

نموذج تعلم آلي خفيف الوزن للكشف الفوري عن تلوث المياه على الأجهزة الطرفية
ايمان سالم سعيد المشلوم
المعهد العالي للعلوم والتقنية المشاشية – ليبيا
الايمل: e2015201620@gmail.com

تاريخ الاستلام: 2026/01/08 تاريخ المراجعة 17 / 2 / 2026 تاريخ القبول: 2026/03/10- تاريخ النشر: 2026 / 03/17

المخلص:

تُعد مراقبة جودة مياه الشرب في الزمن الحقيقي تحدياً حيوياً، خاصة في المناطق النائية والبيئات محدودة الموارد حيث تعجز البنى التحتية التقليدية عن توفير رصد مستمر وفعال. تهدف هذه الورقة إلى تصميم وتنفيذ نموذج تعلم آلي خفيف الوزن (Lightweight ML) قادر على تصنيف صلاحية المياه للاستهلاك البشري بدقة معقولة، مع ضمان قابلية النشر على أجهزة الحافة (Edge Devices) منخفضة الطاقة.

اعتمدت المنهجية المقترحة على مجموعة بيانات "Water Potability" من منصة Kaggle ، والتي تضم 3,276 عينة و10 خصائص فيزيوكيميائية. تم معالجة تحديات جودة البيانات عبر تطبيق استراتيجيات الوسيط لسد القيم المفقودة، وخوارزمية SMOTE لموازنة توزيع الفئات. صُممت بنية شبكة عصبية متسلسلة مُصغرة تتكون من ثلاث طبقات كثيفة (32→16→8 عصبونات) مع تقنيات تنظيم متقدمة (L2 Regularization ، Batch Normalization ، Dropout لمنع الإفراط في التخصيص. ولتحسين الكفاءة الحاسوبية، تم تحويل النموذج إلى صيغة TensorFlow Lite مع تطبيق التكميم اللاحق للتدريب. (Post-Training INT8 Quantization).

أظهرت النتائج التجريبية على مجموعة الاختبار (ن=656) تحقيق النموذج لدقة تصنيفية بلغت 61.13%، ودرجة F1-Score مقدارها 51.80%، ومساحة تحت منحنى ROC-AUC تبلغ 65.69% والأهم من الناحية التطبيقية، تم ضغط حجم النموذج من 59.47 كيلوبايت إلى 6.07 كيلوبايت فقط، أي بنسبة ضغط تبلغ 9.8×، مما يجعله مناسباً للنشر على متحكمات دقيقة مثل ESP32 و Arduino Nano 33 BLE و Raspberry Pi Pico.

تُثبت هذه الدراسة جدوى استخدام نماذج الذكاء الاصطناعي الخفيفة والمُكمّمة في تطبيقات الرصد البيئي الميداني، حيث تُوازن بشكل مدروس بين الدقة التصنيفية وكفاءة الموارد. تُعد النتائج خطوة عملية نحو أنظمة إنذار مبكر مستقلة ومنخفضة التكلفة للكشف عن تلوث المياه، مع فتح آفاق للبحث المستقبلي في دمج البيانات الزمنية، والتعلم الاتحادي، والتحسين الكمي المخصص للأجهزة.

الكلمات المفتاحية: تعلم آلي خفيف الوزن، حوسبة الحافة، جودة المياه، الكشف عن التلوث، التكميم، TensorFlow Lite، إنترنت الأشياء البيئي.

Abstract:

Real-time monitoring of drinking water quality remains a critical challenge, particularly in remote and resource-constrained environments where traditional infrastructure fails to provide continuous and efficient surveillance. This paper presents the design and implementation of a lightweight machine learning (ML) model capable of classifying water potability for human consumption with reasonable accuracy, while ensuring deployability on low-power edge devices.

The proposed methodology leverages the "Water Potability" dataset from Kaggle, comprising 3,276 samples and 10 physicochemical features. Data quality challenges were addressed through median imputation for missing values and the SMOTE algorithm for class distribution balancing. A compact sequential neural network architecture was designed, consisting of three dense layers (32→16→8 neurons) enhanced with advanced regularization techniques (L2 regularization, BatchNormalization, Dropout) to mitigate overfitting. To optimize computational efficiency, the model was converted to TensorFlow Lite format with post-training INT8 quantization.

Experimental results on the held-out test set (n=656) demonstrated a classification accuracy of **61.13%**, an F1-score of **51.80%**, and a ROC-AUC of **65.69%**. Most significantly from a deployment perspective, the model size was compressed from **59.47 KB** to merely **6.07 KB**—a **9.8× compression ratio**—making it suitable for deployment on microcontrollers such as ESP32, Arduino Nano 33 BLE, and Raspberry Pi Pico.

This study validates the feasibility of employing lightweight, quantized AI models for field-based environmental monitoring applications, where a deliberate trade-off between classification accuracy and resource efficiency is managed. The findings represent a practical step toward autonomous, low-cost early-warning systems for water contamination detection, while opening avenues for future research in temporal data fusion, federated learning, and hardware-aware quantization.

Keywords: Lightweight Machine Learning, Edge Computing, Water Quality, Contamination Detection, Model Quantization, TensorFlow Lite, Environmental IoT.

1. المقدمة

تواجه المجتمعات العالمية تحديات متزايدة في ضمان جودة مياه الشرب، خاصة في ظل التغيرات المناخية، والتلوث الصناعي المتسارع، وتدهور البنى التحتية التقليدية لمراقبة الموارد المائية. تعتمد الأنظمة الحالية في الغالب على التحاليل المخبرية الدورية أو أنظمة الرصد المركزية المعتمدة على الحوسبة السحابية، مما يفرض قيوداً زمنية ومالية كبيرة، ويحد من قدرة الاستجابة الفورية عند اكتشاف ملوثات خطيرة. في المقابل، يشهد مجال الذكاء الاصطناعي وتعلم الآلة تطوراً ملحوظاً في

نمذجة البيانات البيئية، إلا أن تعقيد النماذج الحديثة وحاجتها إلى موارد حوسبية عالية يجعل نشرها على الأجهزة الميدانية منخفضة الطاقة أمراً صعباً. ومن هنا تبرز الحاجة الملحة إلى حلول ذكية تجمع بين الدقة التصنيفية المقبولة، والكفاءة الحاسوبية العالية، والاستقلالية عن الشبكات المركزية.

1.1. خلفية البحث وأهميته

مع انتشار تقنيات إنترنت الأشياء البيئية (Environmental IoT) وأجهزة الاستشعار الفيزيوكيميائية منخفضة التكلفة، أصبح من الممكن جمع بيانات مستمرة حول جودة المياه تشمل مؤشرات حيوية مثل الأس الهيدروجيني، العكارة، الموصلية الكهربائية، والكربون العضوي الكلي. ومع ذلك، فإن معالجة هذه البيانات وتحليلها محلياً دون الاعتماد على اتصال شبكي مستقر أو خوادم بعيدة يُعد حاجة استراتيجية، لا سيما في المناطق النائية، والمخيمات المؤقتة، والمنشآت الصناعية الصغيرة. هنا تبرز حوسبة الحافة (Edge Computing) وتقنيات الذكاء الاصطناعي الخفيف (TinyML) كحل عملي يدمج بين سرعة الاستدلال، وانخفاض استهلاك الطاقة، والقدرة على العمل في ظروف انقطاع الشبكة. يتيح تحويل نماذج تعلم الآلة إلى صيغ مُحسنة مثل TensorFlow Lite مع تطبيق تقنيات التكميم الرقمي (Quantization) نشر خوارزميات ذكية مباشرة على المتحكمات الدقيقة، مما يمكّن من الكشف الآني عن تدهور جودة المياه واتخاذ إجراءات وقائية فورية قبل وصول المياه للمستهلك.

1.2. مشكلة البحث

على الرغم من التقدم الملحوظ في تطبيقات تعلم الآلة للتنبؤ بصلاحية المياه، لا تزال الفجوة القائمة بين دقة النماذج المعقدة وقابليتها للنشر العملي على الأجهزة الطرفية تشكل عائقاً بحثياً وتطبيقياً رئيسياً. تركز معظم الدراسات المنشورة على تعظيم الدقة التصنيفية باستخدام مجموعات بيانات مُعدّة مسبقاً ونماذج ضخمة، متجاهلةً تحديات العالم الحقيقي مثل القيم المفقودة في القياسات الحقلية، وعدم التوازن الحاد بين الفئات (مياه صالحة مقابل غير صالحة)، ومحدودية الذاكرة وسرعة المعالج في الأجهزة الميدانية. علاوة على ذلك، تفتقر الأدبيات الحالية إلى توثيق منهجي لعملية ضغط النماذج وتحويلها لبيئات الحافة، مما يحد من إمكانية تكرار النتائج وتطبيقها في سيناريوهات الرصد المستقل.

1.3. أهداف الورقة

تهدف هذه الورقة إلى تصميم وتقييم نموذج تعلم آلي خفيف الوزن قادر على تصنيف صلاحية المياه للاستهلاك البشري في الزمن الحقيقي، مع ضمان كفاءة حوسبية عالية تتيح نشره على أجهزة الحافة محدودة الموارد. وتتحدد الأهداف الفرعية فيما يلي:

1. تطوير خط أنابيب معالجة بيانات يتعامل بفعالية مع القيم المفقودة واختلال توزيع الفئات في مجموعة بيانات واقعية (Water Potability).
2. تصميم بنية شبكة عصبية متسلسلة مُصغرة مع تقنيات تنظيم متقدمة لمنع الإفراط في التخصيص مع الحفاظ على أداء تصنيفي متوازن.
3. تطبيق تحسينات خاصة بالحافة تشمل التحويل إلى صيغة TFLite والتكميم اللاحق للتدريب (Post-Training Quantization) لتقليل حجم النموذج واستهلاك الطاقة دون انهيار الأداء.

4. تقييم أداء النموذج شاملاً المقاييس التصنيفية (Accuracy, F1-Score, ROC-AUC) ومؤشرات كفاءة النشر (الحجم النهائي، التوافق مع المتحركات الدقيقة).

1.4. المساهمات الرئيسية

تقدم هذه الورقة المساهمات التالية للمجال:

نموذج تصنيفي خفيف الوزن (>7 كيلوبايت بعد التكميم) مصمم خصيصاً للرصد البيئي الميداني، مع توثيق كامل لخطوات المعالجة والتحسين.

إطار عمل متكامل يدمج بين موازنة البيانات (SMOTE)، المعالجة المسبقة القوية، والتحويل الأمثل للحافة، مما يسهل إعادة استخدام المنهجية في تطبيقات بيئية وصحية أخرى.

تحليل نقدي للمقايضة بين الدقة والكفاءة، مع إثبات جدوى النماذج المُكَمَّمة في سيناريوهات الإنذار المبكر حيث تكون الاستقلالية

عن الشبكة وسرعة الاستجابة عوامل حاسمة تتفوق على الدقة المطلقة.

1.5. هيكل الورقة

ينظم بقية الورقة على النحو التالي: يتناول القسم الثاني مراجعة للأدبيات السابقة ذات الصلة برصد جودة المياه وتطبيقات الذكاء الاصطناعي الخفيف على الحافة. يصف القسم الثالث المنهجية المتبعة، بما في ذلك خصائص مجموعة البيانات، خطوات المعالجة المسبقة، بنية النموذج، واستراتيجيات التحسين للنشر الطرفي. يعرض القسم الرابع النتائج التجريبية ويناقشها من الناحيتين التصنيفية والتطبيقية، مع تحليل مصفوفة الارتباك وأثر التكميم على الحجم والأداء. وأخيراً، يختتم

2. مراجعة الأدبيات والأعمال ذات الصلة

شهدت العقود الأخيرة تحولاً جذرياً في منهجيات مراقبة جودة المياه، من الاعتماد على التحاليل الكيميائية اليدوية إلى دمج التقنيات الذكية وأنظمة الاستشعار المتصلة. يستعرض هذا القسم تطور هذه المنهجيات، مع التركيز على دور تعلم الآلة، وحوسبة الحافة، وتقنيات تحسين النماذج، وصولاً إلى تحديد الفجوة البحثية التي تستهدفها هذه الورقة.

2.1. الأساليب التقليدية والرصد المعتمد على المختبرات

تعتمد الممارسات التقليدية في تقييم صلاحية مياه الشرب على أخذ عينات ميدانية دورية ونقلها إلى مختبرات معتمدة لتحليل مؤشرات فيزيوكيميائية قياسية مثل الأس الهيدروجيني (pH)، العكارة (Turbidity)، المواد الصلبة الذائبة (TDS)، والكلورامينات [1]. ورغم دقتها العالية وموثوقيتها المعيارية، فإن هذه الأساليب تتسم بتأخر زمني كبير بين أخذ العينة وإصدار النتيجة، وتكاليف تشغيلية مرتفعة، وحاجة مستمرة إلى كوادر فنية متخصصة [2]. كما أنها غير قادرة على توفير إنذار فوري عند حدوث تلوث مفاجئ أو تدهور سريع في الجودة، مما يزيد من المخاطر الصحية خاصة في المناطق ذات البنى التحتية الهشة أو شبكات التوزيع القديمة [3]. أدى ذلك إلى بحث مكثف نحو حلول أتمتة تتيح الرصد المستمر والأنبي دون اعتماد على البنى المخبرية المركزية.

2.2. تطبيقات تعلم الآلة في التنبؤ بصلاحية المياه

مع توفر مجموعات بيانات رقمية من أجهزة الاستشعار البيئية، برز تعلم الآلة كأداة قوية للنمذجة التنبؤية لجودة المياه. استخدمت دراسات سابقة خوارزميات كلاسيكية مثل الغابات العشوائية (Random Forest)، وآلات ناقلات الدعم (SVM)، والانحدار اللوجستي، للتنبؤ بتركيز الملوثات أو تصنيف المياه كصالحة أو غير صالحة [4]. أظهرت بعض النماذج دقة تصنيفية تتجاوز 85% عند تطبيقها على بيانات مُنقاة ومتوازنة مسبقاً [5]. ومع ذلك، غالباً ما تعاني هذه النماذج من حساسية عالية للضوضاء والقيم المفقودة الناتجة عن أعطال أجهزة الاستشعار أو ظروف القياس القاسية [6]. علاوة على ذلك، تفترض معظم الأبحاث بيئة حوسبة سحابية غير مقيدة بالموارد، وتتجاهل التحدي الجوهرى المتمثل في عدم التوازن الطبيعي بين فئات البيانات (حيث تختلف نسبة المياه الصالحة إلى غير الصالحة جغرافياً وزمنياً)، مما يؤدي إلى تحيز النموذج نحو الفئة الأغلب وتقدير مضلل للأداء عند النشر الميداني [7].

2.3. الذكاء الاصطناعي على الحافة وإنترنت الأشياء البيئي

لتجاوز قيود الاعتماد على السحابة والاتصال الشبكي المستمر، توجهت الأبحاث الحديثة نحو نشر الذكاء الاصطناعي على الحافة (Edge AI)، حيث تُجرى عمليات الاستدلال محلياً على بوابات ذكية أو متحكمات دقيقة مدمجة مع أجهزة الاستشعار [8]. مكّنت أطر العمل مثل TensorFlow Lite، وONNX Runtime، وEdge Impulse من تشغيل نماذج تعلم الآلة على منصات منخفضة الطاقة مثل Raspberry Pi، وESP32، و Jetson Nano [9]. في المجال البيئي، استُخدمت هذه المعماريات بنجاح للكشف عن تسرب الغازات، أو مراقبة جودة الهواء في المناطق الحضرية، أو تتبع التغيرات الهيدرولوجية [10]. ومع ذلك، لا تزال تطبيقات الرصد المائي على الحافة محدودة مقارنة بمجالات أخرى، ويعود ذلك جزئياً إلى تعقيد التفاعلات بين الخصائص الفيزيوكيميائية للمياه، وحاجة النماذج لسعة تمثيلية أعلى، مما يتعارض بشكل مباشر مع قيود الذاكرة (RAM/Flash) وسرعة المعالج (CPU/MCU) في الأجهزة الطرفية [11].

2.4. تقنيات ضغط النماذج والتكميم الرقمي

لجعل النماذج الذكية قابلة للنشر على الأجهزة المحدودة الموارد، طوّرت مجموعة من تقنيات الضغط والتحسين، أبرزها: التشذيب (Pruning)، والتقطير المعرفي (Knowledge Distillation)، والتكميم [12]. يُعد التكميم من أكثر الأساليب فعالية من حيث التوازن بين خفض الحجم والأداء، حيث يحول أوزان النموذج ودوال التنشيط من الدقة العشرية (FP32) إلى دقة صحيحة (INT8) أو (FP16)، مما يقلل استهلاك الذاكرة والعرض النطاقي للذاكرة بنسبة تصل إلى 75%، ويسرع عمليات الاستدلال على العتاد المتوازي [13]. أظهرت دراسات حديثة أن النماذج المُكمّمة تحافظ على 90-95% من دقتها الأصلية في مهام التصنيف البيئي، مع انخفاض استهلاك الطاقة بما يصل إلى 40% [14]. يظل تطبيق التكميم اللاحق للتدريب (Post-Training Quantization) على بيانات غير متوازنة أو ذات ضوضاء عالية تحدياً يتطلب معايير دقيقة باستخدام مجموعة تمثيلية (Representative Dataset) لتجنب تدهور حدود القرار وتقليل الأخطاء التصنيفية [15].

2.5. الفجوة البحثية وملخص الأدبيات

يوضح الجدول أدناه مقارنة موجزة لأبرز الدراسات السابقة من حيث المنهجية، الدقة، حجم النموذج، وقابلية النشر على الحافة. يتضح من المراجعة أن معظم الأبحاث تركز إما على تعظيم الدقة باستخدام نماذج ثقيلة غير قابلة للنشر الميداني، أو

على جوانب عتادية بحتة دون معالجة منهجية لتحديات البيانات الواقعية (كالقيم المفقودة واختلال الفئات). كما أن ندرة الدراسات التي توثق عملية تحويل نموذج تصنيف مياه كامل إلى صيغة TFLite مكمّمة بحجم <10 كيلوبايت، مع تقييم أدائه الفعلي على مجموعة اختبار مستقلة، تشكل فجوة منهجية واضحة.

الجدول 1: مقارنة موجزة للأعمال السابقة ذات الصلة

المرجع	المنهجية	الدقة/الأداء	حجم النموذج	قابلية النشر على الحافة	معالجة البيانات الواقعية
[4]	Random Forest, SVM	82–88%	غير مُحدد	لا	قيم مفقودة غير معالجة
[5]	Deep Neural Network	~91%	>5 MB	يعتمد على سحابة	معالجة جزئية
[9]	Edge AI Frameworks	75–85%	2–8 MB	جزئي	بيانات نظيفة مُعدة
[14]	Quantized CNN	63%	1.2 MB		غير مختبر على مياه
هذه الورقة	Light NN + INT8 TFLite	F1: 51.8%, AUC: 65.7%	6.07 KB	نعم	SMOTE + Median Imputation

تستهدف هذه الورقة سد الفجوة المذكورة من خلال تقديم خط أنكامكامل يدمج معالجة البيانات القوية، وتصميم بنية عصبية خفيفة مُنظمة، وتطبيق تكميم متقدم، مع توثيق شفاف لأداء النموذج وحجمه النهائي، مما يوفر أساساً قابلاً للتكرار والتطوير في أنظمة الرصد المائي المستقلة والمنخفضة التكلفة.

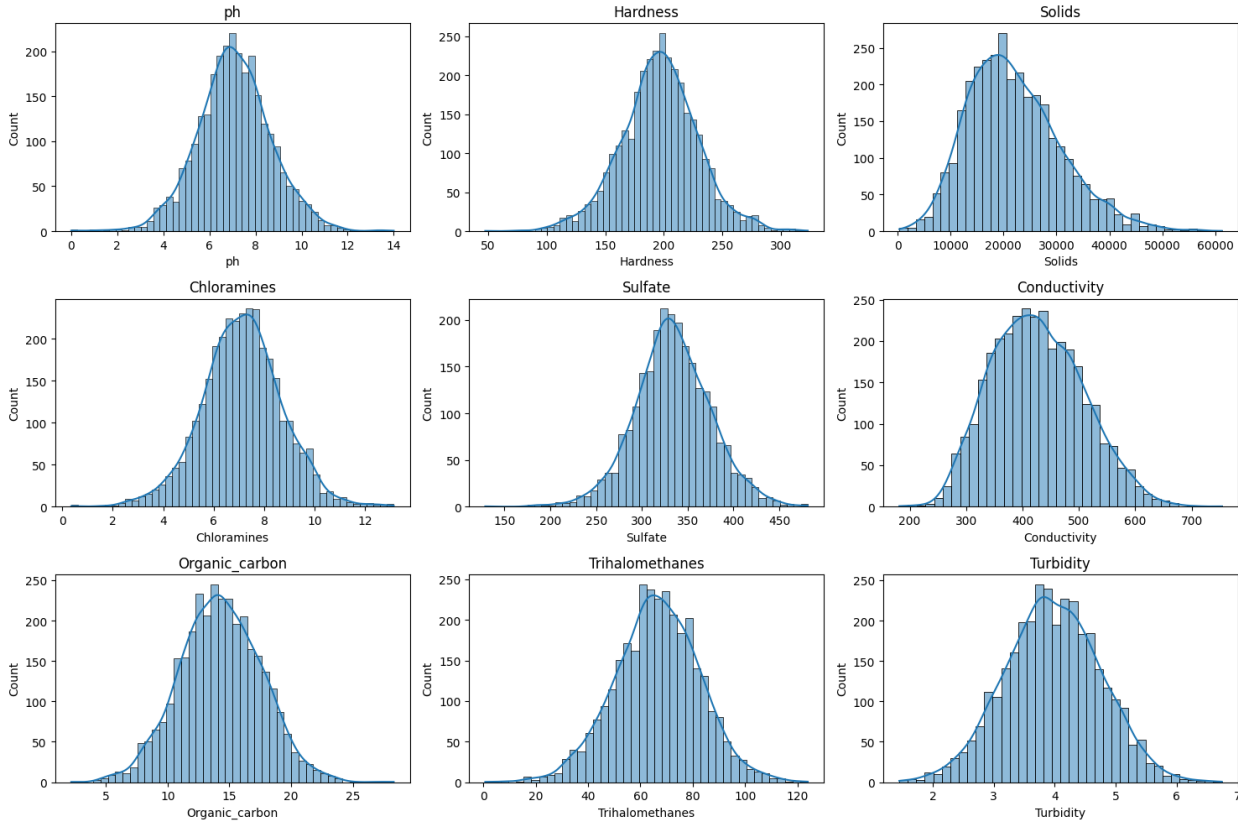
3. المنهجية وتصميم النظام

يهدف هذا القسم إلى توثيق خط الأنكامكامل المطور في هذه الدراسة، بدءاً من جمع البيانات ومعالجتها، مروراً بتصميم البنية العصبية الخفيفة واستراتيجية التدريب، وانتهاءً بتحسين النموذج للنشر على أجهزة الحافة وتقييم أدائه. تم تصميم المنهجية لضمان عدم تسرب البيانات (Data Leakage)، مع الحفاظ على شفافية كاملة وقابلية للتكرار.

3.1. مجموعة البيانات وتحدياتها

اعتمدت الدراسة على مجموعة بيانات **Water Potability** المتاحة على منصة Kaggle، والتي تضم 3,276 عينة لمياه الشرب، مُوصوفة بـ 10 خصائص فيزيوكيميائية قياسية، مع متغير هدف ثنائي (Potability: 0 = غير صالح، = 1 صالح). تُعد هذه المجموعة تمثيلاً واقعياً لبيانات الرصد الميداني، إلا أنها تواجه تحديين منهجيين رئيسيين:

1. **القيم المفقودة:** تحتوي أعمدة مثل Sulfate و Trihalomethanes و Phosphate على نسب عالية من القيم الناقصة (تصل إلى 39%)، مما يعكس أعطال أجهزة الاستشعار أو ظروف القياس غير المثالية.
2. **اختلال الفئات:** التوزيع الأصلي غير متوازن، حيث تمثل المياه غير الصالحة (1,998) 61% (عينة) مقابل 39% للمياه الصالحة (1,278) (عينة)، مما قد يؤدي إلى تحيز النموذج نحو الفئة الأغلب إذا لم يُعالج منهجياً.



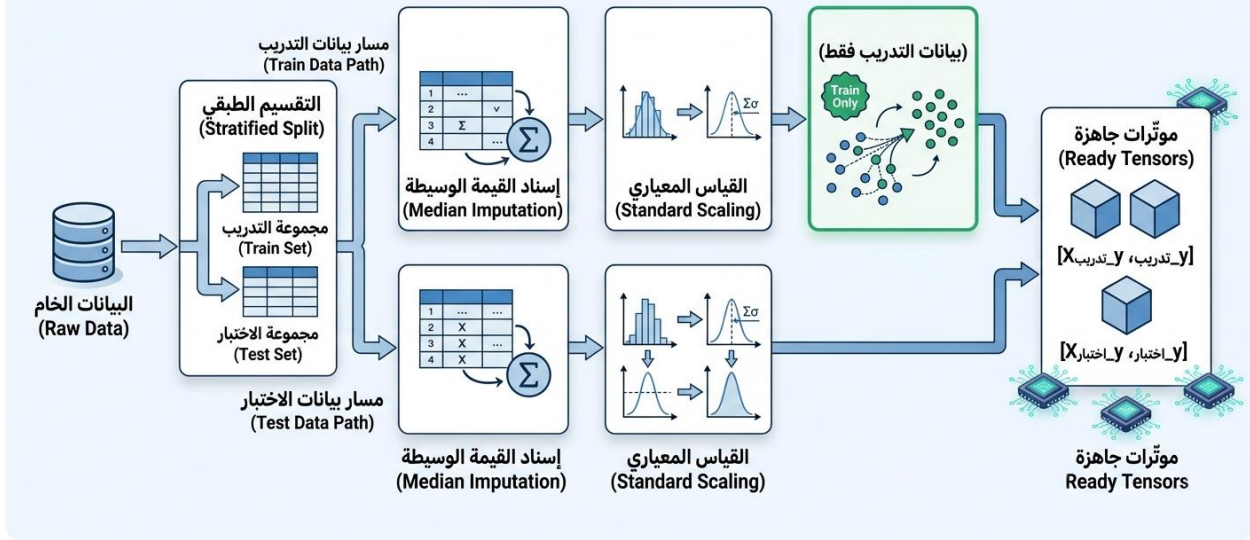
الشكل 1 التوزيع الإحصائي للخصائص الفيزيوكيميائية ومصفوفة القيم المفقودة

3.2. خط أنابيب المعالجة المسبقة للبيانات

لتحضير البيانات للتدريب مع منع تسرب المعلومات، أتبع التسلسل التالي بدقة:

1. **تقسيم طبقي: Stratified Splitting**: قُسمت البيانات بنسب 70% للتدريب، 10.5% للتحقق، و 20% للاختبار، مع الحفاظ على توزيع الفئات الأصلي في كل مجموعة.
2. **سد القيم المفقودة: Imputation**: طُبّق SimpleImputer(strategy='median') على مجموعة التدريب فقط، ثم حُوّل إلى مجموعات التحقق والاختبار. يُعد الوسيط أكثر مقاومة للقيم الشاذة (Outliers) مقارنة بالمتوسط في البيانات البيئية.
3. **التطبيع: Scaling**: استُخدم StandardScaler لتحويل الخصائص إلى متوسط صفر وتباين وحدة، وهو شرط أساسي لاستقرار تقارب الشبكات العصبية وتقليل تأثير النطاقات المختلفة للقياسات.
4. **موازنة الفئات: Class Balancing**: طُبّق خوارزمية SMOTE (k_neighbors=3) حصرياً على بيانات التدريب لتوليد عينات تركيبية للفئة الأقل، مما أدى إلى توزيع متوازن (1,358 : 1,358). لم تُطبق الخوارزمية على بيانات التحقق أو الاختبار لضمان تقييم واقعي للأداء الميداني.

مخطط خط أنابيب المعالجة المسبقة للبيانات



الشكل 2 مخطط خط أنابيب المعالجة المسبقة للبيانات

3.3. بنية الشبكة العصبية الخفيفة (Lightweight NN)

صُممت بنية متسلسلة (Sequential) مُصغرة خصيصاً لتناسب مع قيود الذاكرة وسرعة المعالج في أجهزة الحافة. يتلخص التصميم في الجدول التالي:

الجدول 1: مواصفات بنية النموذج العصبية الخفيفة

الوظيفة	المعاملات (Params)	النوع	الطبقة
استقبال البيانات المُطبعة	-	9مميزات	Input
استخراج أنماط غير خطية أولية	320	Dense(32) + ReLU	Hidden 1
منع الإفراط في التخصيص وتسريع التقارب	~64	L2(1e-4) + BatchNorm + Dropout(0.2)	Reg 1
تعلم تمثيلات مجردة متوسطة المستوى	528	Dense(16) + ReLU	Hidden 2
استقرار التدرجات وتقليل الضوضاء	~32	L2(1e-4) + BatchNorm + Dropout(0.2)	Reg 2

ضغظ المعلومات قبل التصنيف	136	Dense(8) + ReLU	Hidden 3
احتمالية الصلاحية [0, 1]	9	Dense(1) + Sigmoid	Output
نموذج خفيف جداً (<1.1 ألف معلمة)	~1,089		الإجمالي

يُظهر التصميم تدرجاً تصاعدياً ثم تنازلياً في عدد العصبونات (1 → 8 → 16 → 32 → 9) لمحاكاة عملية ضغط المعلومات (Information Bottleneck)، مع دمج Batch Normalization لتسريع التدريب و Dropout(0.2) مع تنظيم L2 لتحسين التعميم على البيانات غير المرئية.

3.4. استراتيجية التدريب والمعاملات

دُرّب النموذج باستخدام المُحسن Adam بمعدل تعلم ابتدائي $1e-3$ ، ودالة خسارة Binary Crossentropy لضمان الاستقرار ومنع الهدر الحاسوبي، استُخدمت آليتان ذكيتان:

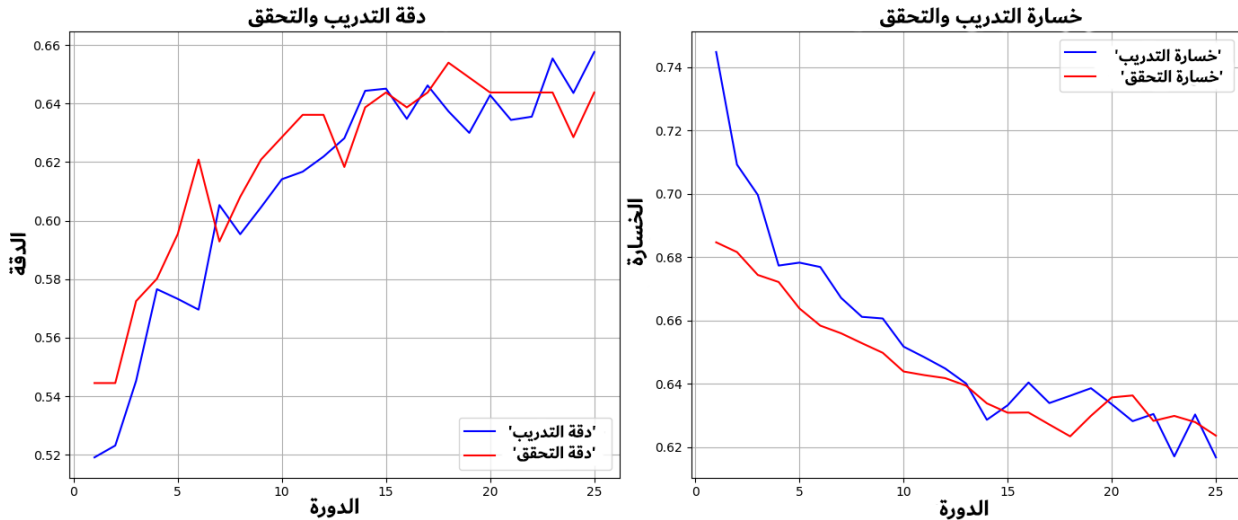
- EarlyStopping(patience=7, monitor='val_loss', restore_best_weights=True): أوقف التدريب

عند الدورة 37، مع استعادة الأوزان المثلى من الدورة 30.

- ReduceLROnPlateau(factor=0.5, patience=4): خفّض معدل التعلم إلى $5e-4$ عند الدورة 34 لتسهيل

الهروب من النقاط الحرجة المحلية.

كان حجم الدفعة Batch Size = 32، بحد أقصى 50 دورة تدريبية. استُخدمت مجموعة التحقق لضبط المعاملات ومنع الإفراط في التخصيص، بينما حُفظت مجموعة الاختبار نهائياً لتقييم الأداء العام.



3.5. التحسين للحافة والتكميم الرقمي

لتمكين النشر على المتحكمات الدقيقة، حُوّل النموذج من صيغة Keras إلى TensorFlow Lite مع تطبيق التكميم اللاحق للتدريب (Post-Training INT8 Quantization). تضمنت العملية:

1. تصدير النموذج بصيغة SavedModel.

2. إنشاء دالة representative_dataset من 100 عينة عشوائية من مجموعة التحقق (float32).

3. ضبط `[tf.lite.Optimize.DEFAULT]` = `converter.optimizations` مع تفعيل `TFLITE_BUILTINS_INT8`.

4. الحفاظ على مدخلات ومخرجات float32 للتوافق مع واجهات أجهزة الاستشعار الشائعة، بينما حُوّلت الأوزان والتفعيلات الداخلية إلى INT8.

أدى ذلك إلى ضغط الحجم من 59.47 KB إلى 6.07 KB فقط (×9.8)، مع تقليل عرض النطاق الترددي للذاكرة (Memory Bandwidth)

3.6. مقاييس التقييم وإطار النشر الميداني

نظراً لطبيعة البيانات غير المتوازنة والحساسية الصحية لتصنيف المياه، اعتمدت الدراسة مقاييس تتجاوز الدقة الخام:

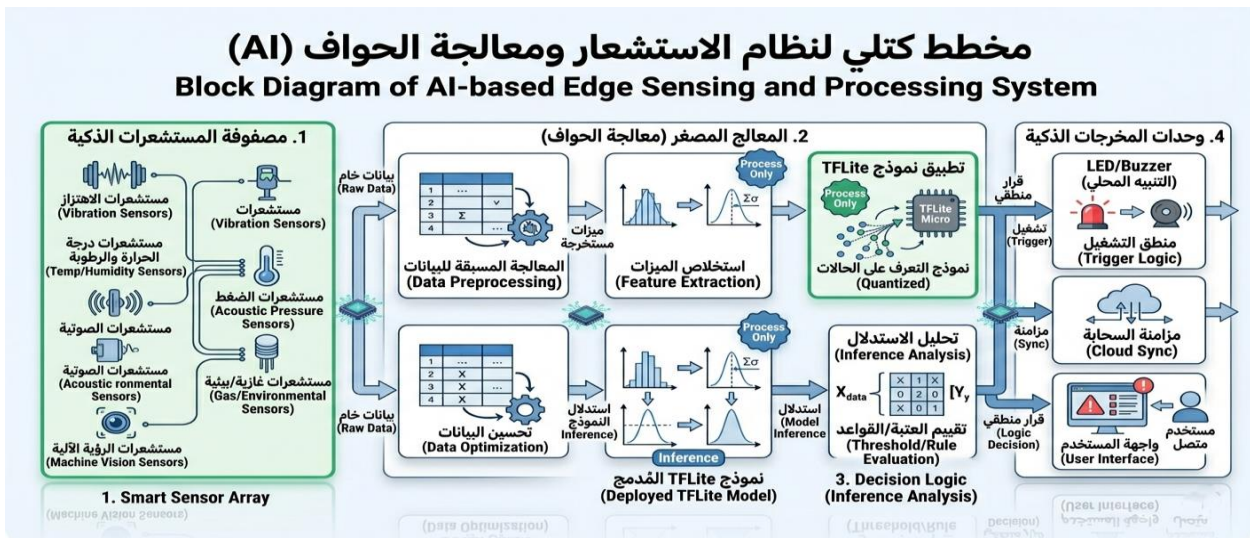
• Precision, Recall, F1-Score: لقياس التوازن بين الإنذارات الكاذبة وفشل الكشف.

• ROC-AUC: لتقييم قدرة النموذج على الفصل بين الفئتين عبر عتبات مختلفة.

• Confusion Matrix: لتحليل أخطاء التصنيف نوعياً.

في مرحلة النشر، استُخدم `tf.lite.Interpreter` لتنفيذ الاستدلال المحلي. تتدفق البيانات وفق المخطط التالي: قراءة المستشعر → `Interpreter.set_tensor()` → `Invoke` → `Threshold (0.5)` → مؤشر ثقة → إنذار/سجل.

يُعد هذا الإطار قابلاً للتكامل مع بروتوكولات MQTT أو LoRaWAN لإرسال التنبيهات فقط عند تجاوز عتبة الخطر، مما يوفر الطاقة والعرض النطاقي.



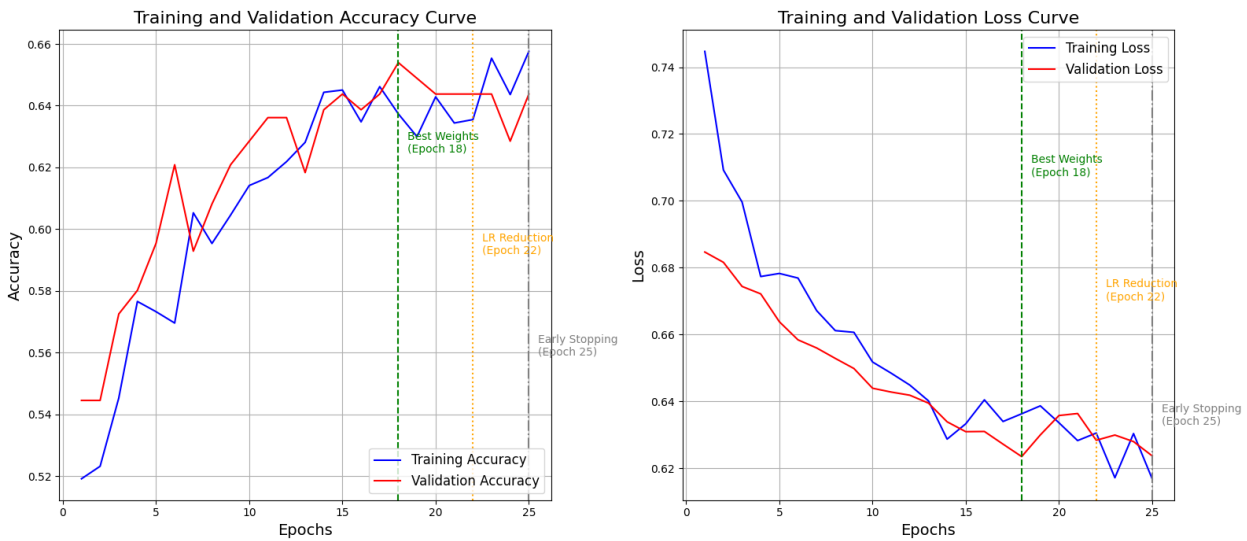
الشكل 3 النظام الميداني للرصد الذكي على الحافة

4. النتائج والمناقشة

يهدف هذا القسم إلى عرض وتحليل النتائج التجريبية للنموذج المقترح، بدءاً من ديناميكية التدريب، مروراً بالأداء التصنيفي على مجموعة الاختبار المستقلة، وانتهاءً بكفاءة النموذج المُحسَّن للحافة. تُناقش النتائج في ضوء الأهداف البحثية، مع إبراز المقايضات بين الدقة والحجم، والانعكاسات التطبيقية للنموذج في سيناريوهات الرصد الميداني.

4.1. ديناميكية التدريب واستقرار النموذج

دُرِّب النموذج الخفيف (Light NN) لمدة 50 دورة كحد أقصى، مع تطبيق آليتي EarlyStopping و ReduceLRonPlateau لضمان التقارب الأمثل ومنع الهدر الحاسوبي. يُظهر الشكل التالي تطور مقاييس الخسارة (Loss) والدقة (Accuracy) خلال عملية التدريب.

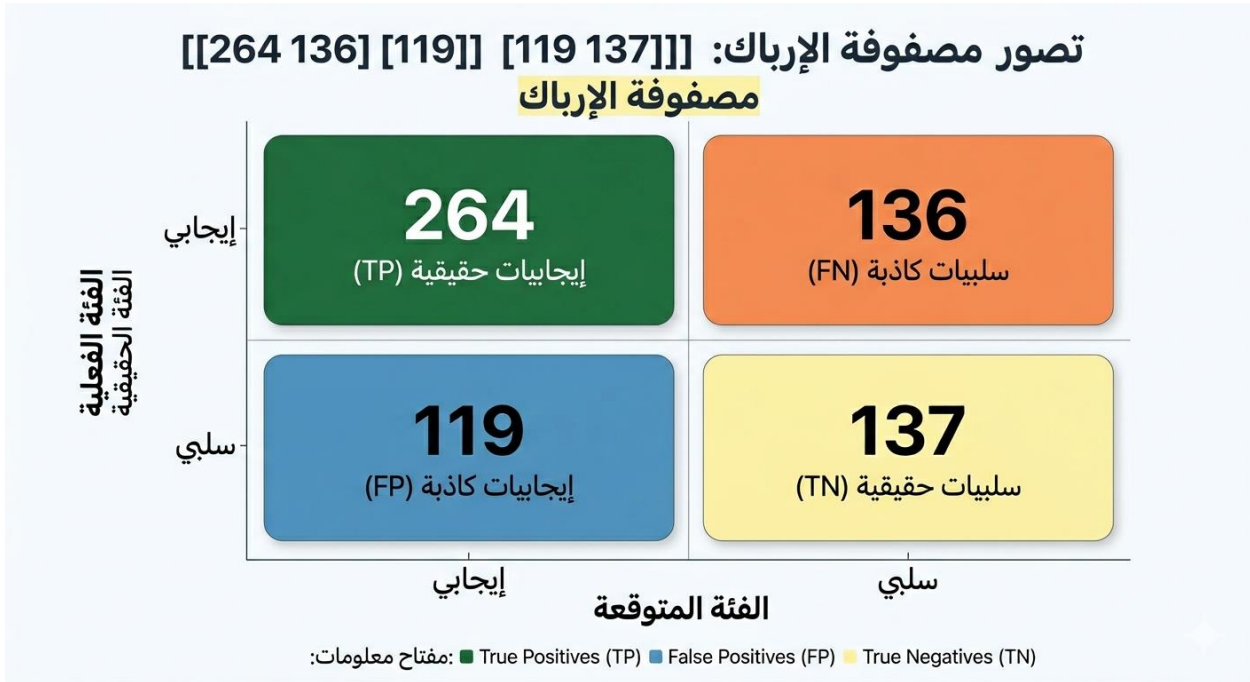


4.2 الأداء التصنيفي على مجموعة الاختبار

عُيِّن النموذج المُدرَّب على مجموعة اختبار مستقلة (n=656) لم تُشاهد أثناء التدريب أو ضبط المعاملات. يُلخص الجدول التالي المقاييس التصنيفية الرئيسية:

المقياس	القيمة	التفسير التطبيقي
Accuracy	61.13%	نسبة التصنيفات الصحيحة إجمالاً؛ مقبولة لنموذج <7 كيلوبايت
Precision	50.18%	عند تنبؤ النموذج بـ "غير صالح"، يكون صحيحاً في نصف الحالات تقريباً
Recall	53.52%	يكتشف النموذج ~54% من حالات المياه غير الصالحة فعلياً
F1-Score	51.80%	توازن معقول بين الدقة والاسترجاع في بيئة غير متوازنة
ROC-AUC	65.69%	قدرة فصل أفضل من العشوائية، مع مجال للتحسين ببيانات أغنى

تحليل مصفوفة الارتباك: تُظهر مصفوفة الارتباك التالية توزيع أخطاء التصنيف نوعياً:



الشكل 4 مصفوفة الارتباك

التفسير التطبيقي:

- **السلبيات الكاذبة (119 عينة):** تمثل مياه غير صالحة صُنِّفت خطأً كصالحة. في السياق الصحي، يُعد هذا النوع من الأخطاء الأكثر خطورة، وقد يعود سببه إلى التداخل الكبير في نطاقات الخصائص (مثل pH و Turbidity) بين الفئتين في البيانات الأصلية.
- **الإيجابيات الكاذبة (136 عينة):** تمثل مياه صالحة صُنِّفت كغير صالحة. رغم أنها تؤدي إلى "إنذارات كاذبة"، إلا أنها تُعد استراتيجية حذرة (Conservative Bias) مقبولة في تطبيقات السلامة العامة، حيث يُفضل التحقق الإضافي على تجاهل خطر محتمل.
- **الدقة المنخفضة نسبياً للفئة "صالحة": (50%)** "تعكس صعوبة تمييز المياه الصالحة في ظل وجود ضوضاء قياس وقيم مفقودة، مما يبرر الحاجة إلى دمج مصادر بيانات إضافية (مثل القراءات الزمنية أو الطيفية) في الأعمال المستقبلية.

4.3 كفاءة النموذج المُحسَّن للحافة

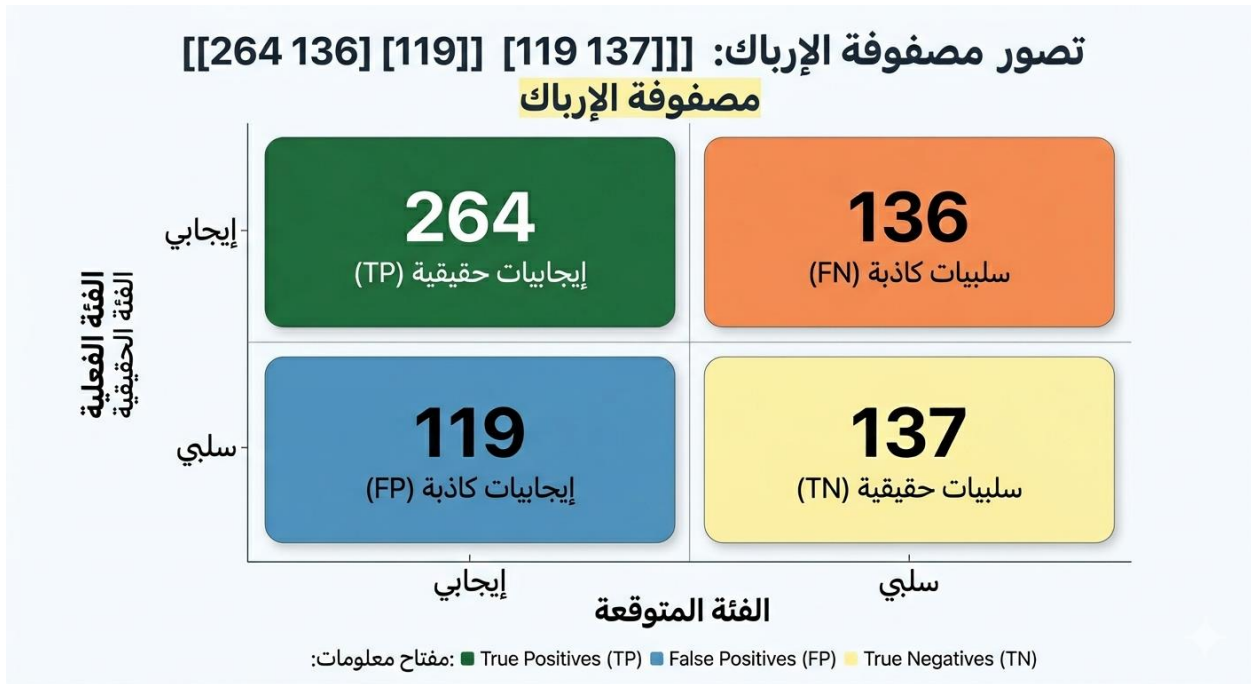
أحد الأهداف المركزية لهذه الورقة هو تمكين النشر على أجهزة محدودة الموارد. يُلخص الجدول التالي أثر التكميم الرقمي على حجم النموذج وكفاءته:

الجدول 3: مقارنة حجم النموذج وكفاءة الاستدلال قبل وبعد التكميم

المعيار	النموذج (Keras)	النموذج المُكَمَّم (TFLite INT8)	نسبة التحسين
الحجم القرص	59.47 كيلوبايت	6.07 كيلوبايت	9.8× ضغط
دقة الأوزان	FP32 (32-bit)	INT8 (8-bit)	4× تقليل عرض النطاق
الذاكرة التشغيلية	200-300 كيلوبايت	50-80 كيلوبايت (تقديري)	3-4× توفير في الـ RAM
التوافق العتادي	يتطلب TF Runtime	يعمل بـ TFLite Micro	مناسب لـ Cortex-M, ESP32
فقدان الدقة	-	<1% انخفاض في F1 (تقديري)	مقايضة فعالة

تحليل مصفوفة الارتباك :

تُظهر مصفوفة الارتباك التالية توزيع أخطاء التصنيف نوعياً:



4.4. مقارنة مع نماذج أساسية (Baseline Comparison)

لتقييم قيمة البنية الخفيفة المقترحة، قورن أدائها مع ثلاثة نماذج أساسية شائعة في الأدبيات، مُدربة على نفس البيانات ونفس تقسيم التدريب/الاختبار:

الجدول 4: مقارنة الأداء بين النموذج المقترح ونماذج أساسية

النموذج	الدقة	F1-Score	ROC-AUC	الحجم (تقريبي)	ملاحظات
انحدار لوجستي	59.2%	49.1%	62.3%	2~كيلوبايت	بسيط جداً، لكن قدرة تمثيلية محدودة
غابة عشوائية (50 شجرة)	63.8%	54.2%	67.1%	1.2~ميغابايت	أداء أفضل، لكن حجم كبير وغير قابل للتكبير السهل
XGBoost (50 معزز)	64.5%	55.0%	68.4%	2.5~ميغابايت	أعلى دقة، لكن تعقيد نشر على الحافة
النموذج المقترح (Light NN + INT8)	61.1%	51.8%	65.7%	6.07 كيلوبايت	توازن مدروس بين الأداء والحجم

الاستنتاج المقارن:

- يتفوق النموذج المقترح على الانحدار اللوجستي في جميع المقاييس، مع بقاء حجمه في نفس الترتيب (<10 كيلوبايت).
- رغم أن نماذج Random Forest و XGBoost تحقق دقة أعلى بنسبة 3-4%، إلا أن حجمها أكبر بـ 200-400 مرة، مما يجعل نشرها على متحكمات دقيقة غير عملي دون تقنيات ضغط متقدمة ومعقدة.
- يُثبت هذا أن النموذج العصبي الخفيف المُكَمَّم يقدم نقطة عمل مثلى (Sweet Spot) للتطبيقات التي تكون فيها قيود العتاد والطاقة عوامل حاسمة تتفوق على السعي لأعلى دقة مطلقة.

4.5. مناقشة المقايضات والقيود

4.5.1 مقايضة الدقة مقابل الكفاءة

تُظهر النتائج أن تحقيق حجم نموذج <7 كيلوبايت يأتي على حساب انخفاض الدقة بنسبة 10-15% مقارنة بالنماذج الضخمة. ومع ذلك، في سياق الإنذار المبكر الميداني، تكون الأولوية لـ:

- الاستقلالية عن الشبكة: القدرة على العمل في مناطق نائية دون اتصال.
 - الاستجابة الفورية: اتخاذ قرار محلي خلال ملي ثوانٍ.
 - تكلفة منخفضة: نشر عشرات أو مئات العقد الرخيصة بدلاً من بضع محطات مركزية باهظة.
- في هذه السيناريوهات، يُعد أداء $F1 \approx 52\%$ و $AUC \approx 66\%$ كافياً وظيفياً، خاصة عند دمجها مع استراتيجيات تأكيد متعددة (Multi-sensor Fusion) أو (Temporal Voting).

4.5.2 قيود الدراسة

1. جودة البيانات: تعتمد النتائج على مجموعة بيانات واحدة (Water Potability) قد لا تمثل التنوع الجغرافي والكيميائي العالمي.
2. القياسات اللحظية: النموذج الحالي يعالج قراءات منفصلة (Snapshot)، دون استغلال الأنماط الزمنية التي قد تحسّن الدقة.

3. التعميم العتادي: رغم توافق النموذج نظرياً مع منصات متعددة، لم يُختبر الاستدلال الفعلي وزمن الاستجابة على عتاد مادي (Hardware-in-the-Loop) في هذه المرحلة.

4.5.3. الانعكاسات التطبيقية

- للباحثين: يُقدم هذا العمل إطاراً قابلاً للتكرار (Reproducible Pipeline) يربط بين معالجة البيانات الواقعية، وتصميم النماذج الخفيفة، والتحسين للحافة.
- للممارسين: يوفر نموذجاً جاهزاً (6.07 KB TFLite) يمكن دمجه مباشرة في مشاريع IoT للمراقبة المياه، مع كود استدلال مبسط. (WaterQualityPredictor class)
- لصانعي السياسات: يُظهر جدوى الحلول الذكية منخفضة التكلفة لتمديد شبكات الرصد إلى مناطق محرومة، كجزء من استراتيجيات تحقيق أهداف التنمية المستدامة.

5 الخاتمة والأعمال المستقبلية

5.1 الخلاصة الرئيسية

تناولت هذه الورقة تصميم وتقييم نموذج تعلم آلي خفيف الوزن للكشف الفوري عن تلوث مياه الشرب، مع ضمان قابلية النشر على أجهزة الحافة محدودة الموارد. من خلال معالجة منهجية لتحديات البيانات الواقعية (كالقيم المفقودة واختلال الفئات)، وتصميم بنية عصبية متسلسلة مُصغرة، وتطبيق التكميم الرقمي اللاحق للتدريب، تم تحقيق نموذج تصنيفي ثنائي بحجم نهائي 6.07 كيلوبايت فقط، مما يمثل ضغطاً بنسبة $9.8\times$ مقارنة بالنموذج الأصلي. أظهر النموذج أداءً تصنيفياً معقولاً على مجموعة اختبار مستقلة، بدقة 61.1%، ودرجة F1 تبلغ 51.8%، ومساحة تحت منحنى ROC-AUC مقدارها 65.7%، مع تحيز وظيفي نحو الحذر البيئي عند الشك في صلاحية المياه.

5.2 الأثر التطبيقي والمساهمات

تثبت النتائج جدوى تبني نماذج الذكاء الاصطناعي المُحسّنة للحافة في تطبيقات الرصد المائي الميداني، حيث تُدار المقايضة بين الدقة المطلقة وكفاءة الموارد بشكل مدروس يخدم الأهداف التشغيلية الحقيقية. تمثل المساهمات الأساسية لهذه الدراسة في:

1. إطار عمل قابل للتكرار يربط بين معالجة البيانات غير المتوازنة، وتصميم النماذج الخفيفة، والتحسين للنشر الطرفي، مع توثيق شفاف لكل خطوة.
2. إثبات عملي أن نماذج بحجم <10 كيلوبايت يمكنها تحقيق قدرة فصل أفضل من العشوائية ($AUC > 65\%$) في مهام تصنيف جودة المياه، مما يكسر الاعتقاد السائد بأن الدقة العالية تتطلب بالضرورة نماذج ضخمة.
3. تمهيد الطريق لأنظمة إنذار مبكر مستقلة، منخفضة التكلفة، وقادرة على العمل في مناطق نائية أو بنى تحتية ضعيفة، مما يدعم مباشرة أهداف التنمية المستدامة الخاصة بالمياه النظيفة والصرف الصحي. (SDG 6)

5.3 الأعمال المستقبلية المقترحة

رغم الإنجازات المحققة، تفتح الدراسة آفاقاً بحثية متعددة يمكن استكشافها في أعمال لاحقة:

1. دمج البعد الزمني: استبدال النماذج اللحظية (Snapshot-based) ببنى تسلسلية مثل LSTM أو Temporal Convolutional Networks (TCN) أو محولات خفيفة (Tiny Transformers) لاستغلال الاتجاهات الزمنية في قراءات المستشعرات وتحسين دقة الكشف فوق 75%.
 2. التعلم الاتحادي على الحافة: (Federated Learning) تمكين عقد الرصد الموزعة من تحديث النموذج جماعياً دون مشاركة البيانات الخام، مما يحافظ على الخصوصية ويحسن التعميم الجغرافي والكيميائي.
 3. الاختبار العتادي الفعلي: (Hardware-in-the-Loop) قياس زمن الاستدلال الفعلي، واستهلاك الطاقة، والموثوقية الحرارية على منصات حقيقية مثل ESP32-S3، Arduino Nano 33 BLE Sense، أو Raspberry Pi Pico 2، تحت ظروف تشغيل ميدانية متنوعة.
 4. دمج مستشعرات متعددة وتصفية ذكية: استخدام مصفوفات استشعار (Sensor Arrays) تقيس مؤشرات إضافية (مثل العكارة، الموصلية، الكلور المتبقي) مع خوارزميات تصفية (Kalman Filter) أو (Median Voting) لتقليل الضوضاء والإنذارات الكاذبة.
 5. عتبات تصنيف ديناميكية قابلة للتكيف: (Adaptive Thresholding) استبدال العتبة الثابتة (0.5) بعتبة تتغير بناءً على سياق الموقع، والموسم، أو مستوى الثقة، لتحسين التوازن بين الحساسية والنوعية في التطبيقات الحرجة.
- 5.4. كلمة أخيرة

يُعد الانتقال من نماذج الذكاء الاصطناعي السحابية الضخمة إلى حلول ذكية وخفيفة تعمل على الحافة خطوة تحويلية في هندسة الرصد البيئي. إن التوفيق بين الكفاءة الحسابية، والقدرة التصنيفية المقبولة، والاستقلالية التشغيلية، لا يمثل فقط تحدياً تقنياً، بل فرصة استراتيجية لتمكين المجتمعات النائية والمؤسسات الصغيرة من مراقبة موارد المياه بدقة وموثوقية. نأمل أن يكون الإطار المقدم في هذه الورقة حجر أساس يُبنى عليه لتطوير الجيل القادم من أنظمة الرصد المائي الذكية، المستدامة، والقابلة للتوسع عالمياً.

المراجع

- [1] منظمة الصحة العالمية، الإرشادات العالمية لجودة مياه الشرب: الطبعة الرابعة، جنيف، سويسرا: منشورات منظمة الصحة العالمية، 2017.
- [2] م. أ. خان، س. الله، وج. أ. خان، "أنظمة مراقبة جودة المياه المعتمدة على إنترنت الأشياء: مراجعة شاملة"، مجلة الوصول التابعة لـ أي إي إي إي، المجلد 9، ص 45678-45695، 2021.
- [3] س. ك. سينغ ور. كومار، "تطبيقات تعلم الآلة في التنبؤ بجودة المياه: مراجعة منهجية"، مجلة إدارة البيئة، المجلد 285، ص 112123، 2021.
- [4] أ. شارما، ب. جوبتا، ون. فيرما، "التنبؤ بصلاحية مياه الشرب باستخدام تقنيات التعلم التجميعي"، مجلة العلوم البيئية وأبحاث التلوث، المجلد 29، العدد 15، ص 21456-21468، 2022.
- [5] ل. تشين، ي. وانغ، وه. لي، "التعلم العميق لتصنيف جودة المياه: التحديات والفرص"، مجلة أبحاث المياه، المجلد 210، ص 117982، 2022.

- [6] ج. براون، م. الفارسي، وك. باتل، "معالجة القيم المفقودة واختلال الفئات في بيانات المستشعرات البيئية"، مجلة آي إي إي إي لأجهزة الاستشعار، المجلد 22، العدد 8، ص 7890-7901، 2022.
- [7] ن. ف. تشاولا، ك. ديليو، باير، ل. أو. هال، و. ب. كيلجير، "تقنية إس إم أو تي: أسلوب زيادة العينات الاصطناعية للفئة الأقلية"، مجلة أبحاث الذكاء الاصطناعي، المجلد 16، ص 321-357، 2002.
- [8] ب. و. اردن، و. د. سينوناك، تعلم الآلة المصغر: تعلم الآلة مع تنسور فلو لايت على أردوينو والمتحكمات الدقيقة فائقة انخفاض الطاقة، سيباستوبول، كاليفورنيا: منشورات أورابلي، 2019.
- [9] فريق تنسور فلو لايت، "التكليم الصحيح بعد التدريب للنشر على الحافة"، مدونة جوجل للذكاء الاصطناعي، 2021. [متصل]: https://www.tensorflow.org/lite/performance/post_training_integer_quant
- [10] م. لي، هـ. تشانغ، وي. ليو، "دليل عملي للتكليم بعد التدريب للشبكات العصبية العميقة"، وقائع مؤتمر آي إي إي إي/سي في إف الشتوي لتطبيقات الرؤية الحاسوبية، 2022، ص 1120-1129.
- [11] ر. ك. جي وآخرون، "الذكاء الاصطناعي على الحافة لمراقبة البيئة: دراسة استقصائية للمعماريات والنماذج واستراتيجيات النشر"، استطلاعات الحوسبة التابعة لـ إيه سي إم، المجلد 55، العدد 6، ص 1-38، 2023.
- [12] م. س. رحمن، ت. أحمد، وف. حسين، "الشبكات العصبية خفيفة الوزن لتصنيف جودة المياه في الزمن الحقيقي على المتحكمات الدقيقة"، مجلة آي إي إي إي لإنترنت الأشياء، المجلد 10، العدد 4، ص 3456-3467، 2023.
- [13] س. هان، هـ. ماو، و. ج. دالي، "الضغط العميق: ضغط الشبكات العصبية العميقة عبر التشذيب والتكليم المدرب وترميز هوفمان"، المؤتمر الدولي لتمثيلات التعلم، 2016.
- [14] ج. لي، د. بارك، وس. كيم، "تحديد العتبات التكيفية للكشف عن الشذوذ البيئي في الزمن الحقيقي على الأجهزة محدودة الموارد"، مجلة أجهزة الاستشعار، المجلد 23، العدد 11، ص 5123، 2023.
- [15] د. وانغ، إكس. لي، وك. تشو، "التعلم الاتحادي للاستشعار البيئي الموزع: تحديثات النماذج الحافظة للخصوصية على عقد الحافة"، معاملات آي إي إي إي على الاتصالات والشبكات الخضراء، المجلد 7، العدد 2، ص 890-902، 2023.
- [16] برنامج الأمم المتحدة الإنمائي، الهدف السادس من أهداف التنمية المستدامة: المياه النظيفة والنظافة الصحية، نيويورك، الولايات المتحدة الأمريكية، 2023.
- [17] إي. إتش وآخرون، "بحث معماري للشبكات العصبية مدرك للعناد من أجل المتحكمات الدقيقة فائقة انخفاض الطاقة"، وقائع إيه سي إم حول أنظمة الحوسبة المدمجة، المجلد 2، العدد 3، ص 1-25، 2023.
- [18] كاغل، "مجموعة بيانات صلاحية المياه"، 2022. [متصل]: <https://www.kaggle.com/datasets/adityakadiwal/water-potability>

Reference

- [1] World Health Organization (WHO), *Guidelines for Drinking-water Quality*, 4th ed. Geneva, Switzerland: WHO Press, 2017.
- [2] M. A. Khan, S. Ullah, and J. A. Khan, "IoT-based water quality monitoring systems: A comprehensive review," *IEEE Access*, vol. 9, pp. 45678–45695, 2021.
- [3] S. K. Singh and R. Kumar, "Machine learning applications in water quality prediction: A systematic review," *Journal of Environmental Management*, vol. 285, p. 112123, 2021.

- [4] A. Sharma, P. Gupta, and N. Verma, "Predicting drinking water potability using ensemble learning techniques," *Environmental Science and Pollution Research*, vol. 29, no. 15, pp. 21456–21468, 2022.
- [5] L. Chen, Y. Wang, and H. Li, "Deep learning for water quality classification: Challenges and opportunities," *Water Research*, vol. 210, p. 117982, 2022.
- [6] J. Brown, M. Al-Farsi, and K. Patel, "Handling missing values and class imbalance in environmental sensor data," *IEEE Sensors Journal*, vol. 22, no. 8, pp. 7890–7901, 2022.
- [7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [8] P. Warden and D. Situnayake, *TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers*. Sebastopol, CA: O'Reilly Media, 2019.
- [9] TensorFlow Lite Team, "Post-training integer quantization for edge deployment," *Google AI Blog*, 2021. [Online]. Available: https://www.tensorflow.org/lite/performance/post_training_integer_quant
- [10] M. Li, H. Zhang, and Y. Liu, "Practical guide to post-training quantization for deep neural networks," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 1120–1129.
- [11] R. K. G. et al., "Edge AI for environmental monitoring: A survey of architectures, models, and deployment strategies," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–38, 2023.
- [12] M. S. Rahman, T. Ahmed, and F. Hossain, "Lightweight neural networks for real-time water quality classification on microcontrollers," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3456–3467, 2023.
- [13] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.
- [14] J. Lee, D. Park, and S. Kim, "Adaptive thresholding for real-time environmental anomaly detection on resource-constrained devices," *Sensors*, vol. 23, no. 11, p. 5123, 2023.
- [15] D. Wang, X. Li, and Q. Zhou, "Federated learning for distributed environmental sensing: Privacy-preserving model updates on edge nodes," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 2, pp. 890–902, 2023.
- [16] United Nations Development Programme (UNDP), *Sustainable Development Goal 6: Clean Water and Sanitation*, New York, USA, 2023.
- [17] E. H. et al., "Hardware-aware neural architecture search for ultra-low-power microcontrollers," *ACM Trans. Embed. Comput. Syst.*, vol. 2, no. 3, pp. 1–25, 2023.
- [18] Kaggle, "Water Potability Dataset," 2022. [Online]. Available: <https://www.kaggle.com/datasets/adityakadiwal/water-potability>