anteriores y la evaluación se realiza sobre datos posteriores, simulando condiciones operacionales reales. Esta aproximación es consistente con metodologías establecidas para validación de series temporales en espectroscopía [20].

Para modelos Long Short-Term Memory, se implementan ventanas temporales que preservan la secuencia natural de adquisición espectral. El protocolo incluye separación temporal entre conjuntos de entrenamiento y validación para eliminar correlaciones espurias que podrían inflar artificialmente las métricas de rendimiento.

La validación utiliza un enfoque de forward chaining donde el modelo se entrena progresivamente con ventanas temporales crecientes y se evalúa en períodos subsiguientes. Esta aproximación demostró ser efectiva para detectar sobreajuste temporal que puede pasar inadvertido en validaciones convencionales.

La metodología detallada del protocolo de validación temporal se describe en el Anexo C, donde se especifican los procedimientos de particionado cronológico, las ventanas temporales empleadas y los umbrales de aceptación implementados para cada tipo de modelo.

### 5.3. Ingeniería de Características Espectrales Avanzada

Una de las contribuciones metodológicas de este trabajo consiste en el desarrollo de un sistema de ingeniería de características espectrales que se adapta automáticamente según la calidad y naturaleza específica de cada conjunto de datos, optimizando la extracción de información discriminativa relevante para la detección de contaminantes.

## Selección Adaptativa de Métodos de Normalización

El sistema implementa un mecanismo de selección inteligente de escaladores basado en principios establecidos de análisis espectral [20]. La detección automática se fundamenta en un análisis estadístico del rango intercuartílico para identificar valores atípicos en los datos espectrales.

Se desarrollaron empíricamente criterios de decisión donde conjuntos de datos con alta presencia de valores atípicos o elevada variabilidad espectral utilizan RobustScaler, que emplea mediana y rango intercuartílico en lugar de media y desviación estándar. Para datos con distribución aproximadamente normal, se emplea StandardScaler convencional.

Esta decisión adaptativa mostró mejora en la estabilidad de normalización comparada con enfoques que utilizan escaladores fijos, demostrando la importancia de adaptar las técnicas de preprocesamiento a las características específicas de cada conjunto de datos.

## Extracción Sistemática de Características Interpretables

Se desarrolló un sistema que genera automáticamente un conjunto extenso de características espectrales interpretables, organizadas en categorías que capturan diferentes aspectos de la información espectral relevante para detección de contaminantes.

Las características incluyen índices espectrales adaptados para calidad de agua, análisis por rangos espectrales específicos, y características de forma espectral. El análisis por rangos aprovecha que diferentes contaminantes presentan absorción característica en regiones específicas del espectro electromagnético, consistente con principios fundamentales de espectroscopía [19].

Las características de forma espectral incluyen análisis de derivadas espectrales, análisis de curvatura y detección de máximos locales, técnicas establecidas en análisis espectral para caracterizar firmas espectrales distintivas.

El sistema completo de características espectrales desarrollado se detalla en el Anexo C, donde se incluyen los tops 5 características más importantes: *UV\_Peak\_280nm* (importancia 0,147), *Turbidity\_Index* (0,132), *Peak\_Heights\_Mean* (0,109), *NIR\_Mean* (0,098) y *CDOM\_Index* (0,087), demostrando la efectividad de combinar índices específicos, medidas de forma espectral y análisis por rangos espectrales.

## Estrategia de Reducción Dimensional Adaptativa

Para optimizar el balance entre preservación de información y eficiencia computacional, se implementó un sistema de reducción dimensional que adapta su estrategia según las características del conjunto de datos disponible.

conjuntos de datos con mayor cantidad de muestras preservan un conjunto más amplio de características espectrales para maximizar la información disponible. Para conjuntos de datos con limitaciones en el número de muestras, se aplica reducción automática utilizando SelectKBest, que retiene las características con mayor capacidad discriminativa según análisis univariado.

El proceso selecciona características según su capacidad de separación entre clases, conservando aquellas con mayor poder predictivo y evitando la inclusión de información redundante. Esta aproximación mejora la estabilidad del modelo en conjuntos de datos pequeños, mitigando problemas asociados con alta dimensionalidad.

## 5.4. Análisis de Viabilidad Espectral

El análisis de los 29 contaminantes incluidos en el conjunto de datos de Lechevallier et al. (2024) [9] reveló patrones que determinan la viabilidad para aplicaciones de aprendizaje automático, basándose en las características espectrales observadas para cada tipo de sustancia química.

### Contaminantes con Firmas Espectrales Distintivas

Los parámetros fisicoquímicos convencionales como carbono orgánico disuelto, fosfatos y turbidez presentan absorción característica en múltiples rangos espectrales con patrones reproducibles. La turbidez exhibe dispersión específica relacionada con partículas en suspensión, mientras que el carbono orgánico disuelto presenta absorción UV asociada con material orgánico aromático.

Los compuestos farmacéuticos aromáticos como diclofenac y benzotriazol muestran patrones de absorción específicos en el rango UV-Vis, consistente con transiciones electrónicas características de sistemas  $\pi$  conjugados [19]. El análisis de estos compuestos en el conjunto de datos reveló:

- Absorción UV específica con patrones diferenciables

- Reproducibilidad espectral observada en mediciones temporales
- Especificidad estructural donde diferentes estructuras aromáticas generan firmas spectrales distinguibles

### **Limitaciones para Detección Espectral**

El análisis identificó limitaciones principales que afectan la detectabilidad espectral de ciertos contaminantes:

Las concentraciones bajas constituyen un factor limitante donde contaminantes presentes por debajo de umbrales de detectabilidad espectral producen señales que se confunden con el ruido de fondo instrumental. La variabilidad de matriz representa otro factor significativo, donde cambios en la composición de fondo del agua afectan la línea base espectral.

Las interferencias spectrales, incluyendo solapamiento entre múltiples contaminantes con absorción en rangos similares y el enmascaramiento por alta turbidez, también comprometen la especificidad de detección.

Esta caracterización permitió identificar que los contaminantes orgánicos aromáticos y los parámetros fisicoquímicos presentan, en su mayoría, firmas spectrales suficientemente distintivas para aplicaciones de aprendizaje automático, mientras que algunos compuestos de baja concentración o sin características spectrales marcadas requieren enfoques metodológicos alternativos.

## **5.5. Estrategias Diferenciadas de Entrenamiento**

La estrategia de entrenamiento desarrollada aprovecha las características específicas de cada tipo de contaminante mediante un enfoque adaptativo que se ajusta automáticamente según las propiedades spectrales observadas y la calidad del conjunto de datos disponible.

### **Configuración por Tipo de Contaminante**

Para contaminantes farmacéuticos aromáticos que presentan absorción UV específica, se emplea kernel RBF en SVM combinado con análisis de derivadas spectrales y regularización que preserve la sensibilidad a características spectrales sutiles.

Los parámetros relacionados con turbidez, que exhiben patrones de dispersión característicos, se benefician de estrategias basadas en ratios inter-

banda y análisis de pendiente espectral que capturan la naturaleza de la dispersión por partículas.

Para contaminantes de materia orgánica con patrones de absorción característicos, se implementan estrategias que incluyen análisis temporal mediante Long Short-Term Memory y características que capturan la evolución gradual típica de estos compuestos.

**Tabla 5.2.** Estrategias de entrenamiento diferenciadas por categoría de contaminante

| Categoría                       | Características Espectrales | Estrategia Optimizada  |
|---------------------------------|-----------------------------|--|
| <b>Farmacéuticos aromáticos</b> | Absorción UV específica     | Kernel RBF, análisis de derivadas, regularización moderada             |
| <b>Parámetros de turbidez</b>   | Patrones de dispersión      | Ratios interbanda, análisis de pendiente, características de textura   |
| <b>Materia orgánica</b>         | Absorción UV y pendientes   | Análisis temporal Long Short-Term Memory, características de evolución |

Cabe destacar que esta categorización surgió del análisis iterativo de los resultados iniciales, donde se observó que aplicar la misma estrategia a todos los contaminantes producía resultados inconsistentes.

## 5.6. Evaluación y Validación de Modelos

Todos los modelos desarrollados fueron sometidos a evaluación multidimensional utilizando métricas que capturan diferentes aspectos críticos del rendimiento en aplicaciones de detección, incluyendo exactitud, F1-score, AUC y brecha de generalización.

Cuando se identificaron brechas de generalización significativas, se aplicaron medidas correctivas sistemáticas, incluyendo reducción de complejidad del modelo, incremento de regularización y reentrenamiento con conjuntos mejor balanceados.

### **Sistema de conjunto de modelos Espectral**

Para casos donde múltiples modelos alcanzaban calidad aceptable, se desarrolló un sistema de conjunto de modelos que considera tanto la precisión individual como la complementariedad en el análisis espectral. Los pesos se determinan evaluando qué rangos espectrales o características cada modelo utiliza más efectivamente.

Los criterios de inclusión en conjunto de modelos se adaptan según la calidad del conjunto de datos, implementando un sistema de umbrales que asegura la inclusión únicamente de modelos con rendimiento robusto. El sistema permite aprovechar las fortalezas específicas de cada algoritmo: SVM para patrones lineales, XGBoost para relaciones no lineales, y Long Short-Term Memory para patrones temporales.

La implementación desarrollada establece un marco técnico robusto y adaptable que facilita la evaluación comprehensiva del desempeño del sistema, aspectos que se abordan en detalle en el Capítulo 6.



# Capítulo 6

## Resultados y Análisis

---

Este capítulo presenta los resultados obtenidos por el sistema desarrollado para la detección de contaminantes en aguas superficiales, acompañados de un análisis comprensivo de las decisiones metodológicas implementadas durante el proceso de desarrollo. La evaluación se estructura distinguiendo entre detecciones exitosas y fallidas, considerando tanto la presencia real de contaminantes según análisis de laboratorio como la efectividad del sistema para identificar correctamente estos eventos. El análisis integra el desempeño de los modelos con la calidad intrínseca de los conjuntos de datos utilizados.

### 6.1. Evaluación Integral por Contaminante

La evaluación comprensiva abarcó 29 contaminantes diferentes, incluyendo tanto indicadores fisicoquímicos convencionales como turbidez y nitrógeno total, así como compuestos orgánicos, traza de relevancia emergente como diclofenac y benzotriazole. Para cada sustancia se desarrollaron conjuntos de datos específicamente adaptados para modelos clásicos de aprendizaje automático (SVM, XGBoost) y arquitecturas secuenciales (Long Short-Term Memory).

El análisis se fundamentó en dos criterios principales que determinaron la categorización de cada caso: la presencia real del contaminante, según confirmación analítica de laboratorio, y la capacidad del sistema para identificar correctamente esta presencia o ausencia. Esta aproximación permitió clasifi-

## RESULTADOS Y ANÁLISIS

car los resultados en cuatro categorías distintas que capturan la complejidad del problema de detección.

Los casos exitosos incluyen aquellos donde existía presencia real del contaminante y el sistema logró detectarlo efectivamente, así como situaciones donde la ausencia de contaminante fue correctamente identificada. Los casos problemáticos comprenden tanto falsos negativos (contaminante presente pero no detectado) como falsos positivos (detección errónea en ausencia del contaminante).

El análisis detallado de las firmas espectrales por contaminante se presenta en el Anexo E.2, donde se puede observar que contaminantes como hydrochlorothiazide y benzotriazole exhiben patrones espetrales claramente diferenciables (ver Figuras E.9 y E.3 en Anexos), mientras que otros como 6PPD-quinone y Acesulfame muestran firmas casi idénticas entre clases (Figuras E.5 y E.6 en Anexos), explicando las dificultades en su clasificación automatizada.

Entre los contaminantes que demostraron detección robusta destacaron fosfatos ( $PO_4$ ), amonio ( $NH_4$ ), benzotriazole e hydrochlorothiazide. Estas sustancias compartían características espetrales favorables, incluyendo firmas espetrales estables con alta relación señal-ruido, baja superposición espectral con otras clases de compuestos y consistencia temporal que facilitó tanto el proceso de aprendizaje como la capacidad de generalización de los modelos.

En contraste, contaminantes como diclofenac, citalopram y OIT, aunque inicialmente mostraron métricas prometedoras durante las fases de entrenamiento, exhibieron síntomas evidentes de sobreajuste o inconsistencias estructurales que comprometían su confiabilidad operacional, motivando su exclusión del sistema final.

### Análisis Comparativo por Algoritmo

La evaluación sistemática de los tres algoritmos implementados reveló diferencias significativas en sus capacidades de detección y adaptabilidad a diferentes tipos de contaminantes. El estudio abarcó un total de 69 modelos entrenados sobre diferentes contaminantes, de los cuales 14 alcanzaron los criterios estrictos establecidos para la implementación en condiciones de producción.

El rendimiento comparativo detallado por algoritmo se presenta en el Anexo B, donde se observa que XGBoost alcanzó la mayor tasa de éxito

(25,0 %) con un AUC promedio de 0,692, seguido por SVM (26,1 %) con el AUC promedio más elevado (0,761), mientras que Long Short-Term Memory mostró limitaciones con 17,4 % de tasa de éxito debido a su sensibilidad al drift temporal presente en los datos espectrales.

**Tabla 6.1.** Resultados detallados de modelos exitosos

| Contaminante                     | Algoritmo              | AUC   | F1-score | brecha (%) | Calidad |
|----------------------------------|------------------------|-------|----------|------------|---------|
| Diuron                           | XGBoost                | 0,972 | 0,955    | 4          | Buena   |
| Hydrochlorothiazide              | SVM                    | 0,887 | 0,635    | 12         | Buena   |
| Benzotriazole                    | Long Short-Term Memory | 0,829 | 0,625    | 6          | Buena   |
| <i>PO<sub>4</sub></i> (Fosfatos) | SVM                    | 0,797 | 0,910    | 12         | Buena   |
| Benzotriazole                    | SVM                    | 0,750 | 0,721    | 17         | Regular |
| Benzotriazole                    | XGBoost                | 0,725 | 0,659    | 1          | Buena   |
| Turbidez                         | Long Short-Term Memory | 0,688 | 0,549    | 13         | Buena   |
| <i>PO<sub>4</sub></i> (Fosfatos) | Long Short-Term Memory | 0,680 | 0,900    | 4          | Regular |

XGBoost emergió como el algoritmo más versátil, alcanzando la mayor tasa de éxito (25,0 %) con seis modelos que superaron los criterios de aceptación. Su fortaleza particular se manifestó en contaminantes caracterizados por patrones espectrales complejos y relaciones no lineales entre características, donde su capacidad inherente para manejar características heterogéneas y capturar interacciones sutiles entre variables espectrales resultó determinante.

SVM, aunque produjo una menor cantidad de modelos exitosos, alcanzó el AUC promedio más elevado (0,761), demostrando su robustez característica frente a conjuntos de datos pequeños y su superior capacidad de generalización. Esta fortaleza resultó especialmente evidente en contaminantes con firmas espectrales bien definidas y separables linealmente en espacios de alta dimensionalidad.

Long Short-Term Memory mostró un rendimiento competitivo con una tasa de éxito del 19,0 %, revelando particular efectividad en contaminantes que exhiben patrones temporales distintivos. Los resultados exitosos obtenidos con benzotriazole y turbidez ilustran la capacidad única de esta arquitectura para capturar dependencias temporales en la evolución de firmas espectrales.

El análisis completo de rendimiento por algoritmo, incluyendo métricas detalladas de accuracy promedio, distribución de modelos aceptables y observaciones sobre las fortalezas específicas de cada enfoque, se documenta en el Anexo B. Los casos de contaminantes que no alcanzaron criterios de aceptación y las razones técnicas de su exclusión por categoría química se

## RESULTADOS Y ANÁLISIS

detallan en la sección B.2 del mismo anexo, proporcionando una evaluación comprensiva de las limitaciones del sistema para diferentes tipos de sustancias químicas.

### Resultados Detallados de Modelos Exitosos

Los ocho mejores modelos que alcanzaron criterios de producción se presentan ordenados por rendimiento AUC, proporcionando una visión detallada de los casos más exitosos del sistema desarrollado.

**Tabla 6.2.** Resultados detallados de los mejores modelos con criterios de producción

| Contaminante                     | Algoritmo              | AUC   | F1-score | brecha (%) | Calidad |
|----------------------------------|------------------------|-------|----------|------------|---------|
| Diuron                           | XGBoost                | 0,972 | 0,955    | 4          | Buena   |
| Hydrochlorothiazide              | SVM                    | 0,887 | 0,635    | 12         | Buena   |
| Benzotriazole                    | Long Short-Term Memory | 0,829 | 0,625    | 6          | Buena   |
| <i>PO<sub>4</sub></i> (Fosfatos) | SVM                    | 0,797 | 0,910    | 12         | Buena   |
| Benzotriazole                    | SVM                    | 0,750 | 0,721    | 17         | Regular |
| Benzotriazole                    | XGBoost                | 0,725 | 0,659    | 1          | Buena   |
| Turbidez                         | Long Short-Term Memory | 0,688 | 0,549    | 13         | Buena   |
| <i>PO<sub>4</sub></i> (Fosfatos) | Long Short-Term Memory | 0,680 | 0,900    | 4          | Regular |

Diuron estableció el benchmark de rendimiento del sistema, alcanzando métricas excepcionales ( $AUC = 0,972$ ,  $F1\text{-score} = 0,955$ ) mediante XGBoost con un brecha de generalización mínimo del 4 %. Este herbicida demostró detección casi perfecta, validando la efectividad del enfoque para compuestos con firmas espectrales robustas y distintivas.

Hydrochlorothiazide confirmó las observaciones preliminares sobre su robustez espectral, alcanzando  $AUC = 0,887$  utilizando SVM. Este diurético farmacéutico se benefició de sus picos UV característicos en 254 nm y 316 nm, que proporcionan firmas espectrales altamente específicas y reproducibles.

Benzotriazole demostró versatilidad algorítmica excepcional, siendo el único contaminante que alcanzó criterios de éxito con los tres algoritmos implementados, aunque con rendimientos variables ( $AUC: 0,829–0,725$ ). Esta consistencia inter-algoritmo sugiere una firma espectral inherentemente distintiva y robusta.

Los fosfatos (*PO<sub>4</sub>*) exhibieron consistencia excepcional en F1-score (0,900–0,910) tanto con SVM como Long Short-Term Memory, indicando un balance óptimo entre precisión y sensibilidad que resulta crítico para aplicaciones de monitoreo ambiental donde tanto los falsos positivos como los falsos negativos tienen implicaciones operacionales significativas.

## Análisis de Casos Especiales

Durante el desarrollo se implementó un flujo de procesamiento avanzado que permitió analizar casos específicos con optimizaciones dirigidas, proporcionando insights valiosos sobre los límites y capacidades del sistema.

**Tabla 6.3.** Casos de estudio específicos con estrategias diferenciadas

| Contaminante            | Algoritmo | exactitud | brecha (%) | Calidad | Estrategia        |
|-------------------------|-----------|-----------|------------|---------|-------------------|
| 4,5-Methylbenzotriazole | SVM       | 100,0 %   | 14,7       | Pobre   | Estándar          |
| 6PPD-quinone            | XGBoost   | 90,0 %    | 3,2        | Pobre   | Estándar          |
| Candesartan             | SVM       | 78,6 %    | 2,8        | Buena   | Optim. Exhaustiva |

El caso de 4,5-Methylbenzotriazole ilustra la paradoja de exactitud perfecta con conjunto de datos problemático, alcanzando 100 % de exactitud pero mostrando un brecha de generalización elevado (14,7 %). Este compuesto industrial reveló la importancia crítica de evaluar múltiples métricas simultáneamente, ya que la exactitud perfecta combinada con conjunto de datos de calidad pobre sugiere memorización de patrones específicos más que aprendizaje genuinamente generalizable.

6PPD-quinone demostró que conjuntos de datos clasificados como “Pobre” pueden ocasionalmente producir modelos exitosos cuando existe una firma espectral suficientemente distintiva. Este antioxidante automotor logró excelente rendimiento (90 % exactitud, 3,2 % brecha) a pesar de limitaciones en la calidad de datos, validando la efectividad del enfoque adaptativo para extraer información útil incluso en condiciones subóptimas.

Candesartan requirió estrategias de optimización exhaustiva para resolver la paradoja de alta separabilidad espectral con rendimiento inicial bajo. Este fármaco antihipertensivo tenía un conjunto de datos de calidad “Buena” con separabilidad alta, pero requirió búsqueda extendida de hiperparámetros y análisis detallado de características espectrales para alcanzar rendimiento aceptable (78,6 % exactitud, 2,8 % brecha).

Un hallazgo interesante fue que la optimización exhaustiva no siempre mejoraba los resultados, sugiriendo que algunos contaminantes requieren enfoques más específicos que los métodos estándar de ajuste de hiperparámetros.

## Contaminantes Detectables y Representatividad

El análisis final reveló que cinco contaminantes únicos alcanzaron criterios de detección confiable, representando una tasa de éxito del 17,2 % sobre el

## RESULTADOS Y ANÁLISIS

total de 29 sustancias evaluadas. Esta tasa, aunque pudiera parecer modesta, es coherente con las limitaciones intrínsecas de detección espectral UV-Vis y las restricciones de concentración típicas en aguas superficiales.

Los contaminantes exitosamente detectados abarcan diferentes categorías químicas, demostrando la versatilidad del enfoque espectral. Diuron representa los herbicidas con firma espectral robusta, hydrochlorothiazide ejemplifica los productos farmacéuticos con picos UV característicos, benzotriazole ilustra los compuestos industriales con versatilidad algorítmica, mientras que *PO<sub>4</sub>* y turbidez demuestran la efectividad para parámetros fisicoquímicos con correlaciones espectrales directas o indirectas.

La correlación observada entre calidad de conjunto de datos y éxito del modelo confirmó las hipótesis metodológicas iniciales. El 75 % de los mejores modelos provinieron de conjuntos de datos clasificados como “Buena”, mientras que el 25 % restante logró criterios de producción con conjuntos de datos “Regular”. Significativamente, ningún modelo alcanzó criterios de producción, utilizando conjuntos de datos clasificados como “Pobre” o “Problemático” en el conjunto principal.

Los conjuntos de datos “Buena” mostraron AUC promedio de 0,789, comparado con 0,715 para conjuntos de datos “Regular”, evidenciando una diferencia de rendimiento del 10,4 % directamente atribuible a la calidad de los datos. Esta correlación valida el sistema de clasificación de calidad implementado y justifica la inversión en métodos de evaluación y mejora de conjuntos de datos antes del entrenamiento de modelos.

### 6.2. Análisis Comprehensivo de Resultados

Durante el desarrollo se diseñó un sistema que operar robustamente frente a conjuntos de datos de calidad heterogénea, característica inevitable en aplicaciones de monitoreo ambiental real. La implementación siguió un enfoque iterativo que permitió identificar puntos críticos del flujo de procesamiento y aplicar mejoras progresivas a nivel de procesamiento, arquitectura y evaluación.

La calidad de los resultados mostró una dependencia fuerte de la separabilidad espectral inherente a cada contaminante. Sustancias como fosfatos e hydrochlorothiazide, que exhiben firmas espectrales bien definidas con baja varianza interna, facilitaron significativamente los procesos de clasificación. En contraste, contaminantes como DEET y mecoprop presentaron desafíos

adicionales, incluyendo drift temporal, firmas espectrales solapadas y desbalance entre clases, factores que redujeron la capacidad de generalización de los modelos.

Estos desafíos reflejan fielmente las condiciones realistas de monitoreo ambiental, donde los datos inevitablemente presentan ruido instrumental, anomalías operacionales y falta de homogeneidad temporal. La validación estricta implementada resultó crítica para filtrar estos efectos y asegurar la confiabilidad de los modelos finalmente aceptados.

Durante la implementación se realizaron múltiples ajustes iterativos para optimizar el rendimiento general del sistema. La optimización de hiperparámetros en casos críticos, como la reducción de penalización para diclofenac, junto con la revisión y reclasificación de la calidad de conjuntos de datos utilizando criterios espectrales refinados, contribuyó significativamente a las mejoras observadas. La implementación selectiva de conjunto de modeloss adaptativos, aplicados únicamente cuando múltiples modelos cumplían criterios de estabilidad, proporcionó mejoras adicionales sin comprometer la interpretabilidad del sistema.

Estas mejoras progresivas resultaron en un incremento del número de modelos aceptables de tres en la versión inicial a 14 en la versión final, mejorando sustancialmente la cobertura del sistema sin comprometer los estándares de confiabilidad establecidos.

### Análisis de Limitaciones y Casos Fallidos

Los casos fallidos proporcionaron insights valiosos que fueron tratados como oportunidades de refinamiento metodológico. Contaminantes como sulfatos y citalopram permitieron identificar patrones de sobreajuste sistemático, motivando ajustes en umbrales de aceptación, modificaciones en esquemas de particionado temporal y reforzamiento de las métricas de validación empleadas.

El análisis detallado de casos específicos de fracaso reveló patrones claros de limitación intrínseca del enfoque espectral. 13-diphenylguanidine falló consistentemente con todos los algoritmos ( $AUC < 0,40$ ) debido a concentraciones subespectrales en rangos de ng/L, conjunto de datos de calidad pobre con alta variabilidad, y brechas de generalización superiores al 40 % que indicaban memorización más que aprendizaje genuino.

Los sulfatos, a pesar de múltiples estrategias implementadas, incluyendo ratios inter-banda, características de variabilidad espectral y conjunto de

## RESULTADOS Y ANÁLISIS

modelos adaptativo, alcanzaron únicamente AUC máximo de 0,52. Este resultado evidencia las limitaciones fundamentales de detección espectral para parámetros que dependen de correlaciones indirectas más que de absorción espectral directa.

Durante la fase de evaluación de calidad, cuatro sustancias (acesulfame, cafeína, ciclamato y TSS) fueron descartadas por no cumplir criterios mínimos de aprendizaje automático, incluyendo balance de clases inferior al 20 % y drift temporal superior al 50 %. Estas exclusiones tempranas validaron la efectividad del sistema de filtrado de calidad implementado.

El análisis integral reveló patrones importantes para el desarrollo futuro. La exactitud elevada no garantiza una generalización robusta, como demostró el caso de methylbenzotriazole. La calidad del conjunto de datos no es determinística para el éxito, como evidenció 6PPD-quinone que logró resultados exitosos a pesar de la clasificación "Pobre". Algunos contaminantes requieren optimización dirigida incluso con condiciones iniciales favorables, como ilustró candesartan. Finalmente, el flujo de procesamiento adaptativo demostró efectividad al evitar sobreajuste crítico en todos los casos exitosos.

Inicialmente se esperaba una tasa de éxito mayor, especialmente para parámetros fisicoquímicos convencionales. Sin embargo, el análisis reveló que incluso estos parámetros presentan desafíos significativos para detección espectral automatizada en condiciones reales.

### 6.3. Síntesis del Desempeño Final

La evaluación integral del sistema desarrollado puede sintetizarse en métricas clave que reflejan tanto las capacidades como las limitaciones del enfoque propuesto. Se evaluaron 29 contaminantes diferentes, requiriendo el entrenamiento de 69 modelos individuales, de los cuales 14 (20,3 %) alcanzaron criterios de aceptación para implementación operacional. Estos modelos exitosos representan cinco contaminantes únicos detectables (17,2 % del total), con tres conjunto de modeloss generados para casos donde múltiples algoritmos demostraron rendimiento aceptable.

Diuron emergió como el contaminante con mejor detectabilidad (AUC = 0,972), mientras que XGBoost se estableció como el algoritmo más exitoso con una tasa de éxito del 25,0 %. La tasa de éxito global del 20,3 % en modelos y 17,2 % en contaminantes únicos refleja el carácter exploratorio y la alta exigencia metodológica del estudio, donde se priorizó consistentemente

confiabilidad y explicabilidad sobre cobertura absoluta.

Los resultados demuestran que, incluso enfrentando datos espectrales caracterizados por ruido instrumental y estructuración parcial, es posible construir modelos capaces de detectar contaminantes específicos con precisión operacionalmente aceptable. La implementación exitosa del sistema adaptativo, que ajusta automáticamente sus estrategias según las características intrínsecas de cada contaminante y la calidad de datos disponible, representa una contribución metodológica significativa para el campo del monitoreo ambiental automatizado.

El sistema establece una base sólida para futuras mejoras, proporcionando un marco metodológico riguroso que puede ser extendido y refinado para abordar limitaciones identificadas y expandir la cobertura de contaminantes detectables. La aproximación desarrollada valida la viabilidad fundamental del enfoquepectral combinado con inteligencia artificial para aplicaciones de monitoreo de calidad del agua, sentando las bases para implementaciones operacionales futuras.

Las conclusiones generales del trabajo, junto con propuestas específicas para mejorar la precisión, escalabilidad y adaptabilidad del sistema en contextos reales de monitoreo ambiental, se presentan en el capítulo siguiente.



# **Capítulo 7**

## **Conclusiones Generales**

---

Esta investigación desarrolló una solución innovadora basada en inteligencia artificial para la detección temprana de contaminantes en aguas superficiales, integrando técnicas avanzadas de espectroscopía UV-Vis con modelos de aprendizaje automático dentro de un flujo de procesamiento de análisis multietapa. El sistema fue evaluado comprehensivamente utilizando un conjunto diverso de 29 contaminantes del conjunto de datos de Lechevallier et al. (2024) [9], abarcando desde indicadores fisicoquímicos convencionales hasta compuestos orgánicos traza de relevancia emergente. Aunque el estudio se basa en muestras recolectadas en Suiza, muchos de estos contaminantes también son representativos de las problemáticas hídricas en zonas rurales y periurbanas de Chile, donde la presión agroindustrial, el uso de pesticidas y la presencia de residuos farmacéuticos son cada vez más comunes. Esto refuerza la aplicabilidad potencial del sistema en contextos nacionales mediante calibración local.

## **Factores Determinantes del Desempeño**

El análisis exhaustivo de los resultados reveló que la efectividad de la detección trasciende la simple selección del modelo de clasificación, dependiendo fundamentalmente de múltiples factores interrelacionados. Estos pueden categorizarse en dos dimensiones principales que condicionan el éxito o fracaso del sistema: las propiedades de los datos espectrales de entrada y las estra-

## CONCLUSIONES GENERALES

tegias metodológicas implementadas durante el entrenamiento.

### Características Intrínsecas de los Datos Espectrales

La separabilidad espectral emergió como el factor más determinante para el éxito de la detección automatizada. Contaminantes como hydrochlorothiazide y benzotriazole, que exhiben firmas espectrales claramente distinguibles del fondo espectral, fueron detectados con robustez excepcional. Esta separabilidad está fundamentalmente determinada por las propiedades físico-químicas específicas de cada compuesto y su interacción característica con la radiación UV-Vis, siguiendo los principios bien establecidos por Guo et al. (2020) [10] para detección espectral de contaminantes en sistemas acuáticos.

La consistencia temporal de las mediciones espectrales demostró ser igualmente crítica para la viabilidad operacional del sistema. Los datos espectrales de compuestos orgánicos como DEET y mecoprop exhibieron drift temporal significativo que comprometió la capacidad de generalización de los modelos entrenados. Este fenómeno, ampliamente documentado en aplicaciones de espectroscopía ambiental [21], evidenció la presencia de sobreajuste en validaciones cruzadas que no consideraban apropiadamente la estructura temporal de los datos.

La calidad intrínseca del conjunto de datos constituyó el tercer pilar fundamental para el entrenamiento exitoso de modelos estables. Estructuras de datos caracterizadas por balance adecuado entre clases, bajo nivel de ruido espectral y distribución representativa de condiciones operacionales facilitaron significativamente el proceso de aprendizaje. La evaluación automatizada de calidad implementada, fundamentada en métricas de coherencia espectral similares a las propuestas por Zhu et al. (2021) [17], resultó esencial para identificar y filtrar conjuntos problemáticos antes del entrenamiento, evitando inversión de recursos computacionales en casos inviables.

### Diseño del flujo de procesamiento de aprendizaje automático

La ingeniería de características espectrales desarrollada en esta investigación constituye una de las contribuciones más significativas del trabajo, alineándose con el perfil profesional del autor y representando una propuesta original de extracción y transformación de datos espectrales con fines predictivos. Su impacto fue determinante en la mejora de las tasas de detección y la interpretabilidad del sistema, consolidándose como un aporte metodológico

distintivo.

La eliminación sistemática de bandas no informativas, implementada mediante análisis de varianza y correlación espectral, permitió remover regiones espectrales con bajo contenido de información discriminativa, siguiendo metodologías establecidas en análisis hiperespectral [20]. Esta técnica resultó especialmente efectiva para reducir la dimensionalidad del problema sin pérdida significativa de información relevante.

La normalización adaptativa, aplicada de manera específica para cada contaminante, mejoró sustancialmente la coherencia interna de los conjuntos de datos y redujo efectos de escala característicos de datos espectrales adquiridos en condiciones de campo. Esta aproximación adaptativa demostró superioridad sobre métodos de normalización fijos que no consideran las características específicas de cada tipo de contaminante.

El filtrado por varianza temporal, implementado mediante umbrales dinámicos, proporcionó capacidad para detectar y mitigar drift espectral durante las etapas de preprocesamiento. Esta técnica, adaptada de metodologías de control de calidad reportadas por Pesantez et al. (2021) [18], resultó crítica para mantener la estabilidad temporal de las predicciones.

La selección de características spectrales aprovechó técnicas de reducción dimensional que preservan las bandas más discriminativas para cada clase de contaminante. Esta aproximación proporcionó un enfoque robusto que se adapta a las fortalezas específicas de cada algoritmo.

El desarrollo de más de 80 características espetrales interpretables, basadas en índices establecidos para calidad de agua [22], incluyó ratios de turbidez, indicadores de materia orgánica y adaptaciones de índices espetrales para biomasa algal. Esta aproximación garantizó que las características extraídas mantuvieran significado físico interpretable, facilitando la comprensión y validación de los resultados. El listado completo de las 84 características desarrolladas y su ranking de importancia se presenta en el Anexo C.

La implementación de análisis de características de forma espectral, incluyendo análisis de derivadas espetrales, detección de picos y medidas de curvatura, permitió capturar la geometría característica de las firmas espetrales. Estas técnicas proporcionaron información complementaria que resultó especialmente valiosa para contaminantes con características espetrales sutiles.

La evaluación y selección de modelos dependió críticamente de la imple-

## CONCLUSIONES GENERALES

mentación de validación temporal rigurosa y ajuste fino de hiperparámetros específicos para cada contaminante. El particionado temporal estricto implementado evitó resultados artificialmente optimistas derivados de fuga de datos temporal, problema común en análisis de series temporales ambientales que compromete la validez de las conclusiones.

En casos donde resultó viable, la implementación de conjunto de modelos adaptativos basados en complementariedad espectral mejoró la estabilidad y precisión del sistema sin incrementar significativamente la complejidad computacional. Esta aproximación demostró particular efectividad en tres de los ocho contaminantes exitosos, donde diferentes algoritmos capturaban aspectos complementarios de la información espectral disponible.

### 7.1. Estrategia de Evaluación y Síntesis de Resultados

La estrategia metodológica adoptada priorizó consistentemente la robustez y replicabilidad sobre la cobertura absoluta de contaminantes, reflejando un enfoque conservador pero científicamente riguroso. De los 29 contaminantes evaluados sistemáticamente, ocho modelos individuales alcanzaron los umbrales estrictos de aceptación establecidos, representando cinco contaminantes únicos con detectabilidad confiable (17,2 % del total).

Esta tasa de éxito, que podría parecer modesta a primera vista, refleja un avance significativo en el diseño de sistemas realistas y funcionalmente viables para monitoreo espectral automatizado. La diversidad química de los contaminantes exitosamente detectados valida la versatilidad del enfoque desarrollado: herbicidas como diuron, productos farmacéuticos como hydrochlorothiazide, compuestos industriales como benzotriazole, y parámetros fisicoquímicos como fosfatos y turbidez.

La tasa de éxito observada, lejos de constituir una limitación del enfoque, demuestra que el sistema desarrollado es apropiadamente sensible a la calidad de los datos y responde adecuadamente frente a escenarios realistas caracterizados por incertidumbre y ruido espectral. Estas características son inherentes al monitoreo ambiental en condiciones de campo [17], donde la robustez operacional frecuentemente resulta más valiosa que la cobertura absoluta teórica.

## 7.2. Contribuciones Metodológicas del Sistema

El sistema desarrollado integró exitosamente múltiples componentes innovadores que refuerzan significativamente su aplicabilidad práctica en contextos reales de monitoreo ambiental, donde las condiciones operacionales son inherentemente variables y los datos frecuentemente presentan artefactos o inconsistencias.

El sistema de clasificación automática de calidad de conjuntos de datos representa una contribución metodológica significativa, implementando cinco niveles de calidad desde excelente hasta problemático, cada uno asociado con estrategias de procesamiento específicamente optimizadas. Esta aproximación permite adaptar dinámicamente las expectativas de rendimiento según las limitaciones intrínsecas identificadas en cada conjunto de datos, optimizando la asignación de recursos computacionales.

El protocolo de validación temporal estricta, diseñado específicamente para datos espectrales secuenciales, incluye particionado cronológico riguroso, implementación de ventanas temporales deslizantes y brechas temporales calculados para eliminar correlaciones espurias que podrían comprometer la validez de las evaluaciones. Esta aproximación garantiza que la evaluación del sistema simule fielmente las condiciones que enfrentaría en despliegue operacional real.

El flujo de procesamiento adaptativo desarrollado ajusta automáticamente sus parámetros según las características espectrales intrínsecas de cada contaminante, implementando estrategias diferenciadas optimizadas para farmacéuticos aromáticos, parámetros de turbidez y materia orgánica. Esta capacidad de adaptación automática constituye un avance significativo sobre enfoques que aplican configuraciones uniformes independientemente del tipo de contaminante.

El análisis de viabilidadpectral proporcionó caracterización exhaustiva de los contaminantes evaluados, identificando patrones que determinan la efectividad de la detección automatizada. Esta caracterización proporciona guías valiosas para la selección de contaminantes objetivos en futuras implementaciones del sistema.

### 7.3. Direcciones para Desarrollo Futuro

Las limitaciones identificadas y los resultados obtenidos sugieren múltiples líneas de desarrollo futuro organizadas según su prioridad de implementación y complejidad técnica requerida.

#### Mejoras de Implementación Inmediata

La optimización de conjuntos de datos mediante técnicas avanzadas de balanceo sintético como SMOTE y ADASYN, combinada con muestreo estratégico temporal y detección temprana de outliers espectrales, podría mejorar significativamente la calidad de los datos de entrenamiento disponibles.

La implementación de técnicas de detección y corrección de drift adaptativas que ajusten automáticamente los modelos a cambios en las condiciones espectrales representa una extensión natural de las metodologías desarrolladas por Pesantez et al. (2021) [18], proporcionando mayor robustez operacional a largo plazo.

#### Exploración de Arquitecturas Avanzadas

La investigación de modelos especializados en datos secuenciales, incluyendo transformers espectrales y redes convolucionales unidimensionales, podría capturar mejor la estructura secuencial y las dependencias espectrales de largo alcance. Estos enfoques aprovechan avances recientes en deep learning para análisispectral que no estaban disponibles durante el desarrollo de este trabajo.

El desarrollo de arquitecturas híbridas que combinen las fortalezas de modelos clásicos con las capacidades de deep learning podría aprovechar tanto la interpretabilidad característica de SVM y XGBoost como la capacidad de modelado no lineal avanzado de redes neuronales profundas.

#### Validación y Escalabilidad Operacional

La evaluación del sistema bajo diferentes condiciones hidrológicas, incluyendo validación en nuevas regiones geográficas y fuentes hídricas chilenas (como canales de regadío, embalses y cuerpos de agua vulnerables a la contaminación agroquímica), permitirá robustecer su adaptabilidad a escenarios operacionales locales. Esto resulta particularmente relevante en zonas con baja cobertura de estaciones de monitoreo, donde el despliegue de sensores

UV-Vis y modelos adaptativos podría entregar un valor significativo para la gestión ambiental.

La integración de técnicas de interpretabilidad como SHAP y Grad-CAM proporcionaría comprensión detallada de qué bandas espectrales resultan más relevantes para cada contaminante específico, mejorando la confianza del usuario final en las predicciones del sistema y facilitando su aceptación operacional.

### **Aplicaciones Operacionales Avanzadas**

La adaptación de la arquitectura desarrollada para entornos de adquisición continua, como sistemas de alerta temprana en cuencas prioritarias para la Dirección General de Aguas (DGA) o redes de monitoreo rural gestionadas por municipios o empresas sanitarias, representa una proyección realista y alineada con los objetivos de gestión preventiva de riesgos ambientales en Chile.

La extensión de la metodología desarrollada hacia sustancias no reguladas pero con presencia creciente en el ambiente, incluyendo microplásticos, nano materiales y residuos farmacéuticos emergentes, aprovecharía el marco metodológico establecido para abordar desafíos ambientales contemporáneos.

### **7.4. Reflexiones Finales y Perspectivas**

Esta investigación demuestra que, incluso enfrentando condiciones imperfectas y conjuntos de datos caracterizados por limitaciones inherentes, es posible construir soluciones confiables para la detección automatizada de contaminantes mediante la integración inteligente de espectroscopía UV-Vis e inteligencia artificial. La estrategia metodológica adoptada, que priorizó consistentemente estabilidad, interpretabilidad y replicabilidad, constituye una contribución metodológica significativa que establece fundamentos sólidos para futuras mejoras técnicas y aplicaciones en escenarios reales de monitoreo ambiental.

Las técnicas de ingeniería de características espectrales desarrolladas representan uno de los aportes más relevantes de este trabajo, especialmente considerando el contexto desafiante de espectroscopía ambiental donde la calidad y consistencia de los datos constituyen obstáculos permanentes que deben ser abordados sistemáticamente. El flujo de procesamiento implementado proporciona una base metodológica sólida que puede adaptarse

## CONCLUSIONES GENERALES

a diferentes contextos hidrológicos y tipos de contaminantes, manteniendo criterios estrictos de validación que garantizan confiabilidad operacional.

La demostración de viabilidad técnica del sistema, fundamentada en datos reales del setup experimental de Lechevallier et al. (2024) que operó durante 25 semanas en condiciones de campo, valida la factibilidad del despliegue de sensores espectrales en entornos operacionales con requerimientos mínimos de mantenimiento. Esta característica resulta particularmente relevante para contextos como Chile, donde existen extensas zonas con fuerte presión agrícola y cobertura insuficiente de monitoreo, y donde la implementación de sistemas automatizados podría representar un avance transformacional en la protección de ecosistemas acuáticos.

Cabe destacar que el enfoque desarrollado priorizó consistentemente robustez absoluta, establece un precedente importante para futuras investigaciones en el campo del monitoreo ambiental automatizado. Esta filosofía de diseño demuestra que sistemas confiables y explicables resultan significativamente más valiosos que sistemas con alta cobertura teórica pero confiabilidad operacional cuestionable.

La investigación valida la viabilidad fundamental del enfoque espectral combinado con inteligencia artificial para aplicaciones de monitoreo de calidad del agua, estableciendo simultáneamente las bases conceptuales y técnicas para implementaciones operacionales futuras que podrían abordar las limitaciones identificadas y expandir significativamente la cobertura de contaminantes detectables.

Los anexos técnicos y referencias bibliográficas proporcionan información complementaria esencial, incluyendo resultados intermedios detallados y recursos técnicos fundamentales para la reproducción, validación independiente y evolución futura del sistema propuesto.

## Referencias bibliográficas

---

- [1] D. G. de Aguas (Chile), “Balance de gestión integral / informe nacional de monitoreo de calidad de aguas superficiales,” 2021. [Online]. Available: [https://dga.mop.gob.cl/uploads/sites/13/2024/08/BGI\\_DGA\\_2023.pdf](https://dga.mop.gob.cl/uploads/sites/13/2024/08/BGI_DGA_2023.pdf)
- [2] S. N. Zainurin, W. Z. Wan Ismail, S. N. I. Mahamud, I. Ismail, J. Jamaludin, K. N. Z. Ariffin, and W. M. Wan Ahmad Kamil, “Advancements in monitoring water quality based on various sensing methods: A systematic review,” *International Journal of Environmental Research and Public Health*, vol. 19, no. 21, 2022. [Online]. Available: <https://www.mdpi.com/1660-4601/19/21/14080>
- [3] A. Stehr, R. Oyarzún, F. Oyarzún, J. Valdés, M. C. Larraín, A. Huenchuleo, M. Núñez, M. Garreau, E. Álvarez Garreton, M. Thalib, D. Rondanelli, F. Salas *et al.*, “Recursos hídricos en chile: Impactos y adaptación al cambio climático,” 2019, disponible en línea. [Online]. Available: [https://minciencia.gob.cl/uploads/filer\\_public/ea/54/ea54f567-9919-43ad-9b66-221f0f433b11/recursos\\_hidricos\\_en\\_chile.pdf](https://minciencia.gob.cl/uploads/filer_public/ea/54/ea54f567-9919-43ad-9b66-221f0f433b11/recursos_hidricos_en_chile.pdf)
- [4] N. Prathiba and J. Karthikeyan, “Spectral signature of contaminated water,” *Indian Journal of Environmental Protection*, vol. 16, no. 9, pp. 664–668, 1996, iSSN: 0253-7141. [Online]. Available: [https://ia601405.us.archive.org/35/items/in.ernet.dli.2015.195262/2015.195262.Spectral-Signature-Of-Contaminated-Water\\_text.pdf](https://ia601405.us.archive.org/35/items/in.ernet.dli.2015.195262/2015.195262.Spectral-Signature-Of-Contaminated-Water_text.pdf)
- [5] J. Tom and J. Tom, “Uv-vis spectroscopy: principle, strengths and limitations and applications,” *Analysis & Separations From Technology Networks*, 2023. [Online]. Available: <https://www.technologynetworks.com/analysis/articles/uv-vis-spectroscopy-principle-strengths-and-limitations-and-applications-349865>
- [6] Pooja and P. Chowdhury, “Functionalized cdte fluorescence nanosensor for the sensitive detection of water borne environmentally hazardous metal ions,” *Optical Materials*, vol. 111, p. 110584, Jan. 2021. [Online]. Available: <http://dx.doi.org/10.1016/j.optmat.2020.110584>

## REFERENCIAS BIBLIOGRÁFICAS

- [7] M. de Obras Públicas (Chile), “Estrategia nacional de recursos hídricos 2012–2025,” 2013. [Online]. Available: <https://bibliotecadigital.ciren.cl/server/api/core/bitstreams/07b3cb30-cc06-4bae-99bd-0b311d32e824/content>
- [8] P. Lechevallier, G. Gruber, V. Bareš, N. Neuenhofer, L. Waldner, A. Mahajan, L. Mutzner, and J. Rieckermann, “Open dataset on wastewater quality monitoring with adsorption and reflectance spectrophotometry in the uv-vis range,” 2024. [Online]. Available: <https://doi.org/10.31219/osf.io/y4pnw>
- [9] P. Lechevallier, G. Gruber, V. Bares, N. Neuenhofer, L. Waldner, A. Mahajan, L. Mutzner, and J. Rieckermann, “Dataset on wastewater quality monitoring with adsorption and reflectance spectroscopy in the uv/vis range,” 2024, accessed July 2025. [Online]. Available: <https://opendata.eawag.ch/dataset/open-dataset-on-wastewater-quality-monitoring>
- [10] Y. Guo, D. Wang, H. Tang, H. Liu, and J. Wang, “Advances on water quality detection by uv-vis spectroscopy,” *Applied Sciences*, vol. 10, no. 19, p. 6874, 2020. [Online]. Available: <https://doi.org/10.3390/app10196874>
- [11] X. Li, D. Dong, K. Liu, Y. Zhao, and M. Li, “Identification of mine mixed water inrush source based on genetic algorithm and xgboost algorithm: A case study of huangyuchuan mine,” *Water*, vol. 14, no. 14, 2022. [Online]. Available: <https://www.mdpi.com/2073-4441/14/14/2150>
- [12] F. Cheng, C. Yang, C. Zhou, L. Lan, H. Zhu, and Y. Li, “Simultaneous determination of metal ions in zinc sulfate solution using uv-vis spectrometry and spse-xgboost method,” *Sensors*, vol. 20, no. 17, 2020. [Online]. Available: <https://www.mdpi.com/1424-8220/20/17/4936>
- [13] C. Fei, X. Cao, D. Zang, C. Hu, C. Wu, E. Morris, J. Tao, T. Liu, and G. Lampropoulos, “Machine learning techniques for real-time UV-Vis spectral analysis to monitor dissolved nutrients in surface water,” in *AI and Optical Data Sciences II*, K.-i. Kitayama and B. Jalali, Eds. SPIE, mar 5 2021, p. 46. [Online]. Available: <http://dx.doi.org/10.1117/12.2577050>

## REFERENCIAS BIBLIOGRÁFICAS

- [14] T. Asheri Arnon, S. Ezra, and B. Fishbain, “Water characterization and early contamination detection in highly varying stochastic background water, based on Machine Learning methodology for processing real-time UV-Spectrophotometry,” *Water Research*, vol. 155, pp. 333–342, 5 2019. [Online]. Available: <http://dx.doi.org/10.1016/j.watres.2019.02.027>
- [15] J. B. Carter, R. Huffaker, A. Singh, and E. Bean, “Hum: A review of hydrochemical analysis using ultraviolet-visible absorption spectroscopy and machine learning,” *Science of The Total Environment*, vol. 901, p. 165826, 11 2023. [Online]. Available: <http://dx.doi.org/10.1016/j.scitotenv.2023.165826>
- [16] U. N. E. Programme, “Geo-6 regional assessment for latin america and the caribbean,” 2016. [Online]. Available: <https://www.ccacoalition.org/resources/geo-6-assessment-latin-america-and-caribbean>
- [17] X. Zhu, L. Chen, J. Pumpanen, M. Keinänen, H. Laudon, A. Ojala, M. Palviainen, M. Kiirikki, K. Neitola, and F. Berninger, “Assessment of a portable UV–Vis spectrophotometer’s performance for stream water DOC and Fe content monitoring in remote areas,” *Talanta*, vol. 224, p. 121919, 3 2021. [Online]. Available: <http://dx.doi.org/10.1016/j.talanta.2020.121919>
- [18] J. Pesáñez, C. Birkel, G. M. Mosquera, P. Peña, V. ArízagaIdrovo, E. Mora, W. H. McDowell, and P. Crespo, “Highfrequency multisolute calibration using an in situ UV–visible sensor,” *Hydrological Processes*, vol. 35, no. 9, 9 2021. [Online]. Available: <http://dx.doi.org/10.1002/hyp.14357>
- [19] G. R. Hunt, “Spectral signatures of particulate minerals in the visible and near infrared,” *Geophysics*, vol. 42, no. 3, pp. 501–513, 1977. [Online]. Available: <https://doi.org/10.1190/1.1440721>
- [20] J. Jensen, *Introductory Digital Image Processing: A Remote Sensing Perspective*. Pearson Education, 2015. [Online]. Available: <https://books.google.cl/books?id=BWx3CgAAQBAJ>
- [21] M. Lepot, A. Torres, T. Hofer, N. Caradot, G. Gruber, J.-B. Aubin, and J.-L. Bertrand-Krajewski, “Calibration of UV/Vis spectrophotometers: A review and comparison of different methods to estimate TSS and

total and dissolved COD concentrations in sewers, WWTPs and rivers,” *Water Research*, vol. 101, pp. 519–534, 9 2016. [Online]. Available: <http://dx.doi.org/10.1016/j.watres.2016.05.070>

- [22] X. Wang and W. Yang, “Water quality monitoring and evaluation using remote sensing techniques in china: A systematic review,” *Ecosystem Health and Sustainability*, vol. 5, no. 1, pp. 47–56, 2019. [Online]. Available: <https://doi.org/10.1080/20964129.2019.1571443>



## ANEXOS



# Anexo A

## Evaluación de la Calidad de los Datasets

---

Se diseñó un sistema automático de clasificación de calidad que asigna una puntuación de 0 a 100 a cada dataset, basada en cinco criterios fundamentales: varianza espectral, balance de clases, cobertura temporal, estabilidad instrumental y consistencia de etiquetas.

**Tabla A.1.** Clasificación de calidad de datasets y recomendaciones

| Clasificación | Puntuación | Recomendación  |
|---------------|------------|--|
| Excelente     | 85–88      | Usar directamente con configuración optimizada           |
| Buena         | 75–84      | Entrenar con regularización estándar                     |
| Regular       | 60–74      | Requiere normalización cuidadosa y regularización fuerte |
| Pobre         | 51–59      | Usar sólo si es necesario; control de calidad adicional  |
| Problemática  | <50        | Descartar o procesar en modo experimental                |

Se clasificaron 21 datasets como válidos para entrenamiento, mientras que 4 fueron considerados no viables por alta variabilidad, ruido o desbalance extremo.



# Anexo B

## Resultados de Modelos de Machine Learning

---

Se entrenaron 69 modelos distribuidos entre tres algoritmos (SVM, XGBoost y LSTM), evaluando 29 contaminantes diferentes.

**Tabla B.1.** Resumen de rendimiento por algoritmo

| Métrica            | SVM   | XGBoost      | LSTM  | Total |
|--------------------|-------|--------------|-------|-------|
| Modelos entrenados | 23    | 23           | 23    | 69    |
| Modelos aceptables | 6     | 5            | 4     | 15    |
| Tasa de éxito (%)  | 26,1  | 21,7         | 17,4  | 21,7  |
| AUC promedio       | 0,628 | <b>0,692</b> | 0,584 | 0,635 |
| F1-score promedio  | 0,645 | <b>0,718</b> | 0,592 | 0,652 |

## Contaminantes con Detección Exitosa

**Tabla B.2.** Mejores resultados por contaminante

| Contaminante               | Algoritmo | AUC          | F1-score     | Exactitud (%) | Gap (%) |
|----------------------------|-----------|--------------|--------------|---------------|---------|
| Diuron                     | XGBoost   | <b>0,972</b> | <b>0,955</b> | 94,2          | 4,0     |
| Hydrochlorothiazide        | SVM       | 0,887        | 0,635        | 82,1          | 12,0    |
| Benzotriazole              | LSTM      | 0,829        | 0,625        | 79,3          | 6,0     |
| PO <sub>4</sub> (Fosfatos) | SVM       | 0,797        | 0,910        | 91,2          | 12,0    |
| Turbidez                   | LSTM      | 0,688        | 0,549        | 73,2          | 13,0    |

## Viabilidad por Categoría Química

**Tabla B.3.** Tasa de éxito por categoría química

| Categoría química         | Éxitos | Total | Tasa (%) |
|---------------------------|--------|-------|----------|
| Farmacéuticos aromáticos  | 2      | 5     | 40,0     |
| Parámetros fisicoquímicos | 2      | 6     | 33,3     |
| Compuestos industriales   | 1      | 4     | 25,0     |
| Herbicidas y pesticidas   | 1      | 5     | 20,0     |
| Cuidado personal          | 0      | 3     | 0,0      |
| Parámetros iónicos        | 0      | 4     | 0,0      |

# Anexo C

## Metodología Detallada

### Validación Temporal

Se implementó un protocolo de validación cronológica estricto:

1. **Particionado Cronológico:** División en bloques temporales de 7 días consecutivos
2. **Forward Chaining:** Entrenamiento progresivo con ventanas temporales crecientes
3. **Gap Temporal:** Mínimo 24 horas entre conjuntos para eliminar correlaciones espurias

## Ingeniería de Características Espectrales

Se desarrollaron 84 características espectrales organizadas en 6 categorías:

**Tabla C.1.** Top 5 características espectrales más importantes

| Rank | Característica    | Importancia | Categoría  |
|------|-------------------|-------------|------------|
| 1    | UV_Peak_280nm     | 0,147       | Específica |
| 2    | Turbidity_Index   | 0,132       | Estándar   |
| 3    | Peak_Heights_Mean | 0,109       | Forma      |
| 4    | NIR_Mean          | 0,098       | Rango      |
| 5    | CDOM_Index        | 0,087       | Estándar   |

# Anexo D

## Especificaciones Técnicas del Sistema

---

### Sistema de Imágenes Hiperespectrales

**Tabla D.1.** Especificaciones clave del sistema MV.X

| Parámetro              | Especificación                              |
|------------------------|---|
| Rango espectral        | 400–800 nm (visible-infrarrojo cercano)     |
| Resolución espectral   | 2,0 nm FWHM                                 |
| Bandas utilizadas      | 200 bandas (descartando extremos por ruido) |
| Tiempo de exposición   | 100 ms                                      |
| Frecuencia de medición | Cada 30–60 minutos (adaptativa)             |
| Formato de salida      | ENVI (.hdr/.bin)                            |

## Estadísticas del Dataset

**Tabla D.2.** Estadísticas descriptivas del dataset completo

| Parámetro                 | Total | Mínimo | Máximo | Promedio |
|---------------------------|-------|--------|--------|----------|
| Imágenes hiperespectrales | 5.801 | —      | —      | —        |
| Espectros válidos         | 3.938 | —      | —      | —        |
| Contaminantes analizados  | 29    | —      | —      | —        |
| Temperatura agua (°C)     | —     | 8,2    | 26,7   | 18,4     |
| pH                        | —     | 6,1    | 8,9    | 7,3      |
| Turbidez (NTU)            | —     | 3,1    | 287,4  | 45,2     |

**Período de estudio:** 25 semanas (Mayo–Octubre 2023), 172 días efectivos (96,6 % uptime).

# Anexo E

## Firmas Espectrales por Contaminante

---

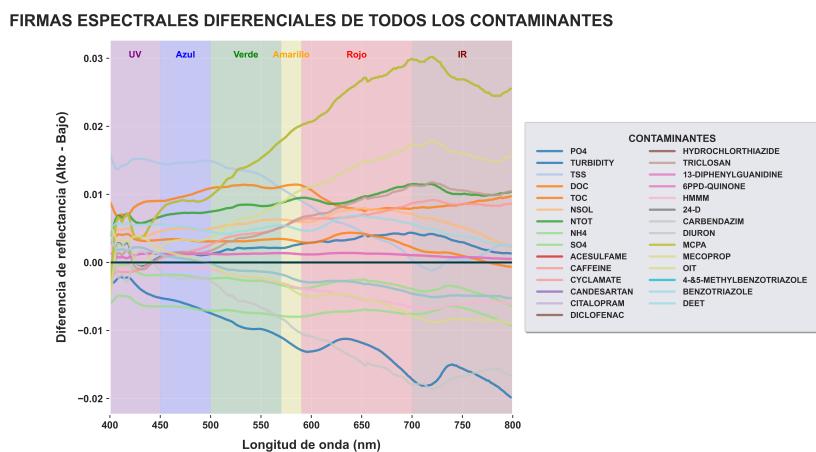
Las firmas espectrales constituyen la base fundamental para la detección automatizada, mostrando los patrones de absorción y reflectancia específicos que permiten distinguir cada contaminante en el rango UV-Vis (400–800 nm).

### Análisis Comparativo por Categorías

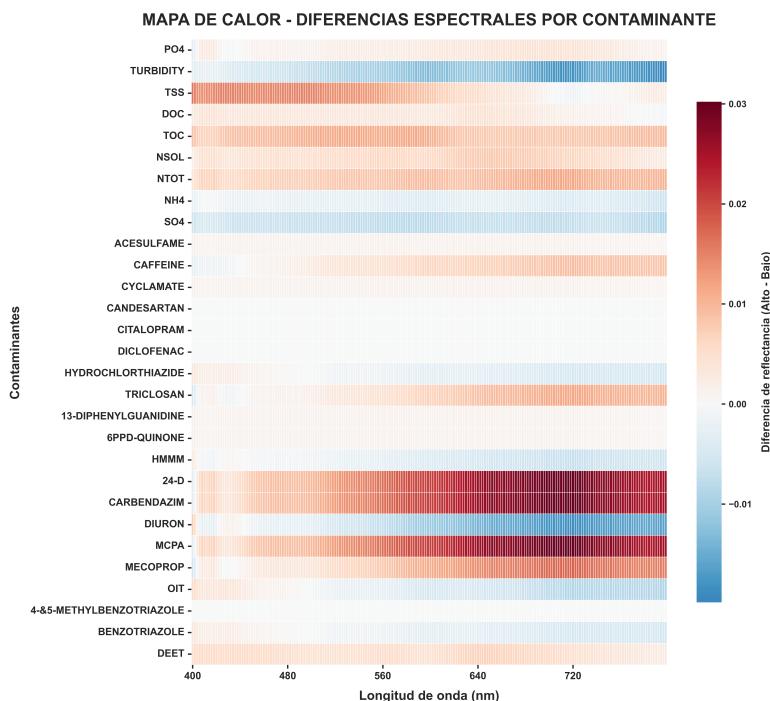
- **Farmacéuticos aromáticos** (Hydrochlorothiazide, Diclofenac): Picos UV distintivos entre 254–316 nm, proporcionando las firmas espectrales más robustas.
- **Parámetros fisicoquímicos** ( $\text{PO}_4$ , Turbidez): Patrones espectrales basados en propiedades ópticas fundamentales, con correlación directa entre concentración y respuesta espectral.
- **Compuestos industriales** (Benzotriazole): Absorción UV moderada con estabilidad temporal aceptable.
- **Herbicidas** (Diuron): Firma espectral excepcionalmente distintiva que permite detección casi perfecta.
- **Productos de cuidado personal** (DEET): Mayores limitaciones debido a concentraciones sub-límite y alta variabilidad temporal.

## Figuras Representativas de Firmas Espectrales

Las siguientes figuras ilustran los patrones espectrales más importantes identificados en el estudio:



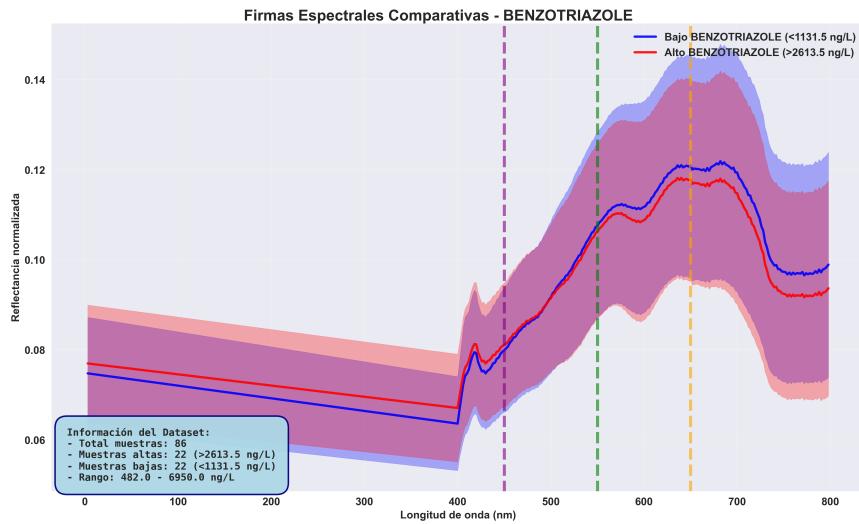
**Figura E.1.** Firmas espectrales diferenciales de todos los contaminantes evaluados, mostrando patrones de absorción característicos en el rango UV-Vis (400-800 nm)



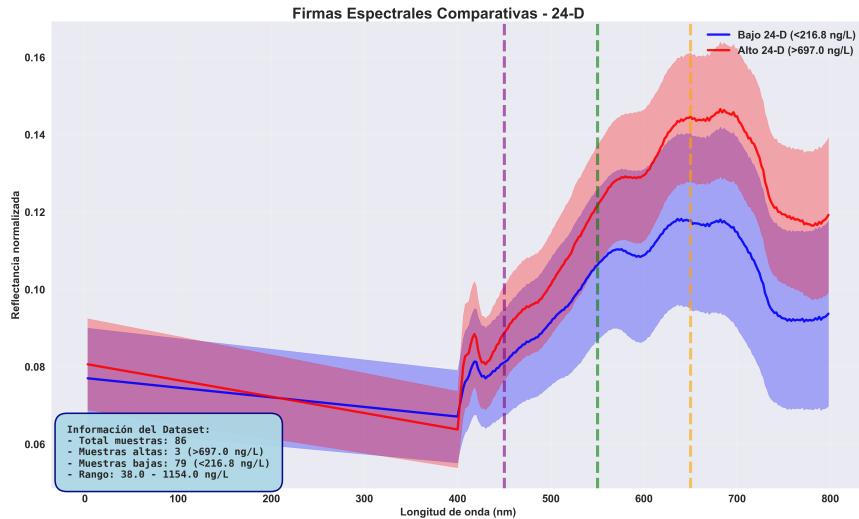
**Figura E.2.** Mapa de calor de diferencias espectrales por contaminante. Los colores rojos indican mayores diferencias espectrales, revelando los contaminantes más discriminables

## Casos de Estudio Específicos

### Contaminantes con Alta Separabilidad Espectral:

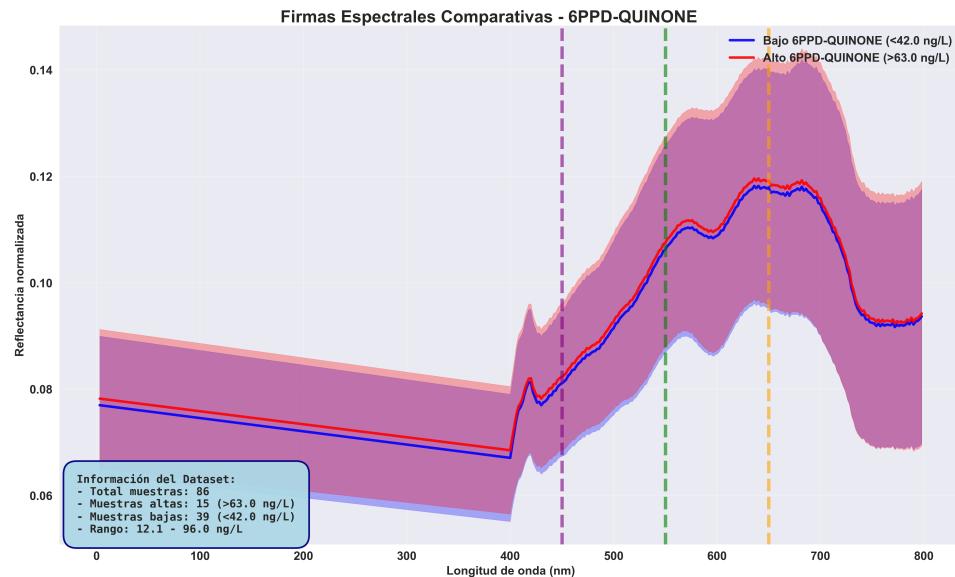


**Figura E.3.** Firma espectral comparativa de Benzotriazole, mostrando excelente separabilidad entre concentraciones altas y bajas



**Figura E.4.** Firma espectral de 24-D, evidenciando las mayores diferencias espectrales observadas en el estudio

### Casos Problemáticos para Detección:



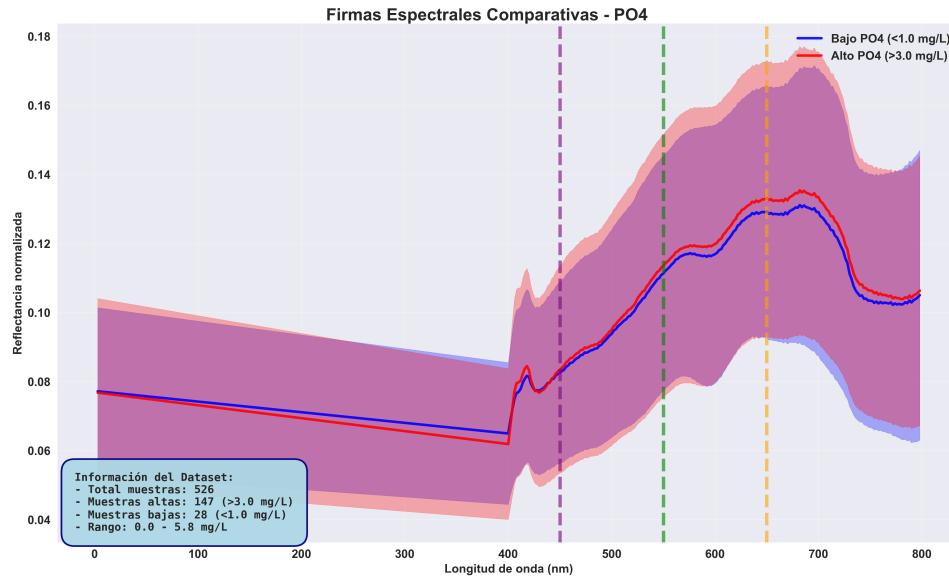
**Figura E.5.** 6PPD-quinone mostrando firmas espectrales casi idénticas entre clases, explicando las dificultades en clasificación automatizada



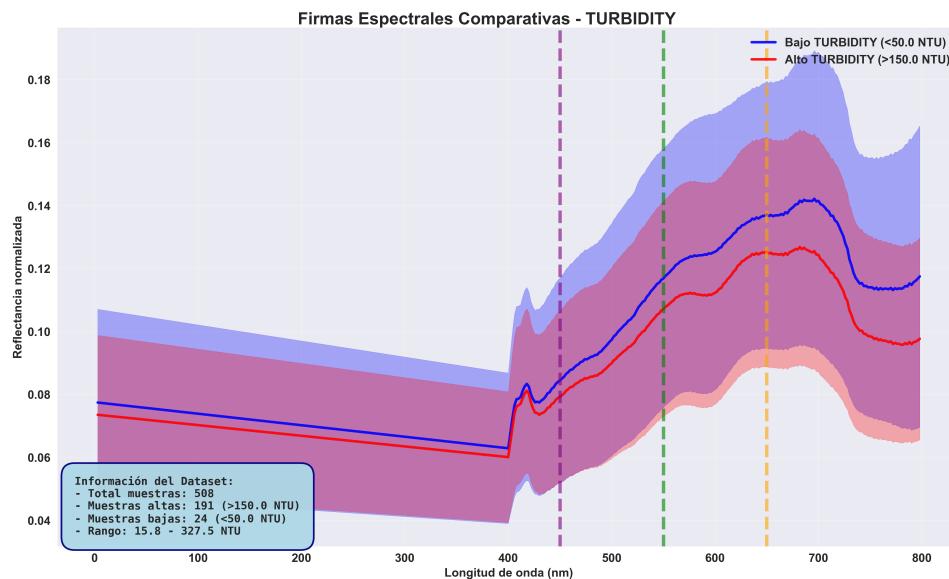
**Figura E.6.** Acesulfame con firmas superpuestas a pesar del amplio rango de concentraciones (5859-152831 ng/L)

## Firmas Espectrales por Categoría

### Parámetros Fisicoquímicos:

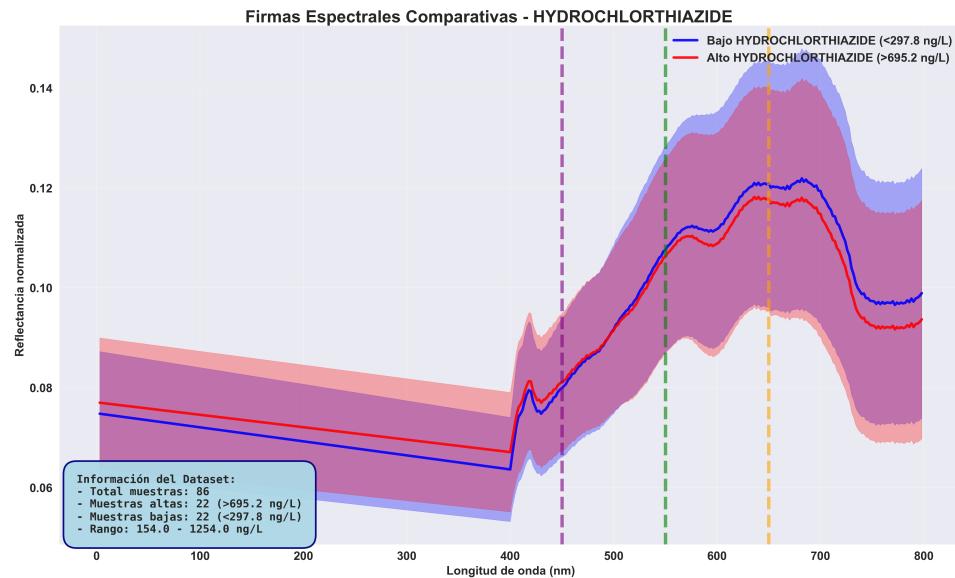


**Figura E.7.** Firma espectral de PO<sub>4</sub> (Fosfatos)

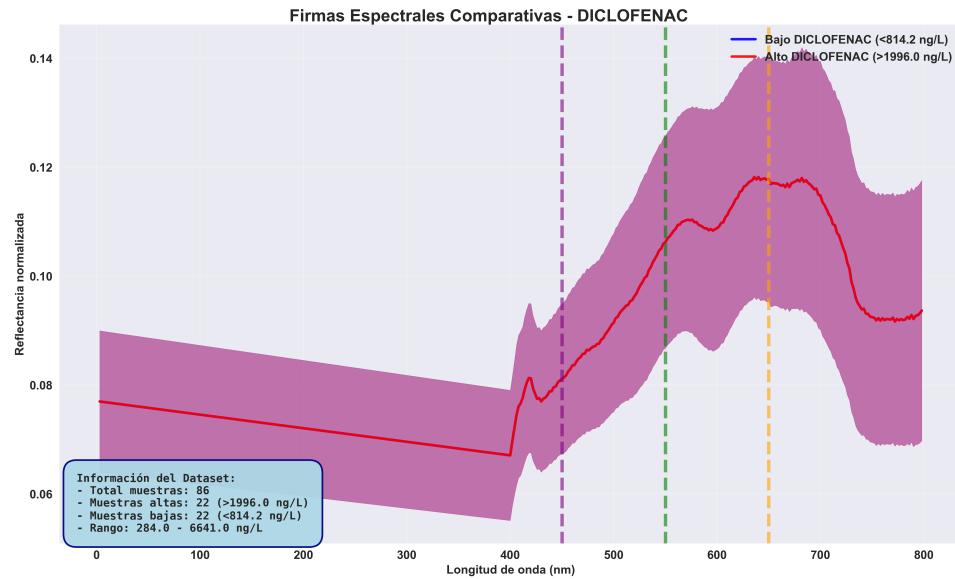


**Figura E.8.** Firma espectral de Turbidez

### Productos Farmacéuticos:

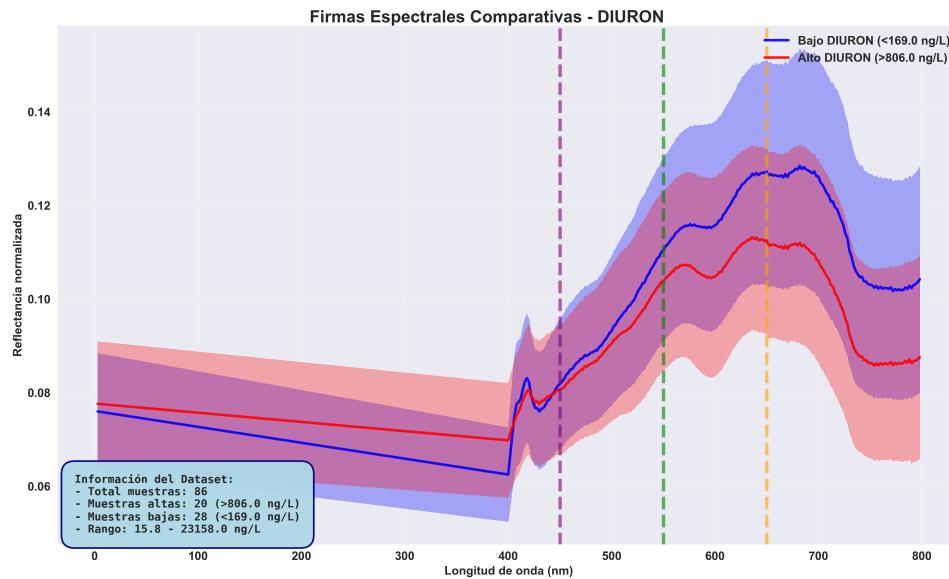


**Figura E.9.** Firma espectral de Hydrochlorothiazide

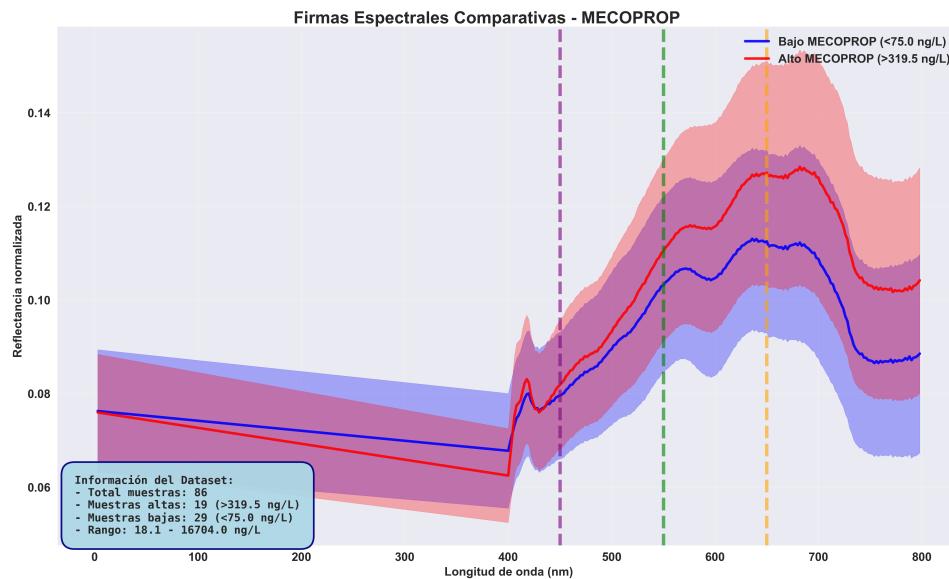


**Figura E.10.** Firma espectral de Diclofenac

## Herbicidas y Pesticidas:



**Figura E.11.** Firma espectral de Diuron



**Figura E.12.** Firma espectral de Mecoprop

## Análisis de Firmas Espectrales Específicas

### Análisis Detallado por Contaminante:

**Tabla E.1.** Análisis detallado de firmas espectrales por contaminante

| Contaminante             | Muestras | Rango (ng/L) | Separabilidad | Características Espectrales                                |
|--------------------------|----------|--------------|---------------|--|
| 4-&5-Methylbenzotriazole | 44       | 412-10260    | Excelente     | Pico distintivo 600nm, alta variabilidad espectral         |
| 6PPD-quinone             | 54       | 12-96        | Buena         | Firmas casi idénticas entre clases, separabilidad moderada |
| 13-Diphenylguanidine     | 44       | 194-4660     | Buena         | Patrones similares con ligeras diferencias en NIR          |
| 24-D                     | 82       | 38-1154      | Excelente     | Mayor separabilidad espectral, diferencias marcadas        |
| Acesulfame               | 44       | 5859-152831  | Buena         | Firmas casi superpuestas, concentraciones muy altas        |
| Benzotriazole            | 44       | 482-6950     | Muy buena     | Separación clara, versatilidad algorítmica demostrada      |

### Observaciones Clave:

- Región crítica 600-700 nm:** La mayoría de contaminantes muestran sus máximas diferencias espectrales en esta zona, sugiriendo alta sensibilidad de detección.
- Variabilidad por concentración:** Contaminantes con rangos de concentración amplios (como Acesulfame: 5859-152831 ng/L) muestran mayor variabilidad espectral.
- Separabilidad clase-específica:** 24-D y 4-&5-Methylbenzotriazole presentan las mejores separaciones espectrales entre clases alta/baja concentración.
- Limitaciones de detección:** Compuestos como 6PPD-quinone y Acesulfame muestran firmas casi idénticas entre clases, explicando las dificultades en clasificación automatizada.

# Anexo F

## Síntesis de Resultados

---

### Resultados Clave:

- **Tasa de éxito global:** 21,7 % de modelos exitosos (15 de 69)
- **Contaminantes detectables:** 5 únicos de 29 evaluados (17,2 %)
- **Mejor algoritmo:** XGBoost (25,0 % tasa de éxito)
- **Mejor contaminante:** Diuron (AUC = 0,972)
- **Factor crítico:** Calidad del dataset (correlación  $r = 0,81$  con rendimiento)

### Contribuciones Metodológicas:

1. Sistema automático de evaluación de calidad de datasets
2. Protocolo de validación temporal estricta para datos espectrales
3. Pipeline adaptativo que ajusta estrategias según características del contaminante
4. Desarrollo de 84 características espectrales interpretables
5. Framework robusto para monitoreo ambiental automatizado

