UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS

CENTRO DE INVESTIGACIÓN EN INGENIERÍAS Y CIENCIAS APLICADAS

Modelado Y Análisis Numérico De Un Sensor De Guía De Onda Tipo Slot Para La Identificación De Contaminantes Presentes En El Agua Mediante Técnicas De Aprendizaje Automático

TESIS PROFESIONAL PARA OBTENER EL GRADO DE:

MAESTRÍA EN SUSTENTABILIDAD ENERGÉTICA

PRESENTA:

Lic. Luis Arturo Tapia Alarcón

DIRECTO: DR. Braulio Misael Villegas Martínez

CO-DIRCTORA: Dra. Gabriela Hernández Luna

CUERNAVACA, MORELOS OCTUBRE DEL AÑO 2026

Instituto de Investigación en Ciencias Básicas y Aplicadas

Resumen

En este trabajo se desarrolló una metodología basada en el modelado numérico de una guía de onda tipo *slot* y técnicas de aprendizaje automático para la identificación de contaminantes presentes en agua. A partir del índice de refracción de soluciones acuosas, se resolvió la respuesta modal de la guía de onda y se calcularon descriptores ópticos y optoelectrónicos relacionados con la constante de propagación, el confinamiento modal, la sensibilidad espectral, el volumen efectivo del *slot* y la interacción electromagnética con el medio analizado. Estos parámetros fueron utilizados como variables de entrada para entrenar y evaluar modelos de clasificación supervisada.

Se compararon árboles de decisión, Random Forest, XGBoost y un perceptrón multicapa para clasificar F^-, H_3AsO_4, KNO_3 y $NaCl$. En condiciones limpias, los modelos de ensamble y el perceptrón multicapa alcanzaron desempeños prácticamente perfectos, con métricas por clase cercanas o iguales a 1.00 y valores de AUC de 1.00. Sin embargo, la evaluación bajo perturbaciones gaussianas evidenció diferencias importantes en robustez. Random Forest y XGBoost mostraron una degradación abrupta al introducir ruido en las variables de entrada, mientras que el árbol de decisión mantuvo como principal limitación la confusión entre KNO_3 y $NaCl$.

El perceptrón multicapa presentó el mejor equilibrio entre exactitud, capacidad discriminante y estabilidad operacional. Su degradación frente al ruido fue progresiva y más controlada que la de los modelos basados en árboles, conservando valores de AUC superiores a 0.90 incluso bajo niveles elevados de perturbación. Los resultados demuestran que los descriptores ópticos derivados de una guía de onda tipo *slot* contienen información suficiente para diferenciar contaminantes en agua y que el aprendizaje automático puede integrarse como una herramienta efectiva para el desarrollo de sensores fotónicos inteligentes. Como trabajo futuro, se propone validar

experimentalmente el sistema, ampliar el conjunto de contaminantes y optimizar la arquitectura del sensor y del modelo de clasificación.

Abstract

This work developed a methodology based on the numerical modeling of a *slot* waveguide and machine learning techniques for the identification of water contaminants. Using the refractive index of aqueous solutions, the modal response of the waveguide was solved, and optical and optoelectronic descriptors were computed, including the propagation constant, modal confinement, spectral sensitivity, effective *slot* volume, and electromagnetic interaction with the analyzed medium. These physically meaningful parameters were used as input features to train and evaluate supervised classification models.

Decision Tree, Random Forest, XGBoost, and Multilayer Perceptron models were compared for the classification of F^-, H_3AsO_4, KNO_3, and $NaCl$. Under clean conditions, the ensemble models and the Multilayer Perceptron achieved nearly perfect performance, with per-class metrics close to or equal to 1.00 and AUC values of 1.00. However, the evaluation under Gaussian perturbations revealed substantial differences in robustness. Random Forest and XGBoost exhibited an abrupt performance degradation when noise was introduced into the input variables, whereas the Decision Tree remained mainly limited by the confusion between KNO_3 and $NaCl$.

The Multilayer Perceptron achieved the best balance between accuracy, discriminative capability, and operational stability. Its degradation under noise was progressive and more controlled than observed in tree-based models, maintaining AUC values above 0.90 even under high perturbation levels. These results demonstrate that optical descriptors derived from a *slot* waveguide contain sufficient information to distinguish water contaminants and that machine learning can be effectively integrated into the development of

intelligent photonic sensors. Future work should focus on experimental validation, expansion of the contaminant dataset, and optimization of both the sensor architecture and the classification model.

AGRADECIMIENTOS

Quiero agradecer a mi asesor, el Dr. Braulio Misael Villegas Martínez, por su confianza, su apoyo y su paciencia durante todo este proceso. Gracias a su gestión poder presentar mi trabajo en el Latin American Optics and Photonics Conference (LAOP), en Bogotá, Colombia, además contar con una beca por parte del RSE US. También gracias a que me puso en contacto con un investigador español he conseguido realizar una pasantía en una empresa española, además de la confianza que ha tenido en mí para presentar en congresos internacionales los trabajos que hemos desarrollado. Sus consejos y su guía me han ayudado a crecer como investigador y también como persona.

A la Dr. Gabriela Hernández Luna, por su apoyo en la búsqueda de bibliografía para sustentar mi trabajo, el apoyo que me brindo para poder participar en congresos y presentar mi trabajo, y por todas las formas en las que buscaba apoyarme y contribuir en mi investigación.

A mi familia, a mi madre Olivia Magdalena Alarcón Martínez y mi segunda madre Estela Guadalupe Alarcón Martínez, por su confianza a ciegas, su apoyo incondicional y su infinito amor, quienes me han apoyado tanto emocional como económicamente en este proceso.

A mi prometida Sahiril Fernando Rodríguez Fuentes, mi novia, mi mejor amiga, con quien siempre he podido hablar francamente de mi miedos e inseguridades, de mis dudas sobre si el camino que llevo es el adecuado, por sus consejos y su guía, por su apoyo y amor.

A mi tío el Dr. Arturo Alarcón Martínez, por siempre confiar en mi y apoyarme desde niño, por aconsejarme y cuidarme.

A mi mamá mita que aún que ya no está conmigo, sé que estaría feliz de ver lo que he avanzado.

A mis sinodales los doctores Roberto Alan Beltran Vargas, Omar Palillero Sandoval, José Antonio Marban Salgado, que con sus comentarios y observaciones me han hecho mejorar mi investigación.

Al Centro de Investigación en Ingeniería y Ciencias Aplicadas (CIICAp), por darme la oportunidad de ser parte de su comunidad y prestarme sus instalaciones durante toda mi carrera y el desarrollo de mi trabajo de tesis.

Y por último quiero agradecer a SECIHT por el apoyo económico brindado en estos dos años de mi maestría a través de la BECA NACIONAL DE ESTUDIOS DE POSGRADO, no. Solicitud: 66df659ff0b81b678081d6f7.

DEDICATORIA

A Magdalena, Estela, Fernanda y Mita.

ÍNDICE GENERAL

Resumen 1

Abstract 2

AGRADECIMIENTOS 4

DEDICATORIA 6

ÍNDICE DE FIGURAS 10

ÍNDICE DE TABLAS 13

1. INTRODUCCIÓN 15

1.1 ANTECEDENTES 15

1.2 PLANTEAMIENTO DEL PROBLEMA 17

1.3 JUSTIFICACIÓN 19

1.4 PREGUNTA DE INVESTIGACIÓN 23

1.5 HIPÓTESIS 23

OBJETIVOS 24

OBJETIVO GENERAL 24

OBJETIVOS PARTICULARES 24

2. MARCO TEORICO 26

2.1 REFRACCIÓN MOLAR E INDICE DE REFRACCIÓN 28

2.1.1 ÍNDICE DE REFRACCIÓN Y POLARIZIBILIDAD 29

2.1.2 RELACIÓN DE LORENTZ-LORENTZ 30

2.1.3. DISPERSIÓN ÓPTICA Y ECUACIÓN DE SELLMEIER 33

2.2 REFLEXIÓN TOTAL INTERNA 36

2.3 CAMPO EVANESCENTE 40

2.4 GUÍAS DE ONDA 42

2.4.1 GUÍA DE ONDA PLANAR 44

2.5 ÁNALISIS MODAL DE UNA GUÍA DE ONDA *SLOT* 45

2.6 ESTRUCTURA DE LA GUÍA DE ONDA *SLOT* Y CONDICIONES DE FRONTERA 48

2.7 FUNDAMENTOS DE APRENDIZAJE AUTOMÁTICO.....55
2.7.1 CLASIFICACIÓN DE MODELOS DE APRENDIZAJE AUTOMÁTICO..... 58
2.7.2 ARBOLES DE DECISIÓN..... 61
2.7.3 BOSQUES ALEATORIOS..... 63
2.7.4 XGBOOST..... 65
2.7.5 REDES NEURONALES ARTIFICIALES..... 67
3. METODOLOGÍA..... 72
3.1 MODELADO Y ANÁLISIS DEL INDICE DE REFRACCIÓN DE SOLUCIONES ACUOSAS.....73
3.1.1 MODELADO ESPECTRAL DEL AGUA PURA..... 73
3.1.2 INCORPORACIÓN DEL EFECTO DE LA CONCENTRACIÓN..... 75
3.1.3 APLICACIÓN AL SISTEMA $NaCl + H2O$..... 77
3.2 MODELADO ÓPTICO DE LA GUÍA DE ONDA *SLOT*.....79
3.2.1 Cálculo modal y obtención de la fracción evanescente..... 80
3.2.2 Cálculo y exportación de las propiedades derivadas..... 81
3.2.1 SOLUCIÓN DE LA ECUACIÓN TRASCENDENTAL PARA EL MODO *TM*..... 82
3.2.2 CÁLCULO DEL FACTOR DE CONFINAMIENTO $\eta slot$..... 87
3.2.3 CÁLCULO DE PARÁMETROS OPTOELECTRÓNICOS DERIVADOS DEL MODO GUIADO..... 90
3.3 METODOLOGÍA DE ENTRENAMIENTO Y EVALUACIÓN DE MODELOS DE APRENDIZAJE AUTOMÁTICO.....94
3.3.1 ANÁLISIS EXPLORATORIO DE DATOS..... 95
3.3.2 METODOLOGÍA DEL ALGORITMO ARBOL DE DECISIÓN..... 101
3.3.3 METODOLOGÍA DEL ALGORITMO BOSQUES ALEATORIOS..... 102
3.3.4 METODOLOGÍA DEL ALGORITMO XGBOOST..... 103
3.3.4 METODOLOGÍA DEL PERCEPTRON MULTICAPA..... 105
4. RESULTADOS Y DISCUSIÓN..... 107
4.1 RESULTADOS ARBOLES DE DECISIÓN.....108
4.2 RESULTADOS BOSQUES ALEATORIOS.....120
4.3 RESULTADOS XGBOOST.....130
5. CONCLUSIONES Y TRABAJO A FUTURO..... 151
5.1 TRABAJO FUTURO.....153

6. REFERENCIAS 154

APENDICES 164

APÉNDICE 1. DEDUCCIÓN MICROSCÓPICA DE LA ECUACIÓN DE SELLMEIER A PARTIR DEL MODELO DE LORENTZ 164

APÉNDICE 2. DERIVACIÓN DE LA ECUACIÓN DE HELMHOLTZ PARA UNA GUÍA DE ONDA PLANA 166

APÉNDICE 3. CONVERSIÓN DE LA FRACCIÓN MÁSICA Y COMPARACIÓN DEL MODELO CON DATOS EXPERIMENTALES 169

A3.1. Conversión de fracción másica a concentración física 169

A3.2. Coeficiente del modelo de índice de refracción 171

A3.3. Relación entre la concentración física y la concentración auxiliar 172

A3.4. Comparación con los datos experimentales de Tan y Huang 173

A3.5. Interpretación de la comparación y alcance de los resultados 174

Anexos 177

Anexo 1 177

Anexo 2 183

Cronograma de actividades 187

ÍNDICE DE FIGURAS

Figura 1.1 Comparación del consumo energético (watts-hora) de distintas metodologías, donde se resalta el bajo consumo de los dispositivos basados en fibras ópticas.	20
Figura 1.2 Comparación de la huella de carbono, mostrando la baja cantidad de emisiones en sensores de fibras ópticas respecto a los métodos tradicionales de laboratorio.	22
Figura 2.1 Esquema general del modelo físico-computacional empleado para generar y analizar los parámetros ópticos utilizados en la clasificación de las soluciones acuosas consideradas.	27
Figura 2.2 Representación esquemática de una molécula rodeada por partículas polarizadas inmersas en un campo eléctrico externo. La polarización del entorno contribuye al campo eléctrico local E_0 experimentado por la molécula central.	31
Figura 0.3 Incidencia, reflexión y transmisión de una onda electromagnética entre dos medios con distintos índices de refracción.	37
Figura 2.4 Propagación de la onda transmitida paralela a la interfaz cuando $\theta_i \theta_c$.	36
Figura 2.5 Reflexión completa de la onda incidente cuando $\theta_i > \theta_c$.	38
Figura 2.6 Representación esquemática del decaimiento exponencial del campo evanescente en el medio de menor índice de refracción (n_t) cuando $\theta_i > \theta_c$.	41
Figura 0.7 Confinamiento de la luz en una guía de onda mediante reflexión total interna. Los rayos con ángulos de incidencia superiores al ángulo crítico permanecen guiados dentro del núcleo, mientras que aquellos que no satisfacen esta condición se acoplan al medio externo y se pierden por radiación.	43
Figura 0.8 Geometría general de una guía de onda planar compuesta por un núcleo de alto índice de refracción rodeado por medios de menor índice que permiten el confinamiento óptico mediante reflexión total interna.	44
Figura 2.9 Geometría de una guía de onda plana de tres capas empleada para el análisis modal.	46
Figura 2.10 Representación conceptual de los modos guiados en una guía de onda tipo *slot*.	47
Figura 2.11 Perfil de índices de refracción y regiones empleadas para el análisis modal de una guía de onda *slot* simétrica.	48
Figura 2.12 Representación esquemática de la estructura jerárquica de un árbol de decisión. El nodo raíz contiene el conjunto inicial de datos, mientras que los nodos internos representan particiones sucesivas del espacio de características y las hojas corresponden a las clases o predicciones finales generadas por el modelo.	61
Figura 2.13 Esquema general del algoritmo de bosques aleatorios. A partir del conjunto de datos original se construyen múltiples árboles de decisión utilizando muestras y subconjuntos de características seleccionados aleatoriamente. La predicción final se obtiene mediante un proceso de votación mayoritaria entre los árboles individuales.	64
Figura 2.14 Esquema conceptual del algoritmo XGBoost. La predicción final se obtiene mediante la suma de las contribuciones individuales de múltiples árboles de decisión entrenados secuencialmente para corregir los errores de los modelos previos.	66
Figura 2.15 Estructura básica de una neurona biológica. Las dendritas reciben señales de entrada, el cuerpo celular procesa la información y el axón transmite la respuesta hacia otras neuronas mediante conexiones sinápticas.	68
Figura 2.16 Modelo matemático de una neurona artificial o perceptrón. Las variables de entrada son ponderadas mediante pesos sinápticos, combinadas	69

linealmente y transformadas mediante una función de activación para producir la salida del modelo.	
Figura 3.1 Índice de refracción del agua en función de la longitud de onda. La línea continua corresponde a los valores calculados mediante la ecuación de Sellmeier, mientras que los marcadores representan los datos experimentales reportados por Daimon.	74
Figura 3.2 índice de refracción del sistema $NaCl + H_2O$ en función de $n(\lambda, C)$ con λ=500nm obtenida mediante el modelo $n(\lambda, C) = n_{H_2O}(\lambda) + kC$.	78
Figura 3.3 Comparación entre los datos experimentales y los resultados obtenidos mediante el modelo desarrollado para solución $NaCl + H_2O$.	81
Figura 3.4 Variación de la constante de propagación modal βen función de la concentración para H_3AsO_4, F^-, KNO_3 y $NaCl$. calculada mediante la solución numérica de la ecuación trascendental *TM* de la guía de onda tipo *slot*. las subfiguras (a-j) corresponden a longitudes de onda representativas comprendidas entre 200 y 1100 nm.	82
Figura 3.5 Sensibilidad modal espectral $(\partial\beta/\partial C)$ En función de la longitud de onda para H_3AsO_4, F^-, KNO_3 Y $NaCl$.	86
Figura 3.6 Variación del factor de confinamiento en el *slot* η_{slot}en función de la concentración para H_3AsO_4, F^-, KNO_3 y $NaCl$. Las subfiguras (A–J) corresponden a longitudes de onda entre 200 y 1100 nm.	88
Figura 3.7 Sensibilidad espectral del factor de confinamiento $\partial\eta_{slot}/\partial C$en función de la longitud de onda para H_3AsO_4, F^-, KNO_3 y $NaCl$.	85
Figura 3.8 Variación del voltaje inducido en la región del *slot* V_{slot}en función de la concentración para los analitos considerados.	91
Figura 3.9 Factor de interacción electromagnética η_{factor}en función de la longitud de onda para los cuatro analitos estudiados. El incremento progresivo de este parámetro indica un aumento de la interacción campo-materia en la región espectral de mayores longitudes de onda.	92
Figura 3.10 Sensibilidad espectral S_λ en función de la longitud de onda.	93
Figura 3.11 Matriz de correlación del conjunto de datos.	97
Figura 3.12 Proyección PCA del conjunto de datos.	95
Figura 3.13 Varianza acumulada explicada por PCA.	99
Figura 4.1 Curva de aprendizaje del clasificador basado en árboles de decisión sin ruido.	110
Figura 4.2 Características más relevantes para la clasificación mediante árboles de decisión.	111
Figura 4.3 Métricas por clase del clasificador de árboles de decisión en el conjunto de prueba limpio.	112
Figura 4.4 Matriz de confusión del modelo de árboles de decisión en el conjunto de prueba limpio.	113
Figura 4.5 Curva ROC multiclase del clasificador de árboles de decisión sin ruido.	114
Figura 4.6 Precisión por clase del modelo de árboles de decisión frente a distintos niveles de ruido.	115
Figura 4.7 Matrices de confusión del modelo de árboles de decisión bajo niveles de ruido de $\mathbf{2.5}\%$, $\mathbf{5.0}\%$, $\mathbf{7.5}\%$ y $\mathbf{10.0}\%$.	116
Figura 4.8 Curvas ROC multiclase del modelo de árboles de decisión bajo niveles de ruido de $\mathbf{2.5}\%$, $\mathbf{5}\%$, $\mathbf{7.5}\%$ y $\mathbf{10.0}\%$.	118
Figura 4.9 Curva de degradación operacional del modelo de árboles de decisión frente al nivel de ruido.	119
Figura 4.10 Curva de aprendizaje del modelo de bosques aleatorios sin ruido.	121
Figura 4.11 Características más relevantes para la clasificación mediante bosques aleatorios.	122

Figura 4.12 Métricas por clase del clasificador de bosques aleatorios en el conjunto de prueba limpio.	123
Figura 4.13 Matriz de confusión del modelo de bosques aleatorios en el conjunto de prueba limpio.	124
Figura 4.14 Curva ROC multiclase del clasificador de bosques aleatorios sin ruido.	125
Figura 4.15 Precisión por clase del modelo de bosques aleatorios frente a distintos niveles de ruido.	126
Figura 4.16 Matrices de confusión del modelo de bosques aleatorios bajo niveles de ruido de $\mathbf{2.5}\%$, $\mathbf{5}.0\%$, $\mathbf{7.5}\%$ y $\mathbf{10}.0\%$.	127
Figura 4.17 Curvas ROC multiclase del modelo de bosques aleatorios bajo niveles de ruido de $\mathbf{2.5}\%$, $\mathbf{5.0}\%$, $\mathbf{7.5}\%$ y $\mathbf{10.0}\%$.	128
Figura 4.18 Curva de degradación operacional del modelo de bosques aleatorios frente al nivel de ruido.	129
Figura 4.19 Curva de aprendizaje del modelo XGBoost sin ruido.	131
Figura 4.20 Diez características más relevantes para la clasificación mediante XGBoost.	132
Figura 4.21 Métricas por clase del clasificador XGBoost en el conjunto de prueba limpio.	133
Figura 4.22 Matriz de confusión del modelo XGBoost en el conjunto de prueba limpio.	134
Figura 4.23 Curva ROC multiclase del clasificador XGBoost sin ruido.	135
Figura 4.24 Precisión por clase del modelo XGBoost frente a distintos niveles de ruido.	136
Figura 4.25 Matrices de confusión del modelo XGBoost bajo niveles de ruido de $\mathbf{2.5}\%$, $\mathbf{5.0}\%$, $\mathbf{7.5}\%$ y $\mathbf{10.0}\%$.	137
Figura 4.26 Curvas ROC multiclase del modelo XGBoost bajo niveles de ruido de $\mathbf{2.5}\%$, $\mathbf{5.0}\%$, $\mathbf{7.5}\%$ y $\mathbf{10.0}\%$.	138
Figura 4.27 Curva de degradación operacional del modelo XGBoost frente al nivel de ruido.	138
Figura 4.28 Curva de pérdida del modelo MLP durante el entrenamiento.	142
Figura 4.29 Curva de precisión del modelo MLP durante el entrenamiento.	142
Figura 4.30 Matriz de confusión del modelo MLP en el conjunto de validación limpio.	143
Figura 4.31 Métricas por clase del modelo MLP en el conjunto de prueba limpio.	144
Figura 4.32 Diez características más relevantes para la clasificación mediante el modelo MLP.	145
Figura 4.33 Curva ROC multiclase del modelo MLP en el conjunto de prueba limpio.	141
Figura 4.34 Curva de aprendizaje del modelo MLP en función del número de muestras.	146
Figura 4.35 Matrices de confusión del modelo MLP bajo niveles de ruido de 2.5 %, 5.0 %, 7.5 % y 10.0 %.	147
Figura 4.36 Precisión por clase del modelo MLP frente a distintos niveles de ruido.	148
Figura 4.37 Curva de degradación operacional del modelo MLP frente al nivel de ruido.	149

ÍNDICE DE TABLAS

Tabla 2.1 Coeficientes de Sellmeier para agua destilada a 20°C.	35
Tabla 2.2 Clasificación de las guías de onda en función de las dimensiones en las que permiten confinar la luz.	43
Tabla 2.3 Analogía entre los elementos de una neurona biológica y una neurona artificial.	69
Tabla 3.1 Propiedades fisicoquímicas empleadas para el cálculo del factor de ajuste $\boldsymbol{k}$ de cada analito.	76
Tabla 3.2 Parámetros geométricos y ópticos de la guía de onda tipo *slot*.	79
Tabla 3.3 Distribución de muestras por clase.	96
Tabla 3.4 Estadísticas de las variables de entrenamiento.	92
Tabla 3.5 Variables con mayor correlación absoluta.	98
Tabla 3.6 Espacio de búsqueda de hiperparámetros empleado durante la optimización bayesiana del clasificador basado en árboles.	102
Tabla 3.7 Hiperparámetros y espacio de búsqueda para la optimización del clasificador de bosques aleatorios.	103
Tabla 3.8 Hiperparámetros y rangos de búsqueda considerados durante la optimización bayesiana del clasificador XGBoost.	104
Tabla 3.9 Hiperparámetros y espacio de búsqueda para la optimización bayesiana de la red neuronal tipo perceptrón multicapa.	106
Tabla 4.1 Hiperparámetros óptimos obtenidos para el clasificador árboles de decisión.	108
Tabla 4.2 Resultados de validación cruzada para el clasificador basado en árboles de decisión.	109
Tabla 4.3 Reporte de clasificación del modelo de árboles de decisión en el conjunto de prueba limpio.	111
Tabla 4.4 Degradación del desempeño global del modelo de árboles de decisión bajo diferentes niveles de ruido.	114
Tabla 4.5 Hiperparámetros óptimos obtenidos para el clasificador basado en bosques aleatorios.	120
Tabla 4.6 Resultados de validación cruzada para el clasificador basado en bosques aleatorios.	120
Tabla 4.7 Reporte de clasificación del modelo de bosques aleatorios en el conjunto de prueba limpio.	123
Tabla 4.8 Degradación del desempeño global del modelo de bosques aleatorios bajo diferentes niveles de ruido.	125
Tabla 4.9 Hiperparámetros óptimos obtenidos para el clasificador XGBoost.	130
Tabla 4.10 Resultados de validación cruzada para el clasificador XGBoost.	130
Tabla 4.11 Reporte de clasificación del modelo XGBoost en el conjunto de prueba limpio.	132
Tabla 4.12 Degradación del desempeño global del modelo XGBoost bajo diferentes niveles de ruido.	135
Tabla 4.13 Resultados de la búsqueda del tamaño de lote para el modelo MLP.	140

Tabla 4.14 Arquitectura general del modelo perceptrón multicapa.	140
Tabla 4.15 Resumen de parámetros del modelo MLP.	141

1. INTRODUCCIÓN

1.1 ANTECEDENTES

El agua constituye uno de los recursos naturales más importantes para el desarrollo humano, la producción agrícola, los procesos industriales y el mantenimiento de los ecosistemas. La disponibilidad de agua potable segura y de calidad es un requisito fundamental para la salud pública, la seguridad alimentaria y el desarrollo sostenible de las sociedades modernas. [1-4] Sin embargo, el crecimiento poblacional, la industrialización, la expansión de las actividades agrícolas y el aumento en la demanda de agua han intensificado la presión sobre las fuentes superficiales y subterráneas, con efectos tanto en la disponibilidad como en la calidad del recurso hídrico. [5-7]

En México, la calidad del agua es una problemática de especial relevancia: factores geológicos y actividades antropogénicas se combinan para favorecer la presencia de contaminantes en fuentes superficiales y subterráneas destinadas al abastecimiento y consumo humano. [8-10] Entre los de mayor interés se encuentran el arsénico, el fluoruro, los nitratos y diversas sales disueltas, cuya presencia en concentraciones elevadas puede generar efectos adversos en la salud humana y el medio ambiente. [11,12] Estas condiciones han motivado el desarrollo de sistemas de monitoreo capaces de proporcionar información confiable y oportuna sobre la composición química del agua.

La determinación de contaminantes en agua se realiza mediante técnicas como la espectrometría de masas con plasma acoplado inductivamente, la espectroscopía de absorción atómica, los sensores electroquímicos y los métodos conductivos. [13-16] Estas técnicas ofrecen elevada sensibilidad y precisión analítica; no obstante, suelen requerir infraestructura especializada, personal capacitado, preparación de las muestras y condiciones de operación controladas, lo que dificulta su implementación en aplicaciones de monitoreo

continuo, remoto o distribuido. Estas limitaciones han impulsado la búsqueda de tecnologías alternativas que permitan realizar mediciones rápidas, remotas y con menores requerimientos operativos.

Dentro de estas alternativas, los sensores ópticos han despertado un interés considerable debido a su elevada sensibilidad, su inmunidad a interferencias electromagnéticas, su capacidad de miniaturización y su potencial de integración en plataformas fotónicas compactas. [19-21] En particular, los sensores basados en guías de onda aprovechan la interacción entre el campo evanescente y el medio circundante para detectar variaciones en el índice de refracción producidas por cambios en la composición química de una muestra. Entre estas estructuras, las guías de onda tipo *slot* resultan especialmente prometedoras, ya que concentran una fracción significativa del campo electromagnético en la región de baja permitividad donde se encuentra el analito, lo que incrementa la interacción entre la luz y el medio de interés.

Los sensores ópticos también ofrecen posibilidades de desarrollo vinculadas con la sustentabilidad energética. En comparación con algunas técnicas instrumentales convencionales, los sistemas fotónicos integrados presentan potencial para operar con menores requerimientos energéticos y para integrarse en dispositivos compactos, características de interés en aplicaciones continuas, distribuidas y de largo plazo, donde el consumo energético es un factor relevante de diseño. [22,23] Esta ventaja, sin embargo, debe evaluarse considerando el sistema completo: la fuente óptica, los elementos de detección, la adquisición de datos y el procesamiento de la información.

Por otra parte, el modelado teórico de estas estructuras permite estudiar su respuesta antes de avanzar hacia la fabricación y la validación experimental. Mediante la solución numérica de las ecuaciones características de la guía de onda y el uso de modelos de dispersión óptica que relacionan la composición de las soluciones acuosas con su índice de refracción, es posible obtener

descriptores ópticos que representan la respuesta de la estructura ante distintos índices de refracción, concentraciones y condiciones de operación.

Los descriptores generados de esta forma conforman un conjunto de datos multivariable en el que pueden existir relaciones no lineales, un escenario en el que el aprendizaje automático resulta una herramienta adecuada. Los avances recientes en inteligencia artificial han permitido construir modelos capaces de identificar patrones complejos en datos experimentales y simulados, con resultados prometedores en aplicaciones relacionadas con la calidad del agua, incluyendo tareas de clasificación, predicción y estimación de contaminantes. [15, 16] Asimismo, diversos estudios han explorado la combinación de algoritmos de aprendizaje automático con sensores ópticos, dispositivos fotónicos y estructuras nanofotónicas para mejorar el procesamiento y la interpretación de su respuesta. [24, 25]

Aunque existen investigaciones enfocadas en el diseño de sensores ópticos y otras orientadas a aplicar algoritmos de aprendizaje automático sobre datos experimentales o sintéticos, la integración de ambas aproximaciones continúa siendo limitada para la clasificación de soluciones acuosas asociadas con arsénico, fluoruro, nitrato y cloruro de sodio a partir de descriptores obtenidos del modelado de una guía de onda tipo *slot*.

1.2 PLANTEAMIENTO DEL PROBLEMA

La presencia de contaminantes como el arsénico, el fluoruro, los nitratos y las sales disueltas compromete el aprovechamiento seguro del agua cuando su naturaleza y concentración afectan la salud humana. Vigilarlos en sitio, de forma continua o distribuida y sin requerir la presencia constante de personal, exige sistemas compactos y de bajos requerimientos operativos. Las técnicas convencionales de laboratorio ofrecen un desempeño analítico adecuado,

pero su infraestructura, su personal especializado y sus procedimientos de laboratorio dificultan su implementación en este tipo de escenarios.

Los sensores ópticos basados en guías de onda tipo *slot* y el aprendizaje automático se perfilan como herramientas complementarias para este fin: las primeras aprovechan la interacción de la luz con el medio circundante para detectar cambios en las propiedades ópticas del agua, y el segundo permite interpretar la respuesta de dispositivos fotónicos. Sin embargo, su integración para clasificar soluciones acuosas con arsénico, fluoruro, nitrato y cloruro de sodio a partir de descriptores obtenidos del modelado de la guía continúa siendo limitada. Avanzar en esta dirección requiere, por un lado, obtener parámetros ópticos sensibles a los cambios del medio y, por otro, disponer de algoritmos que relacionen esas respuestas con el contaminante presente.

La cuestión es si la respuesta óptica calculada de una guía de onda tipo *slot* proporciona información suficiente para distinguir entre los contaminantes cuando varían sus concentraciones y las condiciones de operación, y si las diferencias entre los parámetros obtenidos permiten establecer una clasificación consistente. Por ello, debe evaluarse la capacidad de discriminación del conjunto formado por el modelo físico y los algoritmos de aprendizaje automático.

La presente investigación se limita al modelado analítico mediante la solución numérica de la ecuación trascendental característica de una guía de onda tipo *slot* y de los modelos de dispersión utilizados para calcular el índice de refracción, y a la clasificación computacional de muestras acuosas modeladas con presencia de arsénico, fluoruro, nitrato o cloruro de sodio de forma individual. Con estos resultados se construye un conjunto de datos sintéticos para entrenar y evaluar árboles de decisión, bosques aleatorios, XGBoost y redes neuronales artificiales. No se contempla la fabricación ni la validación experimental del sensor, por lo que los resultados no constituyen por sí mismos la demostración de un sistema de detección operando con muestras reales.

Su aportación consiste en establecer el alcance y las limitaciones de esta relación, evaluando el potencial discriminante de los parámetros ópticos generados mediante simulación e identificando configuraciones que podrían priorizarse en investigaciones experimentales posteriores. A largo plazo, estos conocimientos podrían contribuir a tecnologías de monitoreo útiles para responsables de la gestión del agua y comunidades que requieren información oportuna sobre su calidad.

1.3 JUSTIFICACIÓN

Los métodos empleados para la detección y cuantificación de contaminantes en agua han evolucionado hacia sistemas cada vez más precisos y sensibles. No obstante, muchas de las técnicas analíticas utilizadas actualmente requieren infraestructura especializada, elevados requerimientos energéticos y condiciones controladas de operación, factores que incrementan tanto los costos de análisis como el impacto ambiental asociado a su uso. Por ello, existe un interés creciente en el desarrollo de tecnologías alternativas que permitan realizar mediciones confiables con un menor consumo de recursos.

De acuerdo con la literatura y las especificaciones técnicas de los equipos utilizados para el análisis de contaminantes en agua, los métodos convencionales de laboratorio presentan una demanda energética considerable debido a la complejidad de sus sistemas de excitación, detección y procesamiento. [25-42]. La Figura 1.1 muestra una comparación del consumo energético estimado para diferentes tecnologías analíticas utilizadas en la determinación de contaminantes químicos.

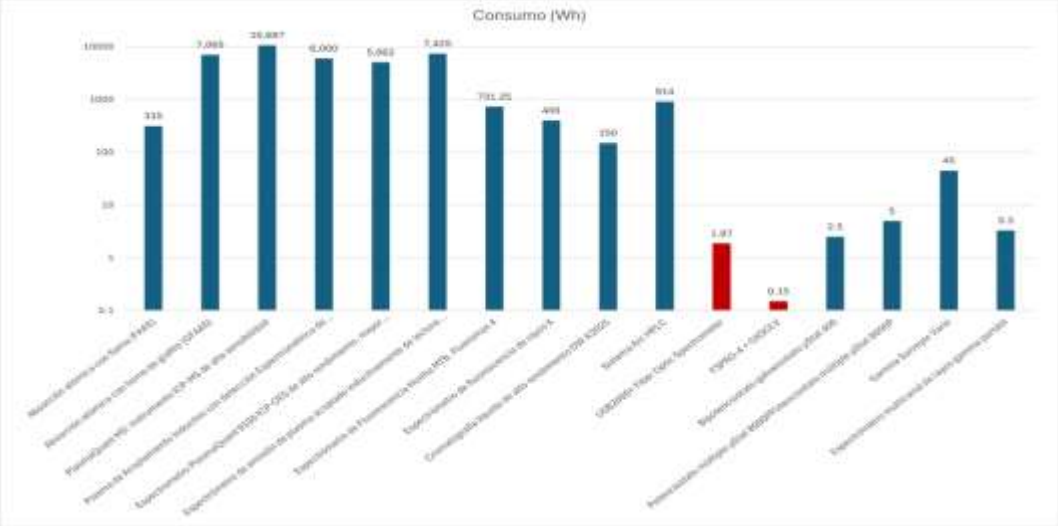


FIGURA 1.1 COMPARACIÓN DEL CONSUMO ENERGÉTICO (WATTS-HORA) DE DISTINTAS METODOLOGÍAS, DONDE SE RESALTA EL BAJO CONSUMO DE LOS DISPOSITIVOS BASADOS EN FIBRAS ÓPTICAS. [25]-[42]

Entre las técnicas de mayor consumo energético se encuentran la espectrometría de masas con plasma de acoplamiento inductivo (ICP-MS) y la espectrometría de absorción atómica con horno de grafito (GFAAS), cuyos requerimientos energéticos alcanzan aproximadamente 10,687 Wh y 7,065 Wh, respectivamente. [26, 28] En contraste, los sistemas fotónicos basados en fibra óptica presentan consumos significativamente menores. Por ejemplo, el espectrómetro USB2000+ registra un consumo aproximado de 1.25 Wh, mientras que el sistema FSPRO-4 + OXSOLV requiere únicamente 0.15 Wh [35, 36].

Es necesario precisar que estos equipos se incluyen como ejemplos de sistemas ópticos basados en fibra óptica y no como dispositivos con capacidades analíticas equivalentes a las técnicas de detección de contaminantes propuestas. El sistema FSPRO-4, en combinación con la sonda OXSOLV, permite medir oxígeno disuelto, pero no identificar ni cuantificar los contaminantes. Por su parte, el USB2000+ es un espectrómetro acoplado a fibra óptica que registra información espectral, su utilización para detectar

sustancias específicas requiere una configuración de medición y una metodología analítica adaptada. En consecuencia, ambos equipos se emplean únicamente como referencias del consumo energético de dispositivos fotónicos, sin atribuirles la capacidad de identificar por sí solos arsénico, fluoruro, nitrato o cloruro de sodio en agua.

Estos valores representan reducciones energéticas superiores al 99.9 respecto a las técnicas convencionales de laboratorio. Particularmente, al comparar el sistema USB2000+ con un equipo ICP-MS se obtiene una reducción aproximada del 99.98%, mientras que el sistema FSPRO-4 + OXSOLV alcanza una disminución cercana al 99.999. Estas cifras se calculan a partir de las especificaciones técnicas de equipos individuales y no constituyen una comparación experimental directa; el consumo de un sistema completo basado en una guía de onda tipo *slot* dependerá además de la fuente óptica, los elementos de detección, la adquisición de datos y el procesamiento de la información. Los sistemas de fibra óptica se emplean aquí como referencia del potencial energético de la tecnología fotónica. La disminución en la demanda energética implica una reducción de las emisiones indirectas asociadas al consumo eléctrico, cuando se utiliza el mismo factor de emisión. Para estimar las emisiones indirectas asociadas al consumo eléctrico, se empleó el factor de emisión del Sistema Eléctrico Nacional 2024, de $0.444 tCO_2e/MWh$, indicado por la SEMARNAT en el documento “Aviso a los establecimientos sujetos a reporte en el Registro Nacional de Emisiones (RENE)”, publicado el 29 de mayo de 2026 [43]. Las emisiones se calcularon multiplicando el consumo energético estimado de cada equipo por dicho factor, con la conversión de unidades correspondiente. Los resultados de este cálculo se presentan en la Figura 1.2.

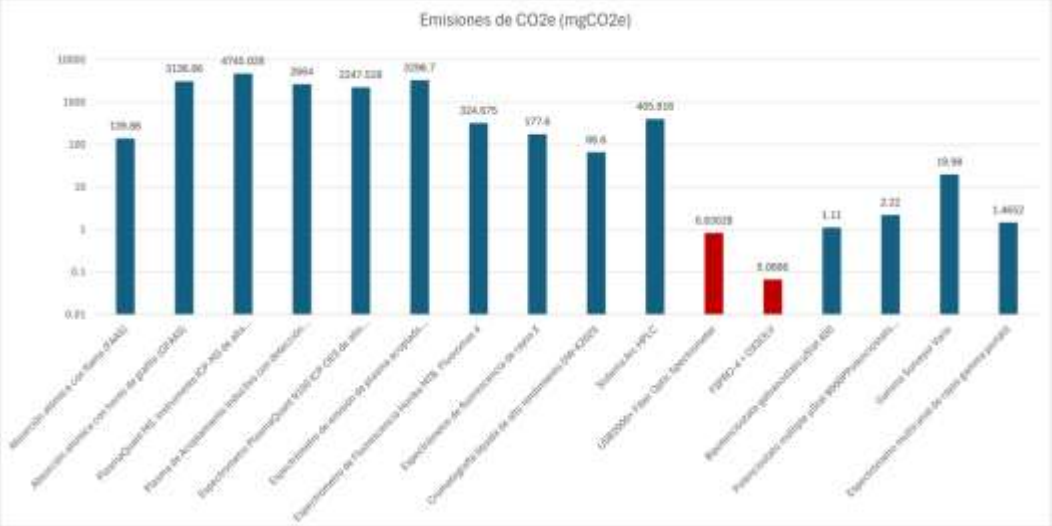


FIGURA 1.2 EMISIONES INDIRECTAS ESTIMADAS ASOCIADAS AL CONSUMO ELÉCTRICO DE LAS TECNOLOGÍAS ANALIZADAS. ELABORACIÓN PROPIA A PARTIR DE LOS CONSUMOS ENERGÉTICOS DE LAS REFERENCIAS [25]–[42] Y DEL FACTOR DE EMISIÓN DEL SISTEMA ELÉCTRICO NACIONAL PUBLICADO POR LA SEMARNAT [43].

Mientras que técnicas como ICP-MS y GFAAS generan aproximadamente 4745 $mgCO_2e$ y 3137 $mgCO_2e$ por operación, respectivamente, los sistemas ópticos presentan emisiones considerablemente menores. El espectrómetro USB2000+ produce alrededor de 0.830 $mgCO_2e$ y el sistema FSPRO-4 + OXSOLV únicamente 0.0666 $mgCO_2e$. En consecuencia, las emisiones asociadas a los sensores ópticos representan menos del 0.02 de aquellas generadas por los métodos convencionales de laboratorio.

Además de las ventajas asociadas a la eficiencia energética y la reducción de emisiones, los sensores ópticos ofrecen características atractivas para aplicaciones de monitoreo continuo, tales como inmunidad a interferencias electromagnéticas, posibilidad de miniaturización, operación remota y capacidad de integración en plataformas fotónicas compactas [35, 37, 38].

En este contexto, la presente investigación propone una metodología computacional que combina el modelado de una guía de onda tipo *slot* con algoritmos de aprendizaje automático para la clasificación de arsénico,

fluoruro, nitrato y cloruro de sodio en agua. La integración de fotónica, modelado físico e inteligencia artificial busca sentar las bases de sensores inteligentes que ofrezcan una alternativa de bajo consumo energético y reducido impacto ambiental frente a las metodologías convencionales de laboratorio, y permite orientar la selección de configuraciones antes de la fabricación, disminuyendo el número de alternativas que deben explorarse experimentalmente.

1.4 PREGUNTA DE INVESTIGACIÓN

¿En qué medida los parámetros ópticos obtenidos mediante el modelado analítico de la guía de onda tipo *slot* permiten clasificar, mediante algoritmos de aprendizaje automático, muestras acuosas con presencia de arsénico, fluoruro, nitrato o cloruro de sodio, considerando diferentes concentraciones y condiciones de operación?

1.5 HIPÓTESIS

Los parámetros ópticos obtenidos mediante el modelado analítico de la guía de onda tipo *slot* contienen información discriminante sobre el tipo de contaminante presente en las muestras acuosas modeladas. En consecuencia, su utilización como variables de entrada permitirá que al menos uno de los algoritmos de aprendizaje automático evaluados clasifique las muestras con una exactitud superior al 90% en diferentes condiciones de operación.

OBJETIVOS

OBJETIVO GENERAL

Desarrollar y evaluar una metodología computacional que integre un modelo teórico de una guía de onda tipo *slot*, resuelto numéricamente, con técnicas de aprendizaje automático, para analizar la capacidad discriminante de los parámetros ópticos en la clasificación de soluciones acuosas con diferentes concentraciones de arsénico, fluoruro, nitrato y cloruro de sodio, con la finalidad de establecer una base para el desarrollo futuro de sistemas fotónicos inteligentes con potencial de operación energéticamente eficiente.

OBJETIVOS PARTICULARES

I. Analizar la dependencia del índice de refracción de las soluciones acuosas respecto de la concentración de arsénico, fluoruro, nitrato y cloruro de sodio mediante modelos de refracción molar y la ecuación de dispersión de Sellmeier.

II. Implementar computacionalmente el modelo teórico de una guía de onda tipo *slot* mediante la solución numérica de su ecuación trascendental característica, con el fin de obtener los parámetros ópticos correspondientes a diferentes condiciones de operación e índices de refracción de las soluciones acuosas.

III. Analizar la distribución del campo electromagnético y su interacción con el medio acuoso, incluida la fracción de energía confinada en la región *slot*, para cuantificar la sensibilidad de los parámetros ópticos de la guía de onda frente a las variaciones del índice de refracción producidas por las concentraciones de los analitos estudiados.

IV. Generar un conjunto de datos sintéticos a partir de los parámetros ópticos obtenidos mediante el modelo teórico implementado computacionalmente para las diferentes concentraciones de arsénico, fluoruro, nitrato y cloruro de sodio consideradas en el estudio.

V. Implementar y entrenar modelos de aprendizaje automático correspondientes a árboles de decisión, bosques aleatorios, XGBoost y redes neuronales artificiales, utilizando el conjunto de datos sintéticos generado mediante el modelo fotónico.

VI. Evaluar y comparar el desempeño de los modelos de aprendizaje automático mediante métricas de clasificación, con el propósito de identificar el algoritmo con mayor capacidad para discriminar entre las clases de soluciones acuosas consideradas.

2. MARCO TEORICO

El modelo físico-computacional desarrollado en esta investigación describe la relación entre la concentración de diferentes contaminantes disueltos en agua y la respuesta óptica de una guía de onda tipo slot. Para ello, la metodología integra fundamentos relacionados con las propiedades ópticas de las soluciones acuosas, la interacción entre la luz y la materia, la propagación electromagnética en medios dieléctricos y el aprendizaje automático. Esta integración permite representar de manera secuencial los procesos que conducen desde una variación en la composición del medio hasta la clasificación de las soluciones consideradas.

La concentración de un analito modifica las propiedades físicas y ópticas de una solución acuosa, particularmente su índice de refracción, el cual también depende de factores como la temperatura y la longitud de onda de operación [60-63]. La relación entre la composición química y el índice de refracción puede estudiarse mediante la polarizabilidad molecular, la refracción molar, la relación de Lorentz-Lorenz y los modelos de dispersión óptica, entre ellos la ecuación de Sellmeier. Estos fundamentos permiten calcular el índice de refracción correspondiente a las diferentes concentraciones de arsénico, fluoruro, nitrato y cloruro de sodio consideradas en este trabajo.

Los valores del índice de refracción constituyen la entrada del modelo teórico de la guía de onda tipo slot, implementado computacionalmente y resuelto mediante métodos numéricos. A partir de la solución de sus ecuaciones características se obtienen parámetros ópticos como el índice efectivo, la constante de propagación, la distribución del campo electromagnético y la fracción de energía confinada en la región slot. Estos parámetros permiten caracterizar la respuesta modal de la estructura frente a las variaciones del índice de refracción producidas por las diferentes concentraciones de los analitos.

Debido a que una variación del índice de refracción no proporciona necesariamente una respuesta selectiva para distinguir por sí sola entre diferentes especies químicas, los parámetros obtenidos mediante el modelo de la guía se utilizan como descriptores para construir un conjunto de datos sintéticos. Posteriormente, este conjunto de datos se emplea para entrenar y evaluar algoritmos de aprendizaje automático correspondientes a árboles de decisión, bosques aleatorios, XGBoost y redes neuronales artificiales. De esta manera, el aprendizaje automático permite analizar de forma conjunta las relaciones entre los descriptores ópticos y determinar su capacidad para discriminar entre las clases de soluciones acuosas consideradas.

La Figura 2.1 resume la arquitectura conceptual del modelo físico-computacional empleado en esta investigación. El esquema representa tres etapas principales: la transducción de las propiedades de la solución en variaciones del índice de refracción; el modelado de la respuesta óptica de la guía de onda; y el procesamiento de los descriptores mediante técnicas de aprendizaje automático. Esta secuencia permite relacionar los fundamentos físicos, matemáticos y computacionales que sustentan la metodología utilizada en los capítulos posteriores.

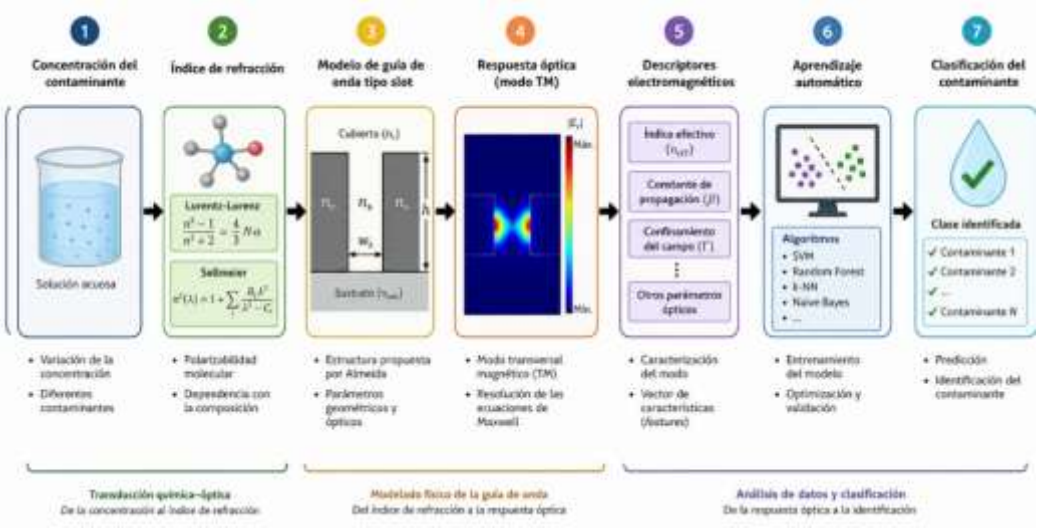


FIGURA 2.1 ESQUEMA GENERAL DEL MODELO FÍSICO-COMPUTACIONAL EMPLEADO PARA GENERAR Y ANALIZAR LOS PARÁMETROS ÓPTICOS UTILIZADOS EN LA CLASIFICACIÓN DE LAS SOLUCIONES ACUOSAS CONSIDERADAS.

De manera complementaria, las tecnologías fotónicas pueden analizarse desde la perspectiva de la sustentabilidad energética. Como se expuso en la justificación, las comparaciones realizadas a partir de la literatura científica y de las especificaciones técnicas disponibles muestran que determinados sistemas ópticos pueden presentar un menor consumo energético operacional y menores emisiones indirectas asociadas al uso de electricidad que algunas técnicas instrumentales convencionales [25-42]. Estos antecedentes no constituyen una medición directa del consumo de la guía tipo slot propuesta, pero permiten contextualizar el potencial de las plataformas fotónicas para el desarrollo futuro de sistemas de monitoreo energéticamente eficientes.

Con el propósito de establecer las bases necesarias para comprender la metodología propuesta, las siguientes secciones desarrollan inicialmente la relación entre la composición de una solución y su índice de refracción mediante los conceptos de polarizabilidad molecular, refracción molar, relación de Lorentz-Lorenz y dispersión óptica. Posteriormente, se presentan los principios de reflexión total interna, campo evanescente y propagación electromagnética en guías de onda dieléctricas. Sobre estos fundamentos se aborda el modelo teórico de la guía de onda tipo slot, incluyendo sus ecuaciones características, la solución del problema modal y la obtención de los parámetros ópticos utilizados como descriptores. Finalmente, se introducen los fundamentos de los algoritmos de aprendizaje automático empleados para el entrenamiento, evaluación y comparación de los modelos de clasificación.

2.1 REFRACCIÓN MOLAR E INDICE DE REFRACCIÓN

La refracción molar constituye una magnitud que relaciona las propiedades microscópicas de la materia, particularmente la polarizabilidad molecular, con propiedades macroscópicas observables como el índice de refracción. Esta

relación proporciona una base teórica para describir cómo la composición, densidad y estructura molecular de un medio influyen en su respuesta óptica.

En esta sección se presentan los fundamentos físicos que vinculan la polarización inducida por un campo electromagnético con el índice de refracción del material. Posteriormente, se desarrolla la relación de Lorentz-Lorenz y los modelos de dispersión óptica basados en la ecuación de Sellmeier, los cuales permiten describir la dependencia espectral del índice de refracción y constituyen la base para el modelado de soluciones acuosas empleadas en este trabajo.

2.1.1 ÍNDICE DE REFRACCIÓN Y POLARIZIBILIDAD

La interacción entre la radiación electromagnética y la materia constituye el fundamento de numerosos fenómenos ópticos. Cuando una onda electromagnética se propaga a través de un material dieléctrico, el campo eléctrico asociado induce desplazamientos en las cargas ligadas de los átomos y moléculas, generando una redistribución de la densidad electrónica. Como consecuencia, se produce una polarización inducida que modifica la respuesta óptica del medio [51].

Una de las magnitudes macroscópicas que describe esta interacción es el índice de refracción, n, definido como la razón entre la velocidad de la luz en el vacío y su velocidad de propagación en el medio:

$$n = \frac{c}{v}, \qquad (1)$$

donde, c, es la velocidad de la luz en el vacío y v es la velocidad de propagación de la radiación electromagnética en el material.

A nivel molecular, la acción del campo eléctrico induce un momento dipolar μ_{ind} en cada molécula. Para un material isótropo, este momento dipolar inducido es proporcional al campo eléctrico local E_{loc} experimentado por la molécula:

$$\mu_{ind} = \alpha_{\mathrm{pol}} E_{loc}, \qquad (2)$$

donde α_{pol} , es la polarizabilidad molecular, magnitud que cuantifica la capacidad de deformación de la nube electrónica bajo la acción de un campo eléctrico externo. Cuanto mayor sea la polarizabilidad de una molécula, mayor será la polarización inducida y, en consecuencia, más significativa será su contribución a las propiedades ópticas del medio [51, 52].

La polarizabilidad molecular puede expresarse en función del volumen de polarizabilidad, α'_{pol}, mediante la relación

$$\alpha_{pol} = 4\pi\epsilon_0 \alpha'_{pol}, \qquad (3)$$

donde ϵ_0 representa la permitividad eléctrica en el vacío [51].

La dependencia entre polarizabilidad e índice de refracción constituye la base para comprender cómo cambios en la composición química de una sustancia modifican su respuesta óptica. En sistemas líquidos y soluciones acuosas, la incorporación de especies químicas altera la polarizabilidad efectiva del medio, produciendo variaciones medibles en el índice de refracción. Esta relación resulta particularmente relevante en aplicaciones de detección óptica, donde el índice de refracción puede emplearse como una variable sensible a cambios de concentración y composición química [53].

2.1.2 RELACIÓN DE LORENTZ-LORENTZ

La relación de Lorentz-Lorenz establece un vínculo entre la polarizabilidad molecular, una propiedad microscópica de la materia, y el índice de refracción, una magnitud macroscópica observable. Esta formulación constituye uno de los modelos fundamentales para describir la respuesta óptica de medios dieléctricos homogéneos e isotrópicos [51].

La derivación de esta relación considera que una molécula inmersa en un medio no experimenta únicamente el campo eléctrico aplicado externamente, sino también la contribución de las moléculas vecinas polarizadas, como se observa en la Figura 2.1.

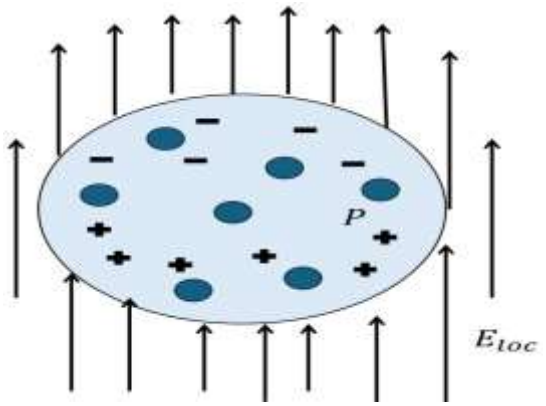


FIGURA 2.2 REPRESENTACIÓN ESQUEMÁTICA DE UNA MOLÉCULA RODEADA POR PARTÍCULAS POLARIZADAS INMERSAS EN UN CAMPO ELÉCTRICO EXTERNO. LA POLARIZACIÓN DEL ENTORNO CONTRIBUYE AL CAMPO ELÉCTRICO LOCAL $\boldsymbol{E_0}$ EXPERIMENTADO POR LA MOLÉCULA CENTRAL [51].

Bajo estas condiciones, el campo eléctrico local puede aproximarse mediante,

$$E_{loc} = E + \frac{P}{3\epsilon_0}, \qquad (4)$$

donde E es el campo eléctrico macroscópico, P la polarización del medio y $P/3\epsilon_0$ el denominado campo de Lorentz.

La polarización macroscópica depende de la densidad numérica de moléculas presentes en el medio, la cual se expresa como

$$N = \frac{N_A \rho}{M}, \qquad (5)$$

donde $N_A = 6.022x10^{23}mol^{-1}$ es el número de Avogadro, ρ la densidad del material y M su masa molar. Por otra parte, la polarización total del medio puede escribirse como

$$P = N\mu_{ind}, \tag{6}$$

dado que la polarización P, esta relacionado con el campo eléctrico y la permitividad relativa de la forma

$$P = (\epsilon_r - 1)\epsilon_0 E. \tag{7}$$

Para medios no magnéticos, la permitividad relativa y el índice de refracción se relacionan mediante

$$n^2 = \epsilon_r, \tag{8}$$

sustituyendo las ecuaciones (5)–(8) en la expresión del campo local y utilizando la definición de polarizabilidad dada por la ecuación (2), se obtiene la relación de Lorentz-Lorenz:

$$\frac{n^2 - 1}{n^2 + 2} = \frac{N_A \rho \alpha_{pol}}{3M\epsilon_0}. \tag{9}$$

Esta ecuación establece que el índice de refracción depende directamente de la polarizabilidad molecular y de la densidad del medio. A partir de la ecuación 9 se define la refracción molar R_m dada por,

$$R_m = \frac{n^2 - 1}{n^2 + 2}\frac{M}{\rho}, \tag{10}$$

o de forma equivalente,

$$R_m = \frac{N_A \alpha_{pol}}{3\epsilon_0}. \tag{11}$$

La refracción molar representa una medida de la contribución óptica de una molécula al índice de refracción del sistema. Debido a su dependencia con la polarizabilidad molecular, esta magnitud permite relacionar propiedades estructurales y electrónicas de las especies químicas con parámetros ópticos observables [53].

Este parámetro permite entender mejor la interacción entre la luz y la materia por ejemplo en soluciones o mezclas con múltiples componentes $MX + H_2O$ o $M_2X + H_2O$, ya que en presencia de un campo eléctrico surgen de reordenamientos estructurales e interacciones intermoleculares, que alteran la polarizabilidad electrónica y la eficiencia de empaquetamiento. Por ejemplo, el exceso de volumen molar en estos sistemas se correlaciona con desviaciones del comportamiento ideal de mezcla, lo que destaca el papel de las interacciones disolvente-soluto en la modulación del índice de refracción molar.

El índice de refracción de las soluciones también presenta relaciones lineales con las concentraciones molares de soluto, como se demuestra en sistemas $NaCl + H_2O$ o $K_2SO_4 + H_2O$ [54]. Esta dependencia permite aplicaciones de detección óptica de contaminantes mediante mediciones precisas de cambios en el índice de refracción. La respuesta del medio al campo eléctrico de la onda luminosa está relacionada con la polarización inducida en el material, la cual contribuye a determinar sus propiedades ópticas. La respuesta específica también varía con la naturaleza química de los contaminantes, ya que las especies con mayor capacidad electrónica o hiperpolarizabilidad, muestran mayores cambios inducidos por el campo [54].

2.1.3. DISPERSIÓN ÓPTICA Y ECUACIÓN DE SELLMEIER

La relación de Lorentz-Lorenz permite establecer una conexión entre la polarizabilidad molecular y el índice de refracción de un medio. Sin embargo, esta formulación no describe explícitamente la dependencia espectral del índice de refracción. En general, el índice de refracción n varía con la frecuencia o la longitud de onda de la radiación incidente, fenómeno conocido como dispersión óptica [55].

La relación de Lorentz–Lorenz, presentada en la ecuación (9), establece el vínculo entre el índice de refracción y la polarizabilidad promedio α_{pol} de un medio denso. No obstante, las deducciones previas se formulan bajo el supuesto de una respuesta estática o de frecuencia fija del campo eléctrico. En general, el índice de refracción depende de la frecuencia de la radiación electromagnética incidente, fenómeno conocido como dispersión óptica [56].

Desde la electrodinámica clásica, la dispersión puede describirse mediante el modelo de Lorentz, en el cual los electrones ligados se consideran osciladores armónicos sometidos a la acción de un campo electromagnético externo. En regiones espectrales alejadas de las resonancias de absorción, la respuesta del material puede expresarse mediante la ecuación de Sellmeier. La deducción microscópica detallada de esta formulación, partiendo del modelo de Lorentz, se presenta en el Apéndice 1.

La forma general de la ecuación de Sellmeier puede escribirse como

$$n(\lambda) = 1 + \sum_{i=1}^{m} \frac{A_i \lambda^2}{\lambda^2 - C_i}, \qquad (12)$$

donde n es el índice de refracción, λ la longitud de onda, A_i los coeficientes asociados a la intensidad efectiva de cada oscilador y $C_i = \lambda_i^2$ los parámetros relacionados con las longitudes de onda características de absorción. El parámetro m representa el número de contribuciones consideradas.

Los coeficientes A_i y C_i se determinan experimentalmente mediante técnicas ópticas de alta precisión, tales como la refractometría y la elipsometría espectroscópica [55]. Físicamente, los términos asociados a menores longitudes de onda representan contribuciones de resonancias electrónicas en la región ultravioleta, mientras que los términos de mayor longitud de onda modelan contribuciones vibracionales e iónicas localizadas en el infrarrojo.

El agua constituye uno de los medios más estudiados desde el punto de vista de la dispersión óptica debido a su relevancia en aplicaciones ambientales, biológicas y fotónicas. Entre los modelos disponibles, la formulación propuesta por Daimon y Masumura [56] permite reproducir experimentalmente el índice de refracción del agua destilada en el intervalo comprendido entre $0.18178\mu m$ y $1.12895\mu m$ mediante una ecuación de Sellmeier de cuatro términos:

$$n(\lambda) = 1 + \frac{A_1\lambda^2}{\lambda^2 - C_1} + \frac{A_2\lambda^2}{\lambda^2 - C_2} + \frac{A_3\lambda^2}{\lambda^2 - C_3} + \frac{A_4\lambda^2}{\lambda^2 - C_4}. \qquad (13)$$

Los valores experimentales de los coeficientes utilizados en este trabajo se presentan en la Tabla 2.1.

TABLA 2.1 COEFICIENTES DE SELLMEIER PARA AGUA DESTILADA A 20°C. [56]

Coeficiente B_i	Valor	Coeficiente C_i	valor
A_1	0.5684027565	C_1	0.005101829712
A_2	0.1726177391	C_2	0.01821153936
A_3	0. 02086189578	C_3	0.02620722293
A_4	0.1130748688	C_4	10.69792721

La implementación de estos coeficientes permite estimar el índice de refracción con incertidumbres del orden de 10^{-5}, lo que resulta adecuado para aplicaciones de modelado óptico y diseño de sensores.

La dependencia térmica del índice de refracción ha sido analizada en diversos estudios. En [57] se muestra que las variaciones de temperatura modifican las frecuencias de resonancia efectivas del medio, mientras que en [56, 57] se reporta que cambios en temperatura y densidad producen variaciones medibles en la respuesta óptica del agua. No obstante, para intervalos moderados alrededor de la temperatura ambiente, las variaciones del índice de refracción son del orden de 10^{-3}, Por lo tanto, en este trabajo se adopta

una temperatura de $20°C$ para los cálculos del índice de refracción y se emplean los coeficientes de Sellmeier reportados por Daimon y Mansura [56] a dicha temperatura, sin incorporar variaciones térmicas en el modelo.

En conjunto, la ecuación de Sellmeier proporciona una descripción espectral del índice de refracción del agua que, combinada con el formalismo de refracción molar desarrollado en la sección anterior, permite modelar el comportamiento óptico de soluciones acuosas que contienen arsénico, fluoruro, nitrato y cloruro de sodio. Esta formulación constituye la base para la estimación de las propiedades ópticas empleadas posteriormente en el modelado de la guía de onda tipo *slot* propuesta en este trabajo.

2.2 REFLEXIÓN TOTAL INTERNA

La reflexión total interna constituye uno de los fenómenos fundamentales que permiten el confinamiento de la luz en fibras ópticas, resonadores y guías de onda integradas. Este fenómeno ocurre cuando una onda electromagnética se propaga desde un medio de mayor índice de refracción hacia otro de menor índice y el ángulo de incidencia supera un valor crítico determinado por las propiedades ópticas de ambos medios [58].

Considerando una interfaz plana entre dos medios con índices de refracción n_i y n_t, la relación entre los ángulos de incidencia y transmisión está dada por la ley de Snell

$$n_i \sin\theta_i = n_t \sin\theta_t, \qquad (14)$$

donde θ_i representa el ángulo de incidencia y θ_t el ángulo de transmisión. La Figura 2.3 muestra la reflexión y transmisión de una onda luminosa al incidir sobre la interfaz entre dos medios.

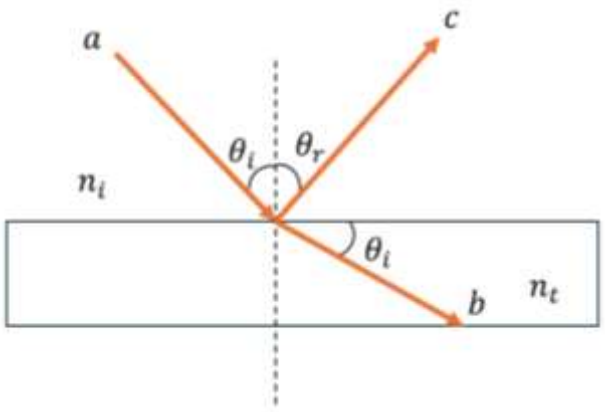


FIGURA 2.3 INCIDENCIA, REFLEXIÓN Y TRANSMISIÓN DE UNA ONDA ELECTROMAGNÉTICA ENTRE DOS MEDIOS CON DISTINTOS ÍNDICES DE REFRACCIÓN [59].

Cuando la luz se propaga desde un medio de mayor índice de refracción hacia otro de menor índice $n_i > n_\mathrm{t}$, el ángulo de transmisión aumenta conforme se incrementa el ángulo de incidencia. Existe un valor particular denominado ángulo crítico θ_c, para el cual el rayo transmitido se propaga paralelamente a la interfaz $\theta_t = 90°$. A partir de la ley de Snell se obtiene

$$n_1 \sin\theta_c = n_2 \ . \qquad (15)$$

La Figura 2.3 ilustra la condición correspondiente al ángulo crítico. Para ángulos de incidencia superiores a este valor, no existe propagación de energía hacia el segundo medio y toda la potencia óptica permanece confinada en el medio de mayor índice de refracción, fenómeno conocido como reflexión total interna, representado en la Figura 2.4.

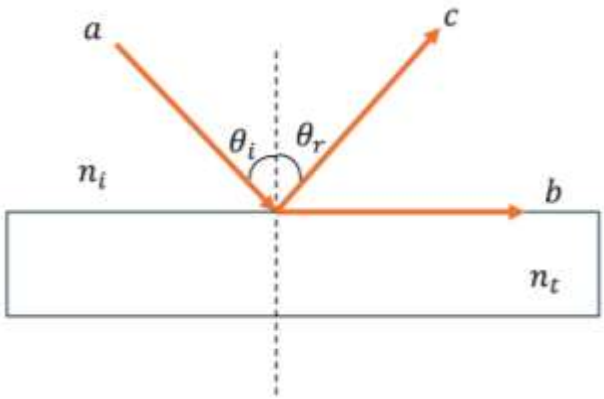


FIGURA 2.4 PROPAGACIÓN DE LA ONDA TRANSMITIDA PARALELA A LA INTERFAZ CUANDO $\boldsymbol{\theta_i}$, $\boldsymbol{\theta_c}$ [59].

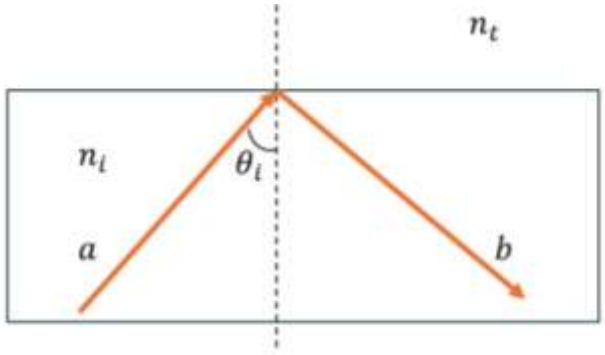


FIGURA 2.5 REFLEXIÓN COMPLETA DE LA ONDA INCIDENTE CUANDO $\boldsymbol{\theta_i > \theta_c}$. [59]

Aunque la óptica geométrica permite describir el fenómeno mediante trayectorias de rayos, una descripción más rigurosa requiere considerar la naturaleza electromagnética de la luz. A partir de las ecuaciones de Maxwell y de las condiciones de continuidad para los campos eléctricos y magnéticos tangenciales en la interfaz, se obtienen los coeficientes de reflexión y transmisión de Fresnel para las polarizaciones transversal eléctrica (TE) y transversal magnética (*TM*) [60].

Para una interfaz dieléctrica no magnética, los coeficientes de reflexión pueden expresarse como

$$r_{TM} = \frac{n_2 \cos\theta_i - n_1 \cos\theta_t}{n_2 \cos\theta_i + n_1 \cos\theta_t} \ , \tag{16}$$

$$r_{TM} = \frac{n_2 \cos\theta_i + i n_1 B}{n_2 \cos\theta_i - i n_1 B} = e^{i\phi_{TM}} \ . \tag{17}$$

Estas expresiones muestran que la amplitud reflejada depende tanto de la polarización como del contraste de índices de refracción entre los medios.

Cuando $\theta_i > \theta_c$, la ley de Snell conduce a valores complejos para $\cos\theta_t$, indicando que el campo transmitido es evanescente y no transporta energía neta en la dirección normal a la interfaz. Para dos medios dieléctricos sin pérdidas, los módulos de los coeficientes de reflexión satisfacen

$$\left[|r_s| = |r_p| = 1, \right]$$

donde r_s y r_p corresponden a las polarizaciones perpendicular y paralela al plano de incidencia, respectivamente. Por tanto, las reflectancias cumplen ($R_s = R_p = 1$), lo que implica que toda la energía incidente se refleja hacia el medio de mayor índice de refracción.

Sin embargo, la ausencia de propagación de energía en el segundo medio no significa que los campos electromagnéticos desaparezcan en la interfaz. El análisis electromagnético muestra que una fracción del campo penetra una corta distancia dentro del medio de menor índice en forma de una onda evanescente cuya amplitud decrece exponencialmente con la distancia a la interfaz. Este fenómeno resulta de particular interés en sensores ópticos basados en guías de onda, ya que permite que el campo electromagnético interactúe con sustancias presentes en el entorno externo sin abandonar el medio guía.

Por esta razón, la reflexión total interna constituye el mecanismo físico que hace posible el confinamiento y la detección óptica en estructuras fotónicas integradas. El comportamiento y las propiedades de la onda evanescente generada bajo estas condiciones se analizan con mayor detalle en la siguiente sección.

2.3 CAMPO EVANESCENTE

Como se describió en la sección anterior, la reflexión total interna ocurre cuando una onda electromagnética se propaga desde un medio de mayor índice de refracción hacia otro de menor índice $n_i > n_t$ y el ángulo de incidencia supera el ángulo crítico $\theta_i > \theta_c$. Bajo estas condiciones, no existe una onda transmitida propagante en el segundo medio y la energía permanece confinada en el medio de mayor índice de refracción [60].

Sin embargo, la aplicación de las condiciones de frontera derivadas de las ecuaciones de Maxwell muestra que los campos electromagnéticos no desaparecen abruptamente en la interfaz. En su lugar, una fracción del campo penetra en el medio de menor índice formando una onda evanescente cuya amplitud decrece exponencialmente con la distancia perpendicular a la interfaz [60-62].

La dependencia espacial de este campo puede expresarse como

$$E(y) = E_0 e^{-\alpha y}, \qquad (18)$$

donde E_0 es la amplitud del campo en la interfaz y α representa la constante de decaimiento. La Figura 2.5 muestra la distribución espacial característica del campo evanescente generado durante la reflexión total interna. Aunque no existe una onda transmitida propagante en el medio de menor índice de refracción, el campo electromagnético penetra una distancia finita dentro de

dicho medio y su amplitud decrece exponencialmente conforme aumenta la distancia respecto a la interfaz.

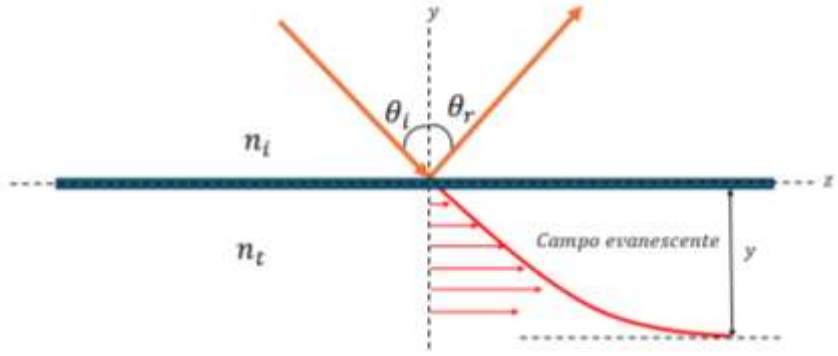


FIGURA 2.6 REPRESENTACIÓN ESQUEMÁTICA DEL DECAIMIENTO EXPONENCIAL DEL CAMPO EVANESCENTE EN EL MEDIO DE MENOR ÍNDICE DE REFRACCIÓN ($\boldsymbol{n_t}$) CUANDO $\boldsymbol{\theta_i > \theta_c}$ [60].

Esta constante se encuentra relacionada con la componente normal del vector de onda en el medio de menor índice mediante

$$k_{yt} = k_0\sqrt{n_t^2 n_i^2 \sin^2\theta_i}\ , \qquad (19)$$

donde $k_0 = 2\pi/\lambda$ es el número de onda en el vacío.

Cuando $\theta_i > \theta_c$, el término bajo la raíz cuadrada es negativo y, por tanto, k_{yt} adquiere un valor imaginario. En consecuencia, la solución transmitida deja de ser una onda propagante y se transforma en una función exponencial decreciente, característica del campo evanescente [60]. Aunque este campo no transporta energía hacia el interior del medio de menor índice, sí mantiene una interacción electromagnética con el entorno cercano a la interfaz. Debido a esta propiedad, cualquier variación en las características ópticas del medio externo puede modificar la distribución del campo y alterar los parámetros de propagación de la onda guiada.

Este fenómeno constituye el principio físico fundamental de numerosos sensores ópticos integrados. En particular, las guías de onda tipo *slot* aprovechan la elevada intensidad del campo evanescente para incrementar la interacción entre la luz y el medio analizado, mejorando la sensibilidad frente a cambios en el índice de refracción producidos por diferentes especies químicas presentes en solución [62, 63].

2.4 GUÍAS DE ONDA

Una guía de onda óptica es una estructura dieléctrica diseñada para confinar y dirigir la propagación de la luz a lo largo de una trayectoria definida. Este confinamiento se logra mediante el fenómeno de reflexión total interna, el cual ocurre cuando la radiación se propaga desde un medio con mayor índice de refracción hacia otro de menor índice y el ángulo de incidencia supera el ángulo crítico descrito en la sección anterior [59].

El principio de funcionamiento puede ilustrarse mediante una capa dieléctrica de índice de refracción $n_2$rodeada por medios con índice de refracción inferior n_1, donde $n_2 > n_1$. Bajo estas condiciones, los rayos que se propagan dentro de la región de mayor índice experimentan reflexiones sucesivas en las interfaces, permaneciendo confinados dentro de la estructura. En consecuencia, la energía óptica puede transportarse a lo largo de grandes distancias con pérdidas reducidas [60-61].

Cuando el ángulo interno de propagación θes mayor que el ángulo crítico,

$$\theta_c = \sin^{-1}\left({}^{1}/_{n}\right), \qquad (20)$$

la reflexión total interna impide que la radiación se propague hacia el medio circundante. No obstante, como se discutió en la sección anterior, una pequeña fracción del campo electromagnético se extiende hacia la región de

menor índice en forma de campo evanescente. Este fenómeno constituye la base física de numerosos sensores ópticos, ya que permite que la luz interactúe con sustancias presentes en el entorno de la guía de onda [61-65].

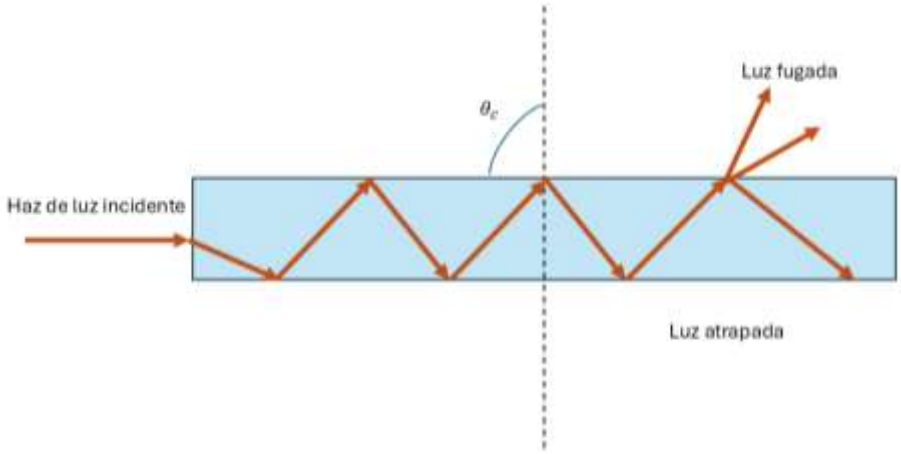


FIGURA 2.7 CONFINAMIENTO DE LA LUZ EN UNA GUÍA DE ONDA MEDIANTE REFLEXIÓN TOTAL INTERNA. LOS RAYOS CON ÁNGULOS DE INCIDENCIA SUPERIORES AL ÁNGULO CRÍTICO PERMANECEN GUIADOS DENTRO DEL NÚCLEO, MIENTRAS QUE AQUELLOS QUE NO SATISFACEN ESTA CONDICIÓN SE ACOPLAN AL MEDIO EXTERNO Y SE PIERDEN POR RADIACIÓN [58].

Las guías de onda ópticas pueden clasificarse de acuerdo con el número de dimensiones en las que se produce el confinamiento de la radiación. Esta clasificación determina las características de propagación de los modos ópticos y las posibles aplicaciones del dispositivo. La Tabla 2.2 resume las configuraciones más comunes.

TABLA 2.2 CLASIFICACIÓN DE LAS GUÍAS DE ONDA EN FUNCIÓN DE LAS DIMENSIONES QUE PERMITEN CONFINAR LA LUZ [61].

Dimensiones del confinamiento de la luz	Clasificación de guía de onda óptica
1D	Guía de onda planar
2D	Canal de guía de onda
	Fibra óptica
3D	Cristales fotónicos

2.4.1 GUÍA DE ONDA PLANAR

La guía de onda planar constituye la estructura más simple utilizada para el confinamiento y propagación de la luz en dispositivos fotónicos integrados. Está formada por una capa dieléctrica de espesor finito y mayor índice de refracción, denominada núcleo o película guía, rodeada por dos medios con índices de refracción inferiores denominados cubierta y sustrato [60,61].

Bajo estas condiciones, la radiación óptica puede permanecer confinada dentro del núcleo mediante reflexiones totales internas sucesivas en ambas interfaces. Sin embargo, como se discutió en la sección anterior, una fracción del campo electromagnético penetra en los medios adyacentes en forma de campo evanescente, fenómeno fundamental para numerosas aplicaciones de sensado óptico [62, 63]. La geometría general de una guía de onda planar se muestra en la Figura 2.7, donde n_p representa el índice de refracción del núcleo, mientras que n_{sup} y n_{inf} corresponden a los índices de refracción de la cubierta y el sustrato, respectivamente.

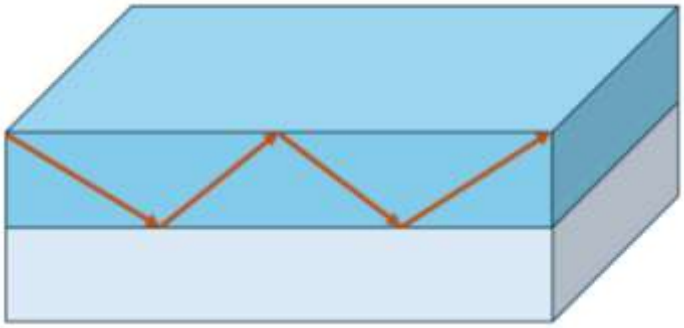


FIGURA 2.8 GEOMETRÍA GENERAL DE UNA GUÍA DE ONDA PLANAR COMPUESTA POR UN NÚCLEO DE ALTO ÍNDICE DE REFRACCIÓN RODEADO POR MEDIOS DE MENOR ÍNDICE QUE PERMITEN EL CONFINAMIENTO ÓPTICO MEDIANTE REFLEXIÓN TOTAL INTERNA [54].

Para garantizar el confinamiento óptico debe cumplirse la condición

$$n_p > n_{sup}, \qquad n_p > n_{inf}, \tag{21}$$

la cual asegura la existencia de reflexión total interna en ambas interfaces del núcleo [60-62].

La guía de onda planar constituye el modelo fundamental para el análisis modal de estructuras fotónicas integradas. A partir de esta configuración es posible desarrollar modelos más complejos, incluyendo guías de onda de canal, resonadores integrados y guías de onda tipo *slot*, las cuales presentan una mayor interacción entre el campo electromagnético y el medio circundante debido al confinamiento reforzado del campo óptico [58-61].

2.5 ÁNALISIS MODAL DE UNA GUÍA DE ONDA *SLOT*

Las guías de onda ópticas permiten el confinamiento y transporte de energía electromagnética mediante el fenómeno de reflexión total interna. Entre las diferentes configuraciones existentes, la guía de onda plana constituye el modelo más simple para el estudio de la propagación modal y proporciona las bases teóricas para el análisis de estructuras más complejas, como las guías de onda tipo *slot*.

Considérese una guía de onda plana formada por una película dieléctrica de índice de refracción n_f, espesor h, una cubierta superior de índice n_c y un sustrato inferior de índice n_s, como se muestra en la Figura 2.8. Para que exista confinamiento óptico es necesario que el índice de refracción del núcleo sea mayor que el de los medios circundantes, es decir,

$$n_f > n_c, \qquad n_f > n_s \tag{22}$$

Bajo estas condiciones, la radiación óptica puede propagarse a lo largo de la estructura mediante modos guiados.

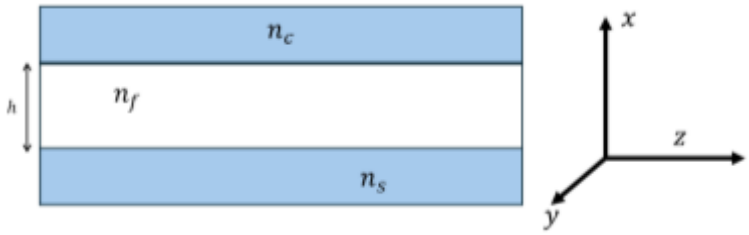


FIGURA 2.9 GEOMETRÍA DE UNA GUÍA DE ONDA PLANA DE TRES CAPAS EMPLEADA PARA EL ANÁLISIS MODAL [58].

La propagación de los campos electromagnéticos en medios dieléctricos homogéneos y libres de cargas puede describirse mediante las ecuaciones de Maxwell. A partir de estas ecuaciones es posible obtener la ecuación de onda para el campo eléctrico y, posteriormente, la ecuación de Helmholtz. La derivación matemática completa se presenta en el Apéndice 2.

Considerando una dependencia armónica temporal para el campo eléctrico,

$$E(x,t) = E(x)e^{-i\omega t}, \tag{23}$$

la ecuación de onda puede expresarse en la forma de Helmholtz

$$\nabla^2 E + k^2 E = 0, \tag{24}$$

donde $k = \frac{2\pi n}{\lambda}$ corresponde al número de onda del medio.

Para una onda que se propaga en la dirección longitudinal z, la solución puede representarse como

$$E(x,z) = E(x)e^{-i\beta z}, \tag{25}$$

donde β es la constante de propagación modal.

Sustituyendo esta expresión en la ecuación de Helmholtz se obtiene una ecuación diferencial ordinaria que describe la distribución transversal del campo eléctrico,

$$\frac{d^2E(x)}{dx^2} + (k^2 - \beta^2)E(x) = 0\,, \tag{26}$$

Esta ecuación constituye la base para el análisis modal de guías de onda dieléctricas.

En una guía de onda óptica la energía electromagnética se propaga en forma de modos guiados. Un modo puede definirse como una distribución espacial de los campos electromagnéticos que conserva su forma transversal durante la propagación y satisface simultáneamente las ecuaciones de Maxwell y las condiciones de frontera impuestas por la estructura.

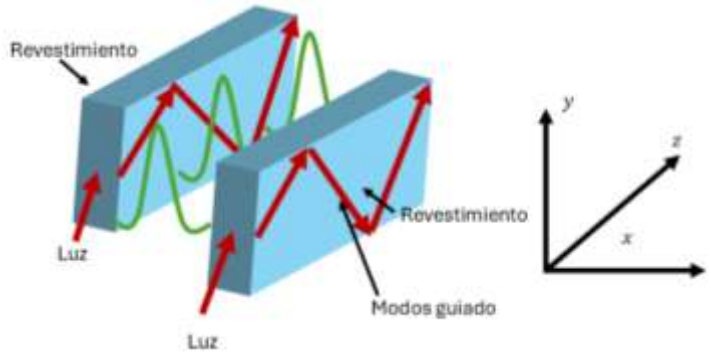


FIGURA 2.10 REPRESENTACIÓN CONCEPTUAL DE LOS MODOS GUIADOS EN UNA GUÍA DE ONDA TIPO *SLOT*.

El número de modos soportados por una guía depende de la longitud de onda de operación, de las dimensiones geométricas de la estructura y del contraste de índices de refracción entre el núcleo y los medios circundantes. En particular, las guías de onda tipo *slot* presentan una elevada concentración del campo electromagnético dentro de la región de bajo índice de refracción, lo que incrementa la interacción entre la luz y el medio analizado.

La ecuación (26) describe la distribución transversal del campo eléctrico asociada a un modo guiado en una estructura dieléctrica. Su solución depende directamente del perfil espacial del índice de refracción y de las condiciones de frontera impuestas por la geometría de la guía de onda.

En el caso de estructuras multicapa, como las guías de onda tipo *slot* empleadas en esta investigación, la solución de la ecuación diferencial debe realizarse de forma independiente en cada región del dispositivo y

posteriormente acoplarse mediante las condiciones de continuidad de los campos electromagnéticos en las interfaces. Este procedimiento conduce a una ecuación trascendental cuyas raíces permiten determinar los modos guiados y sus respectivas constantes de propagación. El desarrollo de este análisis se presenta en la siguiente sección.

2.6 ESTRUCTURA DE LA GUÍA DE ONDA *SLOT* Y CONDICIONES DE FRONTERA

Una vez obtenida la ecuación diferencial que gobierna la distribución transversal del campo eléctrico, es posible extender el análisis a estructuras de mayor complejidad. En particular, las guías de onda tipo *slot* se caracterizan por la presencia de una región central de bajo índice de refracción confinada entre dos regiones de alto índice, lo que favorece una elevada concentración del campo electromagnético dentro de la ranura, incrementando la interacción entre la radiación óptica y el medio analizado [68,69].

Para una guía de onda *slot* simétrica, el perfil del índice de refracción puede dividirse en cinco regiones, como se muestra en la Figura 2.10.

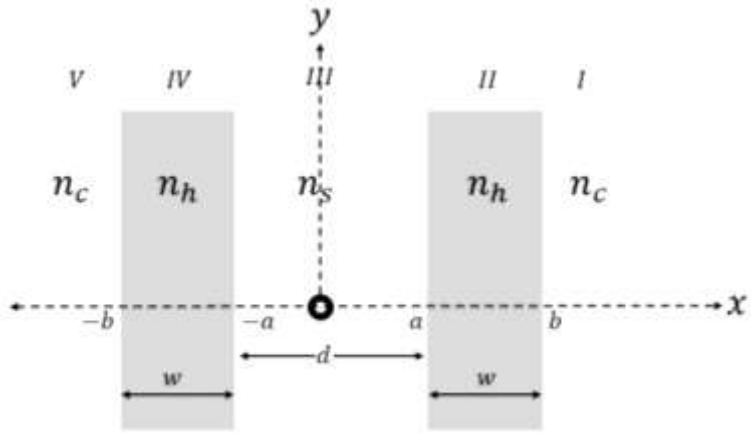


FIGURA 2.11 PERFIL DE ÍNDICES DE REFRACCIÓN Y REGIONES EMPLEADAS PARA EL ANÁLISIS MODAL DE UNA GUÍA DE ONDA *SLOT* SIMÉTRICA [68].

La distribución transversal del campo eléctrico debe satisfacer la ecuación de Helmholtz en cada una de las regiones de la estructura. Dependiendo del valor local del índice de refracción, las soluciones presentan comportamiento oscilatorio o evanescente.

En las regiones de alto índice de refracción del núcleo, donde el índice local es mayor que el índice efectivo del modo guiado, el campo presenta un comportamiento oscilatorio en la dirección transversal. En cambio, en la región central de bajo índice de refracción, slot, y en los revestimientos, donde el índice local es menor que el índice efectivo, el campo presenta un comportamiento evanescente transversal. En estas regiones, la solución de la ecuación de Helmholtz puede expresarse mediante funciones exponenciales o hiperbólicas, dependiendo de la región y de la forma de representar el campo.

Las soluciones para el campo eléctrico transversal $E_y(x)$pueden expresarse como:

- Región I $(x > b)$: Corresponde al revestimiento exterior del lado positivo del eje x. El campo presenta un comportamiento evanescente y decae exponencialmente al alejarse de la guía.

$$E_{y,I}(x) = A_1 e^{-\gamma_c (x-b)}. \qquad (27)$$

- Región II $(a < x < b)$: Corresponde a la región de alto índice de refracción del núcleo situada en el lado positivo del eje x. En ella, el campo presenta un comportamiento oscilatorio.

$$E_{y,II}(x) = A_2 \cos\big(k_h (x-a)\big) + A_3 \sin\big(k_h (x-a)\big). \qquad (28)$$

- Región III $(-a < x < a)$: Corresponde a la región central de bajo índice de refracción denominada slot. Bajo las condiciones de guiado

consideradas, el campo presenta un comportamiento evanescente transversal y puede expresarse mediante funciones hiperbólicas.

$$E_{y,II}(x) = A_4 \cosh(\gamma_s x) + A_5 \sinh(\gamma_s x) \ . \qquad (29)$$

- Región IV $(-b < x < -a)$: Corresponde a la región de alto índice de refracción del núcleo situada en el lado negativo del eje x. En ella, el campo presenta un comportamiento oscilatorio.

$$E_{y,IV}(x) = A_6 \cos\big(k_h(x+a)\big) + A_7 \sin\big(k_h(x+a)\big). \qquad (30)$$

- Región $V \ \ (x < -b)$: Corresponde al revestimiento exterior del lado negativo del eje x. El campo presenta un comportamiento evanescente y decae exponencialmente al alejarse de la guía hacia menos infinito.

$$E_{y,v}(x) = A_8 e^{-\gamma_s(x+b)}. \qquad (31)$$

donde k_h representa la constante de propagación transversal en las regiones de alto índice de refracción, mientras que γ_c y γ_s corresponden a las constantes de decaimiento evanescente en los revestimientos superior e inferior, respectivamente.

Las constantes transversales asociadas al núcleo (k_h), al revestimiento (γ_c) y al sustrato (γ_s) se definen a partir de la constante de propagación (β) y de los índices de refracción de cada región de la forma:

$$\gamma_c^2 = \beta^2 - \left(\frac{\omega}{c} n_c\right)^2, \qquad (32)$$

$$\gamma_s^2 = \beta^2 - \left(\frac{\omega}{c} n_s\right)^2, \qquad (33)$$

$$k_h^2 = \left(\frac{\omega}{c} n_h\right)^2 - \beta^2, \qquad (34)$$

donde $k_0 = \omega/c$ es el número de onda en el vacío, con ω la frecuencia angular de la onda y c la velocidad de la luz en el vacío. Los modos TM guiados satisfacen $\beta > \omega n_s/c\,, \omega n_c/c$, lo que garantiza el decaimiento evanescente fuera de la región guiante. La formulación modal empleada se basa en el análisis de modos transversal magnético (TM) en estructuras *slot* dieléctricas simétricas, ampliamente estudiadas debido a su capacidad para concentrar energía electromagnética en regiones de bajo índice de refracción [70].

Al imponer, en las interfaces ubicadas en $x = \pm a$ y $x = \pm b$, la continuidad de las componentes tangenciales de los campos eléctrico y magnético y de la componente normal del desplazamiento eléctrico, en ausencia de cargas y corrientes superficiales libres, se relacionan los coeficientes de las soluciones en las cinco regiones de la guía. Al considerar la simetría del modo fundamental y el decaimiento del campo en los revestimientos exteriores, se obtiene la siguiente expresión para la distribución modal transversal:

$$\phi(x) = A \begin{cases} \dfrac{1}{n_c^2}\left[cosh(\gamma_s a)\, cos(k_h W) + \dfrac{\gamma_s}{k_h}\left(\dfrac{n_h}{n_s}\right)^2 sinh(\gamma_s a)\, sin(k_h W)\right] e^{-\gamma_c (x-b)},\ x > b \\ \dfrac{1}{n_h^2} cosh(\gamma_s a)\, cos[k_h (x-a)] + \dfrac{\gamma_s}{n_s^2 k_h} sinh(\gamma_s a)\, sin[k_h (x-a)]\,,\ a < x < b \\ \dfrac{1}{n_s^2} cosh(\gamma_s x)\,,\ -a < x < a \\ \dfrac{1}{n_h^2}\left[cosh(\gamma_s a)\, cos\big(k_h (x+a)\big) + \dfrac{\gamma_s}{k_h}\left(\dfrac{n_h}{n_s}\right)^2 sinh(\gamma_s a)\, sin\big(k_h (x+a)\big)\right],\ -b < x < -a \\ \dfrac{1}{n_c^2}\left[cosh(\gamma_s a)\, cos(k_h W) + \dfrac{\gamma_s}{k_h}\left(\dfrac{n_h}{n_s}\right)^2 sinh(\gamma_s a)\, sin(k_h W)\right] e^{\gamma_c (x+b)},\ x < -b \end{cases} \tag{35}$$

donde A es una constante de normalización, n_s corresponde al índice de refracción de la región *slot*, n_h al índice de las regiones de alto índice de refracción, n_c al índice del revestimiento, $W = b - a$ representa el espesor de cada región de alto índice. La solución anterior corresponde al modo *TM* fundamental simétrico [68].

Para determinar los coeficientes desconocidos y los valores permitidos de la constante de propagación β, es necesario imponer las condiciones de continuidad de los campos electromagnéticos en cada una de las interfaces. Para modos TM, estas condiciones exigen la continuidad del campo eléctrico tangencial y de la componente normal del desplazamiento eléctrico.

La aplicación sistemática de estas condiciones genera un sistema homogéneo de ecuaciones lineales cuya solución no trivial únicamente existe cuando el determinante del sistema es igual a cero. Esta condición conduce a la ecuación trascendental que gobierna la propagación modal de la guía de onda *slot*.

Para el modo TM fundamental, dicha ecuación trascendental adopta la forma:

$$\arctan\left[\frac{{n_h}^2\gamma_c}{{n_c}^2 k_h}\right] + \arctan\left[\frac{{n_h}^2\gamma_s}{{n_s}^2 k_h}\tanh(\gamma_s a)\right] + m\pi = k_h(b-a), \qquad (36)$$

donde m denota el orden del modo y a es la mitad del ancho de la ranura, de modo que su separación es $d = 2a$. Las constantes transversales γ_c, γ_s y k_h están dadas por las ecuaciones (32)–(34).

Al sustituir las ecuaciones (32)–(34) en la ecuación (36) y considerar únicamente el modo fundamental ($m = 0$), la ecuación trascendental puede escribirse de forma:

$$\arctan\left[\frac{{n_h}^2\sqrt{\beta^2-\left(\frac{\omega}{c}n_c\right)^2}}{{n_c}^2\sqrt{\left(\frac{\omega}{c}n_h\right)^2-\beta^2}}\right] + \arctan\left[\frac{{n_h}^2\sqrt{\beta^2-\left(\frac{\omega}{c}n_s\right)^2}}{{n_s}^2\sqrt{\left(\frac{\omega}{c}n_h\right)^2-\beta^2}}\tanh\left(\frac{d}{2}\sqrt{\beta^2-\left(\frac{\omega}{c}n_s\right)^2}\right)\right] = W\sqrt{\left(\frac{\omega}{c}n_h\right)^2-\beta^2}\,. \qquad (37)$$

Entre estos parámetros destaca el factor de confinamiento óptico dentro de la región *slot*, definido como la fracción de potencia óptica confinada en la zona sensible de la guía,

$$\eta = \frac{P_{slot}}{P_{total}} = \frac{n_s^2\int_{-a}^{a}\left|E_{s,x}\right|^2 dx}{\sum_j n_j^2\int_j\left|E_{j,x}\right|^2 dx}, \qquad (38)$$

donde P_{slot} representa la potencia óptica contenida en la región de la ranura y P_{total} la potencia total transportada por el modo guiado. A partir de la constante de propagación obtenida para cada longitud de onda a través de la frecuencia angular en Eq. (37), también es posible determinar la velocidad de grupo v_g, la cual describe el transporte de energía a través de la guía de onda [70]. Esta magnitud se obtiene a partir de la relación de dispersión como:

$$v_g = \left(\frac{d\beta}{d\omega}\right)^{-1} \quad , \tag{39}$$

donde la derivada se evalúa numéricamente utilizando los datos dependientes de la longitud de onda. Suponiendo propagación estacionaria del modo guiado, la potencia óptica confinada dentro de la región *slot* puede expresarse como:

$$P_{slot} = P_{in}\eta, \tag{40}$$

donde P_{in} representa la potencia óptica de entrada asociada al modo guiado. A partir de esta cantidad, la densidad de energía electromagnética dentro de la región sensible se estima mediante [70]:

$$u_{slot} = \frac{P_{in} \cdot \eta}{v_g}. \tag{41}$$

La respuesta óptica obtenida se transformó posteriormente en una magnitud eléctrica efectiva a partir de la densidad de energía electromagnética. Considerando las dimensiones transversales de la región de sensado, el campo eléctrico efectivo dentro de la ranura E_{eff} se estimó como [69]:

$$E_{eff} = \sqrt{\frac{2u_{slot}}{\varepsilon_0 n_s^2}}. \tag{42}$$

Una vez determinado el campo eléctrico efectivo, la potencia eléctrica disipada dentro del medio iónico se estimó mediante:

$$P_{elec} = \sigma \left| E_{eff} \right|^2 V_{vol}, \qquad (43)$$

donde V_{vol} corresponde al volumen efectivo de interacción de la región de sensado. La conductividad eléctrica de la solución analizada se define como:

$$\sigma = \Lambda \cdot C_{mol/L}, \qquad (44)$$

donde Λ representa la conductividad molar de la especie iónica considerada [65]. Esta magnitud proporciona un descriptor adicional asociado a la respuesta eléctrica de la guía de onda cargada con el analito.

Se evaluó a partir de la variación del índice modal efectivo respecto a la concentración del analito [60]:

$$S_\lambda = \frac{\partial n_{eff}}{\partial C} \quad , \qquad (45)$$

Adicionalmente, se calculó el factor de interacción electromagnética ξ [61], [70]:

$$\xi = \frac{\int_{slot} |E|^2 dA}{\int_{total} |E|^2 dA}. \qquad (46)$$

Los parámetros físicos obtenidos, incluyendo β, η, v_g, E_{eff}, P_{elec}, S_λ y ξ, conforman una firma electromagnética multidimensional característica de cada especie química analizada. Este enfoque permite relacionar directamente la respuesta modal de la guía de onda con las propiedades fisicoquímicas del medio circundante, proporcionando un conjunto de características relevantes para el entrenamiento de los modelos de aprendizaje automático desarrollados en esta investigación. La utilización de estructuras *slot* resulta particularmente atractiva debido a su elevada sensibilidad frente a pequeñas perturbaciones

del índice de refracción y a su potencial para aplicaciones de monitoreo ambiental y sensado químico integrado [72, 73].

2.7 FUNDAMENTOS DE APRENDIZAJE AUTOMÁTICO

En la sección anterior se establecieron las bases físicas y matemáticas necesarias para modelar la interacción entre la radiación electromagnética y una guía de onda tipo *slot* empleada como elemento sensor. A partir de las ecuaciones de Maxwell, la teoría de propagación modal y la solución de la ecuación trascendental del sistema, es posible obtener parámetros ópticos tales como la constante de propagación β, el índice efectivo n_{eff}, el factor de confinamiento óptico η, la distribución espacial del campo electromagnético y otras magnitudes relacionadas con la sensibilidad del dispositivo frente a variaciones del índice de refracción del medio analizado.

En aplicaciones reales de sensado, la relación entre estos parámetros ópticos y la concentración de los analitos de interés suele ser altamente no lineal y depender simultáneamente de múltiples variables físicas. Como consecuencia, la identificación directa de contaminantes a partir de las respuestas ópticas del sensor puede convertirse en un problema complejo de clasificación y reconocimiento de patrones. En este contexto, las técnicas de aprendizaje automático han demostrado ser herramientas eficaces para extraer información relevante de grandes conjuntos de datos y construir modelos predictivos capaces de generalizar su comportamiento ante nuevas observaciones [74,75].

El aprendizaje automático constituye una rama de la inteligencia artificial orientada al desarrollo de algoritmos capaces de aprender relaciones subyacentes a partir de datos, sin necesidad de definir explícitamente todas las reglas que gobiernan el sistema [74,76]. Desde una perspectiva formal, el

objetivo consiste en aproximar una función que relacione un conjunto de variables de entrada $\mathbf{x}$ con una salida deseada y, minimizando una medida de error y maximizando simultáneamente la capacidad de generalización del modelo frente a datos no observados durante el entrenamiento [74].

En el contexto de esta investigación, las variables de entrada corresponden a parámetros físicos obtenidos mediante el modelado de la guía de onda *slot* para diferentes concentraciones de arsénico, fluoruro, nitrato y cloruro de sodio en agua. Dichos parámetros conforman un conjunto de características que contienen información sobre la respuesta electromagnética del dispositivo y que pueden emplearse para entrenar algoritmos capaces de identificar automáticamente el contaminante presente en la muestra. La calidad de los modelos de aprendizaje automático depende no solo de la arquitectura seleccionada, sino también de la representatividad de los datos utilizados durante el entrenamiento. Por esta razón, los datos sintéticos generados mediante el modelo físico desarrollado en los capítulos previos constituyen la base para la construcción de modelos predictivos que permitan establecer relaciones entre las propiedades ópticas de la guía y la composición química del medio analizado.

Diversos estudios han demostrado la viabilidad de integrar técnicas de aprendizaje automático con sensores fotónicos y sistemas de monitoreo de la calidad del agua, permitiendo abordar tareas de clasificación, detección de contaminantes y análisis multivariable [22-24]. Asimismo, los modelos basados en árboles de decisión, métodos de ensamble y redes neuronales artificiales se han utilizado para resolver problemas de clasificación y regresión en los que las variables de entrada presentan relaciones complejas [75, 76]

En este contexto, Zhou et al. [77] desarrollaron una plataforma nanofotónica que combina espectroscopia Raman mejorada por superficie con algoritmos de aprendizaje automático para detectar y cuantificar contaminantes orgánicos en agua. Por otra parte, Singh et al. [78] utilizaron simulaciones por elementos

finitos y modelos de aprendizaje automático para predecir propiedades ópticas de guías de onda tipo slot. Ambos trabajos ilustran aplicaciones complementarias del aprendizaje automático en el análisis de sistemas fotónicos, aunque difieren en sus objetivos y en la naturaleza de los datos empleados. Más directamente relacionado con la clasificación de líquidos, Das et al. [79] estudiaron una guía de onda dieléctrica tipo slot en el régimen de terahercios y emplearon algoritmos de clasificación a partir de mediciones de transmisión. Asimismo, Vishwaraj et al. [80] exploraron el uso de modelos de aprendizaje automático para la clasificación de modos y la predicción del índice efectivo en estructuras fotónicas basadas en guías de onda con rejillas de Bragg. Liu et al. [81], por su parte, propusieron una arquitectura basada en fibra óptica microestructurada y resonancia de plasmones superficiales, incorporando una red neuronal para estimar pérdidas de confinamiento y explorar aplicaciones relacionadas con la detección de contaminantes en agua.

En conjunto, estos antecedentes muestran distintas formas de aplicar el aprendizaje automático al análisis de señales ópticas, la predicción de propiedades de estructuras fotónicas y la clasificación de líquidos. En este contexto, la presente investigación aborda un enfoque complementario, basado en la integración del modelado numérico de una guía de onda tipo slot con técnicas de aprendizaje automático para la clasificación de cuatro categorías de interés para la calidad del agua: fluoruro, ácido arsénico, nitrato y cloruro de sodio. Por lo tanto, la contribución metodológica de este trabajo consiste en establecer un procedimiento que vincula el cálculo de la respuesta óptica de la guía de onda con la generación de descriptores optoelectrónicos y la evaluación comparativa de distintos algoritmos de clasificación. Además de analizar su desempeño en condiciones sin perturbaciones, se estudia su comportamiento ante perturbaciones gaussianas en las variables de entrada, lo que permite comparar la estabilidad de los modelos frente a variaciones en los descriptores. De esta manera, el trabajo no se limita a evaluar la capacidad

de clasificación de los datos generados, sino que también examina la robustez de los algoritmos dentro del escenario computacional considerado. Esta integración proporciona una base metodológica para el desarrollo posterior de sistemas fotónicos inteligentes y para futuras investigaciones orientadas a la validación experimental del enfoque propuesto.

Por ello, en las siguientes secciones se presentan los fundamentos teóricos de los algoritmos empleados en esta investigación. En particular, se describen los principales paradigmas del aprendizaje automático, así como los modelos de árboles de decisión, bosques aleatorios, XGBoost y redes neuronales artificiales, los cuales serán utilizados posteriormente para evaluar la capacidad de clasificación de contaminantes a partir de los parámetros ópticos generados mediante el modelo de guía de onda *slot*.

2.7.1 CLASIFICACIÓN DE MODELOS DE APRENDIZAJE AUTOMÁTICO

Los algoritmos de aprendizaje automático pueden clasificarse en diferentes categorías de acuerdo con la forma en que utilizan la información disponible durante el proceso de entrenamiento. Aunque existen diversas taxonomías en la literatura, la clasificación más ampliamente aceptada distingue tres paradigmas principales: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo [74,75].

La selección de un paradigma depende de la naturaleza del problema, del tipo de información disponible y de los objetivos específicos de la aplicación. En el contexto de esta investigación, el problema consiste en identificar contaminantes presentes en muestras de agua a partir de parámetros ópticos obtenidos mediante el modelado de una guía de onda tipo *slot*. Debido a que las muestras empleadas para el entrenamiento cuentan con etiquetas

asociadas al contaminante correspondiente, el problema se aborda mediante técnicas de aprendizaje supervisado.

Aprendizaje supervisado

El aprendizaje supervisado se basa en un conjunto de entrenamiento compuesto por pares ordenados $\mathbf{x}$, y, donde $\mathbf{x}$ representa el conjunto de variables de entrada y y la salida deseada o etiqueta asociada [74], [75].

El objetivo consiste en construir una función

$$f: X \rightarrow Y \ , \qquad (47)$$

y capaz de aproximar la relación existente entre las variables de entrada y la salida observada. Durante el entrenamiento, el algoritmo ajusta sus parámetros internos para minimizar una función de error que cuantifica la diferencia entre las predicciones generadas y los valores reales.

Los problemas supervisados suelen dividirse en dos categorías principales:

- Clasificación, cuando la variable de salida corresponde a un conjunto finito de clases.
- Regresión, cuando la variable de salida es continua.

Debido a que en esta investigación se busca identificar la presencia de arsénico, fluoruro, nitrato o cloruro de sodio, el problema se formula como una tarea de clasificación multiclase.

Aprendizaje no supervisado

En el aprendizaje no supervisado únicamente se dispone de observaciones de entrada $\mathbf{x}$, sin etiquetas asociadas que indiquen una respuesta correcta [76, 84].

El objetivo consiste en identificar estructuras ocultas, relaciones o patrones presentes en los datos. Entre las tareas más comunes se encuentran:

- Agrupamiento (*clustering*), cuyo propósito es organizar las muestras en grupos con características similares.
- Reducción de dimensionalidad, utilizada para representar la información en espacios de menor dimensión conservando la mayor cantidad posible de variabilidad.

Estos métodos son ampliamente utilizados en análisis exploratorio de datos, compresión de información y visualización de conjuntos de datos de alta dimensión.

Aprendizaje por refuerzo

El aprendizaje por refuerzo difiere de los enfoques anteriores porque el proceso de aprendizaje se basa en la interacción de un agente con un entorno dinámico [82]. En este paradigma, el agente observa el estado del sistema s_t, ejecuta una acción a_t y recibe una recompensa r_t como consecuencia de dicha acción. El objetivo consiste en aprender una política de decisión que maximice la recompensa acumulada a largo plazo.

Formalmente, este marco suele describirse mediante Procesos de Decisión de Markov (MDP), donde el agente debe equilibrar continuamente la exploración de nuevas acciones y la explotación del conocimiento previamente adquirido [82]. Aunque el aprendizaje por refuerzo ha demostrado resultados sobresalientes en problemas de control, robótica y optimización secuencial, no forma parte de los métodos considerados en esta investigación.

En función de la disponibilidad de datos etiquetados y de los objetivos del sistema propuesto, los algoritmos empleados en este trabajo pertenecen al paradigma de aprendizaje supervisado. En las siguientes secciones se describen los fundamentos teóricos de los árboles de decisión, los bosques

aleatorios, XGBoost y las redes neuronales artificiales, utilizados para la clasificación de contaminantes a partir de los parámetros ópticos generados mediante el modelo físico de la guía de onda *slot*.

2.7.2 ARBOLES DE DECISIÓN

Los árboles de decisión constituyen una de las técnicas de aprendizaje supervisado más utilizadas en problemas de clasificación y regresión debido a su simplicidad conceptual, interpretabilidad y capacidad para modelar relaciones no lineales entre variables de entrada y salida [83, 86].

Un árbol de decisión puede representarse como una estructura jerárquica compuesta por nodos y ramas. Cada nodo interno corresponde a una condición aplicada sobre una variable de entrada, mientras que las ramas representan los posibles resultados de dicha condición. La estructura general de un árbol de decisión se muestra en la Figura 2.11.

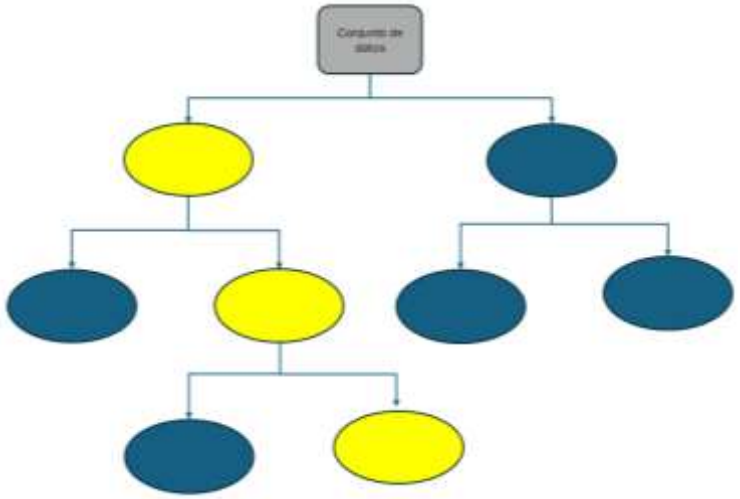


FIGURA 2.12 REPRESENTACIÓN ESQUEMÁTICA DE LA ESTRUCTURA JERÁRQUICA DE UN ÁRBOL DE DECISIÓN. EL NODO RAÍZ CONTIENE EL CONJUNTO INICIAL DE DATOS, MIENTRAS QUE LOS NODOS INTERNOS REPRESENTAN PARTICIONES SUCESIVAS DEL ESPACIO DE CARACTERÍSTICAS Y LAS HOJAS CORRESPONDEN A LAS CLASES O PREDICCIONES FINALES GENERADAS POR EL MODELO. [83]

El proceso de clasificación comienza en el nodo raíz, que contiene el conjunto completo de entrenamiento. Posteriormente, el algoritmo realiza divisiones sucesivas sobre las variables de entrada con el propósito de generar subconjuntos cada vez más homogéneos respecto a la variable objetivo. Este procedimiento continúa hasta alcanzar nodos terminales o hojas, donde se asigna la clase final correspondiente. El proceso de clasificación consiste en recorrer el árbol desde el nodo raíz hasta una hoja, evaluando secuencialmente las condiciones establecidas en cada nivel de la estructura [74].

El objetivo del algoritmo es particionar iterativamente el conjunto de datos de entrenamiento en subconjuntos cada vez más homogéneos respecto a la variable objetivo. Para ello, en cada nodo se selecciona la variable y el umbral que proporcionan la mejor separación entre las clases presentes en los datos [74,75].

Uno de los criterios más utilizados para evaluar la calidad de una partición es el índice de Gini, definido como

$$G = 1 - \sum_{i=1}^{K} p_i^2 \ , \qquad (48)$$

donde p_i representa la proporción de muestras pertenecientes a la clase i y K es el número total de clases. Un valor de $G = 0$ indica que todas las muestras del nodo pertenecen a una única categoría, mientras que valores mayores reflejan una mayor mezcla entre clases [83].

Otro criterio ampliamente utilizado es la entropía de Shannon,

$$H = -\sum_{i=1}^{K} p_i \log_2(p_i) \ , \qquad (49)$$

la cual cuantifica el grado de incertidumbre asociado a la distribución de clases dentro de un nodo. Durante el entrenamiento, el algoritmo busca maximizar la ganancia de información, es decir, la reducción de entropía obtenida después

de realizar una partición [74,76]. La principal ventaja de los árboles de decisión radica en su capacidad para generar modelos fácilmente interpretables, permitiendo identificar de manera explícita cuáles variables influyen con mayor relevancia en el proceso de clasificación. Asimismo, pueden manejar relaciones no lineales y no requieren supuestos estrictos sobre la distribución estadística de los datos [75]. Sin embargo, estos modelos presentan una elevada sensibilidad a pequeñas variaciones en los datos de entrenamiento, lo que puede conducir a fenómenos de sobreajuste (*overfitting*) cuando el árbol crece excesivamente [75]. Esta limitación motivó el desarrollo de métodos de ensamble más robustos, como los bosques aleatorios y los algoritmos de *boosting*, los cuales serán descritos en las siguientes secciones.

2.7.3 BOSQUES ALEATORIOS

Los bosques aleatorios (*Random Forests*) constituyen una técnica de aprendizaje supervisado basada en métodos de ensamble (*Ensemble Learning*), propuesta por Breiman como una extensión de los árboles de decisión convencionales con el objetivo de mejorar la capacidad de generalización y reducir el riesgo de sobreajuste [86]. El principio fundamental consiste en construir múltiples árboles de decisión a partir de subconjuntos aleatorios de los datos de entrenamiento y combinar posteriormente sus predicciones para obtener una respuesta final.

A diferencia de un árbol de decisión individual, cuya estructura puede verse fuertemente influenciada por pequeñas variaciones en los datos de entrenamiento, la idea fundamental consiste en construir un conjunto de árboles de decisión independientes durante la fase de entrenamiento. Cada árbol se genera utilizando una muestra aleatoria de los datos obtenida mediante remuestreo con reemplazo (*Bootstrap Sampling*), así como un subconjunto aleatorio de variables predictoras seleccionado en cada nodo de

partición. Esta estrategia introduce diversidad entre los árboles y disminuye la correlación existente entre ellos, permitiendo obtener modelos más estables y con menor varianza que un árbol individual [86]. La estructura general de un bosque aleatorio se muestra en la Figura 2.12.

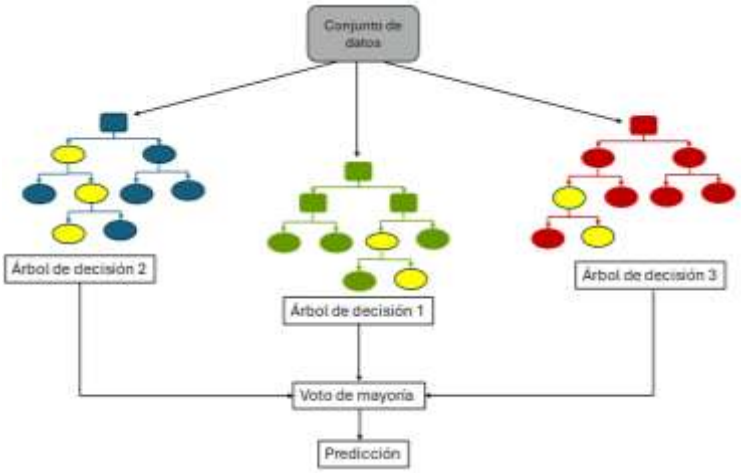


FIGURA 2.13 ESQUEMA GENERAL DEL ALGORITMO DE BOSQUES ALEATORIOS. A PARTIR DEL CONJUNTO DE DATOS ORIGINAL SE CONSTRUYEN MÚLTIPLES ÁRBOLES DE DECISIÓN UTILIZANDO MUESTRAS Y SUBCONJUNTOS DE CARACTERÍSTICAS SELECCIONADOS ALEATORIAMENTE. LA PREDICCIÓN FINAL SE OBTIENE MEDIANTE UN PROCESO DE VOTACIÓN MAYORITARIA ENTRE LOS ÁRBOLES INDIVIDUALES [86].

A partir del conjunto de datos original se generan múltiples árboles de decisión entrenados de manera independiente. Posteriormente, las predicciones individuales son combinadas mediante un mecanismo de votación mayoritaria para problemas de clasificación o mediante el promedio de las salidas para problemas de regresión. La decisión final corresponde a la respuesta agregada del conjunto de modelos.

Matemáticamente, si un bosque está compuesto por Nárboles de decisión y cada árbol produce una predicción $h_i(x)$para una muestra x, la salida final del clasificador puede expresarse como

$$H(x) = \text{mode}\{h_1(x), h_2(x), \dots, h_N(x)\}, \qquad (50)$$

donde $\mathrm{mode}(\cdot)$ representa la clase más frecuentemente predicha por el conjunto de árboles [84].

Una de las principales ventajas de los bosques aleatorios es su capacidad para manejar conjuntos de datos con un gran número de variables sin requerir procesos complejos de selección de características. Además, son relativamente resistentes al ruido experimental y permiten estimar la importancia relativa de cada variable dentro del proceso de clasificación [75], [80]. Estas propiedades han favorecido su aplicación en diversas áreas científicas, incluyendo bioinformática, sistemas de detección, monitoreo ambiental y análisis de calidad del agua [17].

Sin embargo, debido a que el modelo final está compuesto por numerosos árboles, la interpretabilidad resulta menor en comparación con un árbol de decisión individual. Aunque es posible analizar la importancia de las variables, la relación exacta entre las entradas y la decisión final suele ser más difícil de visualizar [75].

2.7.4 XGBOOST

El algoritmo Extreme Gradient Boosting (XGBoost) es un método de aprendizaje supervisado basado en la técnica de *gradient boosting*, desarrollado con el objetivo de mejorar la precisión predictiva y la eficiencia computacional de los modelos basados en árboles de decisión [69]. A diferencia de los bosques aleatorios, donde los árboles se construyen de forma independiente y sus predicciones se combinan posteriormente, XGBoost genera los árboles de manera secuencial, de modo que cada nuevo árbol busca corregir los errores cometidos por los modelos previamente construidos [83, 87].

La idea fundamental consiste en aproximar la función objetivo mediante la suma de múltiples árboles de decisión débiles (*weak learners*), donde cada árbol adicional aporta una corrección incremental al modelo existente. De esta manera, la predicción final puede expresarse como

$$\hat{y} = \sum_{k=1}^{n} f_k\,(x), \qquad (51)$$

donde $\hat{y}$ representa la salida estimada por el modelo, $f_k(x)$ corresponde a la contribución del k-ésimo árbol de decisión y $n$es el número total de árboles empleados durante el proceso de entrenamiento [87].

La estructura general del algoritmo se presenta en la Figura 2.13

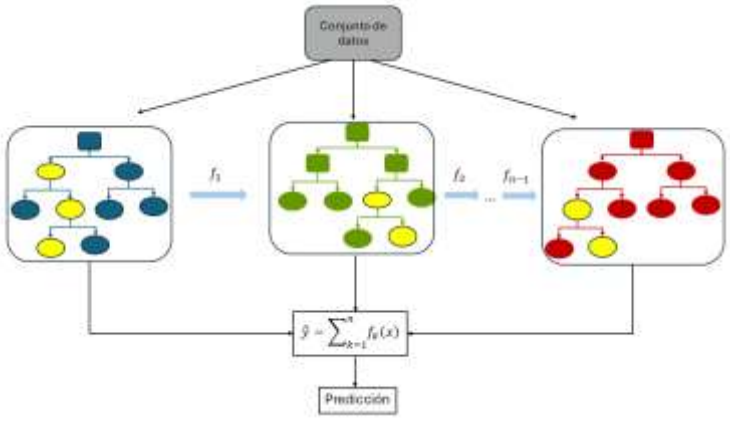


FIGURA 2.14 ESQUEMA CONCEPTUAL DEL ALGORITMO XGBOOST. LA PREDICCIÓN FINAL SE OBTIENE MEDIANTE LA SUMA DE LAS CONTRIBUCIONES INDIVIDUALES DE MÚLTIPLES ÁRBOLES DE DECISIÓN ENTRENADOS SECUENCIALMENTE PARA CORREGIR LOS ERRORES DE LOS MODELOS PREVIOS [87].

A partir del conjunto de datos de entrenamiento se construye un primer árbol que genera una estimación inicial. Posteriormente, cada nuevo árbol se entrena utilizando la información residual generada por los árboles anteriores, permitiendo corregir progresivamente los errores de predicción hasta obtener una aproximación más precisa de la función objetivo.

Durante el entrenamiento, XGBoost minimiza una función objetivo compuesta por dos términos: una función de pérdida que cuantifica el error de predicción y un término de regularización que controla la complejidad del modelo [88]. Esta formulación puede expresarse de manera general como

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{n} \Omega(f_k), \qquad (52)$$

donde $l(y_i, \hat{y}_i)$ representa la función de pérdida asociada a cada observación y $\Omega(f_k)$ es el término de regularización aplicado a cada árbol del modelo. La inclusión de este segundo término permite reducir el riesgo de sobreajuste y mejorar la capacidad de generalización sobre datos no observados [75, 76].

Una de las principales ventajas de XGBoost es su capacidad para modelar relaciones altamente no lineales entre variables de entrada y salida, manteniendo al mismo tiempo una elevada eficiencia computacional. Asimismo, incorpora mecanismos de regularización, manejo automático de valores faltantes y paralelización de operaciones, características que han contribuido a su amplio uso en problemas de clasificación y regresión de alta complejidad [74,87]. Su capacidad para capturar interacciones complejas entre múltiples variables lo convierte en una herramienta adecuada para el análisis de sistemas multivariables con comportamientos no lineales.

2.7.5 REDES NEURONALES ARTIFICIALES

Las redes neuronales artificiales (RNA) constituyen una familia de modelos computacionales inspirados en el funcionamiento del sistema nervioso biológico. Su principal característica es la capacidad de aprender relaciones complejas entre variables de entrada y salida a partir de ejemplos, sin necesidad de definir explícitamente las ecuaciones que describen el fenómeno bajo estudio [89,90].

Estos modelos han sido ampliamente utilizados en problemas de clasificación, regresión, reconocimiento de patrones y procesamiento de señales debido a su capacidad para aproximar funciones altamente no lineales [92, 93]. El desarrollo de las redes neuronales artificiales surge a partir de la observación del funcionamiento de las neuronas biológicas. Como se muestra en la Figura 2.14, una neurona recibe información a través de las dendritas, procesa las señales en el cuerpo celular y transmite una respuesta mediante el axón hacia otras neuronas conectadas por sinapsis.

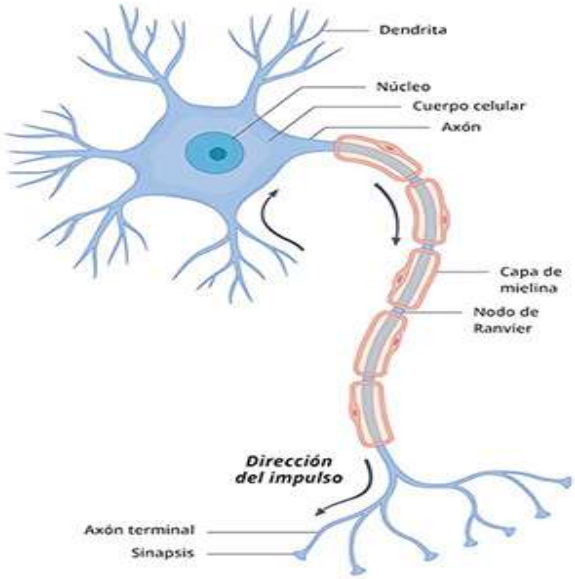


FIGURA 2.15 ESTRUCTURA BÁSICA DE UNA NEURONA BIOLÓGICA. LAS DENDRITAS RECIBEN SEÑALES DE ENTRADA, EL CUERPO CELULAR PROCESA LA INFORMACIÓN Y EL AXÓN TRANSMITE LA RESPUESTA HACIA OTRAS NEURONAS MEDIANTE CONEXIONES SINÁPTICAS [91].

A partir de esta analogía biológica se desarrolló el modelo matemático conocido como perceptrón, representado en la Figura 2.15.

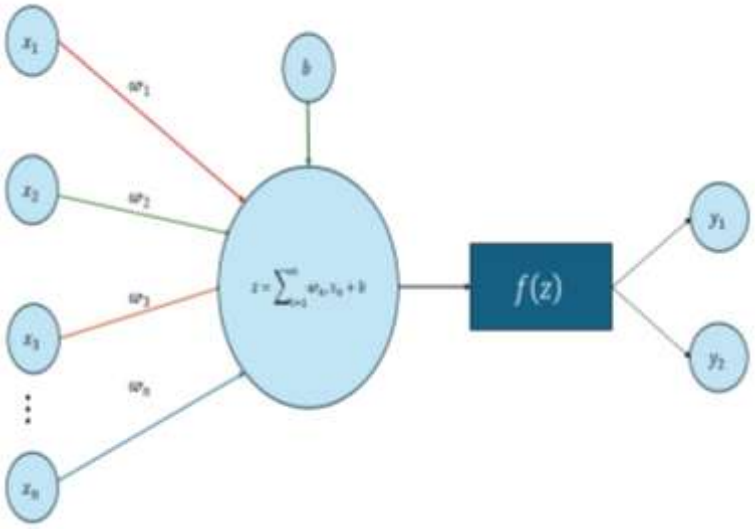


FIGURA 2.16 MODELO MATEMÁTICO DE UNA NEURONA ARTIFICIAL O PERCEPTRÓN. LAS VARIABLES DE ENTRADA SON PONDERADAS MEDIANTE PESOS SINÁPTICOS, COMBINADAS LINEALMENTE Y TRANSFORMADAS MEDIANTE UNA FUNCIÓN DE ACTIVACIÓN PARA PRODUCIR LA SALIDA DEL MODELO [92].

La correspondencia conceptual entre una neurona biológica y una neurona artificial se resume en la Tabla 2.3.

TABLA 2.3 ANALOGÍA ENTRE LOS ELEMENTOS DE UNA NEURONA BIOLÓGICA Y UNA NEURONA ARTIFICIAL [92].

Neurona biológica	Función	Equivalente en el perceptrón	Función del modelo
Dendritas	Entradas que reciben la información a través de impulsos eléctricos.	Entradas $\overrightarrow{x_n} = x_1, x_2, x_3, \dots x_{n-1}$	Características de los datos de entrada.
Núcleo	Procesa la información de entrada.	Suma ponderada $z = \sum_{i=1}^{n} w_n, x_n + b$	Realiza la combinación lineal según su importancia relativa.
Axón	Transmite la información procesada hacia los botones sinápticos.	Función de activación $f(z)$	Transmite la señal procesada tras aplicar la función de activación.

Botones sinápticos	Regulan la transmisión de los neurotransmisores con los que se hace la sinapsis.	Salidas $\hat{y}_j$	Es la salida del modelo, donde se indica la predicción dependiendo si se activa o no la neurona.
Sinapsis	Proceso encargado de transmitir la información a las neuronas subsecuentes.	Pesos, $w_1, w_2, w_3, \dots w_n$ y sesgo, b	Se autoajustan durante el entrenamiento para optimizar la precisión del modelo.

Matemáticamente, la operación realizada por una neurona artificial consiste en calcular una combinación lineal de las variables de entrada:

$$z = \sum_{i=1}^{n} w_i \, x_i + b, \qquad (53)$$

donde x_i representa las variables de entrada, w_i los pesos asociados a cada conexión y b el término de sesgo (***bias***), responsable de desplazar el umbral de activación de la neurona [72,74].

Posteriormente, el resultado se transforma mediante una función de activación:

$$\hat{y} = f(z) \, . \qquad (54)$$

La función de activación introduce no linealidad en el modelo, permitiendo representar relaciones complejas que no pueden ser descritas mediante transformaciones lineales simples [84, 94].

Las redes neuronales modernas están formadas por múltiples neuronas organizadas en capas. Una arquitectura típica se compone de una capa de entrada, una o más capas ocultas y una capa de salida.

Cuando la información fluye únicamente desde la entrada hacia la salida, la arquitectura recibe el nombre de Perceptrón Multicapa (Multilayer Perceptron,

MLP) [84]. Diversos estudios han demostrado que los MLP poseen capacidad de aproximación universal, es decir, pueden aproximar cualquier función continua con una precisión arbitraria siempre que dispongan de un número suficiente de neuronas ocultas [94].

El proceso de entrenamiento consiste en ajustar iterativamente los pesos y sesgos de la red para minimizar una función de pérdida que cuantifica el error entre las predicciones del modelo y los valores reales observados. Este procedimiento se realiza mediante el algoritmo de retropropagación del error (*backpropagation*), combinado con técnicas de optimización basadas en descenso por gradiente [84].

De manera general, la actualización de los parámetros puede expresarse como

$$w_{t+1} = w_t - \eta \frac{\partial L}{\partial w_t} \quad , \quad (55)$$

donde L representa la función de pérdida, η es la tasa de aprendizaje (*Learning Rate*) y $\frac{\partial L}{\partial w_t}$ corresponde al gradiente calculado durante la etapa de retropropagación [94].

Actualmente existen diversas arquitecturas derivadas de las redes neuronales artificiales, entre las que destacan las redes convolucionales (*Convolutional Neural Networks, CNN*), las redes recurrentes (*Recurrent Neural Networks, RNN*), las redes de memoria a largo plazo (*Long Short-Term Memory, LSTM*) y los modelos basados en mecanismos de atención (*Transformers*) [89], [92]. Sin embargo, debido a que el problema abordado en esta investigación corresponde a la clasificación de vectores de características numéricas obtenidos a partir del modelado físico de una guía de onda tipo *slot*, la arquitectura seleccionada es un Perceptrón Multicapa (MLP), cuya estructura resulta adecuada para el procesamiento de este tipo de datos tabulares.

3. METODOLOGÍA

En este capítulo se describe la metodología empleada para el desarrollo del sistema inteligente de detección de contaminantes propuesto en esta investigación. El procedimiento integra el modelado físico de una guía de onda tipo *slot*, el cálculo de parámetros ópticos y optoelectrónicos derivados de la interacción entre el campo electromagnético junto con el medio analizado, así como la implementación de modelos de aprendizaje automático para la clasificación de diferentes especies químicas presentes en soluciones acuosas.

La metodología se desarrollo en tres etapas principales. En la primera se realizó el modelado espectral del índice de refracción de soluciones acuosas que contienen arsénico, fluoruro, nitrato y cloruro de sodio. Para ello se empleó la ecuación de Sellmeier como modelo de dispersión del agua y un factor de ajuste asociado a las propiedades fisicoquímicas de cada analito, permitiendo obtener índices de refracción dependientes tanto de la longitud de onda como de la concentración.

Posteriormente, los índices de refracción calculados fueron utilizados como parámetros de entrada para el modelado óptico de una guía de onda tipo *slot*. A partir de la solución numérica de la ecuación trascendental correspondiente al modo transversal magnético (*TM*) se obtuvo la constante de propagación modal y la distribución espacial del campo eléctrico. Con esta información se determinó el factor de confinamiento en la región del *slot* y se calcularon diversos parámetros optoelectrónicos relacionados con la interacción entre la radiación guiada y el medio analizado, incluyendo sensibilidad espectral, velocidad de grupo, factor de interacción electromagnética y magnitudes eléctricas equivalentes.

Finalmente, los parámetros obtenidos mediante el modelado físico fueron integrados en un conjunto de datos sintético utilizado para el entrenamiento y evaluación de modelos de aprendizaje automático. Como etapa previa se realizó un análisis exploratorio de datos para identificar relaciones entre variables, evaluar la distribución de las clases y analizar posibles redundancias en las características calculadas. Posteriormente se

implementaron y optimizaron diferentes algoritmos de clasificación, incluyendo árboles de decisión, bosques aleatorios, XGBoost y perceptrones multicapa, con el propósito de determinar su capacidad para identificar contaminantes a partir de los parámetros ópticos y optoelectrónicos generados por el sensor fotónico propuesto.

3.1 MODELADO Y ANÁLISIS DEL INDICE DE REFRACCIÓN DE SOLUCIONES ACUOSAS

El desempeño de los sensores ópticos basados en guías de onda depende directamente de las propiedades ópticas del medio analizado, particularmente de su índice de refracción. Debido a que la presencia de especies químicas disueltas modifica la polarizabilidad efectiva de la solución y, por consiguiente, su respuesta óptica, resulta necesario disponer de un modelo capaz de describir la variación del índice de refracción en función de la longitud de onda y de la concentración del analito.

En esta sección se describe el desarrolló un modelo analítico para estimar el índice de refracción de soluciones acuosas que contienen arsénico, fluoruro, nitrato y cloruro de sodio. El procedimiento combina la ecuación de Sellmeier para agua pura con un factor de ajuste dependiente de las propiedades fisicoquímicas de cada especie química, permitiendo obtener una descripción espectral continua del índice de refracción para diferentes concentraciones.

3.1.1 MODELADO ESPECTRAL DEL AGUA PURA

Como se describió en la Sección 2.1.3, la dispersión óptica del agua puede modelarse mediante la ecuación de Sellmeier de cuatro términos. Para esta investigación se emplearon los coeficientes reportados por Daimon para agua destilada a $20°C$, presentados previamente en la Tabla 2.1 [56].

La implementación numérica se realizó en MATLAB considerando un intervalo espectral comprendido el mismo rango espectral reportado de $182.78nm$ a $1128.95nm$, con

incrementos de $1nm$. Para cada longitud de onda se calculó el índice de refracción correspondiente, obteniéndose una descripción continua de la dispersión óptica del agua pura.

Con el propósito de verificar la correcta implementación del modelo, los valores calculados fueron comparados con los datos experimentales reportados. La Figura 3.1 muestra la concordancia entre ambas series de datos, observándose que el modelo reproduce adecuadamente la dependencia espectral del índice de refracción del agua en el intervalo analizado.

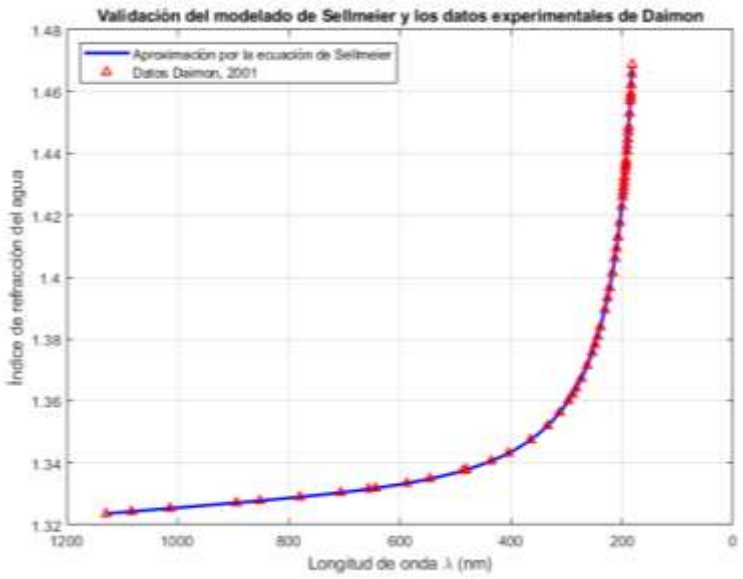


FIGURA 3.1 ÍNDICE DE REFRACCIÓN DEL AGUA EN FUNCIÓN DE LA LONGITUD DE ONDA. LA LÍNEA CONTINUA CORRESPONDE A LOS VALORES CALCULADOS MEDIANTE LA ECUACIÓN DE SELLMEIER, MIENTRAS QUE LOS MARCADORES REPRESENTAN LOS DATOS EXPERIMENTALES REPORTADOS POR DAIMON [56].

Los valores obtenidos de $n(\lambda)$ constituyen la condición base empleada posteriormente para incorporar el efecto de la concentración de los diferentes analitos y generar los índices de refracción de las soluciones acuosas consideradas en esta investigación.

3.1.2 INCORPORACIÓN DEL EFECTO DE LA CONCENTRACIÓN

La presencia de especies químicas disueltas altera la polarizabilidad efectiva del sistema agua-soluto, generando modificaciones en el índice de refracción respecto al correspondiente al agua pura. Diversos estudios han mostrado que, dentro de intervalos moderados de concentración, el índice de refracción de soluciones acuosas presenta una dependencia aproximadamente lineal con la concentración del soluto [53, 54].

Con base en esta consideración, se introdujo un factor de ajuste k asociado a las propiedades fisicoquímicas de cada analito. Dicho factor se obtuvo a partir de la refracción molar y la masa molar de la especie química considerada:

$$k_0 = \left(\frac{R_{soluto}}{M_{soluto}}\right) \times 10^{-3}, \qquad (56)$$

donde R_{soluto} representa la refracción molar y M_{soluto} la masa molar del compuesto. El cociente R_{soluto}/M_{soluto}, una vez expresado en L/mg, es el coeficiente físico K_0, y el factor 10^{-3} de la ecuación (52) equivale a un escalamiento computacional $\alpha = 10^3$, de modo que $k = \alpha K_0$. Este escalamiento no es una conversión adicional de unidades; su deducción para el $NaCl$ se presenta en el Apéndice 3.

La dependencia final del índice de refracción con la longitud de onda y la concentración se expresó mediante:

$$n(\lambda, C) = n_{H_2O}(\lambda) + kC\ , \qquad (57)$$

donde C corresponde al valor de concentración del analito, expresado en mg/L, utilizado en el modelo. En la implementación computacional se introdujo un factor de escalamiento α en el coeficiente de ajuste, de modo que ($k = \alpha K_0$). Este factor se utiliza como parte de la parametrización numérica del modelo y no representa una conversión adicional de unidades. Por tanto, su utilización debe distinguirse de la evaluación del modelo con concentraciones físicas expresadas en mg/L. La dependencia del índice de refracción utilizada para generar los datos sintéticos se expresa por la ecuacion 57, donde C corresponde a la variable de concentración utilizada en la implementación computacional.

El significado y las unidades de esta variable deben mantenerse consistentes con la definición del coeficiente empleado. En particular, el coeficiente escalado no debe interpretarse automáticamente como equivalente al coeficiente físico cuando se utilizan concentraciones obtenidas directamente de datos experimentales. Con el propósito de examinar la respuesta del modelo frente a datos publicados, en el Apéndice 3 se realiza una comparación independiente con los índices de refracción experimentales de soluciones acuosas de NaCl reportados por Tan y Huang [97]. Para dicha comparación, la fracción másica se convierte a concentración física en mg/L mediante la densidad de la solución y se emplea el coeficiente K_0 sin escalar. Esta distinción permite separar la parametrización computacional utilizada para generar los datos sintéticos de la comparación cuantitativa con datos experimentales.

Esta formulación conserva la dependencia espectral descrita por la ecuación de Sellmeier e incorpora una contribución dependiente de la concentración del contaminante. Para la generación de los datos sintéticos, el coeficiente de ajuste (k) se consideró constante dentro del intervalo de concentraciones analizado. Esta aproximación permite representar de manera simplificada la variación del índice de refracción con la concentración, sin describir explícitamente posibles cambios no lineales de las propiedades ópticas de la solución. Los parámetros fisicoquímicos de cada analito utilizados para el cálculo de k se presentan en la Tabla 3.1.

TABLA 3.1 PROPIEDADES FISICOQUÍMICAS EMPLEADAS PARA EL CÁLCULO DEL FACTOR DE AJUSTE $\boldsymbol{k}$ DE CADA ANALITO. [51-53]

Analito	Masa Molar (g/mol)	Refracción Molar (cm^3/mol)
H_3AsO_4	141.94	13.48
F^-	18.998	0.95
KNO_3	101.10	10.81
NaCl	58.44	7.96

3.1.3 APLICACIÓN AL SISTEMA $NaCl + H_2O$

Como primer caso de estudio se consideró una solución acuosa de cloruro de sodio NaCl, debido a que constituye uno de los sistemas electrolíticos más ampliamente estudiados y utilizados como referencia para la validación de modelos de índice de refracción [54]. Los parámetros físicos utilizados para el modelado incluyeron la masa molar y la refracción molar del $NaCl$, así como la densidad y masa molar del agua. A partir de estos valores se calculó el factor de ajuste k correspondiente al sistema $NaCl + H_2O$. Posteriormente se realizó un barrido completo de concentraciones comprendido entre 0 y $1000 mg/L$ como máxima concentración como se establece en la *NOM-127-SSA1-2021* [96], utilizando incrementos de $1 mg/L$. Para cada concentración se calculó el índice de refracción en el intervalo espectral de $200 nm$ a $110 nm$, generando una base de datos compuesta por pares λ, C y su correspondiente valor de índice de refracción. El resultado de este procedimiento fue un conjunto de datos estructurado con tres variables principales, (λ) la longitud de onda, (C) la concentración del analito y (n) índice de refracción calculado para dichas condiciones.

La variación simultánea del índice de refracción con la longitud de onda y la concentración se muestra en la Figura 3.2, donde puede observarse el efecto combinado de la dispersión óptica y del incremento de concentración del soluto.

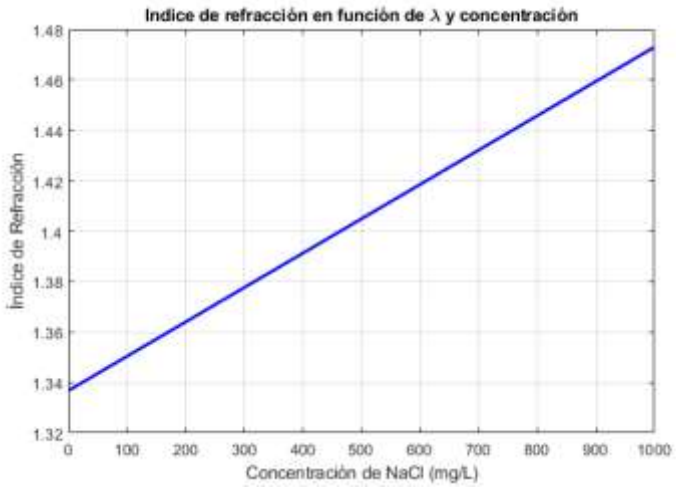


FIGURA 3.2 ÍNDICE DE REFRACCIÓN DEL SISTEMA $\boldsymbol{NaCl + H_2O}$ EN FUNCIÓN DE $\boldsymbol{n(\lambda, C)}$ CON $\boldsymbol{\lambda}$=500NM OBTENIDA MEDIANTE EL MODELO $\boldsymbol{n(\lambda, C) = n_{H_2O}(\lambda) + kC}$.

Como referencia para examinar la tendencia del índice de refracción, los resultados del modelo se compararon con los datos experimentales reportados por Tan y Huang para soluciones acuosas de NaCl [97]. La comparación se presenta en la Tabla A3.1 y en la Figura A3.1 del Apéndice 3. Para calcular los índices modelados utilizados en esta comparación, se convirtió la fracción másica reportada en el artículo a concentración física en mg/L, considerando la densidad de cada solución, y se utilizó el coeficiente $k_0$sin escalar.

Los datos de referencia corresponden a una temperatura de 293.15K y una longitud de onda de $589.3\,nm$, condiciones que se utilizaron para efectuar la comparación. No obstante, las concentraciones experimentales convertidas abarcan aproximadamente de 10,000 a 297,000 mg/L, mientras que el intervalo utilizado para generar el conjunto de datos sintéticos de esta investigación es de 0 a 1000 mg/L. Por esta diferencia de escala, la comparación se interpreta como una referencia del comportamiento general del índice de refracción con la concentración y no como una validación directa del modelo dentro del intervalo de trabajo de la tesis. Los resultados y sus diferencias respecto de los valores experimentales se analizan en el Apéndice 3. La validación del modelo en el intervalo de concentraciones considerado en esta investigación requerirá una

comparación específica dentro de dicho intervalo y, posteriormente, una evaluación experimental.

La metodología descrita para el cloruro de sodio fue posteriormente extendida a las soluciones acuosas de arsénico, fluoruro y nitrato. De esta manera se obtuvo un conjunto completo de índices de refracción dependientes de la longitud de onda y de la concentración, los cuales constituyen la entrada fundamental para el modelado óptico de la guía de onda tipo *slot* desarrollado en la siguiente sección.

3.2 MODELADO ÓPTICO DE LA GUÍA DE ONDA *SLOT*

Una vez obtenidos los índices de refracción en función de la longitud de onda y de la concentración de los analitos, se realizó el modelado óptico de la guía de onda tipo slot. Esta etapa permitió determinar la constante de propagación del modo *TM*, obtener la distribución transversal de su campo eléctrico y cuantificar la fracción del campo confinada en la región ocupada por el analito. A partir de estos resultados se calcularon las propiedades derivadas que integraron el conjunto de datos sintético para la posterior ingeniería de características y el entrenamiento de los modelos de aprendizaje automático.

El procedimiento se organizó en tres etapas consecutivas: solución de la ecuación trascendental para obtener β, reconstrucción del campo eléctrico y cálculo de la fracción evanescente y estimación de la velocidad de grupo, la respuesta eléctrica, el factor de interacción y la sensibilidad. Los parámetros geométricos y ópticos empleados se presentan en la Tabla 3.2.

TABLA 3.2 PARÁMETROS GEOMÉTRICOS Y ÓPTICOS DE LA GUÍA DE ONDA TIPO *SLOT*.

Parámetro		Valor
Índice de refracción del slab	n_s	3.48
Índice de refracción del revestimiento	n_c	2.00
Ancho del *slot*	$2a$	100 nm
Grosor del slab	w_{slab}	450 nm
Rango de longitud de onda	λ	200 - 1100 nm

3.2.1 Cálculo modal y obtención de la fracción evanescente

Para cada par de valores de longitud de onda y concentración se cargaron los parámetros de entrada y se calcularon las constantes necesarias para evaluar la ecuación trascendental del modo *TM* fundamental, presentada en la Sección 2.6. La constante de propagación β se definió como variable simbólica y la ecuación se resolvió mediante vpasolve, utilizando el intervalo de búsqueda comprendido entre $k_0 n_c$ y $k_0 n_{\mathrm{h}}$, de acuerdo con el primer diagrama. El valor obtenido se conservó junto con las condiciones de entrada para el cálculo del campo transversal.

Con el valor de β obtenido se actualizaron las constantes transversales y se determinaron los coeficientes del campo mediante las condiciones de frontera descritas en la Sección 2.6. El perfil de E_{x} se evaluó en una malla de 600 puntos, comprendida entre -0.4 y $0.4\ \mu m$, y se normalizó respecto de su magnitud máxima.

Posteriormente, se integró el cuadrado de la magnitud del campo en el slot y en el dominio transversal para obtener la fracción evanescente, siguiendo la operación indicada en el diagrama. Este procedimiento y la solución modal se repitieron para cada combinación de longitud de onda y concentración. Los resultados se almacenaron en un archivo *CSV* con las columnas λ, C, n, β y ***FE_TM_frac***, utilizado como entrada de la etapa siguiente.

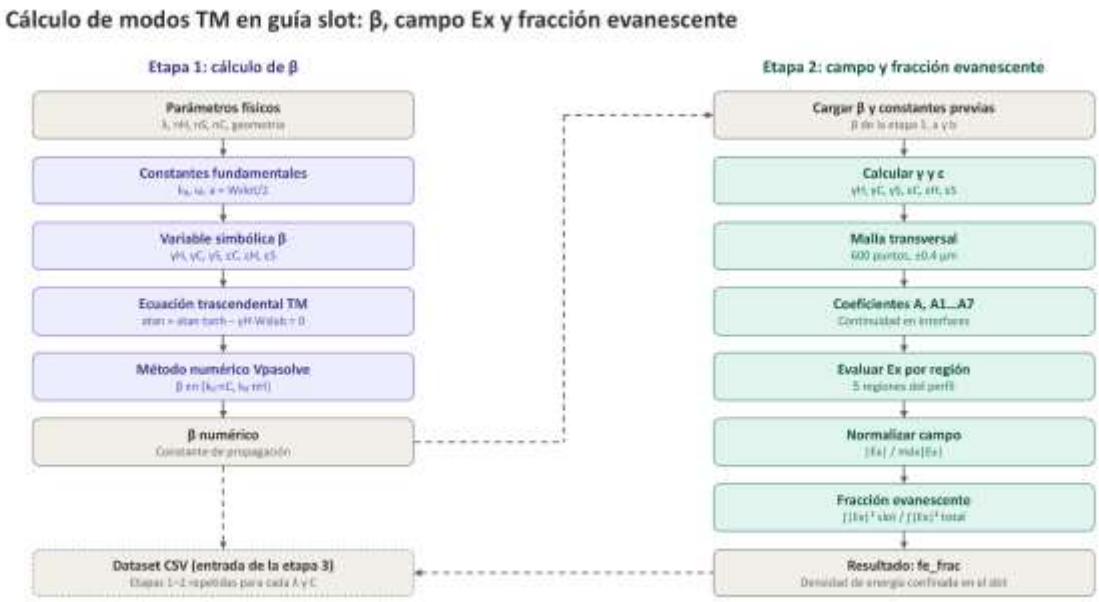


FIGURA 3.3 DIAGRAMA DE FLUJO PARA EL CÁLCULO DE LA CONSTANTE DE PROPAGACIÓN DEL MODO TM, LA DISTRIBUCIÓN TRANSVERSAL DEL CAMPO ELÉCTRICO Y LA FRACCIÓN EVANESCENTE EN EL SLOT.

3.2.2 Cálculo y exportación de las propiedades derivadas

Los datos del archivo modal se ordenaron por concentración y longitud de onda. Para cada concentración se calculó la velocidad de grupo a partir de la derivada de β respecto de la frecuencia angular, con el tratamiento de Savitzky–Golay indicado en el diagrama. A continuación, se evaluaron la magnitud energética en el slot y el campo eléctrico efectivo.

La conductividad se asignó según la concentración: para $C = 0$ se utilizó el valor del agua y, para los demás casos, la relación entre conductividad molar y concentración molar. Con estos resultados se calcularon el voltaje, la resistencia y la potencia eléctrica equivalente. También se obtuvo el factor de interacción mediante la secuencia de operaciones del segundo diagrama.

Para calcular la sensibilidad, los datos se agruparon por longitud de onda y se ordenaron por concentración. La derivada del índice efectivo respecto de C se estimó con *np.gradient*, considerando el espaciado no uniforme. Cuando se dispuso de al menos

siete puntos, se aplicó a la derivada un filtro de *Savitzky–Golay* con ventana de siete puntos y polinomio de grado tres.

Finalmente, se reunieron las cinco columnas de entrada y las cinco propiedades calculadas: velocidad de grupo, voltaje, potencia eléctrica equivalente, factor de interacción y sensibilidad. Para cada analito se exportó el archivo Analito_propiedades_optoelectronicas.csv, destinado a la posterior ingeniería de características y al entrenamiento de los modelos de aprendizaje automático.

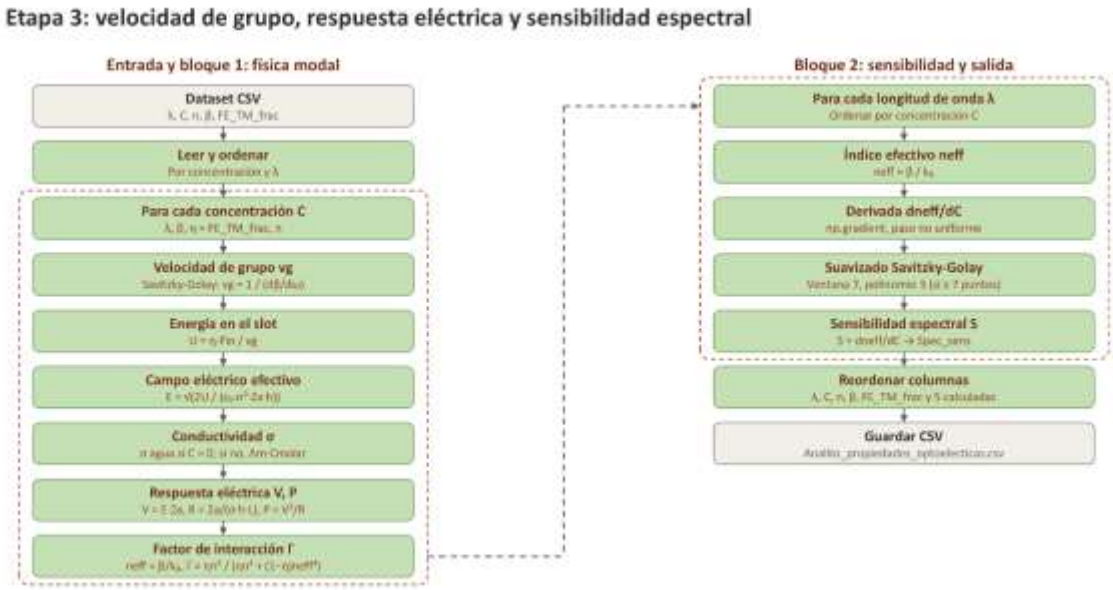


FIGURA 3.4 DIAGRAMA DE FLUJO PARA EL CÁLCULO DE LA VELOCIDAD DE GRUPO, LA RESPUESTA ELÉCTRICA EQUIVALENTE, EL FACTOR DE INTERACCIÓN ELECTROMAGNÉTICA Y LA SENSIBILIDAD, Y SU INTEGRACIÓN EN EL CONJUNTO DE DATOS.

3.2.1 SOLUCIÓN DE LA ECUACIÓN TRASCENDENTAL PARA EL MODO *TM*

La constante de propagación del modo guiado se obtuvo mediante la resolución numérica de la ecuación trascendental *TM* asociada a la estructura *slot*. Debido a que esta ecuación no posee una solución analítica cerrada, se implementó un procedimiento iterativo para localizar la raíz correspondiente al modo fundamental guiado para cada combinación de

longitud de onda e índice de refracción del medio analizado, para ello se utilizaron los parámetros geométricos y ópticos de la guía de onda tipo *slot* resumidos en la Tabla 3.2 se definieron los parámetros ópticos de la estructura y posteriormente se resolvió la ecuación característica dentro del intervalo físicamente permitido para la constante de propagación. Aquellas soluciones complejas o no convergentes fueron descartadas automáticamente, conservándose únicamente los valores reales asociados a modos guiados.

La Figura 3.4 muestra la variación de la constante de propagación modal en función de la concentración para los cuatro analitos propuestos y en el intervalo espectral analizado.

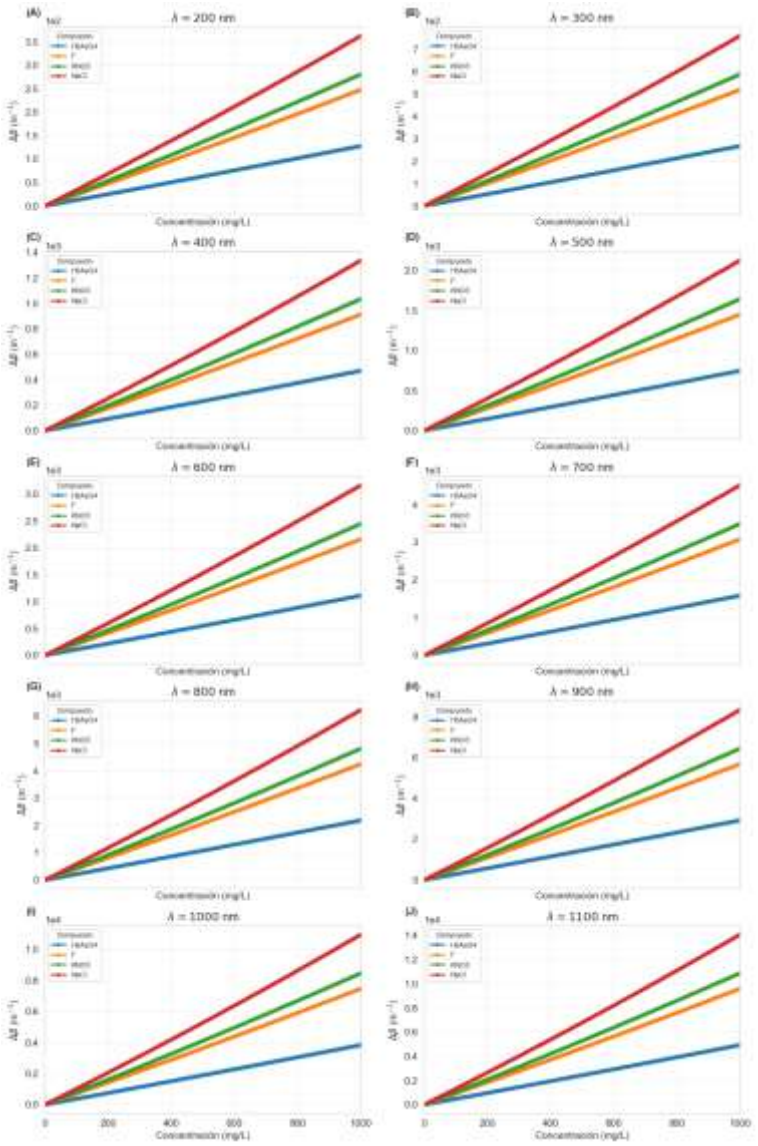


FIGURA 3.4 VARIACIÓN DE LA CONSTANTE DE PROPAGACIÓN MODAL $\boldsymbol{\beta}$ EN FUNCIÓN DE LA CONCENTRACIÓN PARA $\boldsymbol{H_3AsO_4}$, $\boldsymbol{F^-}$, $\boldsymbol{KNO_3}$ Y $\boldsymbol{NaCl}$. CALCULADA MEDIANTE LA SOLUCIÓN NUMÉRICA DE LA ECUACIÓN TRASCENDENTAL *TM* DE LA GUÍA DE ONDA TIPO *SLOT*. LAS SUBFIGURAS (A-J) CORRESPONDEN A LONGITUDES DE ONDA REPRESENTATIVAS COMPRENDIDAS ENTRE 200 Y 1100 $\boldsymbol{nm}$.

En todos los casos se observa una dependencia de tendencia lineal entre la constante de propagación y la concentración del analito. Este comportamiento es consecuencia directa del incremento del índice de refracción efectivo del medio contenido en el *slot*, el cual modifica las condiciones de confinamiento del modo guiado y produce variaciones en la constante de propagación obtenida a partir de la ecuación trascendental.

Asimismo, puede apreciarse que la magnitud de la variación de β depende tanto de la longitud de onda como de la especie química analizada. Para todas las longitudes de onda consideradas, el $NaCl$ presenta la mayor pendiente, seguido por KNO_3, F^- y H_3AsO_4. Esta diferencia está asociada al factor de ajuste k utilizado en el modelado del índice de refracción, el cual incorpora las propiedades fisicoquímicas de cada analito y determina la magnitud de la perturbación óptica inducida en el medio.

También se observa que los valores absolutos de β disminuyen conforme aumenta la longitud de onda, comportamiento consistente con la relación $\beta = n_{\text{eff}} k_0$. Debido a que el número de onda disminuye al incrementar la longitud de onda, la constante de propagación modal presenta valores menores en la región del infrarrojo cercano que en la región visible o ultravioleta.

Con el propósito de analizar la respuesta de la constante de propagación frente a variaciones en la concentración de los analitos, se calculó la sensibilidad modal espectral

$$S_\beta = \frac{\partial \beta}{\partial c}, \tag{58}$$

Este parámetro permite evaluar la capacidad de la guía de onda para traducir pequeñas variaciones en la composición química del medio en cambios detectables de las propiedades modales. La Figura 3.5 presenta la evolución de la sensibilidad modal espectral en función de la longitud de onda para los cuatro analitos considerados.

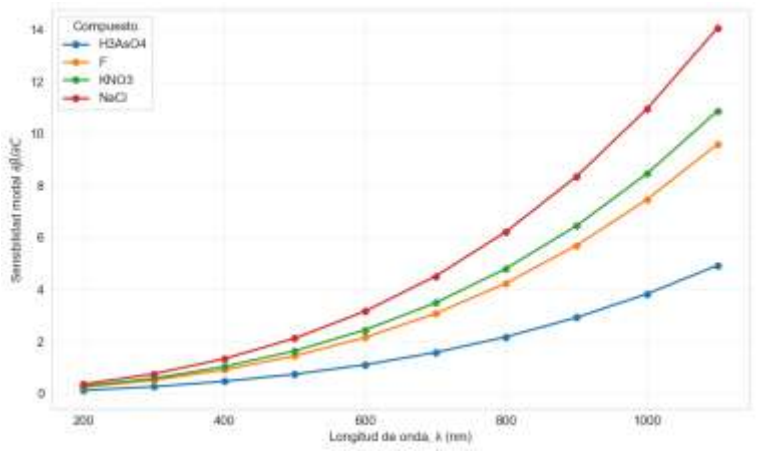


FIGURA 3.5 SENSIBILIDAD MODAL ESPECTRAL ($\partial\beta/\partial C$) EN FUNCIÓN DE LA LONGITUD DE ONDA PARA H_3AsO_4, F^-, KNO_3 Y $NaCl$.

Se observa como la sensibilidad aumenta en todo el intervalo espectral analizado, alcanzando sus valores máximos en la región cercana a 1100 nm. Este comportamiento indica que, para la geometría modelada, las variaciones del índice de refracción inducidas por la concentración producen cambios más pronunciados en la constante de propagación a longitudes de onda elevadas. Esta tendencia es consistente con las diferencias observadas previamente en las pendientes de las curvas de la Figura 3.4 y refleja la distinta magnitud de la perturbación óptica generada por cada especie química sobre el índice de refracción del medio. Estos resultados muestran que la constante de propagación modal constituye un parámetro sensible tanto a la concentración del analito como a la longitud de onda de operación. La relación aproximadamente lineal observada entre β y la concentración permitiendo establecer una correspondencia directa entre las propiedades ópticas del medio contenido en el *slot* y las características de propagación del modo guiado.

3.2.2 CÁLCULO DEL FACTOR DE CONFINAMIENTO η_{slot}

Una vez obtenida la constante de propagación modal β mediante la resolución numérica de la ecuación trascendental TM, se procedió a calcular la distribución espacial del campo eléctrico asociada al modo fundamental de la guía de onda *slot*. Para ello se emplearon las expresiones 33 y 34 de la Sección 2.8 para cada una de las regiones que conforman la estructura y los parámetros geométricos y ópticos de la Tabla 3.2. La solución modal descrita en la sección anterior permitió determinar las amplitudes del campo eléctrico en el revestimiento superior, los slabs y el *slot*. A partir de esta distribución espacial se calculó el factor de confinamiento en el *slot* η_{slot}, definido como la fracción de energía electromagnética confinada dentro de la región sensible de la guía de onda respecto a la energía total del modo guiado.

Este parámetro se obtuvo mediante integración numérica de la intensidad del campo eléctrico a lo largo de toda la sección transversal de la estructura, de acuerdo con la expresión 34. Debido a que el mecanismo de detección propuesto se basa en la interacción del campo evanescente con el analito contenido en el *slot*, el factor de confinamiento constituye una magnitud de interés para evaluar el potencial de sensado de la estructura.

La Figura 3.5 muestra la variación de η_{slot}en función de la concentración para los cuatro analitos estudiados y para diferentes longitudes de onda comprendidas entre 200 y 1100 nm.

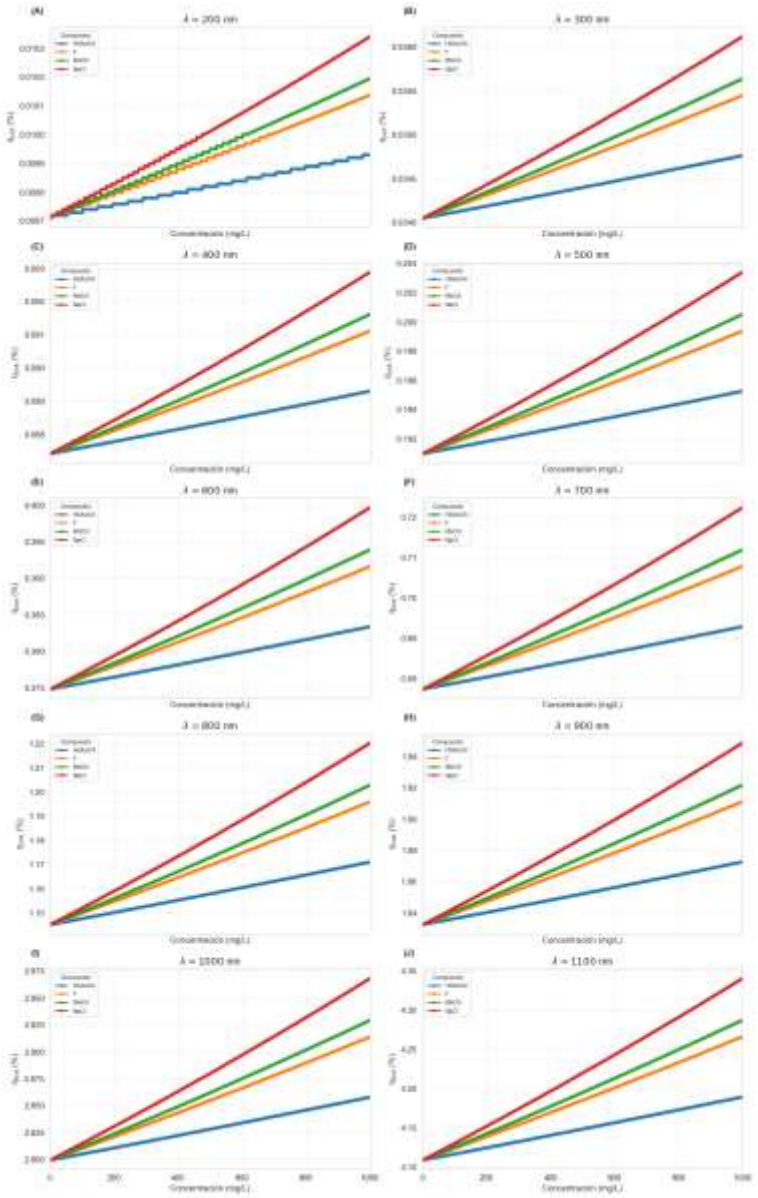


FIGURA 3.6 VARIACIÓN DEL FACTOR DE CONFINAMIENTO EN EL *SLOT* η_{slot} EN FUNCIÓN DE LA CONCENTRACIÓN PARA H_3AsO_4, F^-, KNO_3 Y $NaCl$. LAS SUBFIGURAS (A–J) CORRESPONDEN A LONGITUDES DE ONDA ENTRE 200 Y 1100 NM.

en cada caso se observa un incremento de tendencia lineal del factor de confinamiento conforme aumenta la concentración del analito. Este comportamiento está asociado al incremento del índice de refracción de la solución contenida en el *slot*, el cual modifica las condiciones de propagación del modo guiado y favorece una mayor concentración de energía electromagnética dentro de la región de sensado.

Asimismo, la magnitud del confinamiento presenta una fuerte dependencia espectral. Mientras que para longitudes de onda cercanas a 200 nm el factor de confinamiento es del orden de $10^{-2}\%$, para longitudes de onda próximas a 1100 nm alcanza valores superiores al 4 %. Este resultado indica que el modo TM presenta una interacción progresivamente mayor con el medio contenido en el *slot* conforme aumenta la longitud de onda.

Las diferencias observadas entre analitos siguen la misma tendencia identificada previamente para la constante de propagación modal. Esta diferencia se relaciona con la magnitud de la perturbación óptica introducida por cada especie química sobre el índice de refracción del medio.

Con el propósito de cuantificar la sensibilidad del confinamiento frente a variaciones de concentración, se calculó la derivada espectral del factor de confinamiento respecto a la concentración,

$$\frac{\partial \eta_{slot}}{\partial C}, \tag{59}$$

la cual representa la variación del confinamiento por unidad de concentración del analito.

La evolución espectral de este parámetro se observa en la Figura 3.6.

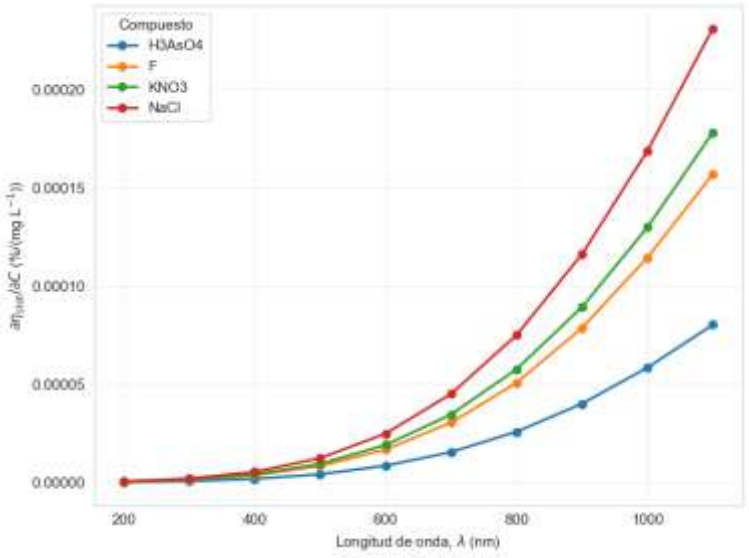


FIGURA 3.7 SENSIBILIDAD ESPECTRAL DEL FACTOR DE CONFINAMIENTO $\partial \eta_{slot}/\partial C$ EN FUNCIÓN DE LA LONGITUD DE ONDA PARA H_3AsO_4, F^-, KNO_3 Y $NaCl$.

Se observa que la sensibilidad del confinamiento aumenta de manera constante en todo el intervalo espectral. Los valores más elevados se obtienen para longitudes de onda superiores a 900 nm, donde pequeñas variaciones en la concentración producen cambios más pronunciados en la fracción de energía confinada dentro del *slot*.

Estos resultados muestran que la constante de propagación modal constituye un parámetro sensible tanto a la concentración del analito como a la longitud de onda de operación. Para cada longitud de onda analizada, la relación aproximadamente lineal observada entre β y la concentración, dentro del intervalo estudiado, permite relacionar las propiedades ópticas del medio contenido en el slot con las características de propagación del modo guiado.

3.2.3 CÁLCULO DE PARÁMETROS OPTOELECTRÓNICOS DERIVADOS DEL MODO GUIADO

Una vez obtenidos la constante de propagación modal y el factor de confinamiento del modo *TM*, se continuo con el cálculo de los parámetros optoelectrónicos complementarios que permiten caracterizar la interacción entre el campo electromagnético guiado y el medio contenido en el *slot*.

Como etapa inicial se estimó la velocidad de grupo v_g, ecuación 35, obtenida a partir de la derivada numérica de la constante de propagación respecto a la frecuencia angular. Debido a que esta magnitud se emplea posteriormente para el cálculo de la energía óptica almacenada en el dispositivo, se aplicó un procedimiento de suavizado basado en filtros de Savitzky-Golay con el propósito de reducir fluctuaciones numéricas asociadas al cálculo de derivadas discretas.

A partir de la velocidad de grupo y del factor de confinamiento previamente calculado, se determinó la energía óptica almacenada en la región del *slot*. Posteriormente, esta energía se empleó para estimar el campo eléctrico efectivo considerando el volumen activo de interacción y las propiedades dieléctricas del medio analizado.

Con el valor del campo eléctrico efectivo se calcularon el voltaje inducido en la región del *slot* V_{slot} y la potencia eléctrica equivalente $P_{electric}$. Para ello se incorporó la conductividad eléctrica de los analitos, estimada a partir de la conductividad molar de cada compuesto y de su concentración correspondiente.

La Figura 3.8 presenta la variación del voltaje inducido en el *slot* para una longitud de onda representativa.

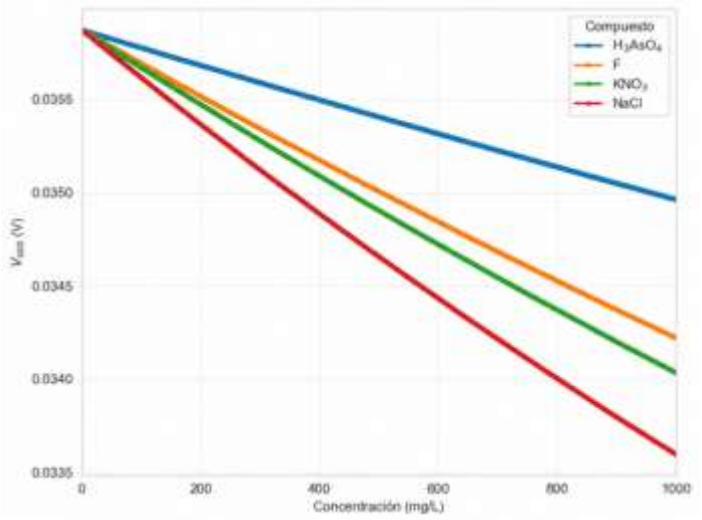


FIGURA 3.8 VARIACIÓN DEL VOLTAJE INDUCIDO EN LA REGIÓN DEL *SLOT* V_{slot} EN FUNCIÓN DE LA CONCENTRACIÓN PARA LOS ANALITOS CONSIDERADOS.

Se observa como V_{slot} disminuye conforme aumenta la concentración del analito. Este comportamiento se encuentra asociado al incremento del índice de refracción del medio, el cual modifica la distribución de energía electromagnética en la región activa del dispositivo y produce una reducción gradual del campo eléctrico efectivo calculado para el modo guiado.

Adicionalmente se calculó el factor de interacción electromagnética η_{factor}, definido como la fracción efectiva de energía dieléctrica asociada al medio contenido en el *slot* respecto a la energía total transportada por el modo guiado. Este parámetro incorpora simultáneamente la información del factor de confinamiento y del índice efectivo del modo, proporcionando una medida de la intensidad de interacción entre el campo electromagnético y el analito.

La evolución espectral de este parámetro se muestra en la Figura 3.9.

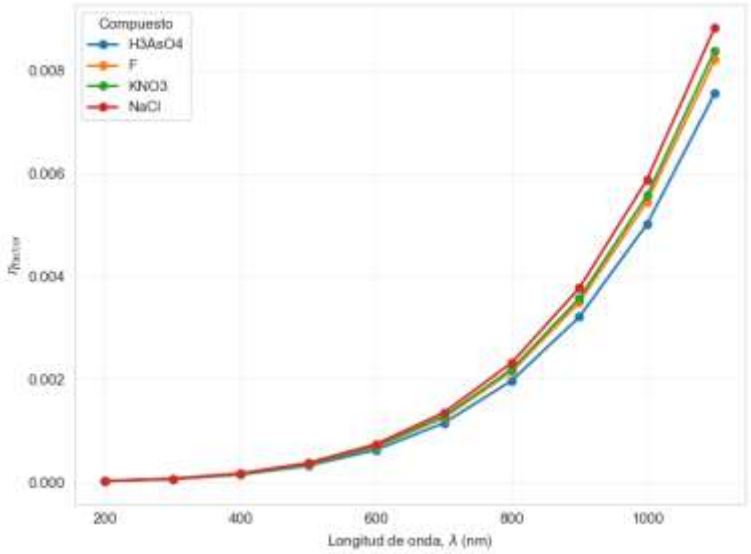


FIGURA 3.9 FACTOR DE INTERACCIÓN ELECTROMAGNÉTICA η_{factor} EN FUNCIÓN DE LA LONGITUD DE ONDA PARA LOS CUATRO ANALITOS ESTUDIADOS. EL INCREMENTO PROGRESIVO DE ESTE PARÁMETRO INDICA UN AUMENTO DE LA INTERACCIÓN CAMPO-MATERIA EN LA REGIÓN ESPECTRAL DE MAYORES LONGITUDES DE ONDA.

Se puede apreciar que η_{factor} aumenta de forma no lineal conforme se incrementa la longitud de onda, alcanzando sus valores máximos en la región cercana a 1100 nm. Este comportamiento indica que, para la geometría propuesta, la interacción entre el campo guiado y el medio analizado se intensifica a longitudes de onda elevadas.

Finalmente, se calculó la sensibilidad espectral, S_λ, este parámetro permite cuantificar la variación del índice efectivo frente a cambios de concentración y constituye una medida directa de la capacidad del dispositivo para transformar perturbaciones químicas en variaciones ópticas detectables.

La Figura 3.10 muestra la evolución de S_λ para los diferentes compuestos considerados. Se observa un incremento monotónico de la sensibilidad conforme aumenta la longitud de onda, alcanzándose los valores máximos en la región del infrarrojo cercano.

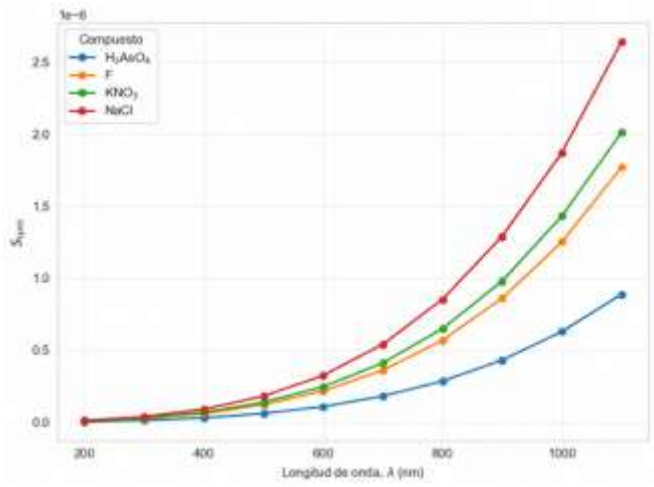


FIGURA 3.10 SENSIBILIDAD ESPECTRAL S_λ EN FUNCIÓN DE LA LONGITUD DE ONDA.

Esta tendencia coincide con el comportamiento observado para η_{factor}, sugiriendo que el aumento de la interacción electromagnética entre el campo guiado y el medio favorece una mayor sensibilidad del sistema frente a cambios de concentración, asimismo, puede apreciarse que las diferencias entre analitos se amplifican progresivamente con la longitud de onda.

Los parámetros obtenidos en esta etapa complementan la descripción modal realizada en las secciones anteriores y permiten caracterizar distintos mecanismos de interacción entre el campo electromagnético y las soluciones analizadas. En conjunto, estas magnitudes constituyen variables derivadas de interés para la construcción del conjunto de datos sintético empleado posteriormente en la etapa de ingeniería de características y entrenamiento de los modelos de aprendizaje automático.

3.3 METODOLOGÍA DE ENTRENAMIENTO Y EVALUACIÓN DE MODELOS DE APRENDIZAJE AUTOMÁTICO

En esta sección se describe la metodología general empleada para el desarrollo, entrenamiento y evaluación de los modelos de aprendizaje automático utilizados en la clasificación de las especies químicas propuestas. Como etapa inicial, se realizó un análisis exploratorio de datos con el fin de examinar la calidad del conjunto de datos, identificar posibles anomalías y comprender la distribución de las variables obtenidas a partir del modelado físico de la guía de onda tipo *slot*.

Posteriormente, se prepararon los datos para su entrada en los algoritmos de clasificación mediante procesos de limpieza, codificación de la variable objetivo e imputación de valores faltantes.

Asimismo, se definieron estrategias de partición de los datos para entrenamiento, validación y prueba, garantizando la conservación de la distribución de clases y la reproducibilidad de los experimentos.

En el caso de los modelos basados en árboles, se empleó una partición de entrenamiento y prueba de 80% y 20%, respectivamente. Para la red neuronal multicapa MLP, el conjunto de datos se dividió en tres subconjuntos: entrenamiento 70%, validación 15% y prueba 15%, utilizando una estrategia estratificada y una semilla aleatoria fija para garantizar la reproducibilidad de los resultados.

Con el propósito de evaluar la robustez de los modelos frente a perturbaciones en las variables de entrada, se generaron versiones perturbadas mediante la adición de ruido gaussiano de media nula y desviación estándar de 0.02. La media nula permite introducir fluctuaciones positivas y negativas sin imponer un sesgo sistemático, mientras que la desviación estándar establece la magnitud de las perturbaciones. El valor de 0.02 se adoptó como una condición de evaluación computacional para analizar la sensibilidad de los modelos ante pequeñas variaciones en sus entradas. Esta perturbación constituye una aproximación simplificada a posibles fluctuaciones de medición y no representa una incertidumbre instrumental determinada experimentalmente.

Todos los modelos fueron implementados mediante un flujo de trabajo de procesamiento orientada a garantizar condiciones homogéneas de entrenamiento. Para los algoritmos basados en árboles se empleó imputación por mediana y escalamiento robusto de las características. En el caso de la red neuronal MLP, se utilizó una normalización mediante StandardScaler, ajustada únicamente con los datos de entrenamiento para evitar fuga de información hacia los conjuntos de validación y prueba.

Finalmente, la selección de hiperparámetros se realizó mediante procedimientos de optimización específicos para cada modelo. Durante esta etapa se empleó validación cruzada y la exactitud como métrica objetivo. Una vez identificada la configuración óptima de cada algoritmo, se entrenó un modelo final utilizando la totalidad del conjunto de entrenamiento y posteriormente se evaluó sobre el conjunto de prueba independiente.

De esta manera, todos los modelos fueron comparados bajo un protocolo experimental consistente, garantizando una evaluación objetiva de su capacidad para clasificar contaminantes presentes en muestras de agua.

3.3.1 ANÁLISIS EXPLORATORIO DE DATOS

Antes iniciar con la configuración de los modelos de aprendizaje automático se realizó un análisis exploratorio de datos con el propósito de caracterizar estadísticamente el conjunto de datos generado, identificar posibles relaciones entre variables, evaluar la presencia de redundancias y analizar la separabilidad de las clases correspondientes a las diferentes especies químicas.

Debido a que el conjunto de datos fue generado sintéticamente, todas las clases contienen el mismo número de muestras. La Tabla 3.3 presenta la distribución de clases por analito.

TABLA 3.3 DISTRIBUCIÓN DE MUESTRAS POR CLASE.

Analito	Número de muestras	Porcentaje (%)
H_3AsO_4	91,091	25.0
F^-	91,091	25.0
KNO_3	91,091	25.0
NaCl	91,091	25.0

Con el fin de identificar la distribución general de las variables, se calcularon medidas descriptivas, valores extremos y asimetría. Los resultados se resumen en la Tabla 3.4.

TABLA 3.4 ESTADÍSTICAS DE LAS VARIABLES DE ENTRENAMIENTO.

clase	count	mean	std	min	25%	50%	75%	max	asimetría
λ	364364	6.50E+02	2.63E+02	2.00E+02	4.20E+02	6.50E+02	8.80E+02	1.10E+03	0
concentración	364364	5.00E+02	2.89E+02	0.00E+00	2.50E+02	5.00E+02	7.50E+02	1.00E+03	0
n_{slot}	364364	1.39E+00	3.84E-02	1.32E+00	1.36E+00	1.38E+00	1.41E+00	1.56E+00	0.659896
β	364364	4.10E+07	2.22E+07	1.88E+07	2.40E+07	3.30E+07	5.16E+07	1.09E+08	1.282301
η	364364	1.04E-02	1.17E-02	9.71E-05	1.07E-03	5.18E-03	1.69E-02	4.34E-02	1.143696
η_{factor}	364364	1.80E-03	2.09E-03	1.63E-05	1.74E-04	8.54E-04	2.88E-03	8.85E-03	1.231419
S_λ	364364	4.72E-07	5.36E-07	3.97E-09	6.45E-08	2.53E-07	7.05E-07	2.64E-06	1.463148
V_{slot}	364364	6.78E-02	4.69E-02	6.92E-03	2.59E-02	5.76E-02	1.05E-01	1.71E-01	0.510316
P_{slot}	364364	9.27E+04	1.90E+05	6.53E-13	2.67E+03	1.57E+04	9.09E+04	1.78E+06	3.815775

Las métricas muestran que las variables presentan escalas significativamente diferentes, abarcando desde magnitudes del orden de 10^{-9} hasta valores superiores a 10^8. Asimismo, algunas variables presentan distribuciones asimétricas elevadas, particularmente aquellas asociadas con potencia e interacción energética. Estas diferencias justifican la necesidad de aplicar procesos de normalización y estandarización antes del entrenamiento de los modelos de aprendizaje automático, especialmente para el modelo de redes neuronales artificiales.

Posteriormente se evaluó la dependencia lineal entre las variables mediante el coeficiente de correlación de Pearson. La matriz de correlación se presenta en la Figura 3.11.

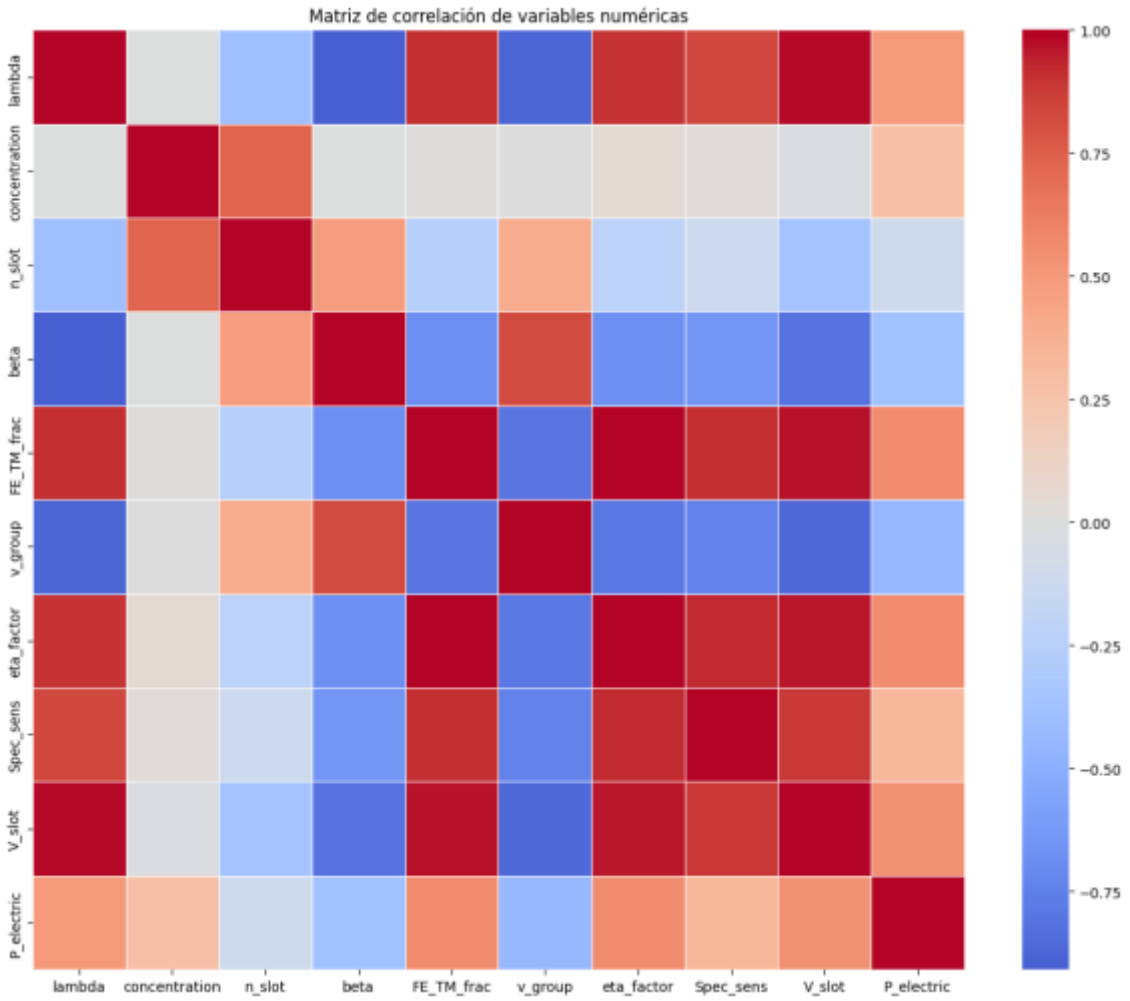


FIGURA 3.11 MATRIZ DE CORRELACIÓN DEL CONJUNTO DE DATOS.

La matriz muestra la existencia de grupos de variables altamente correlacionadas. En particular, se observa una correlación muy elevada entre el factor de confinamiento n_*slot* y el factor de interacción electromagnética eta_factor, así como entre la sensibilidad espectral Spec_sens.

Con el fin de identificar posibles redundancias, se calcularon las correlaciones absolutas más elevadas entre pares de variables. Los resultados se presentan en la Tabla 3.5.

TABLA 3.5 VARIABLES CON MAYOR CORRELACIÓN ABSOLUTA.

Variable 1	Variable 2	Correlación absoluta
η	η_{factor}	0.997640
λ	V_{slot}	0.980845
η	V_{slot}	0.969139
η_{factor}	V_{slot}	0.958600
η_{factor}	S_{λ}	0.920428
η	S_{λ}	0.909369
λ	β	0.908451
λ	η	0.908062

Estos resultados indican que múltiples variables contienen información parcialmente redundante debido a que derivan de los mismos mecanismos físicos de propagación e interacción electromagnética. Sin embargo, se conservaron debido a que los algoritmos de aprendizaje automático pueden aprovechar dichas relaciones durante el proceso de entrenamiento.

Con el objetivo de evaluar la estructura global del conjunto de datos y explorar la separabilidad entre las clases, se aplicó un Análisis de Componentes Principales (PCA). Previamente, las variables fueron estandarizadas para evitar que diferencias de escala afectaran la extracción de componentes.

La Figura 3.12 muestra la proyección bidimensional utilizando las dos primeras componentes principales.

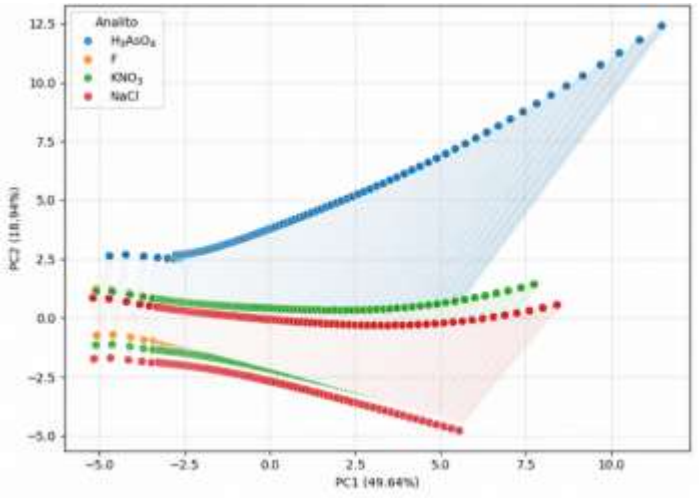


FIGURA 3.12 PROYECCIÓN PCA DEL CONJUNTO DE DATOS.

Puede observarse que las clases presentan regiones diferenciadas dentro del espacio reducido, particularmente para H_3AsO_4 , cuya distribución se observa claramente separada respecto a los demás compuestos. Asimismo, aunque existe cierto solapamiento parcial entre F^- , KNO_3 y $NaCl$ en algunas regiones, las trayectorias generadas por cada analito siguen patrones diferenciables, lo que sugiere la existencia de información suficiente para tareas de clasificación. Las dos primeras componentes principales explican aproximadamente el 68.6% de la variabilidad total del conjunto de datos, distribuyéndose en 49.64% para la primera componente y 18.94% para la segunda.

Complementariamente, la Figura 3.13 presenta la varianza acumulada explicada por el modelo PCA.

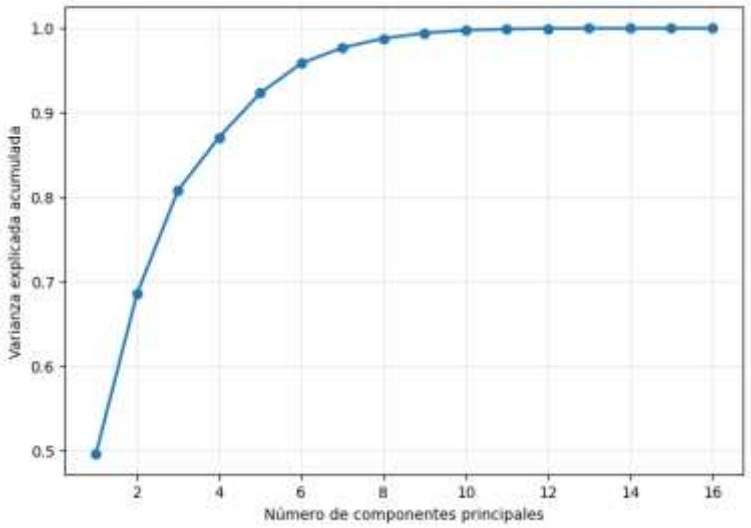


FIGURA 3.13 VARIANZA ACUMULADA EXPLICADA POR PCA.

Se observa que aproximadamente seis componentes permiten explicar más del 95% de la variabilidad total, mientras que diez componentes explican prácticamente la totalidad de la información contenida en el conjunto de datos. Este resultado indica que, aunque el espacio original presenta múltiples variables, gran parte de la información relevante se encuentra concentrada en un subconjunto reducido de dimensiones.

En conjunto, los resultados del análisis exploratorio muestran que el conjunto de datos presenta una distribución balanceada entre clases, una estructura física coherente entre variables y un grado apreciable de separabilidad entre los diferentes contaminantes. Estas características sugieren que el conjunto de datos es adecuado para el entrenamiento y evaluación de modelos de aprendizaje automático orientados a la clasificación de las especies químicas propuestas a partir de parámetros ópticos y optoelectrónicos derivados de una guía de onda tipo *slot*.

3.3.2 METODOLOGÍA DEL ALGORITMO ARBOL DE DECISIÓN

La implementación del algoritmo Decision Tree Classifier se realizó considerando como variable objetivo la identificación de los cuatro analitos propuestos H_3AsO_4, F^-, KNO_3y $NaCl$. El preprocesamiento de los datos, la partición de los conjuntos de entrenamiento y prueba, así como la incorporación de ruido gaussiano, se llevaron a cabo siguiendo el protocolo experimental descrito en la Sección 3.3.

La selección de hiperparámetros se efectuó mediante optimización bayesiana, con el objetivo de maximizar la exactitud promedio obtenida durante el proceso de validación cruzada. Para cada ensayo se construyó un árbol de decisión con una configuración específica de hiperparámetros y posteriormente se evaluó mediante validación cruzada tipo (K)-Fold con cinco particiones.

Durante el proceso de optimización se exploraron diferentes configuraciones relacionadas con la profundidad máxima del árbol (max_depth), el número mínimo de muestras requeridas para efectuar una división interna (min_samples_split), el número mínimo de muestras permitidas en los nodos terminales (min_samples_leaf), el criterio de impureza empleado para seleccionar las particiones (criterion), el parámetro de poda basado en complejidad de costo (ccp_alpha) y la ponderación de clases (class_weight).

La evaluación de cada configuración se realizó calculando la exactitud promedio obtenida en las cinco particiones de validación. Este valor fue utilizado como función objetivo durante el proceso de optimización, permitiendo identificar automáticamente la combinación de hiperparámetros con mejor desempeño. Una vez concluida la búsqueda, se seleccionó la configuración óptima y se entrenó un modelo final utilizando la totalidad del conjunto de entrenamiento. Posteriormente, el clasificador fue evaluado sobre el conjunto de prueba independiente con el propósito de estimar su capacidad de generalización. Adicionalmente, se registraron los resultados obtenidos en cada partición de la validación cruzada, así como la media y desviación estándar de la exactitud, con el fin de analizar la estabilidad estadística del modelo.

La Tabla 3.6 resume el espacio de búsqueda de hiperparámetros considerado durante la optimización del árbol de decisión.

TABLA 3.6 ESPACIO DE BÚSQUEDA DE HIPERPARÁMETROS EMPLEADO DURANTE LA OPTIMIZACIÓN BAYESIANA DEL CLASIFICADOR BASADO EN ÁRBOLES.

Hiperparámetro	Rango explorado
max_depth	3 – 12
min_samples_split	2 – 20
min_samples_leaf	1 – 10
criterion	gini, entropy, log_loss
ccp_alpha	0.0 – 0.05
class_weight	None
Número de pruebas Optuna	50
Validación cruzada	K-Fold (k=5)
Métrica objetivo	Accuracy

3.3.3 METODOLOGÍA DEL ALGORITMO BOSQUES ALEATORIOS

Se implementó un clasificador de bosques aleatorios para evaluar el desempeño de un modelo de ensamble en la identificación de las especies químicas y comparar sus resultados con los obtenidos mediante el árbol de decisión.

Con el fin de garantizar una comparación consistente entre los algoritmos evaluados, se mantuvieron las mismas etapas de preprocesamiento, partición de datos y validación descritas previamente. La selección de hiperparámetros se realizó mediante una búsqueda exhaustiva (*Grid Search*), considerando diferentes combinaciones del número de árboles del ensamble (n_{estimators}), la profundidad máxima de los árboles (max_depth), el número mínimo de muestras por nodo terminal (min_samples_leaf) y el número de variables consideradas en cada división (max_features).

Cada configuración fue evaluada mediante validación cruzada estratificada de cinco particiones, utilizando la exactitud como criterio de selección. Posteriormente, el modelo con el mejor desempeño promedio fue entrenado utilizando la totalidad del conjunto de entrenamiento y sometido a una validación adicional mediante (K)-Fold para analizar la estabilidad de sus resultados.

Finalmente, el modelo optimizado fue evaluado sobre el conjunto de prueba independiente con el propósito de determinar su capacidad de generalización y comparar su desempeño con los demás algoritmos de aprendizaje automático considerados en esta investigación. La Tabla 3.7 presenta el espacio de búsqueda de hiperparámetros utilizado durante el proceso de optimización.

TABLA 3.7 HIPERPARÁMETROS Y ESPACIO DE BÚSQUEDA PARA LA OPTIMIZACIÓN DEL CLASIFICADOR DE BOSQUES ALEATORIOS.

Hiperparámetro	Espacio de búsqueda
(n_{estimators})	100, 200
(max_depth)	10, 20
(min_samples_leaf)	2, 4, 10
(max_features)	sqrt, log2
Método de optimización	Grid Search
Validación durante optimización	Stratified K-Fold (k = 5)
Validación final	K-Fold (k = 5)
Métrica objetivo	Accuracy

3.3.4 METODOLOGÍA DEL ALGORITMO XGBOOST

Como se discutió en la Sección 2.7.4, este algoritmo construye secuencialmente un conjunto de árboles de decisión, donde cada nuevo árbol busca corregir los errores cometidos por los árboles previamente generados mediante la optimización iterativa de una función objetivo.

La selección de hiperparámetros se llevó al igual que los algoritmos anteriores mediante optimización bayesiana. Durante este proceso se exploraron parámetros relacionados con la regularización de los pesos del modelo, la tasa de aprendizaje, la profundidad máxima de los árboles, la proporción de muestras empleadas durante cada iteración de entrenamiento, la proporción de variables utilizadas para construir cada árbol y el peso mínimo requerido en los nodos hijos.

Para cada conjunto de hiperparámetros se construyó un clasificador XGBoost de tipo multiclase utilizando la estrategia *multi:softmax*, permitiendo asignar directamente cada muestra a una de las cuatro categorías consideradas. La evaluación de cada configuración se realizó mediante la métrica de exactitud, empleada como función objetivo durante el proceso de optimización.

Una vez concluida la búsqueda, se seleccionó la combinación de hiperparámetros que produjo el mejor desempeño y se entrenó un modelo final utilizando el conjunto completo de entrenamiento. Posteriormente, dicho modelo fue evaluado sobre el conjunto de prueba independiente con el fin de estimar su capacidad de generalización y comparar su desempeño con los demás algoritmos de clasificación estudiados. El espacio de búsqueda de hiperparámetros considerado durante la optimización del clasificador XGBoost se muestra en la tabla 3.8.

TABLA 3.8 HIPERPARÁMETROS Y RANGOS DE BÚSQUEDA CONSIDERADOS DURANTE LA OPTIMIZACIÓN BAYESIANA DEL CLASIFICADOR XGBOOST.

Hiperparámetro	Intervalo de búsqueda	Hiperparámetro	Intervalo de búsqueda
λ(regularización L2)	(10^{-3}) – 10	colsample_bytree	0.50 – 1.00
α (regularización L1)	(10^{-3}) – 10	min_child_weight	1 – 10
learning_rate	0.01 – 0.10	Número de ensayos	50
max_depth	3 – 12	Función objetivo	multi:sof*TM*ax
subsample	0.50 – 1.00	Métrica de optimización	Accuracy

3.3.4 METODOLOGÍA DEL PERCEPTRON MULTICAPA

La arquitectura de la red neuronal multicapa se definió mediante capas completamente conectadas (*Dense Layers*), utilizando una función de activación ReLU en las capas ocultas y una función Sof*TM*ax en la capa de salida para realizar la clasificación multiclase. Los pesos de la red fueron inicializados mediante el esquema He Normal, adecuado para funciones de activación de tipo ReLU y capaz de favorecer una convergencia más estable durante el entrenamiento.

Con el objetivo de controlar el sobreajuste y mejorar la capacidad de generalización del modelo, se incorporaron diferentes mecanismos de regularización durante el proceso de optimización. Entre ellos se consideraron regularización L2 sobre los pesos sinápticos, normalización por lotes (*Batch Normalization*) y capas de abandono (*Dropout*). La presencia o ausencia de estos mecanismos fue tratada como un hiperparámetro adicional durante la búsqueda de la arquitectura óptima.

La selección de hiperparámetros se realizó mediante optimización bayesiana. Durante este procedimiento se exploraron diferentes configuraciones relacionadas con el número de capas ocultas, el número de neuronas por capa, la intensidad de la regularización L2, la tasa de abandono, la incorporación de capas de normalización por lotes y la tasa de aprendizaje del optimizador.

Para el entrenamiento se empleó el optimizador Adam junto con la función de pérdida *Sparse Categorical Crossentropy*, adecuada para problemas de clasificación multiclase con etiquetas enteras. La métrica utilizada durante el proceso de optimización fue la precisión. A diferencia de los modelos basados en árboles, la evaluación de cada configuración se realizó mediante un esquema de validación cruzada estratificada integrado directamente dentro del proceso de optimización. Para ello se desarrolló una extensión personalizada del optimizador bayesiano capaz de implementar validación cruzada tipo *Stratified K-Fold* con tres particiones, preservando la proporción de muestras de cada clase durante todas las etapas del entrenamiento. La exactitud promedio obtenida en las tres particiones fue utilizada como función objetivo para comparar las diferentes arquitecturas evaluadas.

Con el fin de reducir el tiempo de entrenamiento y evitar procesos innecesarios una vez alcanzada la convergencia, se incorporaron mecanismos de detención temprana (*Early Stopping*) basados en el desempeño de validación. Adicionalmente, se evaluaron diferentes tamaños de lote (batch size) para determinar la configuración más adecuada para el conjunto de datos utilizado.

Una vez concluido el proceso de optimización, se seleccionó la arquitectura con mejor desempeño promedio y se entrenó un modelo final utilizando la totalidad del conjunto de entrenamiento. Posteriormente, dicho modelo fue evaluado sobre el conjunto de prueba independiente con el propósito de estimar su capacidad de generalización y comparar su desempeño con los algoritmos de árboles de decisión, bosques aleatorios y XGBoost.

La Tabla 3.9 resume el espacio de búsqueda de hiperparámetros considerado durante el proceso de optimización del perceptrón multicapa.

TABLA 3.9 HIPERPARÁMETROS Y ESPACIO DE BÚSQUEDA PARA LA OPTIMIZACIÓN BAYESIANA DE LA RED NEURONAL TIPO PERCEPTRÓN MULTICAPA.

Hiperparámetro	Espacio de búsqueda	Hiperparámetro	Espacio de búsqueda
Número de capas ocultas	2 – 4	Función de activación	ReLu
Neuronas capa inicial	64, 128	Fusión de salida	Sof*TM*ax
Regularización L2	$(10^{-5}) - (10^{-3})$	Función de perdida	Sparse Categorical Crossentropy
Batch Normalization	Sí / No	Número máximo de ensayos	10
Dropout	Sí / No	Validación	Stratified K-Fold (k = 3)
Tasa de Dropout	0.2 – 0.5	Métrica objetivo	Precisión
Learning Rate (Adam)	$(10^{-5}) - (5 \times 10^{-4})$		
Batch Size	32, 64, 128		
Optimizador	Adam		

4. RESULTADOS Y DISCUSIÓN

Para evaluar de manera integral el desempeño de los modelos de clasificación se emplearon diversas métricas y gráficos que permiten analizar su precisión, capacidad discriminante y comportamiento durante el entrenamiento.

Cada una de estas métricas facilita una interpretación más completa y rigurosa de los resultados obtenidos.

En primer lugar, la Matriz de Confusión permite visualizar las predicciones del modelo en relación con las clases verdaderas, identificando tanto los aciertos como los errores de clasificación para cada categoría. Esta herramienta es fundamental para detectar patrones específicos de confusión entre clases y comprender qué tan efectivamente el modelo distingue cada una de ellas [84]. Por su parte, la Curva ROC (Receiver Operating Characteristic) y el Área Bajo la Curva (AUC) evalúan la capacidad del clasificador para discriminar entre clases, midiendo su rendimiento a diferentes umbrales de decisión. Esta métrica es especialmente útil para identificar qué tan bien separa el modelo las categorías y para comparar el comportamiento relativo de cada clase [95].

La Curva de Aprendizaje examina cómo evoluciona el error del modelo conforme aumenta el tamaño del conjunto de entrenamiento. Este análisis permite identificar fenómenos como el sobreajuste o el subajuste, y evaluar si la cantidad de datos disponible es suficiente para alcanzar un desempeño estable [95].

Finalmente, las métricas de Precisión, Recall y Puntuación F1 ofrecen una evaluación cuantitativa del desempeño del clasificador tanto a nivel global como por clase, permitiendo identificar desequilibrios o dificultades particulares en la predicción de ciertas categorías [80]. En conjunto, estas herramientas proporcionan un marco sólido para interpretar los resultados del modelo y evaluar su capacidad para clasificar correctamente las muestras del conjunto de datos. A continuación, se presenta el análisis detallado de cada una de ellas.

4.1 RESULTADOS ARBOLES DE DECISIÓN

La optimización bayesiana de hiperparámetros identificó la configuración con mejor desempeño durante la validación cruzada; estos valores se muestran en la tabla 4.1.

TABLA 4.1 HIPERPARÁMETROS ÓPTIMOS OBTENIDOS PARA EL CLASIFICADOR ÁRBOLES DE DECISIÓN.

Hiperparámetros	Valores óptimos
Profundidad máxima	12
Muestras mínimas para división	12
Muestras mínimas por hoja	10
Criterio de división	Gini
Parámetro de poda	2.46×10^{-4}
Ponderación de clases	none

Esta configuración limita la complejidad del árbol, una profundidad máxima controlada, un mínimo de muestras por división y por hoja, y un parámetro de poda por complejidad de costo trabajan juntos para reducir el riesgo de que el modelo memorice patrones específicos del entrenamiento, sin recortar de más su capacidad de separar clases.

La tabla 4.2 resume la validación cruzada con cinco particiones, que sirvió para estimar la estabilidad del clasificador antes de examinar su comportamiento en el conjunto de prueba.

TABLA 4.2 RESULTADOS DE VALIDACIÓN CRUZADA PARA EL CLASIFICADOR BASADO EN ÁRBOLES DE DECISIÓN.

Fold	Precisión %
1	91.31
2	92.00
3	91.61
4	90.43
5	91.47
Promedio	91.36
Desviación estándar	0.0052
Precisión en conjunto de prueba limpio	92.17

El modelo obtuvo una exactitud promedio de 91.36%, con una desviación estándar de 0.52%, lo que indica un comportamiento estable entre las distintas particiones. Además, la exactitud alcanzada en el conjunto de prueba limpio fue de 92.17%, valor cercano al promedio obtenido durante la validación cruzada. Esta concordancia sugiere que el modelo mantiene una capacidad de generalización adecuada cuando las muestras no presentan perturbaciones adicionales.

La Figura 4.1 muestra la evolución del desempeño del clasificador en función del número de muestras de entrenamiento.

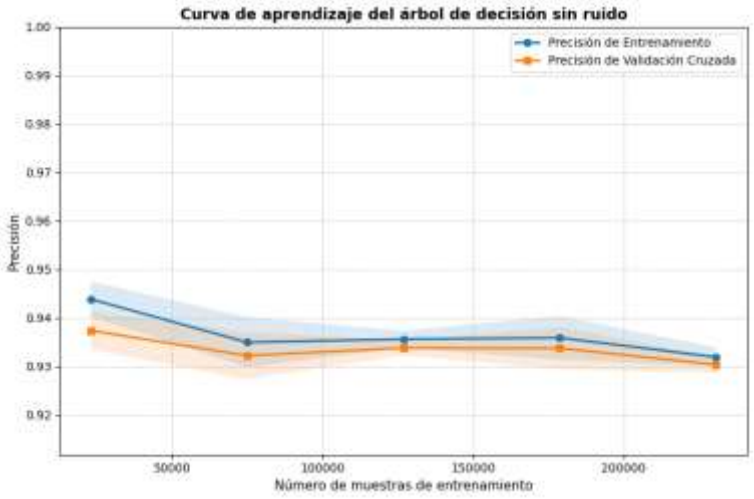


FIGURA 4.1 CURVA DE APRENDIZAJE DEL CLASIFICADOR BASADO EN ÁRBOLES DE DECISIÓN SIN RUIDO.

A lo largo de la curva, las exactitudes de entrenamiento y validación cruzada permanecen próximas, lo que indica una diferencia reducida entre el ajuste al conjunto de entrenamiento y el desempeño sobre datos no vistos durante cada partición. Con $23{,}063$ muestras, el modelo alcanzó una exactitud de entrenamiento de 0.9438 ± 0.0037 y una exactitud de validación cruzada de 0.9374 ± 0.0040. Al utilizar $230{,}630$ muestras, ambos valores disminuyeron ligeramente hasta 0.9319 ± 0.0020 y 0.9303 ± 0.0017, respectivamente, manteniendo una separación mínima entre ambas curvas.

La Figura 4.2 describe la importancia relativa de las variables. El árbol concentró casi toda su capacidad discriminante en unas pocas, P_over_sens encabeza la lista con 0.4160, seguida de sens_P_ratio con 0.3380 y de la concentración con 0.1838. Entre las tres explican cerca del 93.8% de la importancia total.

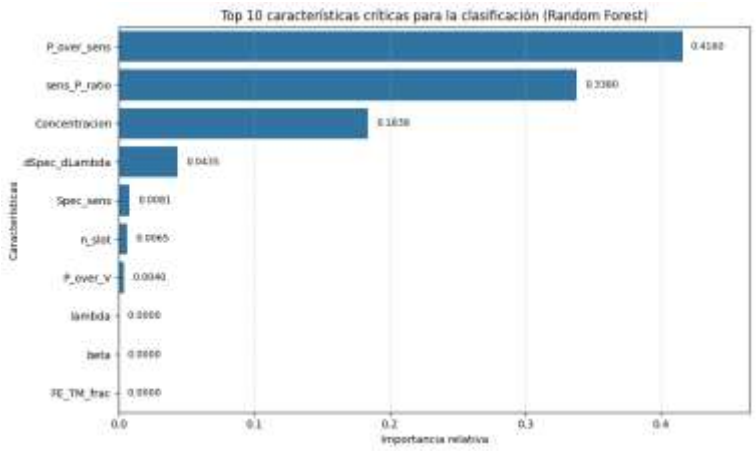


FIGURA 4.2 CARACTERÍSTICAS MÁS RELEVANTES PARA LA CLASIFICACIÓN MEDIANTE ÁRBOLES DE DECISIÓN.

El árbol no repartió la toma de decisiones entre todas las variables por igual, si no que privilegió las relaciones derivadas entre potencia, sensibilidad y concentración. Que variables como λ, β, η_{slot}, V_{slot}, $P_{electric}$ y η_{factor} queden con importancia nula no significa que sean físicamente irrelevantes, su información probablemente quedó absorbida de forma indirecta en las variables derivadas que sí lograron mayor capacidad de separación.

La Tabla 4.3 y la Figura 4.3 resumen el desempeño del modelo por clase sobre el conjunto de prueba limpio. La exactitud global fue de 93%, con un desempeño particularmente alto para las clases F^- y H_3AsO_4. En el caso de F^-, el modelo obtuvo una precisión de 100%, un recall de 99 % y un F1-score de 100%, mientras que para H_3AsO_4 alcanzó una precisión de 99%, un recall de 100% y un F1-score de 100%. Estos resultados indican que ambas clases fueron identificadas con alta efectividad y con un nivel mínimo de confusión respecto a las demás categorías.

TABLA 4.3 REPORTE DE CLASIFICACIÓN DEL MODELO DE ÁRBOLES DE DECISIÓN EN EL CONJUNTO DE PRUEBA LIMPIO.

Clase	Precisión	Recall	F1-score	muestras
F^-	100%	99%	100%	18018
H_3AsO_4	99%	100	100%	18018
KNO_3	93%	78%	85%	18018
$NaCl$	81%	93%	87%	18018

Clase	Precisión	Recall	F1-score	muestras
Precisión			93%	72072
Promedio macro	93%	93%	93%	72072
Promedio ponderado	93%	93%	93%	72072

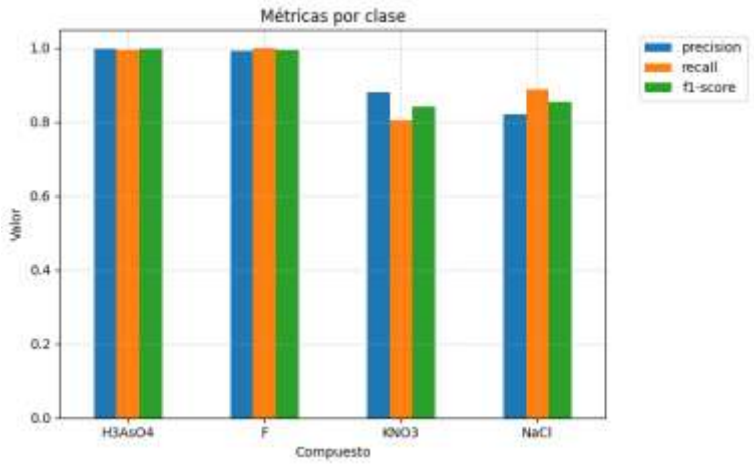


FIGURA 4.3 MÉTRICAS POR CLASE DEL CLASIFICADOR DE ÁRBOLES DE DECISIÓN EN EL CONJUNTO DE PRUEBA LIMPIO.

En contraste, las clases KNO_3 y $NaCl$ presentaron un comportamiento menos equilibrado. Para KNO_3, el modelo alcanzó una precisión de 93%, pero un recall de 78%, lo que indica que una proporción relevante de sus muestras fue clasificada erróneamente como otra clase. $NaCl$ mostró el patrón opuesto: obtuvo un recall elevado de 93%, pero una precisión menor de 81%. Este comportamiento sugiere que el modelo tiende a asignar como $NaCl$ algunas muestras que pertenecen principalmente a KNO_3, lo que evidencia una mayor dificultad para separar ambas clases en el espacio de características utilizado.

La matriz de confusión presentada en la Figura 4.4 confirma el patrón descrito previamente. Las clases F^- y H_3AsO_4 presentan un número mínimo de errores, con $17{,}923$ aciertos de $18{,}018$ muestras para F^- y $18{,}011$ aciertos para H_3AsO_4. Esto evidencia una separación robusta de ambas clases dentro del espacio de características evaluado.

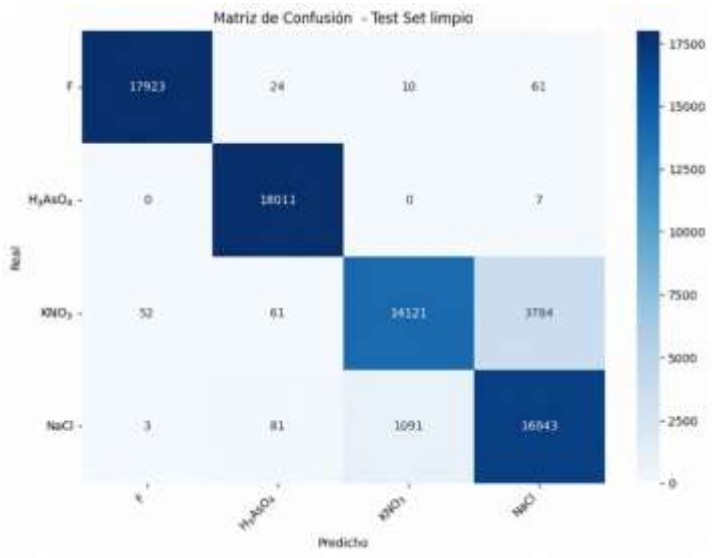


FIGURA 4.4 MATRIZ DE CONFUSIÓN DEL MODELO DE ÁRBOLES DE DECISIÓN EN EL CONJUNTO DE PRUEBA LIMPIO.

La principal fuente de error se concentra en la confusión entre KNO_3 y $NaCl$. En particular, $3{,}784$ muestras reales de KNO_3 fueron clasificadas como $NaCl$, mientras que 1,091 muestras reales de NaCl fueron asignadas a KNO_3. Esta asimetría explica la disminución del recall en KNO_3, ya que una proporción importante de sus muestras no fue recuperada correctamente, así como la reducción de la precisión en $NaCl$, debido a que esta clase recibe muestras que pertenecen a KNO_3.

Las curvas ROC de la Figura 4.5 complementan este análisis al mostrar la capacidad discriminante del modelo en condiciones limpias.

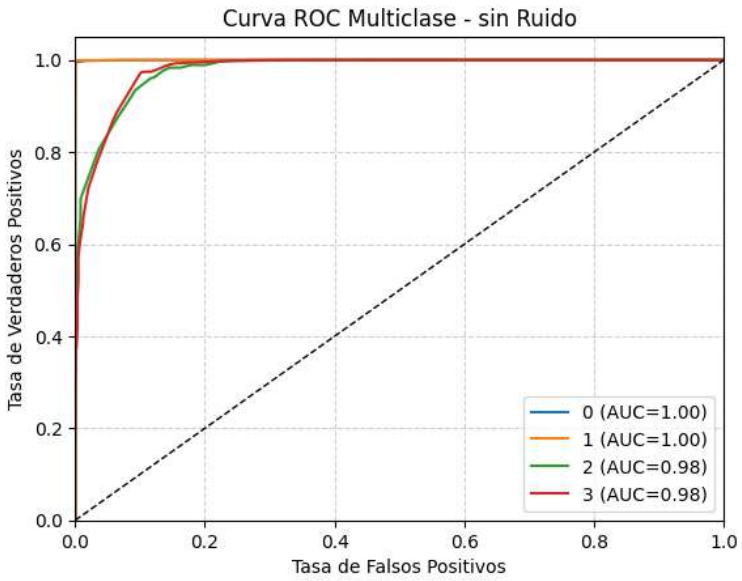


FIGURA 4.5 CURVA ROC MULTICLASE DEL CLASIFICADOR DE ÁRBOLES DE DECISIÓN SIN RUIDO.

Las clases F^- y H_3AsO_4 alcanzan un AUC de 1.00, lo que indica una separación prácticamente perfecta. Por su parte, KNO_3 y $NaCl$ obtienen un AUC de 0.98, valor que sigue representando una discriminación alta, aunque ligeramente inferior a la observada en las otras dos clases.

El siguiente análisis evaluó la robustez del modelo frente a perturbaciones en las variables de entrada. La Tabla 4.4 resume el desempeño global obtenido para los distintos niveles de ruido. En condiciones limpias, el modelo alcanzó una exactitud promedio de 0.9282. Al introducir 2.5% de ruido, la exactitud disminuyó a 0.7172, equivalente a una retención del 77.27. Para niveles superiores de perturbación, la caída fue menos pronunciada, hasta alcanzar una exactitud de 0.7106 con 10.0 de ruido, correspondiente a una retención del 76.56.

TABLA 4.4 DEGRADACIÓN DEL DESEMPEÑO GLOBAL DEL MODELO DE ÁRBOLES DE DECISIÓN BAJO DIFERENTES NIVELES DE RUIDO.

Nivel de ruido	Accuracy promedio	Desviación estándar	Accuracy retenida (%)
0.000	0.928211	1.11×10^{-16}	100.00

Nivel de ruido	Accuracy promedio	Desviación estándar	Accuracy retenida (%)
0.025	0.717205	7.64×10^{-4}	77.27
0.050	0.714236	7.14×10^{-4}	76.95
0.075	0.711988	7.30×10^{-4}	76.71
0.100	0.710638	9.06×10^{-4}	76.56

El salto más fuerte ocurre justo al pasar de datos limpios a 2.5 de ruido, de ahí en adelante la caída es más suave. El árbol es sensible a la primera perturbación de las variables de entrada, pero una vez desplazada su frontera de decisión, seguir aumentando el ruido ya no golpea con la misma fuerza.

La Figura 4.6 muestra cómo cambia la precisión por clase frente al ruido. H_3AsO_4 se mantiene cerca de 0.99 y F^- cerca de 0.96, prácticamente sin verse afectadas. KNO_3 y $NaCl$, en cambio, se desploman: KNO_3 pasa de 0.78 en datos limpios a cerca de 0.52 con 10% de ruido, y $NaCl$ baja de 0.93 a aproximadamente 0.47.

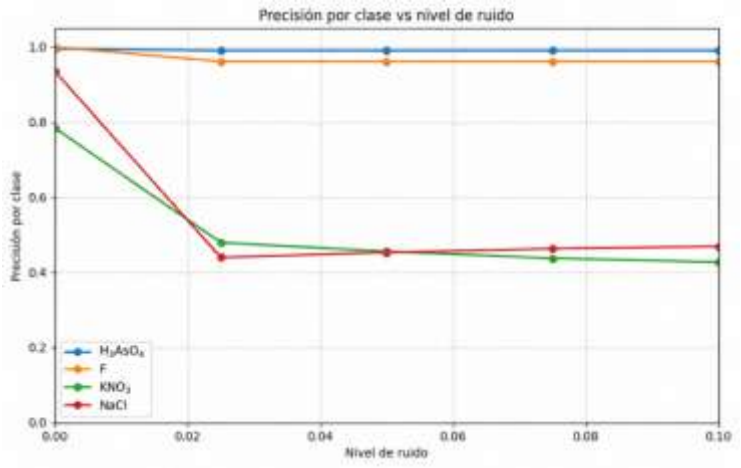


FIGURA 4.6 PRECISIÓN POR CLASE DEL MODELO DE ÁRBOLES DE DECISIÓN FRENTE A DISTINTOS NIVELES DE RUIDO.

Desde el punto de vista de su aplicación, este comportamiento indica que el desempeño obtenido con datos limpios puede disminuir ante pequeñas perturbaciones en las mediciones. Por ello, la confiabilidad del clasificador debe evaluarse a partir del desempeño que conserva bajo ruido, considerando especialmente los casos en que no

identifica correctamente un contaminante presente. La estabilización de la degradación no garantiza por sí misma una clasificación adecuada, pues puede ocurrir en un nivel de desempeño insuficiente para el monitoreo. En consecuencia, estos resultados permiten reconocer una limitación del algoritmo y respaldan la necesidad de validarlo con mediciones experimentales antes de establecer su confiabilidad en un sistema de detección real.

La Figura 4.6 muestra el desempeño por clase bajo los distintos niveles de ruido. Las clases H_3AsO_4 y F^- presentan una degradación mínima, manteniéndose alrededor de 0.99 y 0.96, respectivamente. En contraste, KNO_3 y $NaCl$ muestran una caída considerable: KNO_3 disminuye de 0.78 en condiciones limpias a aproximadamente 0.52 con 10.0% de ruido, mientras que $NaCl$ desciende de 0.93 a cerca de 0.47. Estos resultados evidencian que la robustez del modelo no es homogénea entre clases. Mientras H_3AsO_4 y F^- presentan regiones de decisión más estables, KNO_3 y $NaCl$ son más vulnerables a la perturbación de las características de entrada.

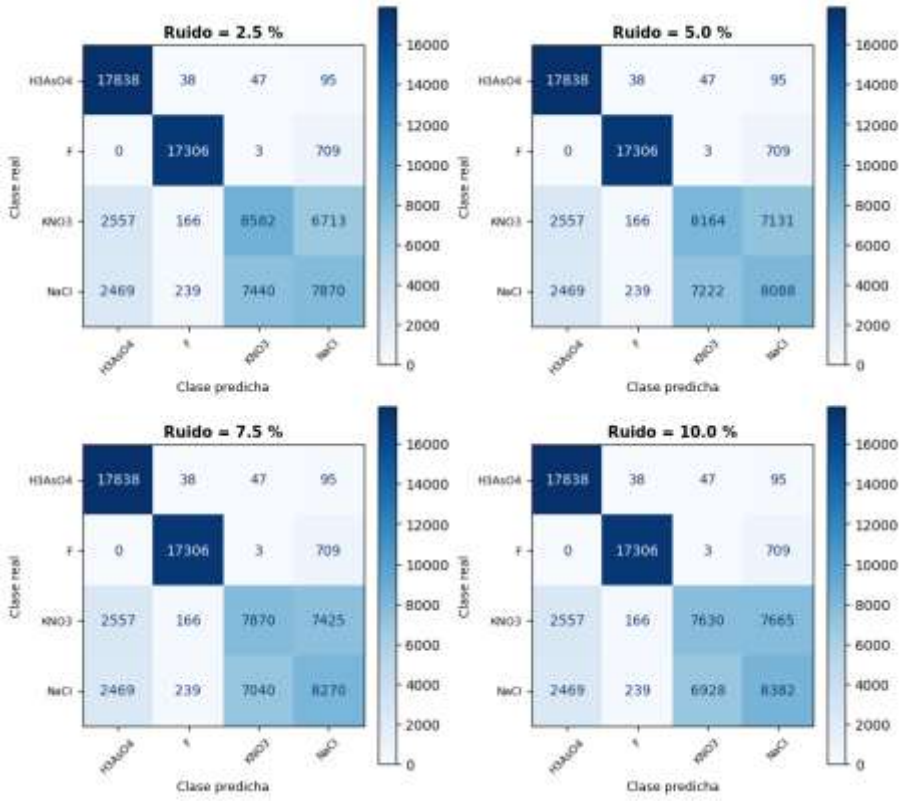


FIGURA 4.7 MATRICES DE CONFUSIÓN DEL MODELO DE ÁRBOLES DE DECISIÓN BAJO NIVELES DE RUIDO DE 2.5%, 5.0%, 7.5% Y 10.0%.

La degradación más evidente se concentra en la separación entre KNO_3 y $NaCl$. En el caso de KNO_3, los aciertos disminuyen progresivamente de $8{,}582$ con 2.5% de ruido a $7{,}630$ con 10.0%. De forma paralela, aumenta el número de muestras de KNO_3 clasificadas como $NaCl$, pasando de $6{,}713$ a $7{,}665$ errores en el mismo intervalo. Este patrón indica que la frontera entre ambas clases es la más sensible a las perturbaciones de entrada.

Para $NaCl$ se observa un comportamiento que podría parecer contradictorio: sus aciertos aumentan de $7{,}870$ a $8{,}382$ al pasar de 2.5% a 10.0% de ruido. Sin embargo, este incremento lejos de una mejora real del desempeño, sino una redistribución de muestras entre KNO_3 y $NaCl$. De hecho, más de $6{,}900$ muestras reales de $NaCl$ son clasificadas como KNO_3 en cada nivel de ruido analizado. Por tanto, las matrices de confusión muestran que la pérdida de desempeño bajo perturbaciones se concentra principalmente en la dificultad del modelo para discriminar entre estas dos clases.

El efecto del ruido también se refleja en las curvas ROC de la Figura 4.8. H_3AsO_4 y F^- mantienen valores de AUC de 0.9055 y 0.9845, respectivamente, en todos los niveles de ruido evaluados. Esto indica que, aunque algunas asignaciones puntuales se modifican, la capacidad discriminante del modelo para estas dos clases permanece estable. En cambio, KNO_3 presenta una degradación progresiva, con una disminución del AUC de 0.6277 con 2.5% de ruido a 0.5940 con 10.0%. $NaCl$ se mantiene en un intervalo entre 0.6755 y 0.6862, sin una caída marcada, aunque con valores claramente inferiores a los observados para H_3AsO_4 y F^-. En conjunto, las curvas ROC confirman que el modelo conserva una capacidad discriminante aceptable para H_3AsO_4 y F^-, pero pierde eficacia para separar KNO_3 y $NaCl$ bajo condiciones ruidosas.

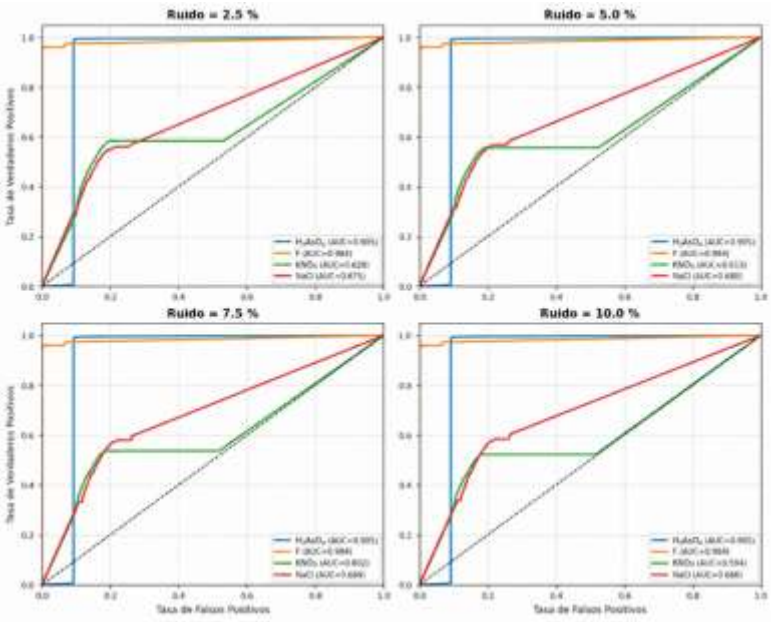


FIGURA 4.8 CURVAS ROC MULTICLASE DEL MODELO DE ÁRBOLES DE DECISIÓN BAJO NIVELES DE RUIDO DE $\mathbf{2.5}\%$, $\mathbf{5}\%$, $\mathbf{7.5}\%$ Y $\mathbf{10.0}\%$.

Finalmente, la Figura 4.9 presenta la degradación operacional del modelo. De acuerdo con los umbrales definidos, el clasificador queda fuera de la zona de operación aceptable desde el primer nivel de ruido evaluado. La exactitud disminuye de 92.82 % en condiciones limpias a valores cercanos al 71 % para todos los niveles de perturbación. Por lo tanto, aunque el modelo presenta un desempeño adecuado con datos limpios, su estabilidad operacional se ve comprometida cuando las variables de entrada incorporan ruido.

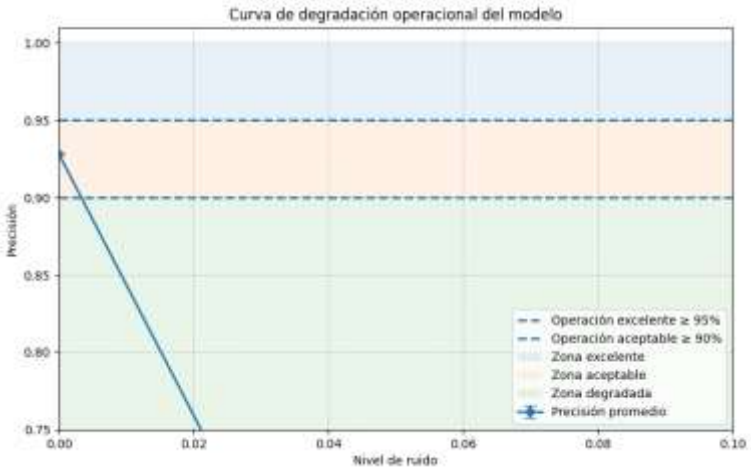


FIGURA 4.9 CURVA DE DEGRADACIÓN OPERACIONAL DEL MODELO DE ÁRBOLES DE DECISIÓN FRENTE AL NIVEL DE RUIDO.

En conjunto, el árbol de decisión constituye un modelo útil como referencia inicial, especialmente por su interpretabilidad y por su capacidad para separar de manera consistente las clases F^- y H_3AsO_4 en condiciones limpias. Sin embargo, su desempeño revela una limitación estructural importante: la frontera de decisión entre KNO_3 y $NaCl$ no es suficientemente estable ante perturbaciones en las variables de entrada. Esta sensibilidad es coherente con la naturaleza particionada del modelo, en la que pequeños desplazamientos de las muestras pueden provocar cambios abruptos en la clase asignada. Por ello, aunque el modelo alcanza una exactitud cercana al 93% en el conjunto de prueba limpio, su degradación bajo ruido reduce su pertinencia para escenarios experimentales donde exista variabilidad instrumental o incertidumbre en las mediciones.

4.2 RESULTADOS BOSQUES ALEATORIOS

El ajuste de hiperparámetros para el modelo de bosques aleatorios por optimización Bayesiana arrojó la configuración que se muestra en la Tabla 4.5.

TABLA 4.5 HIPERPARÁMETROS ÓPTIMOS OBTENIDOS PARA EL CLASIFICADOR BASADO EN BOSQUES ALEATORIOS.

Hiperparámetros	Valores óptimos
Profundidad máxima	20
Número de características por división	sqrt
Muestras mínimas por hoja	2
Número de estimadores	100

Con 100 árboles, una profundidad máxima de 20 y selección aleatoria de características mediante el criterio $sqrt$, el ensamble tiene capacidad suficiente para capturar relaciones no lineales entre variables, mientras el mínimo de muestras por hoja mantiene cierto control sobre la complejidad. Usar múltiples árboles en lugar de uno solo mejoró de forma considerable el desempeño respecto al clasificador individual, al menos en condiciones limpias.

La Tabla 4.6 muestra la validación cruzada con cinco particiones, donde la exactitud se mantuvo al redor de 1.0 en todos los *folds*.

TABLA 4.6 RESULTADOS DE VALIDACIÓN CRUZADA PARA EL CLASIFICADOR BASADO EN BOSQUES ALEATORIOS.

Fold	Precisión %
1	99.98
2	99.99
3	99.98
4	99.97
5	99.98

Fold	Precisión %
Promedio	99.98
Desviación estándar	0.0001
Precisión en conjunto de prueba limpio	100

La validación cruzada mostró un desempeño altamente consistente del modelo. El promedio de precisión fue de 99.98% con una desviación estándar de apenas 0.0001, lo que evidencia una estabilidad muy elevada entre particiones, independientemente de la forma en que se divida el conjunto de entrenamiento. En el conjunto de prueba limpio, la exactitud alcanzó 100%, indicando una capacidad de clasificación prácticamente perfecta en ausencia de perturbaciones por ruido.

La Figura 4.10 muestra la evolución el desempeño conforme crece el número de muestras de entrenamiento. La exactitud de entrenamiento se mantiene cerca de 100% incluso desde los tamaños de muestra más pequeños, mientras que la exactitud de validación cruzada aumenta progresivamente a medida que se incorporan más datos.

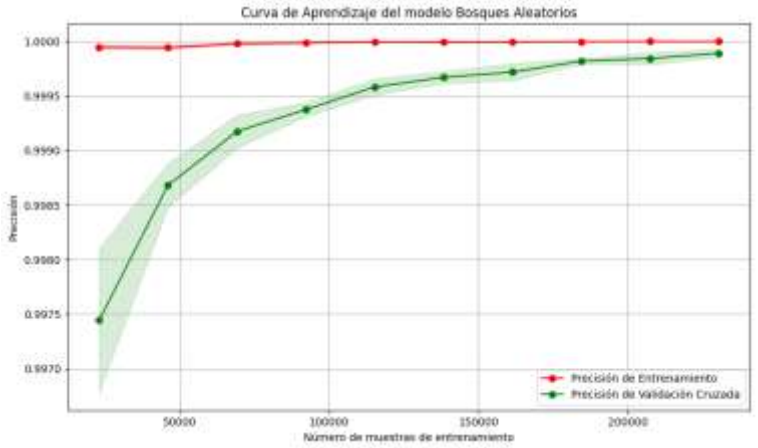


FIGURA 4.10 CURVA DE APRENDIZAJE DEL MODELO DE BOSQUES ALEATORIOS SIN RUIDO.

Con $23{,}063$ muestras, la exactitud de entrenamiento fue de 0.9999 ± 0.0000 y la de validación cruzada de 0.9974 ± 0.0007 al llegar a $230{,}630$ muestras, esta última también alcanzó 0.9999 ± 0.0000. Este comportamiento indica que el modelo se beneficia

claramente del incremento en el volumen de datos y que su desempeño converge hacia una región de alta estabilidad.

La diferencia entre las curvas de entrenamiento y validación cruzada es reducida, particularmente para los mayores tamaños de muestra, por lo que no se observan señales relevantes de sobreajuste en condiciones limpias. No obstante, una exactitud de entrenamiento prácticamente perfecta también sugiere una capacidad de ajuste muy elevada del ensamble, por lo que resulta necesario analizar su comportamiento frente a datos perturbados antes de emitir conclusiones sobre su robustez operacional.

La Figura 4.11 muestra la importancia relativa de las variables dentro del bosque aleatorio. A diferencia del árbol de decisión individual, el ensamble distribuye la importancia entre un mayor número variables, aunque las características derivadas de potencia y sensibilidad siguen siendo las más influyentes.

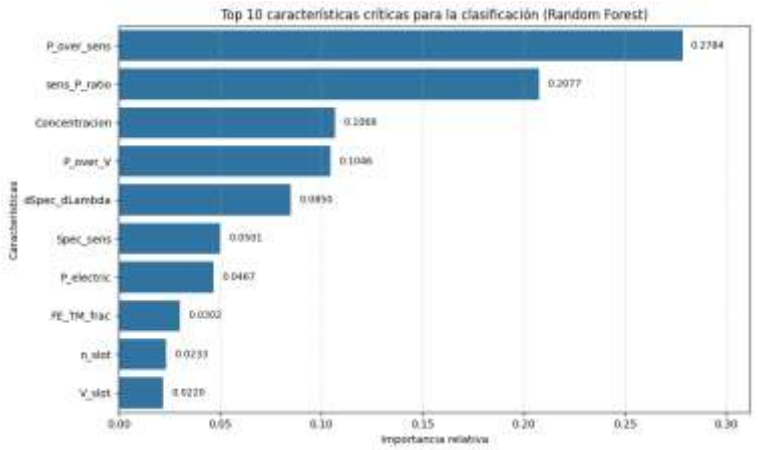


FIGURA 4.11 CARACTERÍSTICAS MÁS RELEVANTES PARA LA CLASIFICACIÓN MEDIANTE BOSQUES ALEATORIOS.

P_over_sens sigue encabezando la lista, con 0.2784, seguida de sens_P_ratio con 0.2077. Después vienen la concentración 0.1068, P_over_V 0.1047 y dSpec_dLambda 0.0850. En conjunto, estos resultados muestran que el modelo se apoya principalmente en relaciones derivadas entre potencia, sensibilidad espectral y concentración para discriminar entre clases. A diferencia del árbol de decisión, en este caso ninguna variable presenta una contribución nula. El ensamble aprovecha información distribuida a lo largo de todo el conjunto de características, incluyendo parámetros ópticos como β, λ, n_{slot}, V_{slot} y η_{factor}. Aunque su peso individual es menor, estas variables pueden aportar

información complementaria en diferentes árboles del bosque, fortaleciendo la capacidad global de discriminación del modelo.

La Tabla 4.7 y la Figura 4.12 muestran el desempeño por clase en el conjunto de prueba limpio. El modelo alcanzó una exactitud global de 100, y las cuatro clases registraron valores de precisión, recall y F1-score de 100%.

TABLA 4.7 REPORTE DE CLASIFICACIÓN DEL MODELO DE BOSQUES ALEATORIOS EN EL CONJUNTO DE PRUEBA LIMPIO.

Clase	Precisión	Recall	F1-score	muestras
F^-	100%	99%	100%	18018
H_3AsO_4	100%	100	100%	18018
KNO_3	100%	100%	100%	18018
$NaCl$	100%	100%	100%	18018
Precisión		100%	100%	72072
Promedio macro	100%	100%	100%	72072
Promedio ponderado	100%	100%	100%	72072

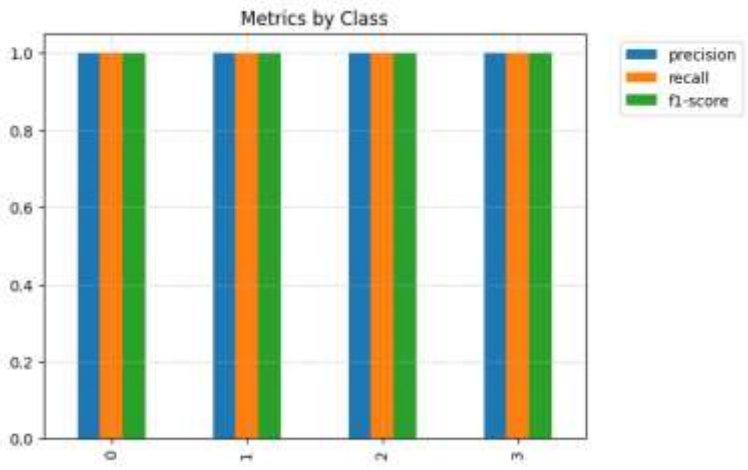


FIGURA 4.12 MÉTRICAS POR CLASE DEL CLASIFICADOR DE BOSQUES ALEATORIOS EN EL CONJUNTO DE PRUEBA LIMPIO.

La coincidencia de estas métricas en 1.00 indica que el bosque aleatorio clasificó correctamente prácticamente todas las muestras del conjunto limpio, superando el

desempeño observado previamente en el árbol de decisión individual bajo las mismas condiciones.

La matriz de confusión de la Figura 4.13 confirma este resultado, ya que las predicciones se concentran casi por completo sobre la diagonal principal.

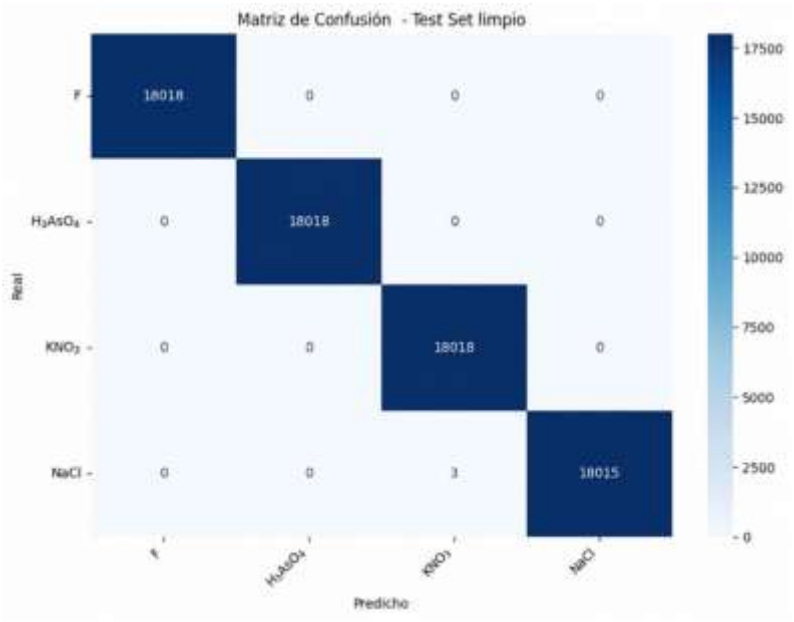


FIGURA 4.13 MATRIZ DE CONFUSIÓN DEL MODELO DE BOSQUES ALEATORIOS EN EL CONJUNTO DE PRUEBA LIMPIO.

Las clases F^-, H_3AsO_4 y KNO_3 tuvieron sus 18,018 muestras clasificadas correctamente en cada categoría, mientras que $NaCl$ registró $18{,}015$ aciertos, con solo tres muestras confundidas con KNO_3 el error en condiciones limpias es prácticamente nulo. Por tanto, el error en condiciones limpias puede considerarse prácticamente nulo.

Las curvas ROC presentadas en la Figura 4.14 refuerzan esta interpretación. Las cuatro clases alcanzan un AUC de 1.00, indicando una separación perfecta en ausencia de ruido.

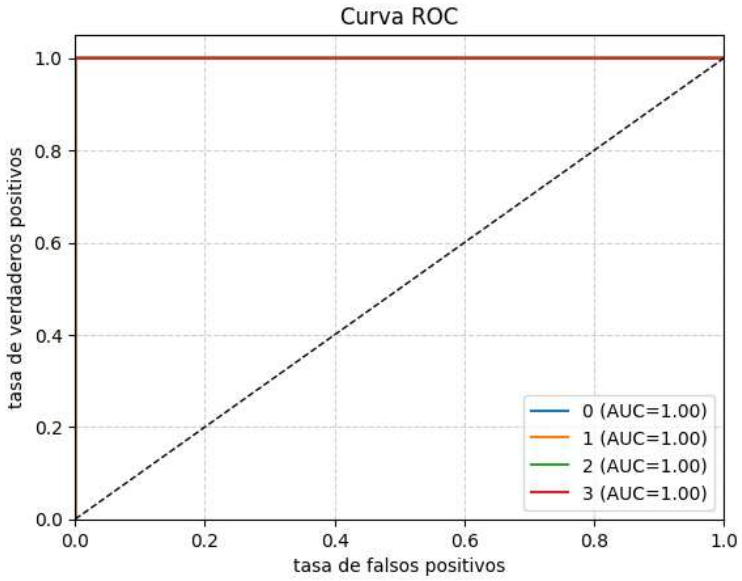


FIGURA 4.14 CURVA ROC MULTICLASE DEL CLASIFICADOR DE BOSQUES ALEATORIOS SIN RUIDO.

Este resultado confirma no solo la correcta asignación de etiquetas, sino también una capacidad discriminante sólida del modelo a lo largo de distintos umbrales de decisión, en concordancia con la matriz de confusión y las métricas por clase.

Posteriormente se evaluó la robustez del bosque aleatorio frente a ruido gaussiano en las variables de entrada. La Tabla 4.8 resume el comportamiento global del modelo bajo estas perturbaciones. Aunque el desempeño sin ruido fue prácticamente perfecto, la incorporación de ruido produjo una caída importante en la exactitud promedio.

TABLA 4.8 DEGRADACIÓN DEL DESEMPEÑO GLOBAL DEL MODELO DE BOSQUES ALEATORIOS BAJO DIFERENTES NIVELES DE RUIDO.

Nivel de ruido	Accuracy promedio	Desviación estándar	Accuracy retenida (%)
0.000	0.999958	0	100
0.025	0.465106	1.058×10^{-4}	46.51

Nivel de ruido	Accuracy promedio	Desviación estándar	Accuracy retenida (%)
0.050	0.441243	1.161×10^{-4}	44.12
0.075	0.427454	1.209×10^{-4}	42.74
0.100	0.419478	1.141×10^{-4}	41.94

La disminución más marcada ocurrió al pasar de datos limpios a 2.5% de ruido, donde la exactitud descendió de 99.99% a 46.51%. A partir de ese punto, la degradación continuó de manera más gradual hasta alcanzar 41.94% con 10.0% de ruido.

La Figura 4.15 muestra el comportamiento de la precisión por clase frente al incremento del ruido.

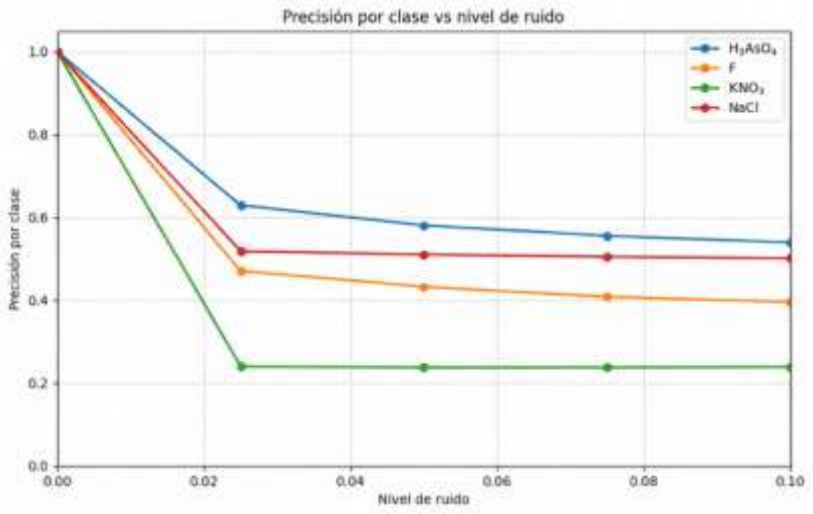


FIGURA 4.15 PRECISIÓN POR CLASE DEL MODELO DE BOSQUES ALEATORIOS FRENTE A DISTINTOS NIVELES DE RUIDO.

En condiciones limpias, las cuatro clases presentan valores cercanos al 100%, sin embargo, al introducir perturbaciones, el desempeño cae de manera abrupta. H_3AsO_4 es la que mejor conserva su desempeño, con precisión cercana a 63% en 2.5% de ruido y 54% en 10% de ruido. $NaCl$ se mantiene alrededor de 52% y baja a 50%. En cambio F^- y KNO_3, muestran una degradación más severa, F^- cae de 100% a cerca de 47% con 2.5% de ruido y a 0.39% con 10%. Finalmente KNO_3 llega hasta valores cercanos a 24% constituyéndose como la clase más afectada por las perturbaciones dentro del ensamble.

Las matrices de confusión bajo distintos niveles de ruido, en la Figura 4.16, muestran con a detalle cómo se redistribuyen los errores.

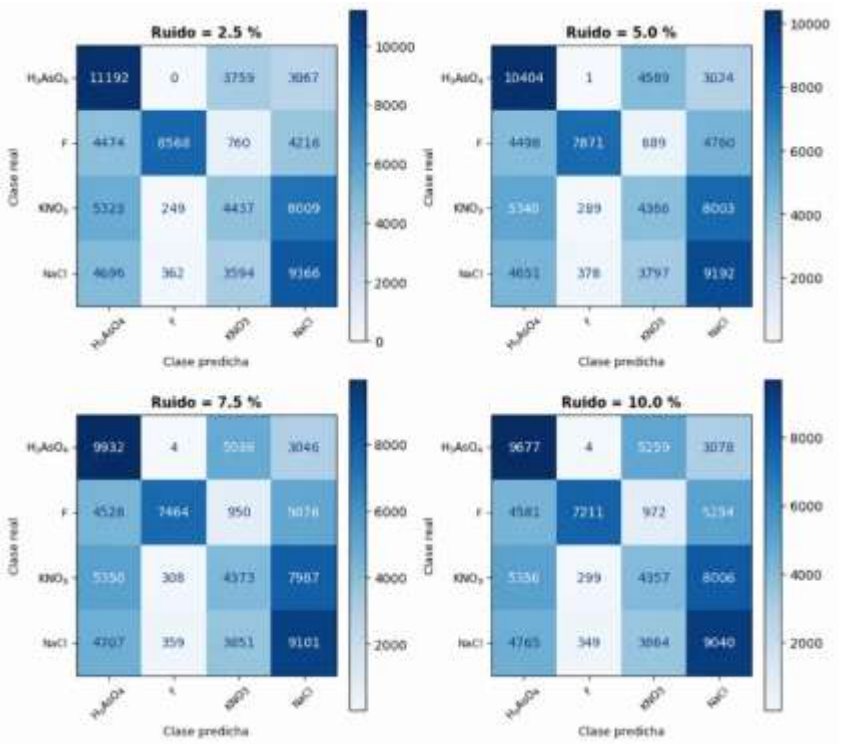


FIGURA 4.16 MATRICES DE CONFUSIÓN DEL MODELO DE BOSQUES ALEATORIOS BAJO NIVELES DE RUIDO DE $\mathbf{2.5}\%$, $\mathbf{5.0}\%$, $\mathbf{7.5}\%$ Y $\mathbf{10.0}\%$.

Con 2.5% de ruido, H_3AsO_4 conserva $11{,}192$ aciertos, aunque parte de sus muestras se va hacia KNO_3 y $NaCl$. En el caso de F^- se registran $8{,}568$ aciertos, con $4{,}474$ muestras clasificadas como H_3AsO_4 y $4{,}216$ como $NaCl$. KNO_3 es la más afectada ya que solo $4{,}437$ muestras quedan bien clasificadas, mientras $8{,}009$ se asignan a $NaCl$ y $5{,}323$ a H_3AsO_4. $NaCl$ acumula 9,366 aciertos, con errores importantes también hacia H_3AsO_4 y KNO_3.

El patrón se acentúa al aumentar el nivel de ruido. Con 10.0%, H_3AsO_4 baja a 9,677 aciertos, F^- a 7,211, KNO_3 a $4{,}357$ y $NaCl$ a $9{,}040$. KNO_3 sigue siendo la más golpeada, con buena parte de sus muestras desplazándose hacia $NaCl$ y H_3AsO_4 . Este comportamiento sugiere que las fronteras de decisión del bosque aleatorio pierden estabilidad principalmente en las regiones donde las características de KNO_3 presentan mayor proximidad con las de otras categorías.

El efecto del ruido también se refleja en las curvas ROC de la Figura 4.17 de un AUC de 1.00 en todas las clases bajo condiciones limpias, los valores disminuyen de forma considerable tan pronto como se introducen perturbaciones.

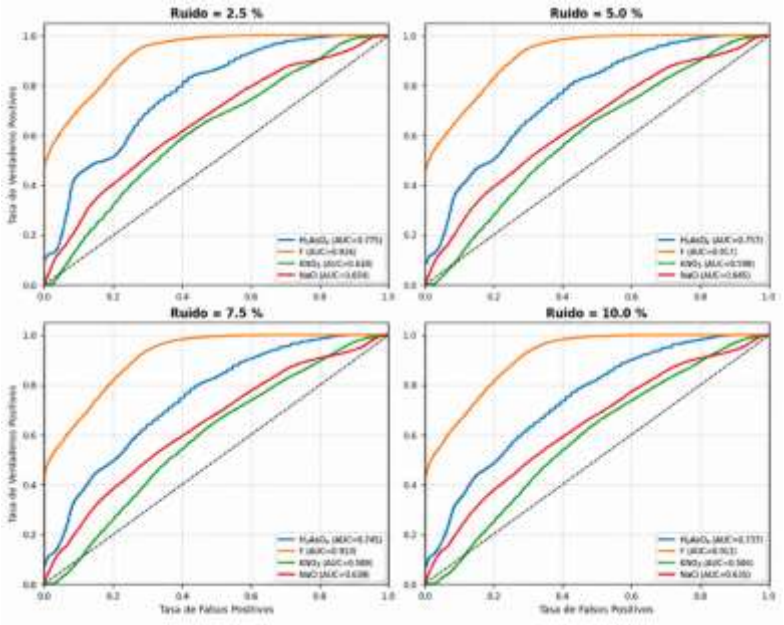


FIGURA 4.17 CURVAS ROC MULTICLASE DEL MODELO DE BOSQUES ALEATORIOS BAJO NIVELES DE RUIDO DE $\mathbf{2.5\%}$, $\mathbf{5.0\%}$, $\mathbf{7.5\%}$ Y $\mathbf{10.0\%}$.

Con 2.5% de ruido, F^- conserva el valor más alto, con un AUC de 0.926, seguido de H_3AsO_4 con 0.775, $NaCl$ con 0.654 y KNO_3 con 0.618. Al incrementar el ruido hasta 10.0%, estos valores descienden a 0.911, 0.737, 0.635 y 0.584, respectivamente. En consecuencia, aunque F^- mantiene una capacidad discriminante relativamente alta, KNO_3 es la clase que pierde mayor separabilidad conforme aumenta la perturbación en las variables de entrada.

Por último, la Figura 4.18 muestra la degradación operacional del modelo. En ausencia de ruido, el clasificador se ubica dentro de la zona de operación excelente, con una exactitud prácticamente igual al 100%. Sin embargo, desde el primer nivel de perturbación el desempeño cae por debajo de los umbrales de operación aceptable. Este resultado indica que, aunque el bosque aleatorio presenta un comportamiento sobresaliente en condiciones limpias, su estabilidad operacional se ve comprometida cuando las variables de entrada incorporan ruido gaussiano.

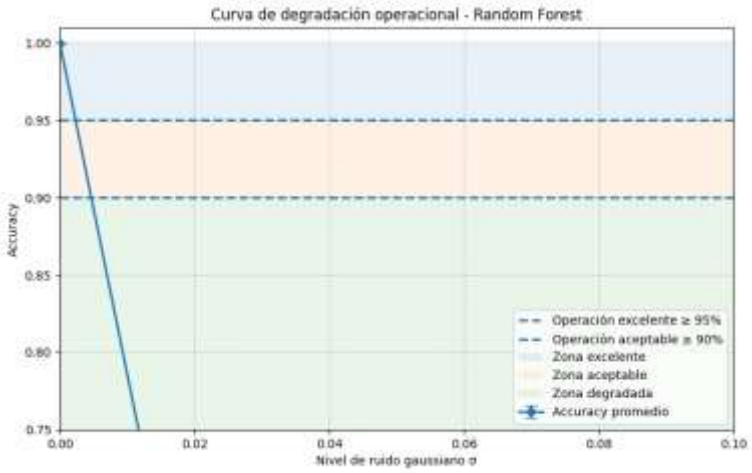


FIGURA 4.18 CURVA DE DEGRADACIÓN OPERACIONAL DEL MODELO DE BOSQUES ALEATORIOS FRENTE AL NIVEL DE RUIDO.

El modelo cae de forma abrupta de una operación excelente a una zona degradada apenas se introduce ruido gaussiano el bosque aleatorio rinde de exitosamente en datos limpios, pero es muy sensible a perturbaciones en las variables de entrada. Que la exactitud caiga de 99.99% a valores cercanos a 42% - 46% deja claro que el modelo depende en gran medida de la estabilidad de las características usadas para clasificar.

El bosque aleatorio representa una mejora clara respecto al árbol de decisión individual cuando se evalúa en condiciones limpias, tanto por su capacidad de clasificación como por el aprovechamiento más distribuido de las variables de entrada. Sin embargo, esta ventaja no se traduce en una robustez equivalente frente a perturbaciones. La caída abrupta del desempeño al introducir ruido revela que el modelo depende de relaciones muy precisas entre las características, por lo que pequeñas variaciones en las mediciones pueden alterar de forma considerable la asignación de clases. En este sentido, aunque el modelo es altamente eficaz bajo condiciones controladas, su sensibilidad al ruido limita su aplicabilidad directa en escenarios experimentales donde la variabilidad instrumental o ambiental forme parte del proceso de medición.

4.3 RESULTADOS XGBOOST

La búsqueda de los hiperparámetros óptimos para el modelo XGBoost por optimización Bayesiana arrojó la configuración que se muestra en la Tabla 4.9.

TABLA 4.9 HIPERPARÁMETROS ÓPTIMOS OBTENIDOS PARA EL CLASIFICADOR XGBOOST.

Hiperparámetros	Valores óptimos
Regularización L2	0.5624436754
Regularización L1	0.0420599542
Tasa de aprendizaje	0.0911168767
Profundidad máxima	5
Submuestreo de filas	0.5486672914
Submuestreo de columnas	0.7024464225
Peso mínimo por hijo	2
Número de estimadores	286

La profundidad máxima a 5 permite limitar la complejidad de cada árbol individual, y la regularización L1/L2 ayuda a controlar el sobreajuste, los parámetros de submuestreo de filas y columnas suman aleatoriedad al entrenamiento. En conjunto, esta combinación da un modelo con mucha capacidad de ajuste, pero con mecanismos de regularización.

La Tabla 4.10 muestra los resultados de la validación cruzada con cinco particiones. En todos los *folds*, la exactitud fue superior a 99.96%, con variaciones mínimas entre particiones.

TABLA 4.10 RESULTADOS DE VALIDACIÓN CRUZADA PARA EL CLASIFICADOR XGBOOST.

Fold	Precisión %
1	99.98
2	99.97
3	99.97

Fold	Precisión %
4	99.96
5	99.98
Promedio	99.97
Desviación estándar	0.0001
Precisión en conjunto de prueba limpio	100

Estos valores indican que, bajo condiciones limpias, el modelo mantiene un desempeño altamente estable sobre distintos subconjuntos del conjunto de entrenamiento. La baja variabilidad entre particiones sugiere que el clasificador no depende de una división particular de los datos para alcanzar un rendimiento elevado.

La Figura 4.19 presenta la curva de aprendizaje del modelo. La exactitud de entrenamiento se mantiene en 1.0000 para todos los tamaños de muestra evaluados, mientras que la exactitud de validación cruzada permanece por encima de 0.9994 desde el primer punto analizado.

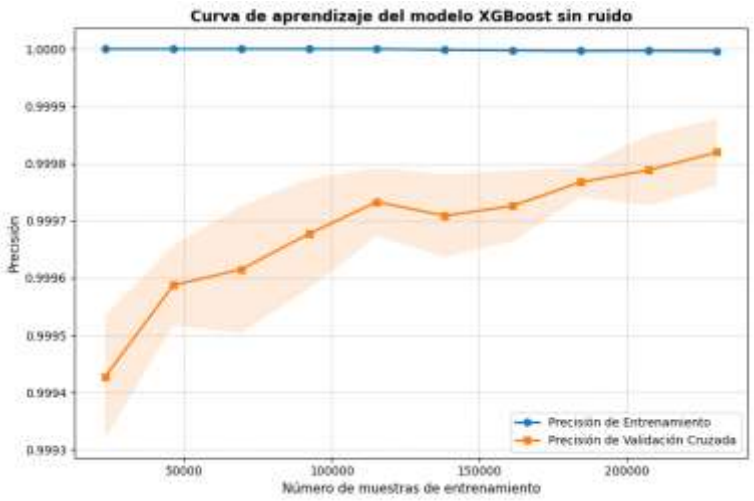


FIGURA 4.19 CURVA DE APRENDIZAJE DEL MODELO XGBOOST SIN RUIDO.

Con $23{,}063$ muestras, la exactitud de entrenamiento fue de 1.0000 y la exactitud de validación cruzada de 0.9995, Al incrementar el conjunto de entrenamiento hasta $230{,}630$ la exactitud de validación cruzada aumentó ligeramente hasta 0.9998. La proximidad entre ambas curvas indica un comportamiento estable en condiciones limpias, sin embargo, la exactitud de entrenamiento perfecta también evidencia una capacidad de ajuste muy elevada. Por ello, resulta necesario analizar el comportamiento del modelo bajo perturbaciones antes de establecer conclusiones sobre su robustez operacional.

La Figura 4.20 describe la importancia relativa de las variables para el modelo XGBoost que concentra gran parte de su capacidad discriminante en dos variables principales, sens_P_ratio y P_over_sens.

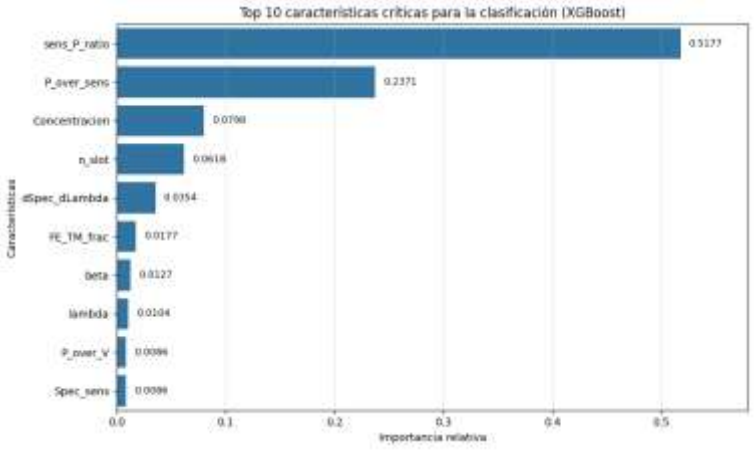


FIGURA 4.20 DIEZ CARACTERÍSTICAS MÁS RELEVANTES PARA LA CLASIFICACIÓN MEDIANTE XGBOOST.

Estos resultados muestran que XGBoost se apoya principalmente en relaciones derivadas entre sensibilidad espectral, potencia y concentración para separar las clases.

La tabla 4.11 y la Figura 4.21 muestran el desempeño por clase en el conjunto de prueba limpio con métricas perfectas para cada clase.

TABLA 4.11 REPORTE DE CLASIFICACIÓN DEL MODELO XGBOOST EN EL CONJUNTO DE PRUEBA LIMPIO.

Clase	Precisión %	Recall %	F1-score %	muestras
F^-	100	100	100	18018
H_3AsO_4	100	100	100	18018
KNO_3	100	100	100	18018

Clase	Precisión %	Recall %	F1-score %	muestras
$NaCl$	100	100	100	18018
Precisión		100	100	72072
Promedio macro	100	100	100	72072
Promedio ponderado	100	100	100	72072

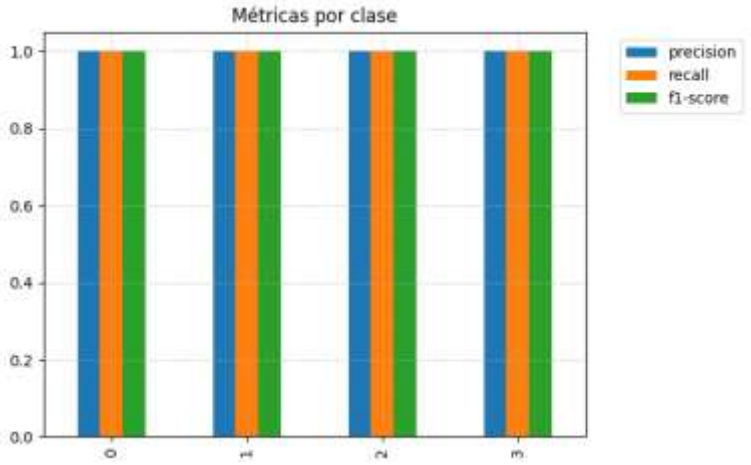


FIGURA 4.21 MÉTRICAS POR CLASE DEL CLASIFICADOR XGBOOST EN EL CONJUNTO DE PRUEBA LIMPIO.

La coincidencia de las métricas por clase en valores cercanos a 1.00 indica que el modelo clasifica correctamente prácticamente todas las muestras del conjunto limpio. Además, no se observa un sesgo evidente hacia alguna categoría específica, ya que el desempeño se mantiene equilibrado entre las cuatro clases evaluadas.

La matriz de confusión de la Figura 4.22 confirma este comportamiento. Las predicciones se concentran casi por completo sobre la diagonal principal, con un número mínimo de errores fuera de ella.

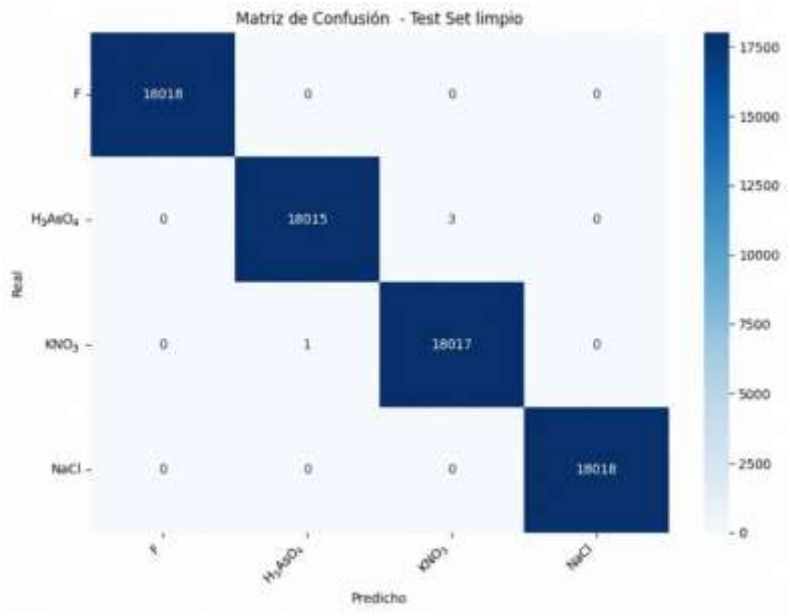


FIGURA 4.22 MATRIZ DE CONFUSIÓN DEL MODELO XGBOOST EN EL CONJUNTO DE PRUEBA LIMPIO.

Las clases F^- y $NaCl$ registran 18,018 aciertos, mientras que H_3AsO_4 registró 18,015 aciertos, con tres muestras confundidas con KNO_3, por su parte KNO_3 tuvo 18,017 aciertos y una sola muestra clasificada como H_3AsO_4. Las curvas ROC multiclase de la Figura 4.23 refuerzan esta interpretación. Las cuatro clases alcanzan un AUC de 1.00, lo que indica una separación altamente efectiva en ausencia de ruido.

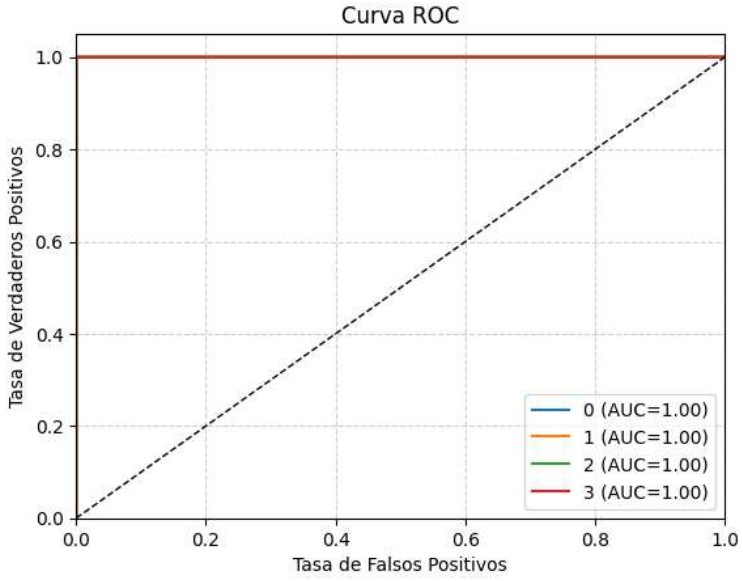


FIGURA 4.23 CURVA ROC MULTICLASE DEL CLASIFICADOR XGBOOST SIN RUIDO.

Un AUC de 1.00 en todas las categorías confirma que el modelo no solo asigna correctamente las etiquetas en el conjunto limpio, sino que también mantiene una alta capacidad discriminante a través de distintos umbrales de decisión. Este resultado es coherente con las métricas por clase y con la matriz de confusión previamente analizada.

El desempeño del modelo XGBoost también se examinó bajo perturbaciones gaussianas en las variables de entrada, con el propósito de determinar si su alta exactitud en condiciones limpias se traducía en estabilidad frente al ruido. Los resultados globales se presentan en la Tabla 4.12. Aunque XGBoost alcanza un desempeño prácticamente perfecto sin perturbaciones, la introducción de ruido produce una caída considerable en la exactitud promedio.

TABLA 4.12 DEGRADACIÓN DEL DESEMPEÑO GLOBAL DEL MODELO XGBOOST BAJO DIFERENTES NIVELES DE RUIDO.

Nivel de ruido	Accuracy promedio	Desviación estándar	Accuracy retenida (%)
0.000	0.999944	0	100
0.025	0.354547	2.41×10^{-3}	35.46
0.050	0.357935	2.31×10^{-3}	35.80
0.075	0.360072	2.10×10^{-3}	36.01
0.100	0.361242	2.13×10^{-3}	36.13

La disminución más abrupta ocurre al pasar de datos limpios a 2.5% de ruido. En este punto, la exactitud promedio desciende de 0.999944 a 0.354547, lo que corresponde a una retención de únicamente 35.46%. Para niveles superiores de ruido, la exactitud aumenta ligeramente hasta alcanzar 0.361242 con 10% de perturbación. Sin embargo, esta variación es reducida y no puede interpretarse como una recuperación real del modelo, sino como una redistribución de las predicciones bajo condiciones ruidosas.

La Figura 4.24 muestra la precisión por clase frente a los distintos niveles de ruido. En condiciones limpias, todas las clases parten de valores cercanos a 1.00, sin embargo, la precisión disminuye de forma abrupta al introducir perturbaciones.

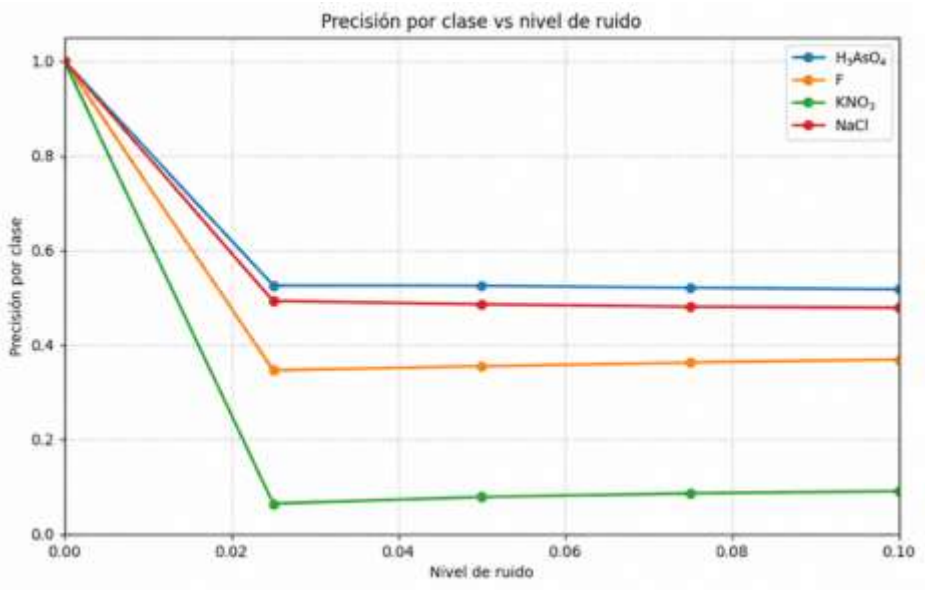


FIGURA 4.24 PRECISIÓN POR CLASE DEL MODELO XGBOOST FRENTE A DISTINTOS NIVELES DE RUIDO.

Las clases H_3AsO_4 y $NaCl$ rondan entre 0.52 y 0.49 respectivamente tras introducir ruido, la clase F^- cae con valores entre 0.35 y 0.37 en los niveles perturbados. La clase más afectada es KNO_3, cuya precisión se hunde hasta cerca de 0.06 con apenas un 2.5% de ruido y sigue por debajo de 0.10 incluso con 10.0% frente a ruido. Este comportamiento indica que el modelo pierde casi por completo la capacidad de asignar de forma confiable las muestras pertenecientes a esta clase bajo condiciones perturbadas.

Las matrices de confusión bajo ruido, presentadas en la Figura 4.25, permiten observar con mayor detalle la redistribución de los errores.

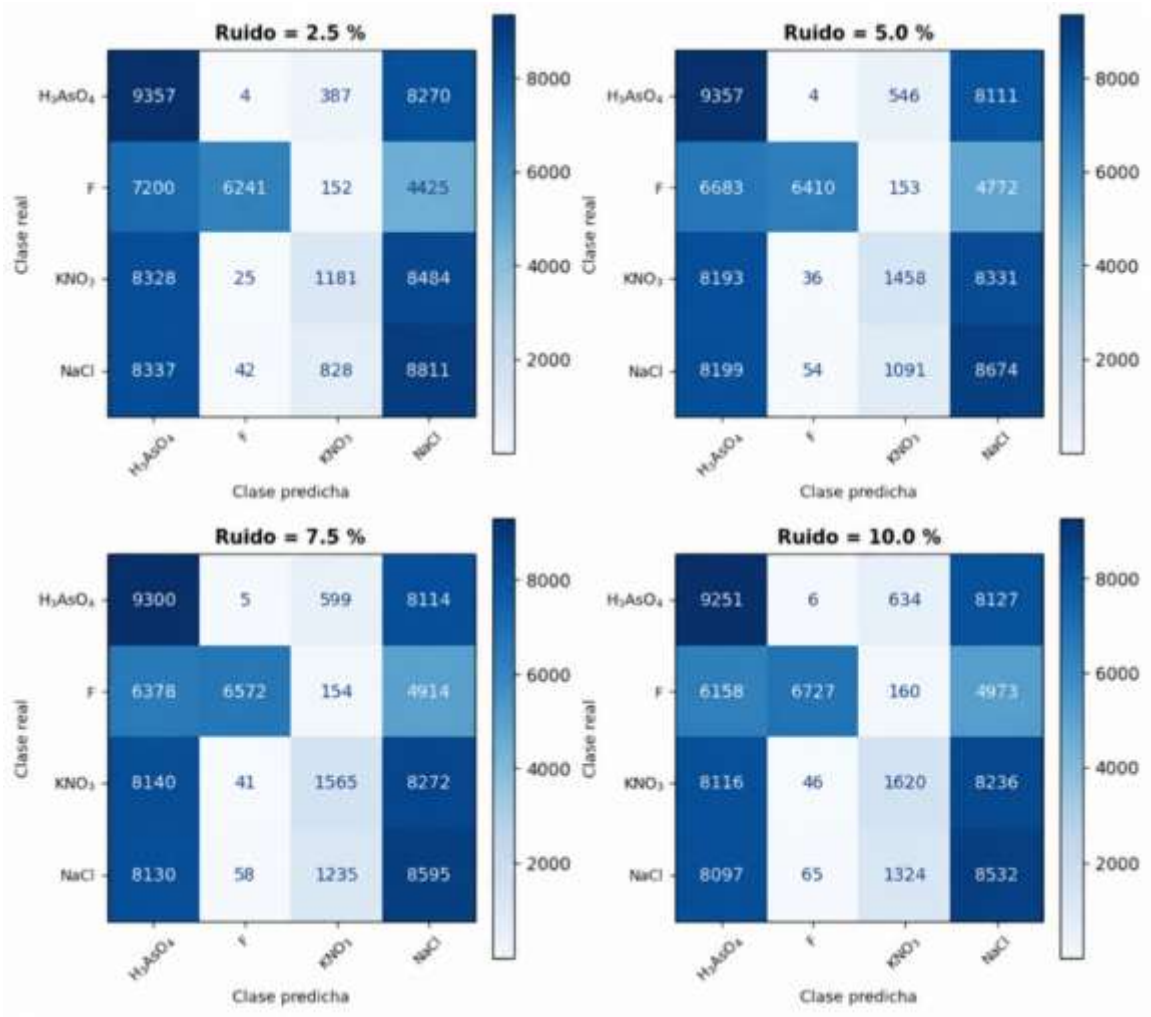


FIGURA 4.25 MATRICES DE CONFUSIÓN DEL MODELO XGBOOST BAJO NIVELES DE RUIDO DE $\mathbf{2.5}\%$, $\mathbf{5.0}\%$, $\mathbf{7.5}\%$ Y $\mathbf{10.0}\%$.

Con 2.5% de ruido, H_3AsO_4 conserva $9{,}357$ aciertos, aunque una proporción importante de sus muestras se desplaza hacia $NaCl$. La mayor degradación se observa en KNO_3 que conserva únicamente $1{,}181$ aciertos de $18{,}018$ muestras, mientras que la mayor parte de sus errores se reparte entre H_3AsO_4 y $NaCl$. $NaCl$, por su parte, acumula $8{,}811$ aciertos, aunque también presenta una confusión importante hacia H_3AsO_4.

Al incrementar el ruido, el patrón general se mantiene. Con 10.0% de perturbación, H_3AsO_4 registra $9{,}251$ aciertos, F^- alcanza $6{,}727$, KNO_3 llega a $1{,}620$ y $NaCl$ obtiene $8{,}532$. Aunque KNO_3 muestra un ligero aumento respecto al nivel de 2.5% de ruido, el número de aciertos continúa siendo muy bajo frente al tamaño total de la clase. Por tanto,

esta categoría constituye el punto más vulnerable del modelo frente a perturbaciones en las variables de entrada.

El efecto del ruido también se refleja en las curvas ROC de la Figura 4.26. A partir de un AUC de 1.00 para todas las clases en condiciones limpias, los valores disminuyen de forma considerable cuando se perturban los datos.

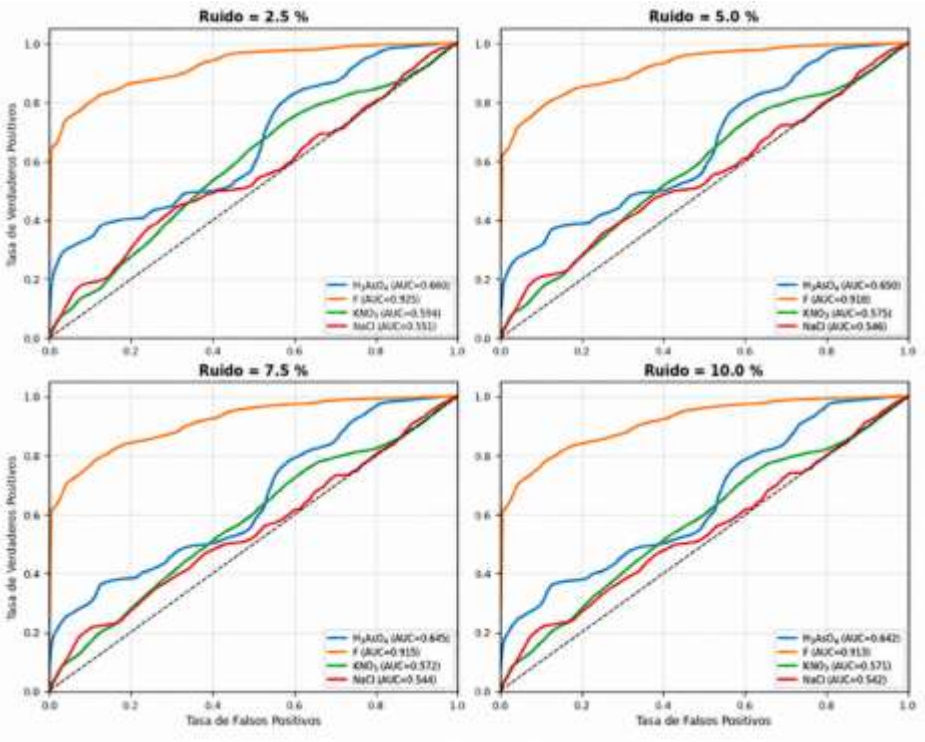


FIGURA 4.26 CURVAS ROC MULTICLASE DEL MODELO XGBOOST BAJO NIVELES DE RUIDO DE $\mathbf{2.5\%}$, $\mathbf{5.0\%}$, $\mathbf{7.5\%}$ Y $\mathbf{10.0\%}$.

Con 2.5% de ruido, F^- conserva el AUC más alto, con un valor de 0.925, seguida de H_3AsO_4, con 0.660. En contraste, KNO_3 y $NaCl$ presentan valores menores, de 0.584 y 0.551, respectivamente. Al aumentar el ruido hasta 10.0%, los AUC descienden a 0.913 para F^-, 0.642 para H_3AsO_4, 0.571 para KNO_3 y 0.542 para $NaCl$. Estos resultados indican que F^- conserva una capacidad discriminante relativamente alta bajo perturbación, mientras que KNO_3 y $NaCl$ se aproximan a un régimen de baja separabilidad.

Por último, la Figura 4.27 presenta la curva de degradación operacional del modelo. En ausencia de ruido, XGBoost se ubica en la zona de operación excelente, con una

exactitud prácticamente igual a 1.00. No obstante, desde el primer nivel de perturbación el desempeño cae abruptamente por debajo de los umbrales aceptables.

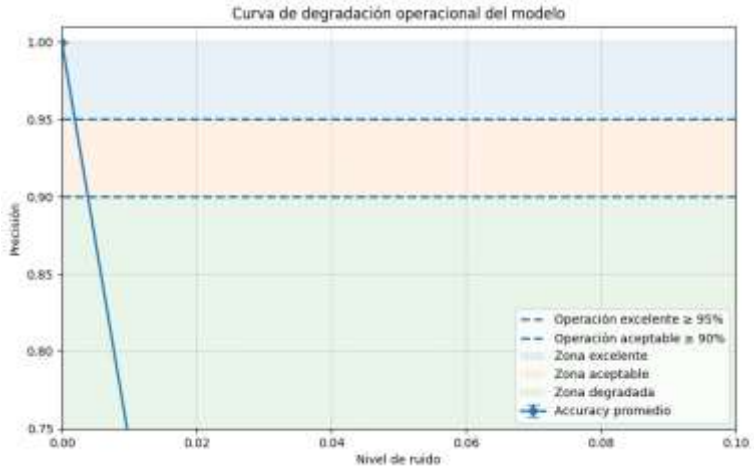


FIGURA 4.27 CURVA DE DEGRADACIÓN OPERACIONAL DEL MODELO XGBOOST FRENTE AL NIVEL DE RUIDO.

La transición directa desde una operación excelente hacia una zona degradada evidencia una pérdida sustancial de estabilidad operacional. La exactitud permanece entre 0.35 y 0.36 para todos los niveles de ruido evaluados, confirmando que el modelo no logra conservar su desempeño cuando las variables de entrada presentan perturbaciones.

En términos generales, XGBoost muestra una capacidad de clasificación sobresaliente en condiciones limpias, con métricas prácticamente perfectas y una separación efectiva entre las cuatro clases. Sin embargo, esta capacidad no se sostiene frente a ruido en las variables de entrada. La fuerte dependencia del modelo respecto a características derivadas específicas, principalmente sens_P_ratio y P_over_sens, parece favorecer un ajuste altamente preciso en datos limpios, pero limitado ante variaciones en las mediciones. Por ello, aunque XGBoost constituye el modelo con mejor desempeño nominal en ausencia de perturbaciones, su baja robustez frente al ruido restringe su aplicabilidad directa en escenarios experimentales donde exista incertidumbre instrumental o variabilidad en los datos ópticos.

4.4 RESULTADOS PERCEPTRON MULTICAPA

Para definir la configuración optima del modelo, se realizó una búsqueda de hiperparámetros por validación cruzada, evaluando distintos tamaños de lote, $batch\ size$. Los resultados se presentan en la Tabla 4.13.

TABLA 4.13 RESULTADOS DE LA BÚSQUEDA DEL TAMAÑO DE LOTE PARA EL MODELO MLP.

Batch size	Precisión promedio en K-fold %
32	91.64
64	94.78
128	93.31

Como se observa en la tabla anterior el $batch\ size$ de 64 obtuvo el mejor desempeño promedio durante la validación cruzada, con una exactitud de 94.78%, por esta razón, dicho valor se seleccionó como configuración final del entrenamiento.

La Tabla 4.14 resume la arquitectura final del perceptrón multicapa. El modelo está constituido por seis capas densas de 256 neuronas, una capa con la función de activación y una capa de salida de cuatro neuronas, correspondiente a las cuatro clases analizadas.

TABLA 4.14 ARQUITECTURA GENERAL DEL MODELO PERCEPTRÓN MULTICAPA.

Bloque del modelo	Descripción	Dimensión de salida	Parámetros
Entrada → Capa oculta 1	Capa densa completamente conectada	256 neuronas	3,584
Capa oculta 2	Capa densa completamente conectada	256 neuronas	65,792
Capa oculta 3	Capa densa completamente conectada	256 neuronas	65,792
Capa oculta 4	Capa densa completamente conectada	256 neuronas	65,792

Bloque del modelo	Descripción	Dimensión de salida	Parámetros
Capa oculta 5	Capa densa completamente conectada	256 neuronas	65,792
Capa oculta 6	Capa densa completamente conectada	256 neuronas	65,792
Función de activación	Activación no lineal aplicada a la última capa oculta	256 neuronas	0
Capa de salida	Capa densa para clasificación multiclase	4 neuronas	1,028

La Tabla 4.15 resume el número de parámetros del modelo. La arquitectura cuenta con $333{,}572$ parámetros entrenables, mientras que el total asciende a $1{,}000{,}718$ al considerar también los parámetros asociados al optimizador.

TABLA 4.15 RESUMEN DE PARÁMETROS DEL MODELO MLP.

Tipo de parámetro	Número de parámetros	Tamaño aproximado MB
Parámetros totales	1,000,718	3.82
Parámetros entrenables	333,572	1.27
Parámetros no entrenables	0	0.00
Parámetros del optimizador	667,146	2.54

El número de capas ocultas y de neuronas por capa proporciona al MLP una capacidad de representación considerable. Con esta estructura, el modelo puede aprender patrones complejos del conjunto de datos sin requerir una arquitectura convolucional o recurrente, ya que las variables de entrada corresponden a descriptores numéricos derivados del modelado óptico y no a datos espaciales o secuenciales. En este sentido, el MLP resulta adecuado para capturar patrones no lineales en un espacio de características tabular como en el que se está trabajando.

La Figura 4.28 muestra la evolución de la función de pérdida durante el entrenamiento. La pérdida disminuye de manera rápida durante las primeras épocas y posteriormente se

estabiliza de forma progresiva. La mejor época se alcanzó en la época 26, con una pérdida mínima de 0.0141.

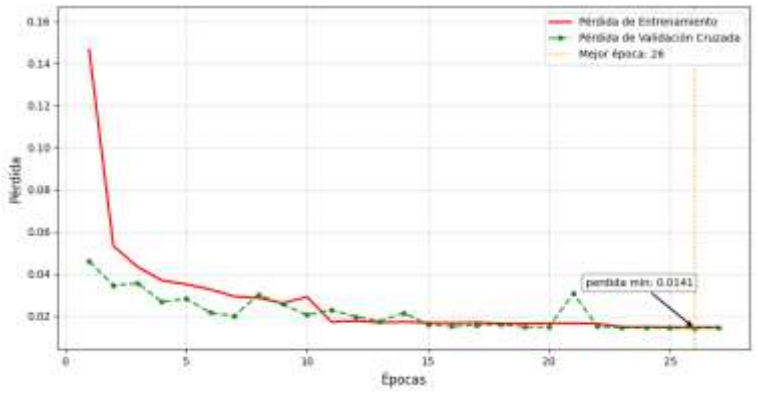


FIGURA 4.28 CURVA DE PÉRDIDA DEL MODELO MLP DURANTE EL ENTRENAMIENTO.

La pérdida de entrenamiento inició alrededor de 0.146 y descendió por debajo de 0.04 durante las primeras épocas. Posteriormente, tanto la pérdida de entrenamiento como la de validación se estabilizaron en un intervalo cercano a 0.014 y 0.016. La separación final entre ambas curvas fue de apenas 0.0004, lo que indica una convergencia estable y sin una brecha relevante entre el ajuste al conjunto de entrenamiento y el desempeño sobre los datos de validación.

La Figura 4.29 presenta la evolución de la exactitud durante el entrenamiento. El desempeño aumenta rápidamente desde las primeras épocas y se mantiene cercano a 1.0000 hacia el final del proceso. La exactitud de validación conserva un comportamiento similar, lo que confirma la estabilidad del entrenamiento.

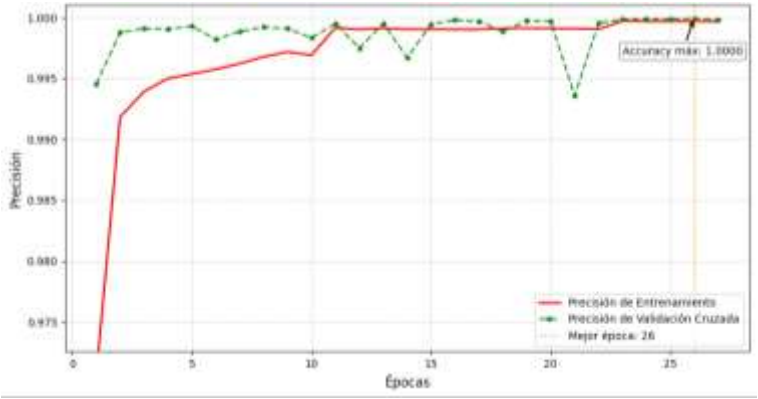


FIGURA 4.29 CURVA DE EXACTITUD DEL MODELO MLP DURANTE EL ENTRENAMIENTO.

El modelo alcanzó una exactitud máxima de 1.0000, con la mejor época nuevamente ubicada en la época 26. La separación final entre la exactitud de entrenamiento y la de validación fue de solo 0.0001, lo que evidencia un ajuste estable y sin muestras de sobreajuste en condiciones limpias.

La matriz de confusión del conjunto de validación limpio, presentada en la Figura 4.30, concentra las predicciones casi por completo sobre la diagonal principal.

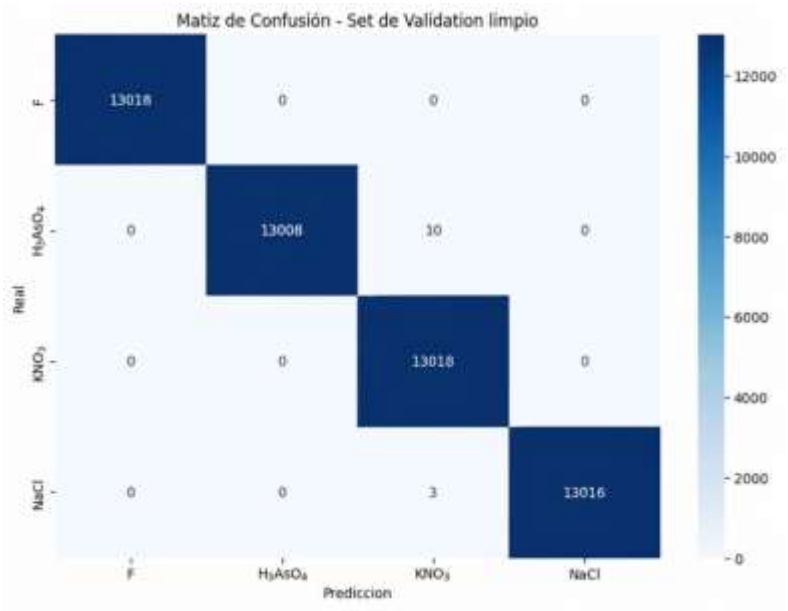


FIGURA 4.30 MATRIZ DE CONFUSIÓN DEL MODELO MLP EN EL CONJUNTO DE VALIDACIÓN LIMPIO.

En el conjunto de validación, la clase F^- tuvo sus $13{,}018$ muestras correctamente clasificadas. H_3AsO_4 registró $13{,}008$ aciertos, con solo 10 muestras clasificadas como KNO_3. La clase KNO_3 alcanzó $13{,}018$ aciertos, mientras que $NaCl$ obtuvo $13{,}016$ aciertos y únicamente tres muestras fueron asignadas erróneamente a KNO_3.

El reporte de clasificación presentado en la Tabla 4.20 y en la Figura 4.31 muestra valores de precisión, *recall* y F1-score de 1.00 para las cuatro clases.

TABLA 4.20 REPORTE DE CLASIFICACIÓN DEL MODELO MLP EN EL CONJUNTO DE PRUEBA LIMPIO.

Clase	Precisión %	Recall %	F1-score %	muestras
F^-	100	100	100	13514
H_3AsO_4	100	100	100	13514

Clase	Precisión %	Recall %	F1-score %	muestras
KNO_3	100	100	100	13514
$NaCl$	100	100	100	13514
Precisión		100	100	54054
Promedio macro	100	100	100	54054
Promedio ponderado	100	100	100	54054

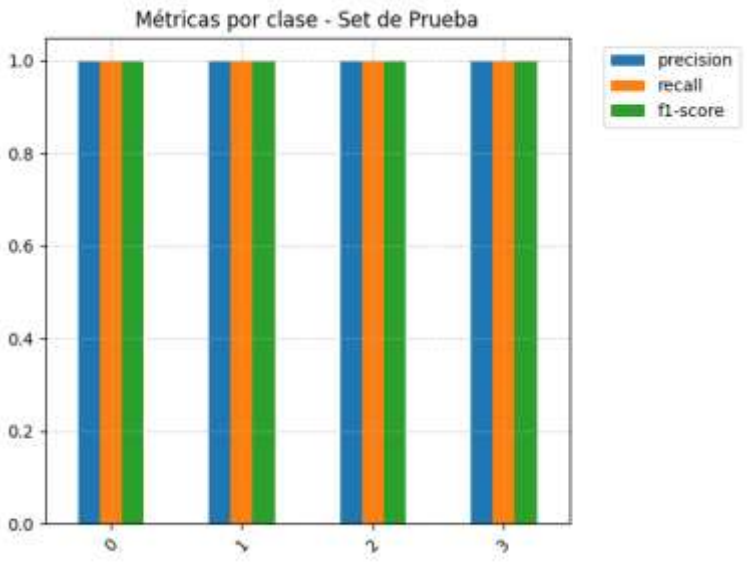


FIGURA 4.31 MÉTRICAS POR CLASE DEL MODELO MLP EN EL CONJUNTO DE PRUEBA LIMPIO.

La coincidencia entre precisión, *recall* y F1-score en todas las clases indica que el MLP clasificó correctamente prácticamente todas las muestras del conjunto de prueba limpio. Además, el soporte balanceado entre clases permite interpretar estos resultados sin la influencia de un desbalance significativo en la distribución de muestras.

La Figura 4.32 presenta la importancia relativa de las características empleadas por el modelo.

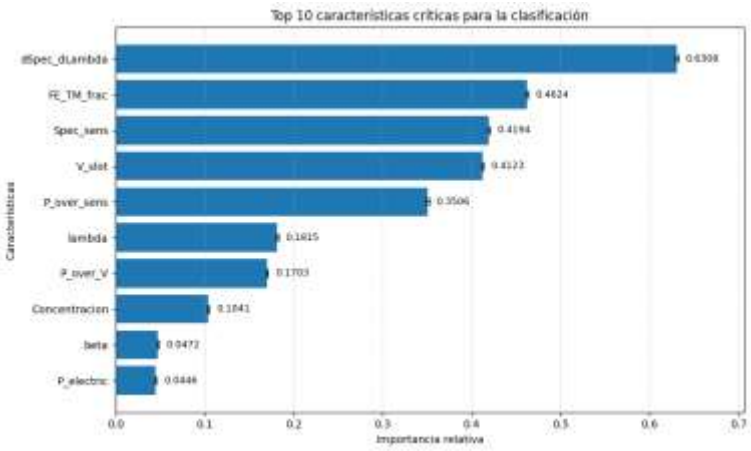


FIGURA 4.32 DIEZ CARACTERÍSTICAS MÁS RELEVANTES PARA LA CLASIFICACIÓN MEDIANTE EL MODELO MLP.

A diferencia de los modelos basados en árboles, donde la importancia se concentró principalmente en P_over_sens y sens_P_ratio principalmente, el MLP asigna mayor relevancia a variables asociadas con la sensibilidad espectral y la distribución modal. La variable dSpec_dLambda encabeza la lista con una importancia media de 0.6308, seguida de FE_*TM*_frac con 0.4624 y S_{spec} con 0.4194. Este comportamiento indica que la red aprovecha de manera más marcada la información óptica relacionada con la variación espectral, el confinamiento modal y la sensibilidad del sistema.

La curva ROC del conjunto de prueba limpio se muestra en la Figura 4.33.

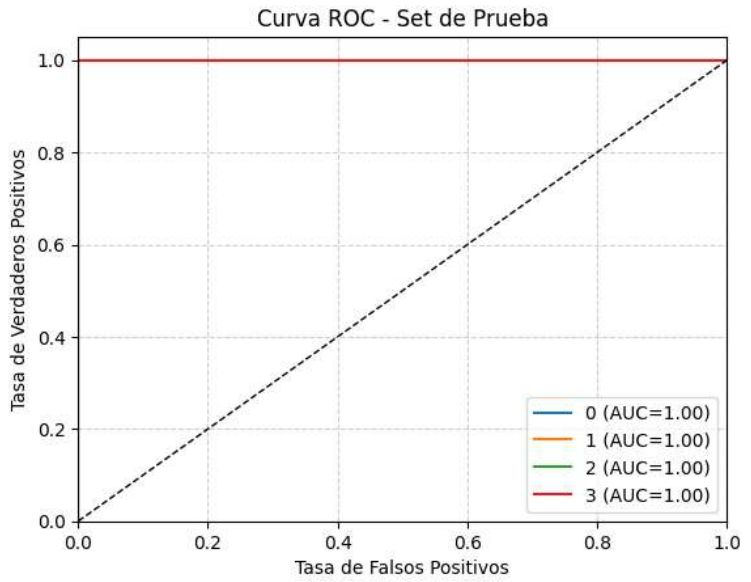


FIGURA 4.33 CURVA ROC MULTICLASE DEL MODELO MLP EN EL CONJUNTO DE PRUEBA LIMPIO.

El AUC de 1.00 confirma una capacidad discriminante prácticamente perfecta en condiciones limpias. Este resultado es coherente con el reporte de clasificación y con la matriz de confusión, ya que el modelo no solo asigna correctamente las etiquetas, sino que también mantiene una separación clara entre las clases a lo largo de distintos umbrales de decisión.

La Figura 4.34 presenta la curva de aprendizaje del modelo en función del número de muestras de entrenamiento.

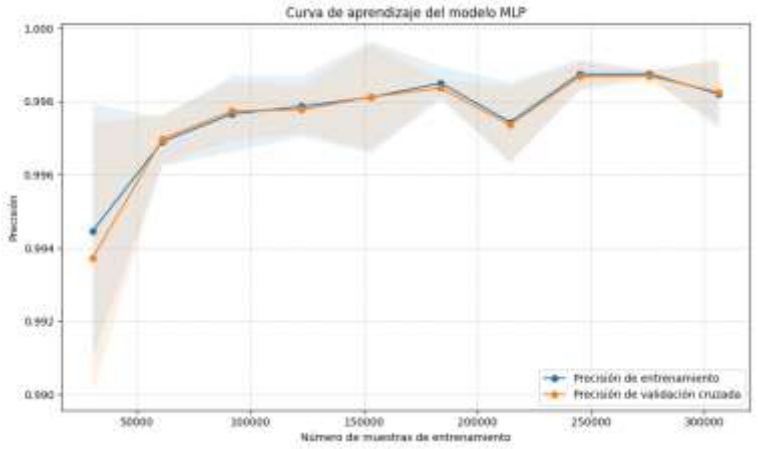


FIGURA 4.34 CURVA DE APRENDIZAJE DEL MODELO MLP EN FUNCIÓN DEL NÚMERO DE MUESTRAS.

Las curvas de entrenamiento y validación cruzada se mantienen próximas en todos los tamaños de muestra. Incluso con $30{,}630$ muestras, ambas métricas superan 0.993, y conforme aumenta el número de datos se estabilizan alrededor de 0.998, con desviaciones estándar reducidas. Este comportamiento indica que la arquitectura generaliza adecuadamente en condiciones limpias y no presenta señales relevantes de sobreajuste.

Siguiendo el mismo criterio de evaluación aplicado a los modelos anteriores, se analizó la respuesta del MLP ante perturbaciones gaussianas en los descriptores ópticos, con el propósito de determinar si su elevada capacidad de clasificación en condiciones limpias se mantenía bajo escenarios menos idealizados. Los resultados globales se presentan en la Tabla 4.23.

Nivel de ruido	Accuracy promedio	Desviación estándar	Accuracy retenida (%)
0.000	0.999741	0	100.00
0.025	0.969482	5.31×10^{-4}	96.97
0.050	0.914178	1.259×10^{-3}	91.44
0.075	0.859857	1.494×10^{-4}	86.01
0.100	0.810971	1.752×10^{-3}	81.12

Las matrices de confusión bajo distintos niveles de ruido se muestran en la Figura 4.35.

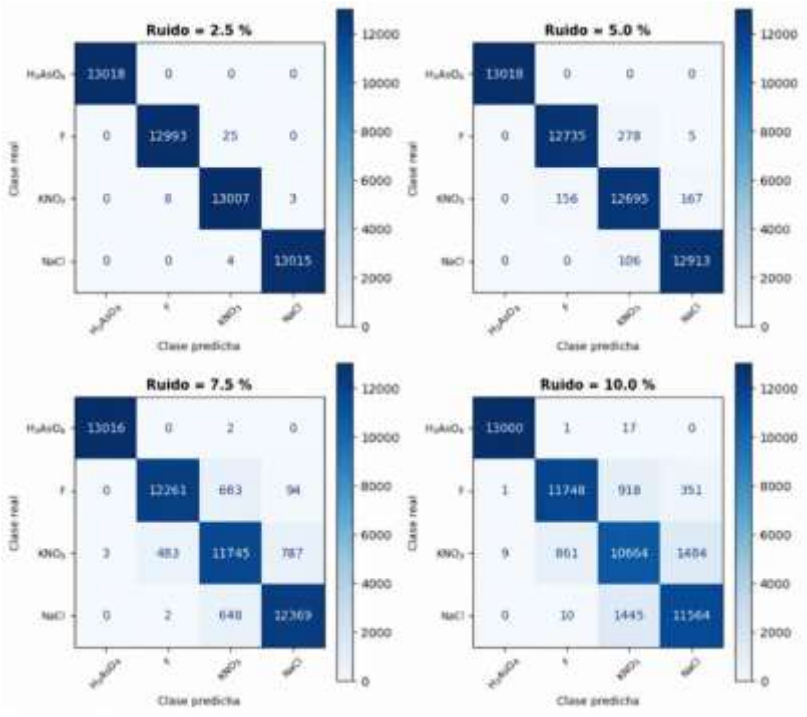


FIGURA 4.35 MATRICES DE CONFUSIÓN DEL MODELO MLP BAJO NIVELES DE RUIDO DE 2.5%, 5.0%, 7.5% Y 10.0%.

Con 2.5% de ruido, el modelo conserva un número elevado de aciertos en todas las clases. Los errores son mínimos y se distribuyen principalmente entre F^-, KNO_3 y $NaCl$, mientras que H_3AsO_4 permanece prácticamente sin afectación. Al aumentar el ruido a 5.0%, los errores comienzan a incrementarse, especialmente para F^-, KNO_3 y $NaCl$, aunque el modelo aún mantiene una clasificación mayoritariamente correcta.

A partir de 7.5% de ruido, la degradación se vuelve más visible. Las confusiones se concentran principalmente entre KNO_3 y $NaCl$, y en menor medida entre F^- y KNO_3. Con 10.0% de ruido, H_3AsO_4 continúa siendo la clase más estable, mientras que KNO_3 presenta la mayor pérdida de aciertos. La confusión principal ocurre entre KNO_3 y $NaCl$, con una redistribución prácticamente simétrica entre ambas clases. Esto indica que la frontera de decisión entre estos dos compuestos es la más sensible a las perturbaciones de entrada.

El comportamiento discriminante del modelo bajo ruido también se evaluó mediante curvas ROC, presentadas en la Figura 4.36.

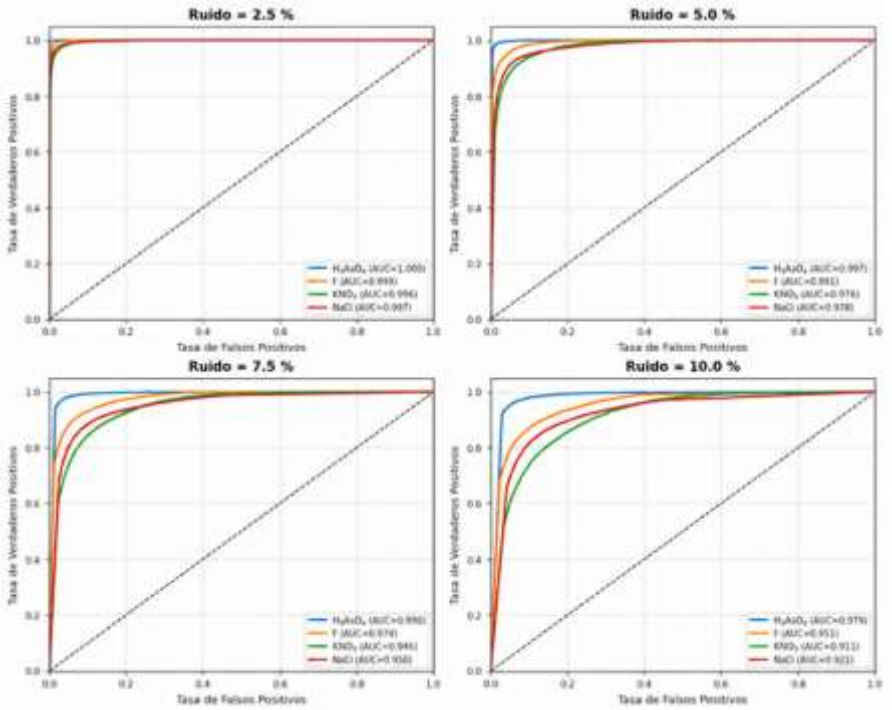


FIGURA 4.36 CURVAS ROC MULTICLASE DEL MODELO MLP BAJO NIVELES DE RUIDO DE 2.5%, 5.0%, 7.5% Y 10.0%.

Con 2.5% de ruido, el modelo mantiene una capacidad discriminante prácticamente perfecta para todas las clases. Este resultado confirma que, bajo perturbaciones leves, el MLP conserva una separación elevada entre las categorías, en concordancia con las métricas por clase y las matrices de confusión obtenidas para este nivel de ruido.

Al incrementar el ruido a 5.0%, los valores de AUC disminuyen de forma moderada, aunque permanecen altos, 0.997 para H_3AsO_4, 0.991 para F^-, 0.976 para KNO_3 y 0.978

para $NaCl$. Esto indica que el modelo aún conserva una capacidad de discriminación sólida, aunque comienzan a observarse diferencias entre clases.

Para niveles de ruido de 7.5% y 10.0%, la pérdida de separabilidad se vuelve más evidente. Con 7.5% de ruido, los valores de AUC fueron 0.990 para H_3AsO_4, 0.974 para F^-, 0.946 para KNO_3 y 0.950 para $NaCl$. Con 10.0 de ruido, estos valores disminuyeron a 0.979, 0.951, 0.911 y 0.921, respectivamente. A pesar de esta degradación progresiva, todos los valores se mantienen por encima de 0.90 incluso en el nivel máximo de ruido evaluado.

Estos resultados indican que el MLP conserva una capacidad discriminante alta frente a perturbaciones en los datos de entrada.

La Figura 4.37 muestra la precisión por clase frente al incremento del ruido.

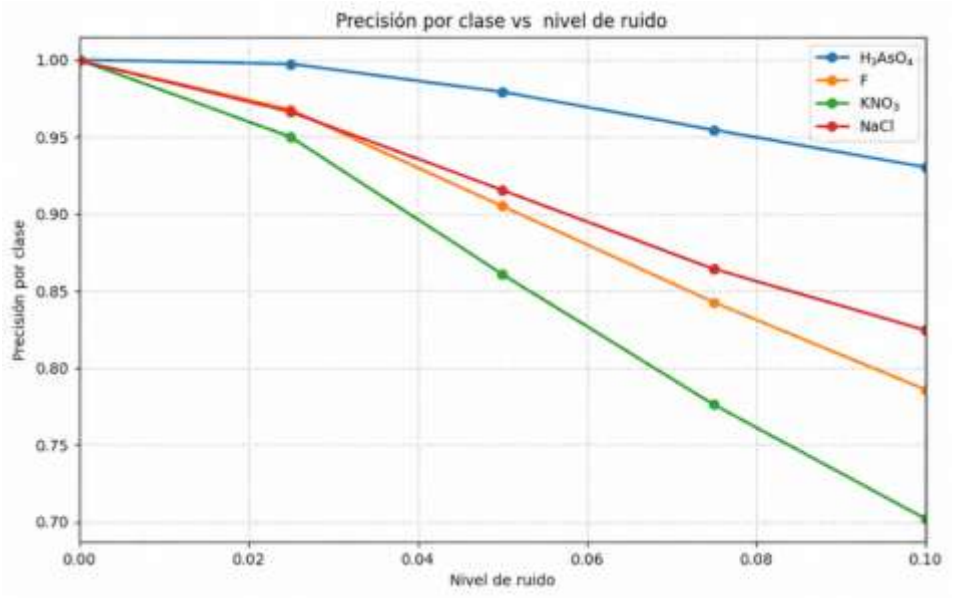


FIGURA 4.37 PRECISIÓN POR CLASE DEL MODELO MLP FRENTE A DISTINTOS NIVELES DE RUIDO.

El H_3AsO_4 es la clase más estable, manteniéndose por encima de 0.93 incluso con 10.0% de ruido. $NaCl$ presenta una degradación moderada, con valores que descienden desde niveles cercanos a 1.00 hasta aproximadamente 0.82. La clase F^- muestra una mayor sensibilidad, alcanzando valores cercanos a 0.79 en el nivel máximo de perturbación. KNO_3 es la clase más afectada, con una precisión cercana a 0.70 bajo 10.0% de ruido.

Este resultado coincide con las matrices de confusión, donde KNO_3 concentra la mayor proporción de errores en condiciones de perturbación alta.

Finalmente, en la Figura 4.38 se presenta la degradación operacional del modelo en función del nivel de ruido.

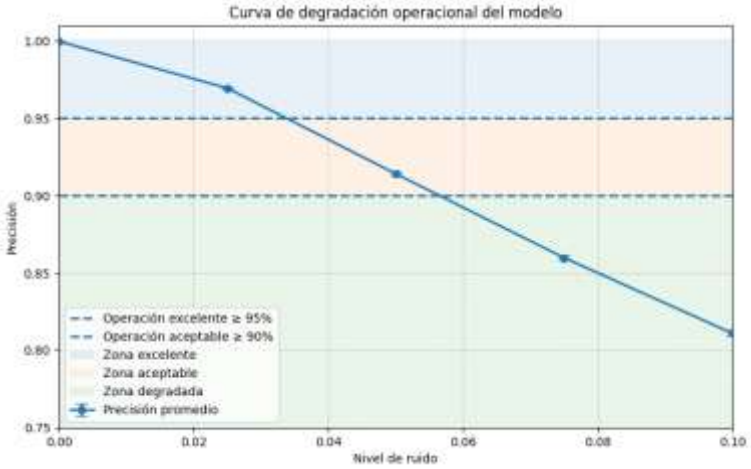


FIGURA 4.38 CURVA DE DEGRADACIÓN OPERACIONAL DEL MODELO MLP FRENTE AL NIVEL DE RUIDO.

El MLP se mantiene en zona de operación excelente en ausencia de ruido y con 2.5% de perturbación. Con 5.0% de ruido permanece dentro de una zona aceptable, mientras que con 7.5% y 10.0% entra en una zona degradada. Aun así, la disminución del desempeño es gradual, en contraste con los modelos basados en árboles, cuyo desempeño se redujo de manera más abrupta desde el primer nivel de perturbación. La exactitud retenida por encima del 90% hasta 5.0% de ruido indica una mayor tolerancia del MLP frente a variaciones en las variables de entrada.

En términos generales, el modelo MLP representa la alternativa más sólida entre los modelos evaluados, no solo por su desempeño prácticamente perfecto en condiciones limpias, sino por la forma gradual en que responde a la incorporación de ruido. A diferencia de los clasificadores basados en árboles, cuya degradación operacional fue más abrupta, la red mantuvo una capacidad discriminante alta incluso bajo perturbaciones moderadas en los descriptores ópticos. Si bien KNO_3 continuó siendo la clase más sensible, la pérdida de desempeño no comprometió de manera inmediata la estabilidad global del modelo. Esto sugiere que la arquitectura neuronal captura relaciones más robustas entre las variables ópticas y las clases de contaminantes, por lo

que resulta la opción más adecuada para escenarios donde las mediciones puedan estar sujetas a variabilidad experimental o incertidumbre instrumental.

5. CONCLUSIONES Y TRABAJO A FUTURO

Los resultados de esta investigación demuestran que los descriptores ópticos derivados del modelado numérico de una guía de onda tipo *slot* poseen capacidad discriminativa suficiente para identificar contaminantes presentes en agua mediante técnicas de aprendizaje automático. La metodología desarrollada permitió vincular el índice de refracción de soluciones acuosas con la respuesta modal de la guía de onda y, a partir de ello, generar parámetros ópticos y optoelectrónicos con significado físico para la clasificación de F^{-}, H_3AsO_4, KNO_3 y $NaCl$.

El análisis exploratorio confirmó que las variables obtenidas presentan distribuciones coherentes con el comportamiento óptico del sistema, así como correlaciones relevantes entre los parámetros electromagnéticos. Además, el balance entre clases permitió evaluar los modelos sin sesgos asociados a una distribución desigual de muestras, favoreciendo una comparación objetiva entre los distintos algoritmos.

En condiciones limpias, los modelos evaluados mostraron una alta capacidad de separación entre clases. Los modelos de ensamble y el perceptrón multicapa alcanzaron desempeños prácticamente perfectos, con métricas por clase cercanas o iguales a 1.00 y valores de AUC de 1.00. Esto confirma que, en ausencia de perturbaciones, los descriptores ópticos propuestos contienen información suficiente para diferenciar los contaminantes analizados.

El árbol de decisión funcionó como una línea base interpretable, aunque su desempeño fue menor que el de los modelos de ensamble y el MLP. Su principal limitación se observó en la separación entre KNO_3 y $NaCl$, donde se concentró la mayor parte de los errores. Esta tendencia sugiere que ambas clases ocupan regiones próximas en el espacio de características, lo que dificulta su discriminación mediante fronteras de decisión discretas. El análisis de importancia de características evidenció diferencias entre familias de modelos. En los clasificadores basados en árboles, la decisión dependió principalmente de relaciones derivadas entre potencia, sensibilidad y concentración, como P_over_sens

y sens_P_ratio. En cambio, el perceptrón multicapa asignó mayor relevancia a variables asociadas con la respuesta espectral, el confinamiento modal y el volumen efectivo del *slot*, como dSpec_dLambda, η, S_{sens}, V_{slot} y P_over_sens. Esto indica que la red neuronal aprovechó una representación más distribuida de la información óptica.

La evaluación bajo perturbaciones gaussianas permitió distinguir entre desempeño ideal y robustez operacional. Aunque Random Forest y XGBoost alcanzaron resultados prácticamente perfectos en condiciones limpias, ambos presentaron una degradación abrupta al introducir ruido en las variables de entrada. Esto indica que sus fronteras de decisión fueron altamente efectivas para datos no perturbados, pero poco estables frente a variaciones pequeñas en los descriptores ópticos. El árbol de decisión también mostró sensibilidad al ruido, manteniendo como principal limitación la confusión entre KNO_3 y $NaCl$.

El perceptrón multicapa presentó el mejor equilibrio entre exactitud, capacidad discriminante y robustez. En condiciones limpias alcanzó un desempeño prácticamente perfecto, mientras que bajo ruido mostró una degradación progresiva y más controlada que los modelos basados en árboles. Incluso en los niveles más altos de perturbación, las curvas ROC conservaron valores de AUC superiores a 0.90 para todas las clases, lo que evidencia una mayor estabilidad en la separación multiclase. Aunque KNO_3 fue la clase más sensible a las perturbaciones, la pérdida de desempeño no comprometió de forma inmediata la estabilidad global del modelo.

En conjunto, los resultados confirman que la combinación de modelado electromagnético y aprendizaje automático constituye una estrategia viable para el desarrollo de sensores fotónicos inteligentes orientados a la identificación de contaminantes en agua. No obstante, también muestran que la selección del modelo no debe basarse únicamente en la exactitud obtenida en datos limpios, sino en su capacidad para conservar el desempeño ante incertidumbre en las variables de entrada. Bajo este criterio, el MLP se perfila como la alternativa más adecuada entre los modelos evaluados, al capturar relaciones no lineales más estables entre los descriptores ópticos y las clases de contaminantes.

5.1 TRABAJO FUTURO

Como continuidad de esta investigación, resulta necesario validar experimentalmente los descriptores propuestos mediante una plataforma fotónica basada en guía de onda tipo *slot*. La incorporación de mediciones reales permitiría cuantificar las diferencias entre el comportamiento simulado y el observado experimentalmente, así como evaluar el efecto de ruido instrumental, variaciones de fabricación, fluctuaciones térmicas y desviaciones en la concentración de los analitos.

También se propone ampliar el conjunto de datos mediante la incorporación de nuevas especies químicas, intervalos de concentración más amplios y condiciones ambientales adicionales, como temperatura, ph o mezclas de contaminantes. Esta extensión permitiría analizar la capacidad de generalización del modelo en escenarios más cercanos a condiciones reales de operación y evaluar su escalabilidad hacia problemas multiclase de mayor complejidad.

De igual forma se considera conveniente estudiar otras configuraciones de guías de onda integradas e incorporar fenómenos adicionales, como pérdidas ópticas, rugosidad superficial, tolerancias de fabricación y efectos termo-ópticos. Esto permitiría incrementar el realismo del modelo y aproximar con mayor precisión la respuesta esperada en un dispositivo fabricado.

Finalmente, la continuidad del trabajo consiste en vincular el modelo de clasificación con una propuesta experimental completa del sensor. Para ello será necesario definir la configuración óptica, los intervalos de longitud de onda más adecuados, la resolución mínima requerida del sistema de medición y las condiciones necesarias para obtener experimentalmente los descriptores empleados por el modelo. De esta manera, el sistema de aprendizaje automático dejaría de operar únicamente como una etapa de clasificación numérica y pasaría a integrarse en una arquitectura de detección fotónica para el monitoreo de contaminantes en agua.

6. REFERENCIAS

[1] V. C. Patil et al., “Assessing agricultural water productivity in desert farming system of Saudi Arabia,” IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 8, no. 1, pp. 284–297, 2015, doi: 10.1109/JSTARS.2014.2320592.

[2] M. Pivokonsky, L. Cermakova, K. Novotna, P. Peer, T. Cajthaml, and V. Janda, “Occurrence of microplastics in raw and treated drinking water,” Sci. Total Environ., vol. 643, pp. 1644–1651, 2018, doi: 10.1016/j.scitotenv.2018.08.102.

[3] S. Bhojwani, K. Topolski, R. Mukherjee, D. Sengupta, and M. M. El-Halwagi, “Technology review and data analysis for cost assessment of water trea*TM*ent systems,” Sci. Total Environ., vol. 651, pp. 2749–2761, 2019, doi: 10.1016/j.scitotenv.2018.09.363.

[4] D. Ghernaout, “Increasing Trends Towards Drinking Water Reclamation from Treated Wastewater,” World J. Appl. Chem., vol. 3, no. 1, p. 1, 2018, doi: 10.11648/j.wjac.20180301.11.

[5] G. Busico et al., “Multivariate statistical analysis to characterize/discriminate between anthropogenic and geogenic trace elements occurrence in the Campania Plain, Southern Italy,” Environ. Pollut., vol. 234, pp. 260–269, 2018, doi: 10.1016/j.envpol.2017.11.053.

[6] J. S. Famiglietti, “The global groundwater crisis,” Nat. Clim. Chang., vol. 4, no. 11, pp. 945–948, 2014, doi: 10.1038/nclimate2425.

[7] N. E. Peters and M. Meybeck, “Water quality degradation effects on freshwater availability: Impacts of human activities,” Water Int., vol. 25, no. 2, pp. 185–193, 2000, doi: 10.1080/02508060008686817.

[8] O. Rubio et al., “Índice de Calidad de Agua (ICA) en la Presa La Boquilla en Chihuahua, México,” Ecosistemas y Recursos Agropecuarios, vol. 1, no. 2, pp. 139–150, 2014.

[9] B. Jiménez-Cisneros, M. T. Armienta, and L. A. Aguilar, El Agua en México: Cauces y Encauces. México, 2010.

[10] P. Ávila, “Diagnóstico de la calidad del agua subterránea en los sistemas municipales de abastecimiento en el Estado de Yucatán, México,” Ingeniería, vol. 8, no. 2, 2004.

[11] L. M. Camacho et al., “Occurrence and trea*TM*ent of arsenic in groundwater and soil in northern Mexico and southwestern USA,” Chemosphere, vol. 83, no. 3, pp. 211–225, 2011, doi: 10.1016/j.chemosphere.2010.12.067.

[12] M. A. Armienta, “Advances in the development of adequate, affordable, and sustainable trea*tm*ent alternatives to drinking water contamination by arsenic, heavy metals, and nitrates in Mexico,” in Advances in Drinking Water Purification for Small Systems and Emerging Issues, 2024, pp. 421–456, doi: 10.1016/B978-0-323-91733-9.00019-2.

[13] S. Ahuja, “Water quality worldwide,” in Handbook of Water Purity and Quality, 2021, pp. 19–33, doi: 10.1016/B978-0-12-821057-4.00003-3.

[14] J. F. Schijven et al., “QMRAspot: A tool for Quantitative Microbial Risk Assessment from surface water to potable water,” Water Res., vol. 45, no. 17, pp. 5564–5576, 2011, doi: 10.1016/j.watres.2011.08.024.

[15] J. Vymazal, “Constructed wetlands for wastewater trea*tm*ent,” Ecol. Eng., vol. 25, no. 5, pp. 475–477, 2005, doi: 10.1016/j.ecoleng.2005.07.002.

[16] A. G. Capodaglio, “Integrated, decentralized wastewater management for resource recovery in rural and peri-urban areas,” Resources, vol. 6, no. 2, Art. no. 22, 2017, doi: 10.3390/resources6020022.

[17] A. Yimit et al., “Thin film composite optical waveguides for sensor applications: A review,” Talanta, vol. 65, pp. 1102–1109, 2005, doi: 10.1016/j.talanta.2004.06.045.

[18] G. Mam*TM*in et al., “A highly sensitive and selective optical waveguide sensor based on a porphyrin-coated ZnO film,” Sensors Actuators A Phys., vol. 309, p. 111918, 2020, doi: 10.1016/j.sna.2020.111918.

[19] G. Mam*TM*in et al., “Porphyrin film/K+ ion-exchanged optical waveguide gas sensor,” Opt. Laser Technol., vol. 128, p. 106260, 2020, doi: 10.1016/j.optlastec.2020.106260.

[20] S. Ghosh and B. M. A. Rahman, “An Innovative Straight Resonator Incorporating a Vertical *Slot* as an Efficient Bio-Chemical Sensor,” IEEE J. Sel. Top. Quantum Electron., vol. 23, no. 2, pp. 132–139, 2017, doi: 10.1109/JSTQE.2016.2630299.

[21] R. E. Dessy, “Waveguides as chemical sensors,” Anal. Chem., vol. 61, no. 19, pp. 1079A–1094A, 1989, doi: 10.1021/ac00194a719.

[22] B. Mourched, M. Abdallah, M. Hoxha, and S. Vrtagic, “Machine-Learning-Based Sensor Design for Water Salinity Prediction: A Conceptual Approach,” Sustainability, vol. 15, no. 14, 2023, doi: 10.3390/su151411468.

[23] S. Chugh, S. Ghosh, A. Gulistan, and B. M. A. Rahman, “Machine Learning Regression Approach to the Nanophotonic Waveguide Analyses,” J. Lightwave Technol., vol. 37, no. 24, pp. 6080–6089, 2019, doi: 10.1109/JLT.2019.2946572.

[24] B. You, Y. Yue, M. Sun, J. Li, and D. Jia, “Design of a Real-Time Salinity Detection System for Water Injection Wells Based on Fuzzy Control,” Sensors, vol. 21, no. 9, 2021, doi: 10.3390/s21093086.

[25] Shimadzu, “Leaching Test of Sb and Ge from PET by Graphite Furnace Atomic Absorption Spectrophotometry (GF-AAS),” Application News No. A645, Jan. 2021.

[26] Shimadzu, “Atomic Absorption Spectrophotometers AA-6880 Series,” Document C122-E060, [En línea]. Disponible en: Archivo PDF (AA-6880 Series.pdf). [Accedido: 1 de diciembre de 2025].

[27] Analytik Jena, “Inductively Coupled Plasma Mass Spectrometry PlasmaQuant MS Elite,” [En línea]. Disponible en: https://www.analytik-jena.com/products/chemical-analysis/elemental-analysis/icp-ms/?gad_source=1&gad_campaignid=355693528&gbraid=0AAAADY_nQifFxEjm5WGemgVSproO*TM*XU&gclid=CjwKCAjwrbfWBhAeEiwAc2e6Pjn57OrEuNhx3EZ1n_ui0hgm4yS3_LeNvS369uBOpZNq-Jj5E4aMHhoCIJ0QAvD_BwE. [Accedido: 1 de diciembre de 2025].

[30] Shimadzu, “ICPMS-2030 Specifications,” Shimadzu, [En línea]. Disponible en: https://www.shimadzu.com/an/products/elemental-analysis/icp-ms/icpms-2030/spec.h*TM*l. [Accedido: 1 de diciembre de 2025].

[31] DRAWELL, “ICP-MS DW-SUPEC7000 Inductively Coupled Plasma Mass Spectrometer,” Brochure, [En línea]. Disponible en: Archivo PDF (ICP-MS DW-SUPEC7000.pdf). [Accedido: 1 de diciembre de 2025].

[32] Analytik Jena, “Manual de instrucciones PlasmaQuant 9100 (Elite) High-Resolution Array ICP-OES,” Edición D (07/2023), Jena, Alemania. Disponible en: https://www.analytik-jena.com/products/chemical-analysis/elemental-analysis/icp-oes/plasmaquant-9200-series/ [En línea]. [Accedido: 1 de diciembre de 2025].

[33] DRAWELL, “ICP-OES MICS Full Spectrum Direct Reading Inductively Coupled Plasma Emission Spectrometer,” Brochure, [Disponible en: https://es.drawellanalytical.com/icp700t-petrochemical-version-inductively-coupled-plasma-emission-spectrometer-2icp-oes-mics-full-spectrum-direct-reading-inductively-coupled-plasma-emission-spectrometer/], [En línea]. [Accedido: 1 de diciembre de 2025]

[34] DRAWELL, “Energy Dispersion X-ray Fluorescence Spectrometer (DW-NP-5010)” Brochure, [Disponible en: https://www.drawellanalytical.com/energy-dispersion-x-ray-fluorescence-spectrometer-dw-np-5010/], [En línea]. [Accedido: 1 de diciembre de 2025]

[35] “Diode Array Detector User Manual,” 2021.03, [Disponible en: https://www.agilent.com/cs/library/usermanuals/public/G7117-DAD-UseMa-en-SD-29000132.pdf?srsltid=AU7gw4Wv-lo1YRNsWTe22Rbh13v8f9n_xH8fYYRCIKhfSvIkdonozLnI], [En línea]. [Accedido: 1 de diciembre de 2025]

[36] Waters, “Sistema Arc HPLC - Especificaciones del instrumento” Ficha Técnica. [Disponible en: https://www.waters.com/nextgen/mx/es/products/chromatography/chromatography-systems/arc-hplc-system.h*TM*l],[En línea]. [Accedido: 1 de diciembre de 2025]

[37] Ocean Optics, “USB2000+ Fiber Optic Spectrometer Installation and Operation Manual,” Document 270-00000-000-02-201107, [Disponible en: https://mmrc.caltech.edu/Biologic%20Stopped%20Flow/Manuals/Ocean%20Optics/USB2000+%20Operating%20Instructions.pdf], [En línea]. [Accedido: 1 de diciembre de 2025].

[38] PyroScience GmbH, “FireSting®-PRO Optical Multi-Analyte Meter USER MANUAL,” Document Version 1.07, [Disponible en: https://www.pyroscience.com/en/downloads/laboratory-devices], [En línea]. [Accedido: 1 de diciembre de 2025].

[39] Metrohm DropSens, “µStat 400 Bipotentiostat/Galvanostat Ref. STAT400,” Brochure, [En línea]. [Accedido: 1 de diciembre de 2025].

[40] Metrohm, “DRP-STAT8000 Potentiostat/Galvanostat/EIS,” Metrohm, [En línea]. [Disponible en: https://www.metrohm.com/es_mx/products/d/rp-s/drp-stat8000.h*TM*l.], [Accedido: 1 de diciembre de 2025]

[41] SatisGeo, “GS-512i Portable Gamma Ray Spectrometer OPERATION MANUAL,” 2015, Brno, República Checa. [Disponible en: https://satisgeo.com/downloads/], [En línea]. [Accedido: 1 de diciembre de 2025]

[42] Vario, “GAMMA SURVEYOR VARIO Espectrómetro geofísico de rayos gamma de 2048 canales para exploración de campo, registro de pozos y uso en laboratorio,” Ficha Técnica, [Disponible en: http://www.gfinstruments.cz/version_es/index.php?cont=surveyor_V_&ear=ov_es], [En línea]. [Accedido: 1 de diciembre de 2025].

[43] Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT), “Aviso Factor de Emisión del Sistema Eléctrico Nacional 2024,” 28 de febrero de 2025. [En línea].

[44] C. Huber, I. Klimant, C. Krause, T. Werner, T. Mayr, and O. S. Wolfbeis, “Optical sensor for seawater salinity,” Fresenius. J. Anal. Chem., vol. 368, no. 2, pp. 196–202, 2000, doi: 10.1007/s002160000493.

[45]

[46] H. Schmidt, H. Wolf, and E. Hassel, “A method to measure the density of seawater accurately to the level of 10−6,” Metrologia, vol. 53, no. 2, p. 770, 2016, doi: 10.1088/0026-1394/53/2/770.

[47] H. Wu, C. Tan, X. Dong, and F. Dong, “Design of a Conductance and Capacitance Combination Sensor for water holdup measurement in oil–water two-phase flow,” Flow Meas. Instrum., vol. 46, pp. 218–229, Dec. 2015, doi: 10.1016/j.flowmeasinst.2015.06.026.

[48] P. Grosso, D. Malardé, M. Le Menn, Z. Wu, and J.-L. M. de B. de la Tocnaye, “Refractometer resolution limits for measuring seawater refractive index,” Opt. Eng., vol. 49, no. 10, p. 103603, Oct. 2010, doi: 10.1117/1.3497571.

[49] Z. Qiu, “A simple optical model to estimate suspended particulate matter in Yellow River Estuary,” Opt. Express, vol. 21, no. 23, pp. 27891–27904, 2013, doi: 10.1364/OE.21.027891.

[50] R. Röttgers, D. McKee, and C. Utschig, “Temperature and salinity correction coefficients for light absorption by water in the visible to infrared spectral region,” Opt. Express, vol. 22, no. 21, pp. 25093–25108, 2014, doi: 10.1364/OE.22.025093.

[51] Atkins, Físico-Química. Oxford University Press.

[52] T. Engel and P. Reid, Physical Chemistry. Pearson.

[53] A. H. Price, Espectroscopia y estructura de moléculas complejas.

[54] W. Alavia, I. Soto, and J. A. Lovera, “Modeling of the refractive index for the systems MX + H2O, M2X + H2O, H3BO3 + MX + H2O, and H3BO3 + M2X + H2O. M = K+, Na+, or Li+ and X = Cl− or SO42−,” Processes, vol. 9, no. 3, 2021, doi: 10.3390/pr9030525.

[55] P. T. Parrish and J. L. Bashford, “Temperature dependence of the refractive index of water based on Sellmeier coefficients,” J. Opt. Technol., vol. 82, no. 4, pp. 245–251, 2014.

[56] M. Daimon and A. Masumura, “Measurement of the refractive index of distilled water from the near-infrared region to the ultraviolet region,” Appl. Opt., vol. 46, no. 18, pp. 3811–3820, 2007, doi: 10.1364/AO.46.003811.

[57] A. H. Harvey, J. S. Gallagher, and J. M. H. Levelt Sengers, “Revised formulation for the refractive index of water and steam as a function of wavelength, temperature and density,” J. Phys. Chem. Ref. Data, vol. 27, no. 4, pp. 761–774, 1998, doi: 10.1063/1.556029.

[58] H. Zappe, Fundamentals of Micro-Optics: Technology, Devices and Applications. Cambridge, U.K.: Cambridge University Press, 2010, doi: 10.1017/CBO9780511781797.

[59] G. Lifante, Integrated Photonics: Fundamentals. Chichester, U.K.: Wiley, 2003, doi: 10.1002/0470861401.

[60] C. R. Pollock and M. Lipson, Integrated Photonics. New York, NY, USA: Springer, 2003, doi: 10.1007/978-1-4757-5522-0.

[61] B. E. A. Saleh and M. C. Teich, Fundamentals of Photonics, 3rd ed. Hoboken, NJ, USA: Wiley, 2019.

[62] C. A. Barrios, “Optical *Slot*-Waveguide Based Biochemical Sensors,” Sensors, vol. 9, no. 6, pp. 4751–4765, Jun. 2009, doi: 10.3390/s90604751.

[63] A. Yimit, A. G. Rossberg, T. Amemiya, and K. Itoh, “Thin film composite optical waveguides for sensor applications: A review,” Talanta, vol. 65, pp. 1102–1109, 2005, doi: 10.1016/j.talanta.2004.06.045.

[64] J. D. Jackson, Classical Electrodynamics, 3rd ed. New York, NY, USA: John Wiley & Sons, 1999.

[65] M. Fox, Optical Properties of Solids, 2nd ed. Oxford, U.K.: Oxford University Press, 2010.

[66] R. W. Boyd, Nonlinear Optics, 4th ed. Cambridge, MA, USA: Academic Press, 2020.

[67] M. Born and E. Wolf, Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light. Cambridge, U.K.: Cambridge University Press, 2019, doi: 10.1017/9781108769914.

[68] V. R. Almeida, Q. Xu, C. A. Barrios, and M. Lipson, “Guiding and confining light in void nanostructure,” Opt. Lett., vol. 29, no. 11, pp. 1209–1211, Jun. 2004, doi: 10.1364/OL.29.001209.

[69] Y. Liu, M. Kong, and Y. Jiang, “Transverse magnetic modes in planar *slot* waveguides,” J. Opt. Soc. Am. B, vol. 32, no. 10, pp. 2052–2060, Oct. 2015, doi: 10.1364/JOSAB.32.002052.

[70] B. M. Villegas-Martínez, H. M. Moya-Cessa, and F. Soto-Eguibar, “Exotic light transition between superluminal and subluminal group velocity in a PT coupled slab waveguide,” Photonics Nanostruct. Fundam. Appl., vol. 54, Art. no. 101128, 2023, doi: 10.1016/j.photonics.2023.101128.

[71] G. P. Agrawal, Lightwave Technology: Components and Devices. Hoboken, NJ, USA: Wiley-Interscience, 2005.

[72] J. Han et al., “Highly Sensitive Liquid M-Z Waveguide Sensor Based on Polymer Suspended *Slot* Waveguide Structure,” Polymers, vol. 14, no. 19, Art. no. 3967, 2022, doi: 10.3390/polym14193967.

[73] S. N. Khonina et al., “Environmental Monitoring: A Comprehensive Review on Optical Waveguide and Fiber-Based Sensors,” Biosensors, vol. 12, no. 11, Art. no. 1038, 2022, doi: 10.3390/bios12111038.

[74] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw Hill, 1997.

[75] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[76] K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT Press, 2012.

[77] J. Zhou, Z. Wu, C. Jin, and J. X. J. Zhang, “Machine learning assisted dual-functional nanophotonic sensor for organic pollutant detection and degradation in water,” *npj Clean Water*, vol. 7, Jan. 2024, Art. no. 3, doi: 10.1038/s41545-023-00292-4.

[78] Y. Singh, S. Jena, and H. Subbaraman, “Parametric optimization of the slot waveguide characteristics using a machine-learning approach,” *Sci. Rep.*, vol. 15, Jul. 2025, Art. no. 24037, doi: 10.1038/s41598-025-07521-5.

[79] P. Das *et al.*, “Liquid characterization using a dielectric slot waveguide and machine learning for sensing in the terahertz regime,” in *Proc. MikroSystemTechnik Kongress 2025*, Duisburg, Germany, Oct. 2025, pp. 135–137, doi: 10.30420/456614086.

[80] N. P. Vishwaraj, C. T. Nataraj, R. P. K. Jagannath, S. Talabattula, and G. R. Prashanth, “Machine learning assisted strip waveguide Bragg gratings design for refractive index-based biosensing applications,” *Optik*, vol. 300, Apr. 2024, Art. no. 171622, doi: 10.1016/j.ijleo.2024.171622.

[81] W. Liu *et al.*, “Adaptive machine learning-empowered SPR sensor composed of micro-structured optical fibers for water pollutant detection,” *Opt. Laser Technol.*, vol. 197, May 2026, Art. no. 114710, doi: 10.1016/j.optlastec.2026.114710.

[82] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[83] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Wadsworth, 1984.

[84] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham, U.K.: Packt Publishing, 2019.

[85] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012, doi: 10.1016/C2009-0-61819-5.

[86] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[87] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[88] J. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[89] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[90] K. Hornik, M. Stinchcombe, and H. White, "Multilayer Feedforward Networks are Universal Approximators," Neural Networks, vol. 2, no. 5, pp. 359–366, 1989, doi: 10.1016/0893-6080(89)90020-8.

[91] NICHD, "Ilustración de las Partes del Sistema Nervioso," en ¿Cuáles son las partes del sistema nervioso? NICHD Español. [En línea]. Actualizado: 17 de octubre de 2019. Disponible en: https://espanol.nichd.nih.gov/salud/temas/neuro/informacion/partes.

[92] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," Neural Networks, vol. 61, pp. 85–117, 2015, doi: 10.1016/j.neunet.2014.09.003.

[93] D. E. Rumelhart, G. E. Hinton, and J. R. Williams, "Learning Representations by Back-Propagating Errors," Nature, vol. 323, pp. 533–536, 1986, doi: 10.1038/323533a0.

[94] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” Nature, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.

[95] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[96] Secretaría de Salud, “Agua para uso y consumo humano. Límites permisibles de la calidad del agua. NOM-127-SSA1-2021,” Diario Oficial de la Federación, 2 de mayo de 2022.

[97] C.-Y. Tan and Y.-X. Huang, “Dependence of Refractive Index on Concentration and Temperature in Electrolyte Solution, Polar Solution, Nonpolar Solution, and Protein Solution,” J. Chem. Eng. Data, vol. 60, no. 10, pp. 2827–2833, 2015, doi: 10.1021/acs.jced.5b00018.

[98] L. A. Tapia-Alarcón and B. M. Villegas-Martinez, “Optoelectronic Dataset for *Slot*-Waveguide-Based Detection of Water Contaminants.” Zenodo, Jun. 18, 2026, doi: 10.5281/zenodo.20753105.

APENDICES

APÉNDICE 1. DEDUCCIÓN MICROSCÓPICA DE LA ECUACIÓN DE SELLMEIER A PARTIR DEL MODELO DE LORENTZ

La respuesta dieléctrica de un medio material ante campos ópticos variables puede describirse, en el marco clásico, mediante la teoría de la dispersión electrónica. El desarrollo presentado a continuación sigue la formulación del oscilador armónico amortiguado expuesta por Jackson [64], complementada con tratamientos para medios dieléctricos transparentes descritos en Fox [65] y Boyd [66].

Considérese un medio dieléctrico homogéneo sometido a un campo eléctrico armónico asociado a una onda electromagnética incidente, cuya dependencia temporal se expresa como

$$\mathbf{E}(t) = \mathbf{E}_0 e^{-i\omega t} \quad , \quad \text{(A1.1)}$$

donde $\mathbf{E}_0$ es la amplitud del campo y ω la frecuencia angular. En el modelo de Lorentz, los electrones ligados al núcleo experimentan tres contribuciones: (1) la fuerza debida al campo eléctrico externo, (2) una fuerza de restitución de tipo armónico y (3) un término de amortiguamiento proporcional a la velocidad. La ecuación de movimiento para un electrón de masa m_e y carga e está dada por [55] [56]:

$$m_e \frac{d^2\mathbf{x}}{dt^2} + m_e\gamma \frac{d\mathbf{x}}{dt} + m_e\omega_0^2\mathbf{x} = -e\mathbf{E}_0 e^{-i\omega t} \quad , \quad \text{(A1.2)}$$

donde $\mathbf{x}$ representa el desplazamiento respecto a la posición de equilibrio, γ es la constante de amortiguamiento y ω_0 la frecuencia natural de resonancia.

Buscando soluciones en estado estacionario de la forma $\mathbf{x}(t) = \mathbf{x}_0 e^{-i\omega t}$, se obtiene para la amplitud

$$\mathbf{x}_0 = \frac{-e/m_e}{\omega_0^2 - \omega^2 - i\gamma\omega}\mathbf{E}_0 \quad , \quad \text{(A1.3)}$$

El momento dipolar inducido por cada oscilador es $p = -ex$. Para una densidad numérica N de electrones, la polarización macroscópica resulta [64]:

$$\mathbf{P} = N\mathbf{p} = \frac{Ne^2/m_e}{\omega_0^2-\omega^2-i\gamma\omega}\mathbf{E}_0 e^{-i\omega t} \ . \quad \text{(A1.4)}$$

En regiones espectrales alejadas de resonancias, donde la absorción es despreciable, puede asumirse $\gamma \approx 0$. En este régimen, la expresión anterior se reduce a [66]:

$$\mathbf{P} = \frac{Ne^2/m_e}{\omega_0^2-\omega^2}\mathbf{E} \ . \quad \text{(A1.5)}$$

Por definición, la polarización se relaciona con la permitividad relativa mediante $\mathbf{P} = \varepsilon_0(\varepsilon_r - 1)\mathbf{E}$. Para medios no magnéticos, donde $\varepsilon_r = n^2(\omega)$, se obtiene [64]:

$$n^2(\omega) - 1 = \frac{Ne^2}{\varepsilon_0 m_e(\omega_0^2-\omega^2)} \ . \quad \text{(A1.6)}$$

En un sistema real con múltiples frecuencias de resonancia, la contribución total se expresa como una suma sobre osciladores efectivos caracterizados por factores de oscilador f_i:

$$n^2(\omega) = 1 + \frac{e^2}{\varepsilon_0 m_e}\sum_{i=1}^{m}\frac{Nf_i}{\omega_i^2-\omega^2} \ . \quad \text{(A1.7)}$$

Para reescribir esta expresión en función de la longitud de onda en el vacío, se utilizan las relaciones

$$\omega = \frac{2\pi c}{\lambda}, \qquad \omega_i = \frac{2\pi c}{\lambda_i} \ . \quad \text{(A1.8)}$$

Sustituyendo, se obtiene

$$n^2(\lambda) = 1 + \sum_{i=1}^{m}\frac{\left(\frac{e^2 N f_i \lambda_i^2}{4\pi^2 c^2 \varepsilon_0 m_e}\right)\lambda^2}{\lambda^2-\lambda_i^2} \ , \quad \text{(A1.9)}$$

definiendo los parámetros

$$A_i = \frac{e^2 N f_i \lambda_i^2}{4\pi^2 c^2 \varepsilon_0 m_e}, \qquad C_i = \lambda_i^2 \ , \quad \text{(A1.10)}$$

se obtiene la forma estándar

$$n^2(\lambda) = 1 + \sum_{i=1}^{m}\frac{A_i\lambda^2}{\lambda^2-C_i} \ . \quad \text{(A1.11)}$$

En el rango óptico (infrarrojo cercano y ultravioleta), es habitual aproximar la respuesta mediante cuatro términos dominantes, asociados a resonancias efectivas en el ultravioleta y el infrarrojo. Bajo esta aproximación, la ecuación de Sellmeier adopta la forma

$$n^2(\lambda) = 1 + \frac{A_1\lambda^2}{\lambda^2 - C_1} + \frac{A_2\lambda^2}{\lambda^2 - C_2} + \frac{A_3\lambda^2}{\lambda^2 - C_3} + \frac{A_4\lambda^2}{\lambda^2 - C_4}\,. \quad \text{(A1.12)}$$

APÉNDICE 2. DERIVACIÓN DE LA ECUACIÓN DE HELMHOLTZ PARA UNA GUÍA DE ONDA PLANA

La propagación de ondas electromagnéticas en medios dieléctricos lineales, homogéneos e isotrópicos puede describirse mediante las ecuaciones de Maxwell [64]. En ausencia de cargas libres y corrientes eléctricas, dichas ecuaciones se escriben como

$$\nabla \cdot \mathbf{E} = 0\,, \quad \text{(A2.1)}$$

$$\nabla \cdot \mathbf{H} = 0, \quad \text{(A2.2)}$$

$$\nabla \times \mathbf{E} - \mu \frac{\partial \mathbf{H}}{\partial t} = 0\,, \quad \text{(A2.3)}$$

$$\nabla \times \mathbf{H}\epsilon \frac{\partial \mathbf{E}}{\partial t} = 0\,. \quad \text{(A2.4)}$$

Aplicando el operador rotacional a la ecuación (35),

$$\nabla \times (\nabla \times \mathbf{E}) = 0, \quad \text{(A2.5)}$$

y utilizando la identidad vectorial

$$\nabla \times (\nabla \times \mathbf{E})\nabla(\nabla \cdot \mathbf{E})\nabla^2\mathbf{E} = \mathbf{0}, \ \text{(A2.6)}$$

se obtiene

$$\nabla(\nabla \cdot \mathbf{E})\nabla^2\mathbf{E} - \mu\frac{\partial}{\partial t}(\nabla \times \mathbf{H}) = 0. \quad \text{(A2.7)}$$

Sustituyendo la ecuación (36),

$$\nabla(\nabla \cdot \mathbf{E})\nabla^2\mathbf{E} - \mu\epsilon\frac{\partial^2\mathbf{E}}{\partial t^2} = 0, \quad \text{(A2.8)}$$

Dado que $\nabla \cdot \mathbf{E} = 0$, se obtiene la ecuación de onda

$$\nabla^2\mathbf{E} - \mu\epsilon\frac{\partial^2\mathbf{E}}{\partial t^2} = 0. \quad \text{(A2.9)}$$

Para resolver esta ecuación se propone una separación de variables

$$E(x,t) = X(x)T(t)\ . \quad \text{(A2.10)}$$

Sustituyendo en la ecuación (41) y dividiendo entre XT,

$$\frac{1}{X}\frac{d^2X}{dx^2} - \mu\epsilon\frac{1}{T}\frac{d^2T}{dt^2} = 0. \quad \text{(A2.11)}$$

Ambos términos deben ser iguales a una constante de separación $-k^2$, obteniéndose

$$\frac{d^2T}{dt^2} + k^2c^2T0 = 0, \quad \text{(A2.12)}$$

y

$$\frac{d^2X}{dx^2} + k^2X = 0, \quad \text{(A2.13)}$$

La solución temporal puede escribirse como

$$T(t) = T_0 e^{-i\omega t}, \quad \text{(A2.14)}$$

mientras que la solución espacial conduce a la ecuación de Helmholtz

$$\nabla^2 E + k^2 E = 0 \, . \quad \text{(A2.15)}$$

Esta ecuación constituye la base para el análisis modal de guías de onda dieléctricas y para la determinación de los modos *TM* empleados posteriormente en el modelado de la guía de onda slot.

APÉNDICE 3. CONVERSIÓN DE LA FRACCIÓN MÁSICA Y COMPARACIÓN DEL MODELO CON DATOS EXPERIMENTALES

A3.1. Conversión de fracción másica a concentración física

Tan y Huang [97] reportan índices de refracción experimentales para distintas soluciones, entre ellas cloruro de sodio ($NaCl$) disuelto en agua. Los datos empleados en esta comparación corresponden a una temperatura de $293.15K$ y una longitud de onda de $589.3nm$. En el artículo, la composición de las soluciones se describe mediante la fracción másica del soluto. La fracción másica w se define como la razón entre la masa de soluto y la masa total de la solución. Si se toma como ejemplo el $NaCl$, se tiene:

$$w = \frac{m_{NaCl}}{m_{solución}} = \frac{m_{NaCl}}{m_{NaCl} + m_{H_2O}}, \quad \text{(A3.1)}$$

La fracción másica es adimensional, ya que sus unidades corresponden a kg de $NaCl$ por kg de solución. Al multiplicar ambos lados por la masa total de la solución, se obtiene:

$$w(m_{NaCl} + m_{H_2O}) = m_{NaCl}, \quad \text{(A3.2)}$$

$$wm_{NaCl} + wm_{H_2O} - m_{NaCl} = 0, \quad \text{(A3.3)}$$

$$m_{NaCl}(w - 1) = -wm_{H_2O}. \quad \text{(A3.4)}$$

Por tanto:

$$\frac{m_{NaCl}}{m_{H_2O}} = \frac{w}{1 - w}. \quad \text{(A3.5)}$$

Esta expresión representa una razón de masas. Por sí sola, no corresponde a una concentración volumétrica, ya que todavía no considera el volumen ocupado por la solución. Para expresar la cantidad de NaCl por unidad de masa de agua, se define la concentración auxiliar C^*, en gramos de NaCl por kilogramo de agua. Debido a que $1kg = 1000g$ se obtiene

$$\mathrm{C}^* = (1000\ \mathrm{g/kg})\frac{\mathrm{w}}{1 - \mathrm{w}}. \quad \text{(A3.6)}$$

La ecuación (A3.6) permite calcular una magnitud útil para describir la composición de la solución. Sin embargo, C^*no debe etiquetarse como mg/L, porque se expresa en gramos de soluto por kilogramo de agua, no en miligramos por litro de solución. Para comparar los datos experimentales con el modelo formulado en términos de concentración volumétrica, es necesario convertir la fracción másica a una concentración física expresada en mg/L. Para ello, se utiliza la densidad de la solución, ρ_{sol}, definida como

$$\rho_{sol} = \frac{m_{solución}}{V_{solución}}. \quad \text{(A3.7)}$$

donde $V_{solución}$ es el volumen total de la solución. Al multiplicar ambos lados por w, se obtiene:

$$w\rho_{sol} = w\frac{m_{solución}}{V_{solución}}. \quad \text{(A3.8)}$$

En términos de unidades, el producto de kg de $NaCl$ por kg de solución y kg de solución por litro de solución equivale a kg de $NaCl$ por litro de solución. Se utilizan las equivalencias $1kg$ de $NaCl = 1000g$ de $NaCl$ y $1g$ de $NaCl = 1000mg$ de $NaCl$. Por tanto:

$$1 = \frac{10^6 \text{ mg de NaCl}}{1 \text{ kg de NaCl}}. \quad \text{(A3.9)}$$

Al aplicar este factor al producto $w\rho sol$, la concentración se expresa como:

$$C_{fis\ (\frac{mg}{L})} = (10^6 \text{ mg/kg})\, w\rho_{sol}. \quad \text{(A3.10)}$$

Que es la concentración física y Esta conversión requiere utilizar la densidad correspondiente a la composición y temperatura de cada solución. Por ello, no es adecuado asignar una única densidad a todas las fracciones másicas cuando se busca realizar una comparación cuantitativa. Por ejemplo, para una solución con $w = 0.01$, correspondiente a una solución de $NaCl$ al 1% en masa, se considera una densidad de $1.00508kg/L$, la concentracion fisica se calcula como

$$C_{fis} = (10^6 \text{ mg/kg})(0.01)(1.00508 \text{ kg/L}) = 10050.8 \text{ mg/L}, \quad \text{(A3.11)}$$

El resultado muestra que una fracción másica del 1% no equivale a una concentración de $10.05 mg/L$. El valor 10.05 se obtiene al expresar numéricamente la concentración en g/l, mientras que la concentración correspondiente en mg/L es $10{,}050.8 mg/L$. Es importante distinguir estas unidades para evitar factores de conversión incorrectos.

A3.2. Coeficiente del modelo de índice de refracción

En la metodología de esta tesis, la dependencia del índice de refracción respecto de la concentración se representa mediante una aproximación lineal

$$\mathrm{n}(\lambda, \mathrm{C}) = \mathrm{n}_{\mathrm{H_2O}} + \mathrm{K}_0 \mathrm{C}^*_{\mathrm{mg/L}}, \quad \text{(A3.12)}$$

cuando C se expresa en mg/L, el coeficiente K_0 debe expresarse en L/mg, de manera que el producto $\mathrm{K}_0 C$ sea adimensional. En el modelo empleado para esta comparacion, $NaCl$, se utiliza

$$\mathrm{K}_0 = 1.3621 \times 10^{-7} L/mg. \quad \text{(A3.13)}$$

Este coeficiente se utiliza sin aplicar el factor de escalamiento α al calcular los indices de refraccion asociados con las concentraciones fisicas obtenidas de los datos de Tan y Huang. A una longitud de onda de $589.3nm$, el indice de refraccion del agua pura calculado con el modelo espectral utilizado en la tesis es aproximadamente

$$\mathrm{n}_{\mathrm{H20}}(589.3nm) \approx 1.333349. \quad \text{(A3.14)}$$

En consecuencia, para cada fracción másica reportada por Tan y Huang [97], el índice de refracción calculado se obtiene mediante

$$\mathrm{n}_{\mathrm{mod}} = 1.333349 + (1.3621 \times 10^{-7} L/mg)\mathrm{C}_{\mathrm{fis}}, \quad \text{(A3.15)}$$

donde $\mathrm{C}_{\mathrm{fis}}$ se introduce en mg/L. La ecuacion anterior se emplea aquí para se emplea aquí para comparar la respuesta del modelo lineal con los índices experimentales publicados. La comparación no implica que los datos experimentales se encuentren dentro del intervalo de concentración utilizado para generar el conjunto de datos sintéticos de la tesis.

A3.3. Relación entre la concentración física y la concentración auxiliar

Si de Eq. A3.6 despegamos w se llega a

$$w = \frac{C^*}{1000+C^*}, \quad \text{(A3.16)}$$

Y sustituyendo esta expresión en la ecuación A3.11, se obtiene

$$C_{fis}\left[\frac{mg}{L}\right] = 10^6\left(\frac{C^*}{1000+C^*}\right)\rho_{sol}[\frac{kg}{L}], \quad \text{(A3.17)}$$

Esta ecuación muestra que la conversión entre ambas magnitudes depende de la composición y de la densidad de la solución. Por lo tanto, C^*no puede utilizarse como sustituto directo de C_{fis} en una ecuación cuyo coeficiente se haya definido para concentraciones en mg/L. En la implementación computacional del modelo se puede introducir un factor de escalamiento para trabajar con valores numéricos de menor magnitud. Sin embargo, dicho factor debe definirse de forma consistente con las unidades del coeficiente y de la concentración. Si se modifica la escala numérica de la concentración, también debe modificarse el coeficiente para conservar el mismo producto adimensional. Por ejemplo, si se define una concentración numérica escalada como

$$C_{esc,num} = \frac{C_{fis[mg/L]}}{1000}. \quad \text{(A3.18)}$$

El valor numérico $C_{esc,num}$ representa la misma concentración expresada numéricamente en g/L. En ese caso, el coeficiente correspondiente debe expresarse en L/g.

$$k_g = 1000k_0. \quad \text{(A3.19)}$$

Así, el producto que determina el incremento del índice de refracción conserva su valor:

$$k_0 C_{fis[mg/L]} = k_g C_{esc,num}. \quad \text{(A3.20)}$$

Esta identidad se debe a la conversión coherente de unidades. No justifica multiplicar el coeficiente por un factor de escala y, al mismo tiempo, utilizar una concentración física

en mg/L sin transformar. Por ello, la comparación de esta sección se realiza directamente con k_0y C_{fis}, ambos expresados en sus unidades correspondientes.

A3.4. Comparación con los datos experimentales de Tan y Huang

Para comparar el modelo con los resultados publicados por Tan y Huang [97], se calcularon las concentraciones físicas a partir de las fracciones másicas y las densidades correspondientes a cada solución. Posteriormente, los índices de refracción modelados se obtuvieron mediante la ecuación (A3.15), utilizando el coeficiente k_0 sin escala. La siguiente Tabla presenta los seis puntos experimentales considerados. Las concentraciones físicas se expresan en mg/L, mientras que los índices de refracción son adimensionales.

Tabla A.3.1 Comparación entre los índices de refracción experimentales reportados por Tan y Huang y los calculados con el modelo lineal para soluciones acuosas de NaCl a 293.15 K y 589.3 nm [97].

Fracción másica w	Concentración Tan y Huang (mg/L)	Índice experimental	Concentración modelado (mg/L)	Índice modelado	Error absoluto	Error relativo (%)
0.01	10 050.8	1.3365	10.1	1.3347	0.0018	0.133
0.05	51 700	1.3434	52.63	1.3404	0.0030	0.224
0.10	107 070	1.3521	111.11	1.3479	0.0042	0.308
0.15	166 275	1.3610	176.47	1.3560	0.0050	0.368
0.20	229 560	1.3702	250	1.3646	0.0056	0.408
0.25	297 200	1.3795	333.33	1.3738	0.0057	0.411

El error absoluto se define como la diferencia en valor absoluto entre el índice experimental y el índice modelado

$$E_{abs=}|n_{exp} - n_{mod}|. \quad \text{(A3.21)}$$

El error relativo porcentual se calcula respecto al valor experimental

$$E_{rel(\%)=} \frac{|n_{exp} - n_{mod}|}{n_{exp}} \text{ x } 100. \text{ (A3.22)}$$

Los resultados muestran que el índice de refracción calculado con el modelo lineal es inferior al valor experimental en los seis puntos. El error absoluto aumenta aproximadamente de 0.0018 a 0.0057, mientras que el error relativo se encuentra aproximadamente entre 0.13% y 0.41%. Esta diferencia sistemática indica que el modelo lineal reproduce de manera aproximada el incremento del índice de refracción con la concentración, pero no reproduce exactamente los valores experimentales. Además, la discrepancia aumenta en el intervalo de concentraciones analizado, lo que es consistente con que una relación lineal de pendiente constante no necesariamente representa con precisión toda la respuesta de una solución concentrada.

A3.5. Interpretación de la comparación y alcance de los resultados

La Figura A3.1a presenta el índice de refracción calculado con el modelo lineal de la tesis, y la Figura A3.1b muestra los índices experimentales reportados por Tan y Huang, ambos a $293.15K$ y $589.3nm$. En las dos figuras el índice de refracción aumenta de forma monótona con la concentración de $NaCl$, y a simple vista ambas curvas son casi rectas. Como las figuras utilizan intervalos de concentración distintos en sus ejes horizontales, no se superponen, y la comparación cuantitativa se realiza con los valores de la Tabla A3.1. En ella, los datos de Tan y Huang pasan de 1.3365 a 1.3795 entre $10{,}050.8$ y $297{,}200mg/L$, mientras que el modelo lineal, en el mismo intervalo, pasa de 1.3347 a 1.3738. El comportamiento es cualitativamente el mismo, y el índice modelado queda por debajo del experimental en los seis puntos, con un error absoluto de 0.0018 a 0.0057 y un error relativo menor a 0.42%.

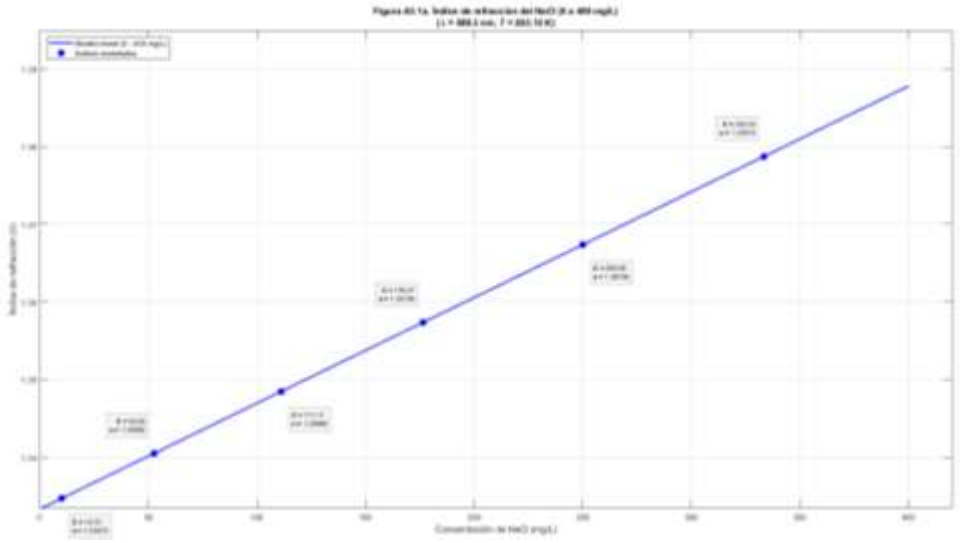


Figura A3.1ª Índice de refracción del NaCl en agua calculado con el modelo lineal de la tesis, n = nH_2O + KC, con K = ($(R_{soluto/M_soluto})X10^{-3}$) L/mg, a 589.3 nm y 293.15 K.

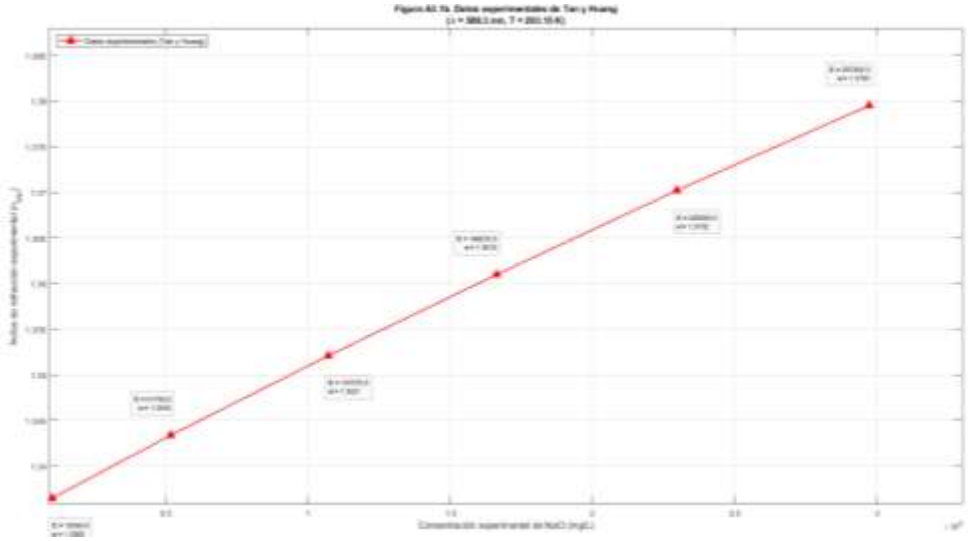


Figura A3.1b Índices de refracción experimentales de soluciones acuosas de NaCl reportados por Tan y Huang a 589.3 nm y 293.15 K, para fracciones másicas de 0.01 a 0.25 [97].

Sin embargo, la pendiente constante del modelo lineal no reproduce completamente la variación observada en los datos experimentales. Con los valores de la Tabla A3.1, la pendiente promedio experimental es de aproximadamente $1.50 \times 10 - 7L/mg$, alrededor de 10 % mayor que el coeficiente $K_0 = 1.3621 \times 10 - 7L/mg$. Además, la pendiente local experimental no es constante, pues disminuye de aproximadamente $1.66 \times 10 - 7L/mg$ en el primer tramo a $1.37 \times 10 - 7L/mg$ en el último. Esto indica una ligera curvatura de

los datos que una recta de pendiente fija no puede capturar. Por esta razón, el error absoluto crece con la concentración, aunque de manera cada vez más lenta: casi no cambia entre los dos últimos puntos (0.0056 y 0.0057), porque en ese tramo la pendiente experimental se aproxima a K_0. También se observa que el modelo ya presenta una diferencia de 0.0018 en el punto de menor concentración. Esta diferencia inicial puede asociarse tanto a la mayor pendiente experimental a concentraciones bajas como a la diferencia entre el valor de n_{H_2O} del modelo espectral y el del agua de referencia utilizada en los experimentos. Este último punto no puede resolverse con los datos disponibles.

Por ello, la proximidad general de ambas tendencias debe interpretarse como una comparación ilustrativa del comportamiento del índice de refracción, y no como una validación experimental del modelo completo del sensor.

También es necesario considerar que las concentraciones de la Tabla A3.1, comprendidas aproximadamente entre $10{,}000$ y $297{,}000 mg/L$, son entre diez y trescientas veces superiores al intervalo de 0 a $1000 mg/L$ utilizado para generar los datos sintéticos de esta tesis. Por tanto, la comparación se realiza en un intervalo de concentración distinto al empleado en el estudio computacional principal. En ese intervalo de trabajo, el modelo predice un incremento total del índice de refracción de apenas $1.4 \times 10 - 4(de 1.33335 a 1.33349)$, valor menor que los errores absolutos de la Tabla A3.1. Por lo tanto, estos errores no pueden extrapolarse ni descartarse para el intervalo de 0 a $1000 mg/L$, donde la desviación entre el modelo y la respuesta real del $NaCl$ podría ser distinta y debe evaluarse de manera específica.

En consecuencia, los resultados de esta tabla no demuestran por sí solos la precisión del modelo dentro del intervalo de trabajo de la tesis ni su desempeño ante muestras reales. Los datos de Tan y Huang se incorporan como una referencia bibliográfica para examinar la tendencia general del índice de refracción de soluciones acuosas de $NaCl$ y para verificar que el orden de magnitud del coeficiente K_0 es coherente con los valores publicados. La validación del modelo dentro del intervalo de concentración de interés, así como su aplicabilidad a muestras reales, requerirá comparaciones específicas en dicho intervalo y, posteriormente, una evaluación experimental.

Anexos

Anexo 1



CERTIFICADO DE PARTICIPACIÓN

LXVIII CONGRESO NACIONAL DE FÍSICA

XL ENCUENTRO NACIONAL Divulgación Científica

smf SOCIEDAD MEXICANA DE FÍSICA

LA SOCIEDAD MEXICANA DE FÍSICA, A.C.
AGRADECE LA PARTICIPACIÓN DE:
(BECADO)

LUIS ARTURO TAPIA ALARCÓN

CENTRO DE INVESTIGACIÓN EN INGENIERÍA Y CIENCIAS APLICADAS

EN EL LXVIII CONGRESO NACIONAL DE FÍSICA
DEL 12 AL 17 DE OCTUBRE DE 2025,
CENTRO DE CONVENCIONES, EDOMEX
TOLUCA, ESTADO DE MÉXICO

DR. ORACIO NAVARRO CHÁVEZ
PRESIDENTE

cnf.smf.mx

GOBIERNO DEL ESTADO DE MÉXICO | ESTADO DE MÉXICO ¡El poder de servir! | Educación Secretaría de Educación Pública | Ciencia y Tecnología

EDUCACIÓN SECRETARÍA DE EDUCACIÓN, CIENCIA, TECNOLOGÍA E INNOVACIÓN | Universidad Autónoma del Estado de México

CERTIFICATE

OF PARTICIPATION

THIS IS TO CERTIFY THAT

Luis Arturo Tapia-Alarcón, Braulio Misael Villegas-Martinez, Gabriela Hernández-Luna, Roberto Alan Beltrán-Vargas, Omar Palillero-Sandoval, José Antonio Marban-Salgado

Universidad de la Cañada

Presented a Paper Titled "Theoretical model for detecting chemical components using Machine Learning approaches" AT THE **6th INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND DATA SCIENCE (ICAIDATS)**

Teotitlán de Flores Magón, Oaxaca, México on November 14, 2025

Dr. Miguel Angel Sánchez Acevedo
President ICAIDATS 2025

Universidad de la Cañada
RECTORÍA

Dra. Margarita Bernabé Pineda
Officer in Charge of the
University of La Cañada

Certificate of Participation



Awarded to

Luis Arturo Tapia Alarcón

For participating in an IBM Quantum event and completing an official Qiskit Fall Fest 2025 event.

11/25/2025

Katie Pizzolato
Vice President, Quantum Algorithms & Scientific Partnerships

Date

IBM

La Sociedad Mexicana de Física atraveś de la Vocalía de Divulgación otorga la presente

CONSTANCIA

A: **Luis Arturo Tapia Alarcón**

Por su participación en la actividad "Computacion cuantica y aplicaciones", desarrollada como parte de las actividades nacionales de divulgación científica del Quantum Day México 2026, llevado a cabo el 14 de abril de 2026.

Dra. Jennifer López Chacón
Vocal de Divulgación de la
Sociedad Mexicana de Física

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

EL TECNOLÓGICO NACIONAL DE MÉXICO

A TRAVÉS DEL CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO

OTORGA EL PRESENTE

RECONOCIMIENTO

A

Luis Arturo Tapia Alarcón | Braulio Misael Villegas Martínez | Gabriela Hernández Luna | Roberto Alan Beltrán Vargas | José Antonio Marban Salgado | Omar Palillero Sandoval

Por su destacada participación con la Contribución: Artículo Revista,
en el área: Inteligencia Artificial, con el Título:

Theoretical Model for Detecting Chemical Components Using Machine Learning Approach

presentada en el marco del 6.° Congreso Internacional de Tecnología y Ciencia Aplicada (CITCA), celebrado en el Centro Nacional de Investigación y Desarrollo Tecnológico del 19 al 22 de Mayo de 2026.

Cuernavaca, Morelos, 22 de Mayo de 2026

EDUCACIÓN | CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
DIRECCIÓN

Lorenzo O. Hernández
Director del TecNM/CENIDET

Noé Alejandro Castro Sánchez
Coordinador General del 6to. CITCA

Cuernavaca, Morelos a 22 de mayo de 2026

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico



2026 año de Margarita Maza

6° CITCA
Congreso Internacional de Tecnología y Ciencia Aplicada

Anexo 2


IBM
2025 Qiskit Fall Fest Mentor
ISSUED TO
Luis Arturo Tapia Alarcón
2025 Qiskit Fall Fest Mentor
IBM
IBM Quantum
Issued on: 26 NOV 2025 | Issued by: IBM
Verify: https://www.credly.com/go/8cWrv5b5

In recognition of the commitment to achieve professional excellence



Luis Arturo Tapia Alarcón

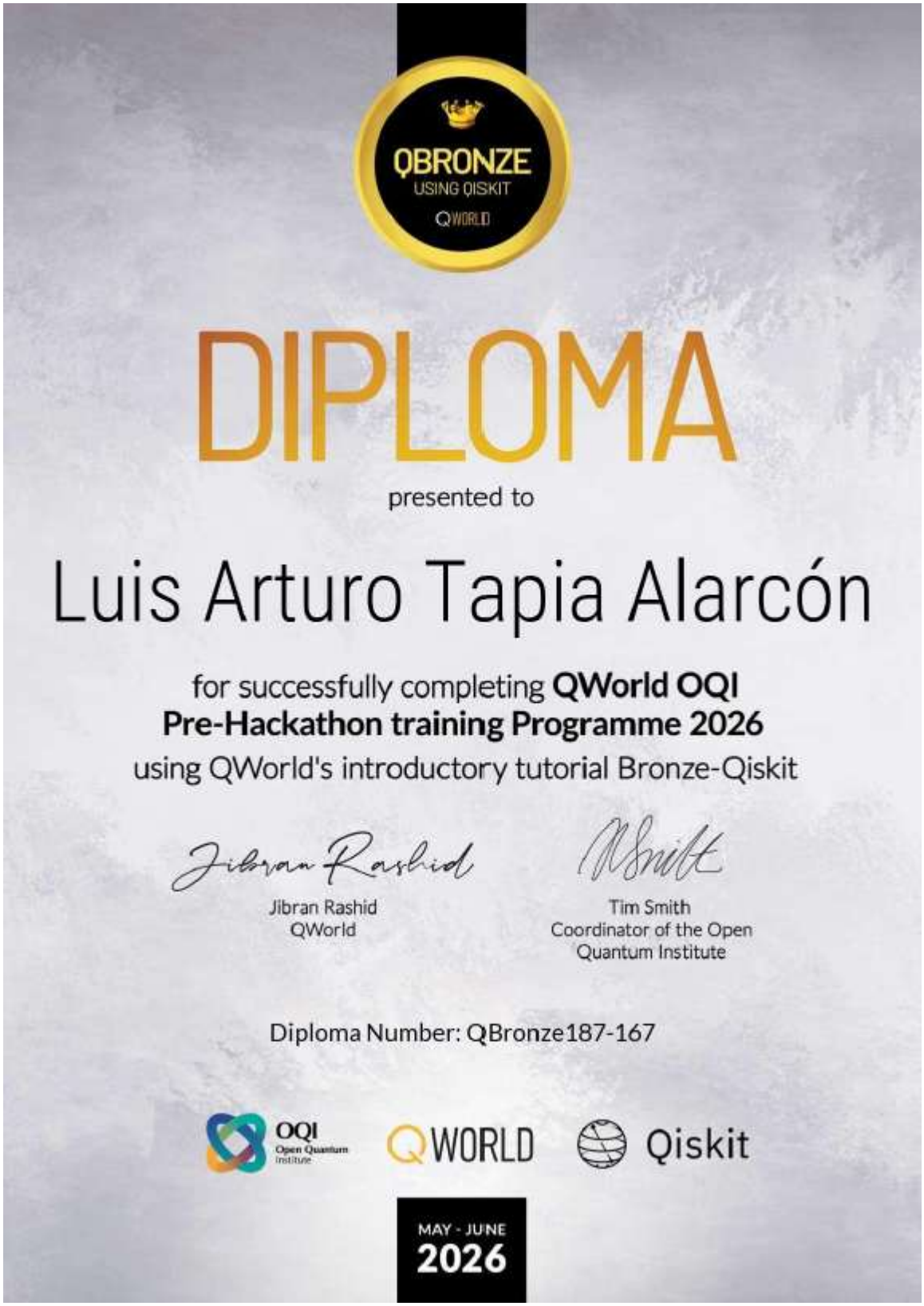
Has successfully satisfied the requirements for:

Basics of Quantum Information

Issued on: Aug 27, 2025
Issued by: IBM

Verify: https://www.credly.com/badges/bf6ae9d5-51aa-424f-80e0-7f325621207d

IBM


QBRONZE
USING QISKIT
QWORLD
DIPLOMA
presented to
Luis Arturo Tapia Alarcón
for successfully completing QWorld OQI
Pre-Hackathon training Programme 2026
using QWorld's introductory tutorial Bronze-Qiskit
Jibran Rashid
QWorld
Tim Smith
Coordinator of the Open
Quantum Institute
Diploma Number: QBronze187-167
OQI
Open Quantum
Institute
QWORLD
Qiskit
MAY - JUNE
2026


QCOBALT
WORKSHOP
QWORLD
DIPLOMA
presented to
Luis Arturo Tapia Alarcón
for successfully completing QWorld OQI Pre-Hackathon training Programme 2026 using QWorld's intermediate-level tutorial Cobalt, focusing on modeling and solving combinatorial optimization problems.
Jibran Rashid
QWorld
Tim Smith
Coordinator of the Open Quantum Institute
Diploma Number: QCobalt10-148
OQI
Open Quantum Institute
QWORLD
MAY - JUNE
2026

Cronograma de actividades

ACTIVIDADES	Semestre 1	Semestre 2	Semestre 3	Semestre 4
	1er año		2º año	
Revisión bibliográfica				
Unidades de aprendizaje				
Detallar que modelos de aprendizaje automático y redes neuronales serán empleados en el análisis y clasificación de los datos sintéticos				
Conceptualizar el diseño del sensor basado en guías de onda				
Implementar un protocolo de generación de datos utilizando el software Matlab				
Codificación de los algoritmos de Machine learning para el reconocimiento y clasificación de los compuestos químicos				
Redacción de tesis				
Redacción de un artículo/Presentación en congreso				
Presentación de examen final				

Cuernavaca, Morelos, a 08 de septiembre de 2026.

DR. J JESÚS ESCOBEDO ALATORRE
ENCARGADO DE DESPACHO DE LA
DIRECCIÓN DEL CIICAp
P R E S E N T E

Atendiendo a la solicitud para emitir DICTAMEN sobre la revisión de la TESIS titulada: Modelado Y Análisis Numérico De Un Sensor De Guía De Onda Tipo Slot Para La Identificación De Contaminantes Presentes En El Agua Mediante Técnicas De Aprendizaje Automático que presenta el estudiante **Luis Arturo Tapia Alarcón**, para obtener el grado de la **Maestría en Sustentabilidad Energética**.

Nos permitimos informarle que nuestro voto es **APROBATORIO.**

Atentamente
Por una humanidad culta

(Se adiciona efirma UAEM)

DR. OMAR PALILLERO SANDOVAL
DR. JOSÉ ANTONIO MARBÁN SALGADO
DR. ROBERTO ALAN BELTRÁN VARGAS
DRA. GABRIELA HERNÁNDEZ LUNA
DR. BRAULIO MISAEL VILLEGAS MARTÍNEZ

PLAZO PARA LA REVISIÓN 20 DÍAS HÁBILES (A PARTIR DE LA FECHA DE RECEPCIÓN DEL DOCUMENTO)

NOTA. POR CUESTION DE REGLAMENTACIÓN LE SOLICITAMOS NO EXCEDER EL PLAZO SEÑALADO, DE LO CONTRARIO LE AGRADECEMOS SU ATENCIÓN Y NUESTRA INVITACIÓN SERÁ CANCELADA.

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

Se expide el presente documento con firma electrónica UAEM, soportada por el certificado vigente a la fecha de su elaboración y con efectos plenos de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS PUBLICADOS en el ÓRGANÓ INFORMATIVO UNIVERSITARIO "ADOLFO MENÉNDEZ SAMARÁ" número 117 de fecha 20 de abril de 2021.

Sello electrónico

GABRIELA HERNANDEZ LUNA | Fecha:2026-09-10 16:00:18 | FIRMANTE

p67m9CbNLhXkW+dNn4DJ8U4jhOHIt0Oj4BAUw/Iq1oZ820HZssfklXhHesUycvZHKm1BPaT6Zv5eqWpi8UyqTjktONISaZ8npWSIjl7dwRxse7ckC1IE/NZNA2i0hoy4aHJm3wnmkwr5iHEPAlv9hJUL4wT9roPdFVMJwKETC/zJoegPo/T6HSHWT2fg+8K/5hIF7I25qtpFzIEVyEJ3N6j2E7QWO6EbdmAn1NJ5MdtIqAXAT/W4dZVtXBp2irvPf63Ci3nrRGV6sr/bZhO/9/ivYWC4dS8AtWEHoaLqDTBmGQg8J2zlHh10JgzaDtLqTyHJK+IUR5vJctAbqpX/Aw==

OMAR PALILLERO SANDOVAL | Fecha:2026-09-15 15:47:24 | FIRMANTE

dnaGMRePFblQFiK2NdhoPP5GVfYc4RD7l9OWokReSWEcwXkLPZy6a/Vdr3wkIy4/BFKLJzaTdCEhxKKw18DdATCRrfGTJgJu8ritQR5eb8uMkxY/SLAtpIOrehDXip5UtMj4Y0bA19PrsjtzMSjL1Qk5KgOJgbXOFY9ZxPLRjcLgfhXVcFp2LTPJdPj1+qmUQL4iOBycpgbGrpFY0DgLGwVNKchlzUsdMX9THOTr617VhpLimEdNTg5gLDr67DFFwyLA/+EtCs02Xn0ZjYo4s/kRrF9agFxGOtQoc8wo5mX6Hm7bglA7lFziyt+rspQwtwnpuUxzjBA4dm5OKHMozA==

ROBERTO ALAN BELTRAN VARGAS | Fecha:2026-09-17 14:15:35 | FIRMANTE

UAG5eMespZ4h5+rorIyNZrDDaDkEgGpWjPnW7jYnMDllTeOEfOAc1d2abzikBpOO2sW8zefs8MD1PvqTiNNsznMZrXSurMiV4COw+cXZ4iQEs4pePpOE8/S4ndeMOPeGxxS3gug8rZ3SMURXk2cT9mBa4HqFlq9moizW3iZTRavzbpnzSc0fALTduFtRebGgwKu2FYecg8LLs1UWwxPkalPHwK0CrCL9HfZXQ5dPftM+TeLLTNZ7oPpl/q/3MAuZ0mrFhXhE5AGKlrFjN0/o83uSeMLXzik0YvT+3FyGD2WgO80Knq5szx0ew5RA0ZR4A560koWmHP2vlDOvtZkotQ==

JOSE ANTONIO MARBAN SALGADO | Fecha:2026-09-25 13:17:22 | FIRMANTE

Nlbm0ayYoLq47WZeYBdZer+YZjrpuFzRoNIhIcHSR59koOD0xt1Gr8N4Izf9puLE/PMxXPqvNSg3v8thn3r8adgcNBmXE38bDYovx7YqTQn81YWDSYAdj6+0Eg/qLLiDsrZnD669SvAI+Nm5EZrSeMOkL4A5+40tfn2spXbZaRhho2H8mmR1f5Hc/oMku5hkW1WQm9Rt4lp3I+jhOE10YjhD6+O8suXMaa+rvbLPi3plDsVm6ZQYxhYfioMtnSQ9Mef241S9dWp+g0fAK4Q9Vbr6gWMJmKw33z4cUGw4RaOMaUFLqxPdwWxzdXM20gL2Zb/xYuH3gtFn3zWkOEq+tQ==

BRAULIO MISAEL VILLEGAS MARTINEZ | Fecha:2026-09-26 06:36:49 | FIRMANTE

tbDvAaoFYOmi1FKbKGf66sz7mow0kLVg7BeKfeheymyFCCUjj9QxsEjW7bS4Ai/bVXSCqcXrow5DM9u/V3MlGUp40tiDKqElm+fjkwJCtw3oudTHtlEATkbZ+iE+k+LVuUAV3/e2VpfXRfxZVF4bxndQ9guxqgSg4j7oUD9MZx3l9u7ZhQLXNbBEWg5MaECnjOhqfwtXLX6ur7pNt/8bvLcJOnnzME7v4xyJJ9MQH7IpUZ6LF+1EtVyHLEw1yBDN+BQjjtQ384h71yb4EJt8Hj+ar2oPYNUfWnohFxmfgKyAyTgVxkmZqPsRoOSoceflGk4jaYTuGE68Qki8D7j6cw==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

nJBREgaNe

https://efirma.uaem.mx/noRepudio/tVyhyh8nb4IoAPwYlFQcAa8o7CImDJ5i