计算机与智能教育技术

2026年9月 第1卷第2期

智能算法辅助SERS环境痕量污染物检测进展

周天宇，游春霞*

江苏师范大学物理与光电工程学院，江苏徐州，221116

摘　要 | 环境水体中痕量污染物具有浓度低、组分复杂和基质干扰明显等特点，复杂SERS光谱的可靠解析成为快速检测中的关键问题。本文围绕智能算法辅助SERS环境检测，梳理机器学习、深度学习及迁移学习在污染物识别、浓度预测和多组分解析中的应用，并比较不同方法的适用边界。现有研究表明，PCA、SVM、RF、PLSR及CNN等方法可利用全谱信息改善分类和定量分析，基于纯组分光谱的迁移策略也有望减少混合样本建库需求，但真实基质、基底与仪器差异、数据泄漏及模型泛化仍限制实际应用。智能算法能够拓展SERS复杂光谱分析能力，后续需加强标准化数据获取、独立外部验证、未知类拒识及便携筛查与实验室确证的衔接。

关键词 | 表面增强拉曼光谱；环境痕量污染物；机器学习；深度学习；迁移学习

Copyright © 2026 by the author(s).

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

https://creativecommons.org/licenses/by-nc/4.0/

1 引言

环境水体中的农药、抗生素、染料和多环芳烃等污染物往往以低浓度、多组分形式存在，其迁移转化和长期暴露效应使水质监测既要关注检测灵敏度，也要兼顾复杂基质中的识别可靠性[1-2]。气相色谱、液相色谱及其与质谱联用技术仍是确证分析的重要手段，但样品萃取、净化和富集过程通常耗时[3]。因此，现场筛查和污染热点定位仍需要能够快速提供分子信息的补充技术。

表面增强拉曼散射（surface-enhanced Raman scattering，SERS）借助金、银等纳米结构附近的局域电磁场放大拉曼信号，同时保留分子振动指纹，具有用样量小、响应快和适于多目标识别的特点[4-6]。但基底热点分布、分子竞争吸附、荧光背景以及仪器和批次差异均会改变谱峰强度或形状，多种污染物共存时的峰重叠又会削弱人工选峰和单峰校准的稳定性。

智能算法的作用由此落在光谱解释环节：利用全谱变量完成类别判别、浓度回归或混合组分解析，而不是替代SERS基底与实验过程。PCA、SVM、RF和PLSR适合样本量有限、变量共线明显的光谱数据；CNN等深度模型则可从原始或轻度预处理光谱中学习局部谱形特征[7-8]。

随着检测对象由单一污染物转向多组分体系，常规监督学习需要覆盖大量组分组合和浓度比例。Tan等[9]提出的迁移学习辅助SERS框架仅以纯化合物光谱训练随机森林模型，再识别并定量未参与训练的混合物，为减少混合样本建库需求提供了方法

基金项目：江苏省大学生创新创业训练计划项目（S202610320059）。

作者简介：周天宇，江苏师范大学学生，本科；游春霞（通讯作者），江苏师范大学物理与光电工程学院副教授，西澳大学访问学者，研究方向为可见光通信和智能信号处理。

文章引用：周天宇, 游春霞. 智能算法辅助 SERS 环境痕量污染物检测进展[J]. 计算机与智能教育技术, 2026, 1(2): 45-55.

https://doi.org/10.68070/f53x4r53

思路；但其受控肉碱体系仍需在环境污染物和真实水样中进一步验证。

本文围绕SERS环境污染检测中的光谱获取、智能分析和结果验证，综述传统机器学习、深度学习及纯组分迁移方法在定性识别、定量预测和多组分解析中的应用，并讨论数据标准化、模型泛化、解释性及便携检测等问题，整体研究框架见图1。

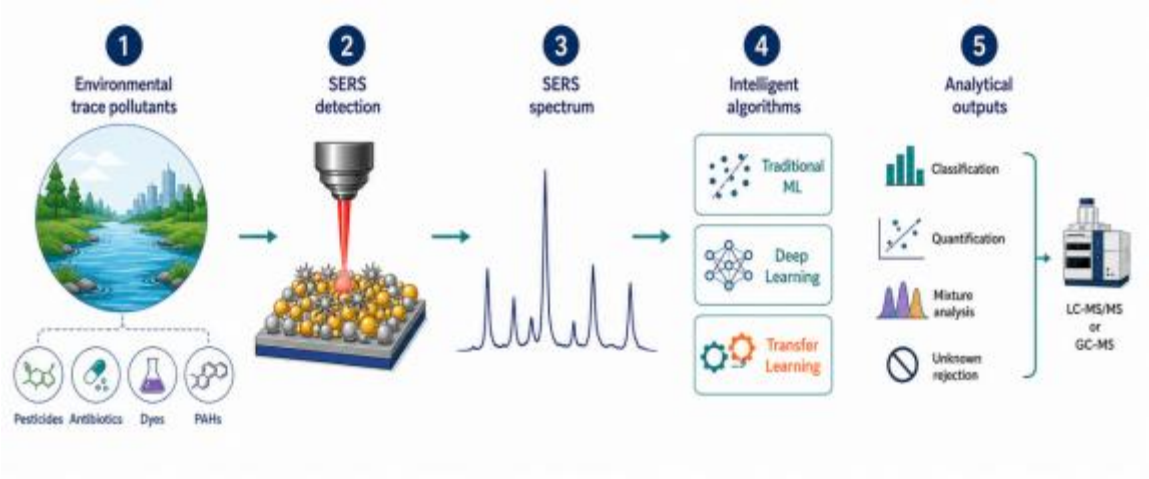

图 1 智能算法辅助 SERS 环境痕量污染物检测的研究框架

2 SERS环境污染检测基础及智能分析流程

2.1 SERS检测原理与优势

拉曼散射记录分子振动引起的非弹性散射光，可提供分子结构信息，但散射截面较小，常规拉曼对痕量组分的响应有限[10]。Fleischmann等[11]在粗糙银电极上观察到吸附吡啶的强拉曼信号，随后研究确认其来源涉及金属纳米结构诱导的表面增强效应[11-12]。

SERS通常由电磁增强和化学增强共同贡献。入射光激发金属纳米结构的局域表面等离激元共振，在纳米间隙、尖端等位置形成高强度局域场；吸附分子与金属表面的电荷转移还可能改变分子极化率[12-13]。其中，电磁增强一般是主要贡献，两类增强机制如图2所示。

这一机制使SERS兼具痕量响应和振动指纹信息，适合农药、抗生素、染料及疏水性有机物的快速筛查[4-6]。但基底和基质差异会造成SERS信号波动，限制不同样品之间的直接比较。

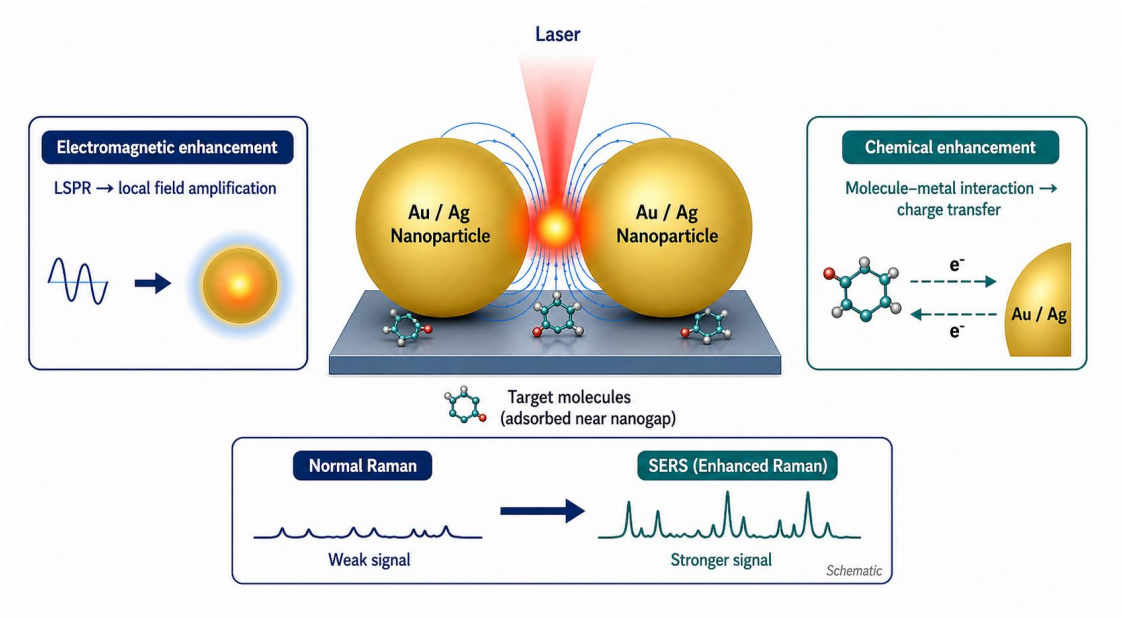

图 2 SERS 电磁增强与化学增强机制示意图

2.2 SERS环境污染检测与数据分析流程

SERS环境污染检测与数据分析流程如图3所示，可概括为水样采集与前处理、目标物富集或选择性捕获、SERS基底作用、光谱采集、数据质控与预处理、模型推断以及结果复核。

低浓度水样通常需要适当的前处理或基底富集，使目标分子接近热点区域。光谱采集阶段应保持主要采集参数一致，并记录必要的基底和质量控制信息。

原始光谱常含荧光基线、宇宙射线尖峰、随机噪声和强度漂移。基线校正、平滑和归一化等预处理可以改善建模条件，但预处理参数应仅依据训练数据确定，并原样用于验证集和外部样品[14]。

分类模型输出类别，回归模型预测浓度，开放集分析还应允许输出“未知”或“超出适用域”。

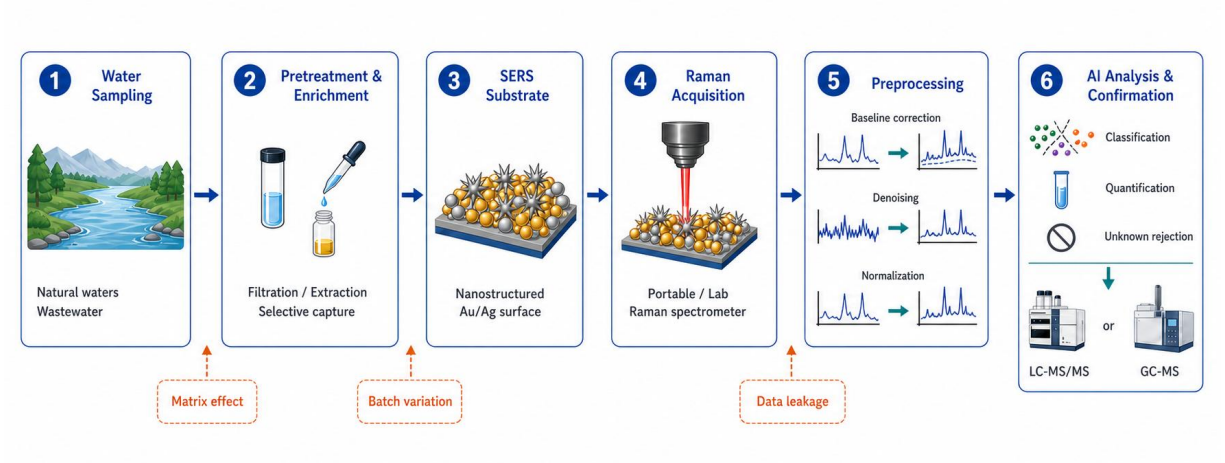

图 3 智能算法辅助 SERS 环境污染物检测与分析流程

2.3 SERS智能分析的主要任务

定性识别关注样品中是否存在目标物及其类别。单一特征峰在背景变化或结构相近化合物之间可能缺乏唯一性，因此分类模型通常利用峰位、峰形和相对强度的组合信息。

定量预测需要建立全谱响应与浓度之间的校准关系。热点分布和竞争吸附会使峰强度偏离简单线性关系，因此常需利用全谱多变量信息建模，其有效范围仍受训练浓度、基底和基质约束。

多组分解析同时涉及组分识别和含量估计。混合光谱并不总是纯组分光谱的线性叠加，组分间可能竞争热点或改变吸附构型。由此形成两类路线：一类直接采集大量混合样本训练模型，另一类尝试复用纯组分知识并向混合体系迁移，后者有助于减少混合训练数据需求。

2.4 环境对象与证据层级

环境SERS研究涉及农药、抗生素、染料及芳香族污染物等对象。不同污染物的富集和吸附特性存在差异，检测模型通常需要针对具体对象建立和校准。

算法证据需区分配制样品、真实水样加标和天然污染水样：前者用于验证方法可行性，后两者更接近实际基质和监测应用。相关拉曼研究或受控混合体系可提供方法参考，但不能直接代表环境水样检测性能。典型环境污染物的SERS检测特点及主要智能分析需求见表1。

表 1 环境典型污染物类型及其 SERS 智能分析需求

对象类型	典型对象	SERS 检测特点	智能分析需求
农药	福美双、有机磷农药等	痕量响应较强，但基质和吸附差异明显	类别识别、浓度预测
抗生素	环丙沙星、多西环素、左氧氟沙星等	结构相近、混合谱峰易重叠	多类别识别、混合比例估计
染料及工业有机物	亚甲蓝、台盼蓝等	拉曼特征较丰富，但共存物可能竞争吸附	分类、定量及未知类检测
多环芳烃	菲、荧蒽等	疏水性强、在水中富集和重复性控制较难	变量筛选、非线性定量
复杂混合体系	多种污染物共存	峰重叠、竞争吸附和未知干扰并存	组分识别、定量解混、未知类拒识

https://doi.org/10.68070/f53x4r53

3 传统机器学习辅助SERS环境污染检测

传统机器学习承担着从SERS光谱到类别或浓度结果的转换。SERS光谱通常具有变量维数高、相邻波数强相关、样本量相对有限以及基底和基质引入非线性波动等特征。针对这些特点，主成分分析（PCA）主要用于探索性降维，支持向量机（SVM）和随机森林（RF）可用于分类或非线性回归，偏最小二乘回归（PLSR）则常用于共线全谱数据的定量建模[7,16-19]。

3.1 主成分分析（PCA）

主成分分析（Principal Component Analysis，PCA）通过线性变换将原始高维变量投影到少量相互正交的主成分，以降低数据维度。对于波数变量高度相关的SERS光谱，PCA可用于观察样本聚类趋势、识别异常光谱并压缩后续监督模型的输入变量[7]。但PCA的优化目标是解释数据方差，而不是最大化污染物类别或浓度的预测能力，因此不宜将得分图上的表观分离等同于分类准确率。

Huang等[20]以废水中加标喹啉为对象比较了两类金纳米结构SERS基底，并利用PCA和机器学习分析目标物与干扰物光谱。商业金纳米柱基底在纯水中的检出限为1.15 μg/L，但在废水中受非目标分子干扰，其检出限恶化至97.5 mg/L。该结果说明PCA和多变量预测模型能够辅助揭示和利用光谱差异，却不能从根本上补偿基底选择性不足。

3.2 支持向量机（SVM）

支持向量机（Support Vector Machine，SVM）通过寻找最大间隔超平面完成分类，并可借助核函数表示非线性决策边界[17-18]。其结构风险最小化思想适合处理SERS研究中常见的高维、小样本问题；用于定量任务时，可采用支持向量回归建立全谱特征与浓度之间的非线性关系。

Weng等[21]比较了稻田水中3种农药的多种分类与回归模型。其中，SVM对phosmet的定量结果最佳，验证集R^2为0.99952，而另外两种农药的最佳模型为PLSR，说明模型性能会随目标物和任务而变化。

3.3 随机森林（RF）

随机森林（Random Forest，RF）通过对训练样本和候选变量进行随机抽样构建多棵决策树，并以投票或平均获得分类、回归结果[19]。该方法能够描述光谱变量与目标响应之间的非线性关系，对变量缩放要求较低，并具有一定的抗噪性和变量筛选能力。

Guo等[22]采用Ag纳米颗粒SERS结合RF模型定量分析水中的菲和荧蒽，并比较了光谱预处理组合及互信息变量筛选的影响。研究共制备31个PAH样品，菲和荧蒽模型在预测集上的R^2分别为0.9780和0.9770。该工作支持RF用于PAH非线性全谱定量，但31个配制水样不足以代表真实环境条件。

3.4 偏最小二乘回归（PLSR）

偏最小二乘回归（Partial Least Squares Regression，PLSR）通过提取同时解释光谱矩阵和浓度响应的潜变量，在降维过程中建立定量关系，适合处理SERS全谱变量多、共线性强的问题[16]。与单一特征峰校准相比，PLSR能够综合多个谱带的信息，但其校准关系通常受训练浓度、基底和基质范围限制。

在稻田水农药研究中，PLSR分别对fonofos和sulfoxaflor获得最佳定量结果，验证集决定系数分别为0.99914和0.99970[21]。Weng等[15]还将波数选择与PLSR结合，用于thiram的SERS定量分析[15]。

3.5 方法比较与评价

从已有环境SERS研究看，算法适用性更多由任务类型和光谱数据结构决定，而不是由模型复杂度决定。PCA主要用于观察样本分布和压缩高维变量；在类别判别或非线性回归中，SVM和RF具有更直接的建模作用；对于共线性明显的全谱定量，PLSR仍是常用方法。不同模型的比较应基于一致的数据划分和评价指标。

Sahin等[23]将纸基Ag纳米颗粒柔性SERS基底与机器学习结合用于农药水溶液区分，但相关验证仍停留在配制体系。上述传统机器学习方法的适用任务、优势及主要局限汇总见表2。

表 2 传统机器学习方法在 SERS 分析中的比较

方法	适用任务	优势	主要局限
PCA	探索性降维、聚类观察、异常识别	直观，可缓解变量共线	最大化总体方差而非类别或浓度预测能力
SVM/SVR	高维小样本分类、非线性定量	小样本条件下具有较强建模能力	对缩放、核函数和参数选择敏感
RF	非线性分类、回归、变量筛选	对变量缩放不敏感，可给出变量重要性	小样本和同源重复谱条件下仍可能过拟合
PLSR	共线全谱定量	结构简洁，适合光谱校准	校准范围外不宜外推，潜变量数需严格选择

4 深度学习辅助SERS环境污染检测

深度学习通过多层非线性变换从光谱中自动学习特征表示，能够减少对人工选峰和手工特征设计的依赖[8,14]。对于峰形重叠、背景波动明显的SERS数据，卷积神经网络（CNN）可提取局部谱带模式，全连接深度网络可联合完成分类和回归，注意力机制则有助于建立不同波数区域之间的全局联系。

4.1 卷积神经网络（CNN）

卷积神经网络（Convolutional Neural Network，CNN）利用共享卷积核扫描一维光谱，在相邻波数范围内提取峰位、峰宽和相对强度等局部特征，并通过多层组合形成更高层次的判别表示。相较于将少数人工选择峰作为输入，CNN可以利用更完整的谱形信息[24-25]。

Li等[26]构建深度学习SERS平台用于环境水中6种有机磷农药检测。ResNet模型分类准确率超过96%，浓度等级预测准确率超过92%，并在农田水、河水和鱼塘水加标实验中完成目标物识别。

Yuan等[27]针对环丙沙星、多西环素和左氧氟沙星混合水溶液，提出CNN与非负弹性网络（NN-EN）结合的分析框架：CNN负责混合物类别识别，NN-EN负责组分比例定量。CNN识别准确率为98.68%。需要限定的是，其核心实验采用已知抗生素配制体系，尚不能代表未知天然水样。

后续研究进一步将该框架用于4个采样点采集的市政湖水加标样品，并用NN-EN预测混合抗生素的组分比例[28]。这一工作提高了环境基质真实性，但仍属于目标种类已知条件下的定向分析，与开放世界中的未知污染物识别存在差距。

4.2 LSTM及序列建模的适用边界

长短期记忆网络（Long Short-Term Memory，LSTM）通过门控结构保持或遗忘序列信息，可缓解传统循环神经网络的梯度消失问题[29]。SERS光谱沿波数轴具有有序结构，但它并非随时间演化的典型时间序列。因此，将LSTM用于SERS时，应证明长程序列建模相较于一维卷积、全连接网络或传统回归确有增益，而不能仅因数据是一维数组就机械套用。

目前直接面向环境污染物SERS检测且具有真实水样验证的LSTM研究明显少于CNN，其在环境SERS中的适用性仍需更多直接证据支持。

4.3 CNN-LSTM融合模型

CNN-LSTM通常先以卷积层提取局部峰形特征，再将压缩后的特征序列输入LSTM以建模较远谱区之间的关系。这种结构在理论上能够结合局部模式识别和序列依赖建模，但参数量和训练复杂度也随之增加。

4.4 全连接深度网络与未知类检测

Zhou等[30]基于AgNP-ZnONR-SNF复合薄膜构建混合SERS数据，并采用全连接网络与KNN进行定性、定量及分布外识别。模型定性和浓度等级分类准确率分别为92.3%和90.8%，未知类识别灵敏度为80%、特异度为89.3%。

该工作将未知类检测纳入SERS分析，但化学样品由蒸馏水配制，且未知类样品包含细菌，尚不能代表天然水体中结构相近的未知有机污染物。

4.5 注意力机制与Transformer

Transformer通过自注意力机制直接计算不同输入位置之间的关联，在理论上有利于捕捉相距较远的谱带组合[31]。然而，Transformer通常需要较大的数据规模和严格正则化，而环境SERS数据常受样本量有限、仪器差异和基底批次漂移制约。目前其在环境水体SERS中的直接原始证据仍较有限。不同深度学习模型的现有证据及环境应用边界汇总见表3。

表 3 深度学习方法的证据基础与环境应用边界

模型	现有环境 SERS 证据	当前边界
CNN	已有配制体系及真实水样研究[26-28]	需防止重复光谱泄漏，并开展跨基质验证
LSTM	真实环境水样直接证据相对有限	光谱并非典型时间序列，需证明相对 CNN 的增益
CNN-LSTM	环境 SERS 系统验证仍不足	参数较多，需与简单基线及消融实验比较
全连接深度网络	已有水中有机物和未知类检测探索[30]	天然水体及结构相近未知物验证不足
Transformer	环境水体 SERS 直接证据较少	数据、计算和外部验证成本较高，尚属发展方向

综上所述，深度学习提升了SERS光谱自动特征学习和混合体系解析能力，其中CNN及深度全连接网络已有环境水样或水溶液研究支撑[26-28,30]；LSTM、CNN-LSTM和Transformer在环境SERS中的直接证据相对有限，更适合作为待验证的发展方向。

5 迁移学习驱动的SERS环境多组分智能检测

当检测对象由单一污染物扩展到复杂混合体系时，监督模型通常需要覆盖不同组分组合和浓度比例的训练数据。若候选物数量增加，组合空间会快

https://doi.org/10.68070/f53x4r53

速扩大，而SERS基底制备、配样和重复测量的成本使穷举式数据采集难以实现。迁移学习的核心价值在于复用源任务中获得的特征或模型，使其适应相关目标任务。

Tan等[9]提出了一种仅用纯组分SERS光谱解析未知混合物的两阶段框架。该研究对象为结构相近的肉碱类生物标志物，而非环境污染物，其价值主要在于为减少混合体系训练数据需求提供方法学思路。

5.1 环境复杂样品中的数据与分布挑战

环境水样除具有多组分组合问题外，基质、基底和仪器变化还会造成光谱分布偏移，因此实验室标准光谱与真实水样光谱并不天然属于相同分布。

传统封闭集分类通常假定测试样品只包含训练阶段已知类别，但真实环境样品可能含有训练库外物质或非目标干扰物。若模型被迫在已知类别中输出一个结果，就可能产生高置信度误判。多组分环境检测因此至少包含三个相互关联的问题：候选污染物是否存在、已识别组分的含量或比例是多少，以及样品是否超出模型适用范围。

5.2 基于纯组分光谱的两阶段自适应迁移框架

Tan等[9]选择L-carnitine（L-car）、O-acetyl-L-carnitine（O-car）、propionyl-L-carnitine（P-car）和butyryl-L-carnitine（B-car）4种结构相近的肉碱分子作为模型体系。研究通过Ag纳米立方体修饰金电极开展电化学SERS测量，外加负电位有助于增强和区分这些光谱相似分子的振动特征。

为明确Tan等[9]所称“迁移学习”的含义，可从源域、目标域、源任务、目标任务及知识复用方式进行说明。该研究以4种纯肉碱SERS光谱作为主要训练数据，可视为源域；未参与训练的二元、三元和四元混合物光谱构成目标域[9]。第一阶段的源任务是利用纯组分与空白光谱训练单组分随机森林分类器，目标任务是在混合谱中判断相应组分是否存在；第二阶段的源任务是利用纯组分间的谱学差异建立组合模型，目标任务则是估计混合物中各组分的摩尔比例和浓度[9]。因此，该框架中的“迁移”主要体现为纯组分谱学判别特征及已训练随机森林模型向混合光谱分析任务的复用，而非典型深度学习中的参数迁移与目标域微调。

该框架包含顺序执行的两个阶段。第一阶段建立4个单组分随机森林分类器（RF-C），分别区分单一肉碱与空白样品，再将纯组分中学习到的分子特异性特征用于判断未知混合谱中相应组分是否存在。完成组分识别后，系统根据第一阶段输出选择第二阶段的相应模型[9]。其两阶段识别与定量流程如图4所示。

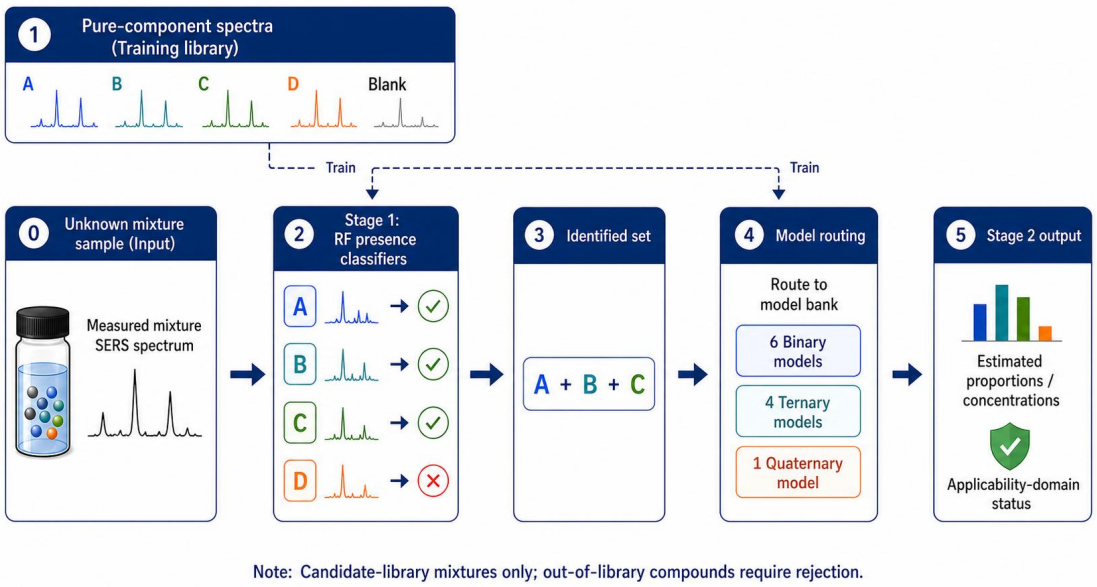

图 4 基于纯组分 SERS 光谱的未知混合物两阶段识别与定量框架

第二阶段不需要预先采集大规模混合训练集。针对4种肉碱可能形成的6种二元组合、4种三元组合和1种四元组合，研究者共建立11个多组分RF分类器，这些模型仍仅使用纯肉碱光谱训练，并利用模型输出概率估计未知混合物中各组分的摩尔比例和浓度[9]。

模型并非依赖单一特征峰，而是综合峰强、峰位和峰形等多个光谱特征形成“特征条码”。

5.3 盲测设计、性能与适用边界

研究使用11种未知二元、三元和四元肉碱混合物验证框架，共包含550条未参与训练的SERS光谱

[9]。结果显示，框架实现100%的组分识别准确率，平均定量误差小于3%。

原文以含L-car、O-car和B-car的未知三元体系说明顺序决策的必要性：第一阶段识别存在组分并排除P-car，第二阶段再调用对应三元模型预测比例；若直接使用四元模型，则会错误预测少量P-car[9]。因此，该方法的优势来自“先识别、后选择组合模型并定量”的自适应流程，而不仅是随机森林算法本身。

尽管结果具有方法学意义，其外推边界仍需明确。首先，4种肉碱共享相似骨架，实验也基于受控的电化学SERS平台，真实水样中污染物的吸附和光谱响应可能更加复杂。其次，研究的候选范围仅包含4种预定义肉碱，所谓“未知混合物”是组成和比例未知，并不等同于数据库外未知物的开放集检测。模型也尚未在天然环境水样、不同基底或不同仪器之间验证。

5.4 面向环境水体检测的可实施路线

若将上述框架用于环境水体，候选污染物数据库除标准谱图外，还需要覆盖不同浓度、基底批次和真实水样中的基质变化。在候选污染物范围明确后，可先利用纯组分模型判断目标物是否存在，再根据识别结果选择相应的定量模型；对于训练库之外的样品，还需结合置信度和适用域判断保留拒识空间。据此，可将智能SERS环境检测的验证证据概括为表4所示的四个层级。

表4 智能 SERS 环境检测的四级验证体系

层级	样品类型	可支持的结论
Ⅰ	纯水或缓冲液标准体系	原理和方法可行
Ⅱ	人工多组分混合体系	受控混合体系适用
Ⅲ	不同来源真实水样加标	真实基质中具有应用潜力
Ⅳ	天然污染水样	支持实际筛查或监测应用

Wang等[32]采用Cycle-GAN实现不同拉曼仪器间的光谱映射，为跨仪器光谱对齐提供了方法参考。为进一步比较已有研究的检测对象、智能分析任务及验证深度，代表性智能SERS环境检测研究汇总见表5。

表 5 代表性智能 SERS 环境检测研究汇总

代表研究	检测对象与样品	算法与任务	代表性结果	验证层级及主要边界
Weng 等[21]	3种农药；稻田水	KNN、SVM、PLSR 等；分类与定量	KNN 分类准确率100%；phosmet 的 SVM 定量 R^2V=0.99952	Ⅲ—Ⅳ；目标物为预定义农药
Guo 等[22]	菲、荧蒽；配制水样	RF；定量预测	两种 PAHs 预测集 R^2分别为0.9780、0.9770	Ⅰ；样本量较小，缺乏真实水样验证
Li 等[26]	6种有机磷农药；农田水、河水、鱼塘水	深度学习；分类及浓度等级预测	分类准确率>96%，浓度等级准确率>92%	Ⅲ；主要为已知目标物加标检测
Yuan 等[28]	3种抗生素混合物；市政湖水	CNN+NN-EN；识别与比例定量	在4个采样点湖水中实现混合比例预测	Ⅲ；候选组分预先已知
Zhou 等[30]	水中有机污染物；蒸馏水体系	全连接网络、KNN；分类、定量及未知类识别	定性准确率92.3%，未知类灵敏度80%	Ⅰ—Ⅱ；未知类验证与天然环境仍有差距
Huang 等[38]	5种磺胺类抗生素及混合物；实际水样	MLP+数据增强；多类别识别	实际水样准确率93%，并以 LC-MS 验证	Ⅲ—Ⅳ；跨基底、跨仪器验证仍不足

6 挑战与未来发展趋势

环境监测中的实际误差同时受到样品前处理、污染物富集、基底状态、仪器响应和数据划分方式等因素影响，因此模型准确率并不能单独代表完整检测方法的可靠性。

6.1 SERS数据与实验流程标准化

SERS信号易受纳米结构、热点分布和分析物吸附状态影响，同一基底表面的均匀性并不代表批次间重现性。Bell等[33]指出，定量SERS还应设置独立验证，并结合回收率和预测误差评价分析性能[33]。

Fornasaro等[34]组织15个实验室开展多仪器定量SERS协作研究，直接评价方法的重现性和真度。结果表明，单一实验室内的重复性不能直接代表跨仪器或跨实验室条件下的稳定性。

光谱预处理应保持一致。Bocklitz等[35]发现，物理意义上合理的预处理组合未必都能获得稳定模型，因此相关参数应在训练阶段确定并固定用于验证。环境SERS数据库还应保存原始光谱及样品来源、基底批次和主要采集条件。

6.2 模型泛化、数据泄漏与适用域

https://doi.org/10.68070/f53x4r53

SERS研究常以光谱条数表示数据量，但模型泛化更取决于独立样品、基底和实验批次数。若同一样品或基底的重复谱被随机分入训练集和测试集，背景及批次特征会同时出现在两侧，造成数据泄漏。Guo等[36]在拉曼化学计量学流程中也强调训练、验证、测试划分及外部验证对模型可靠性的影响。两种数据划分方式的差异如图5所示。

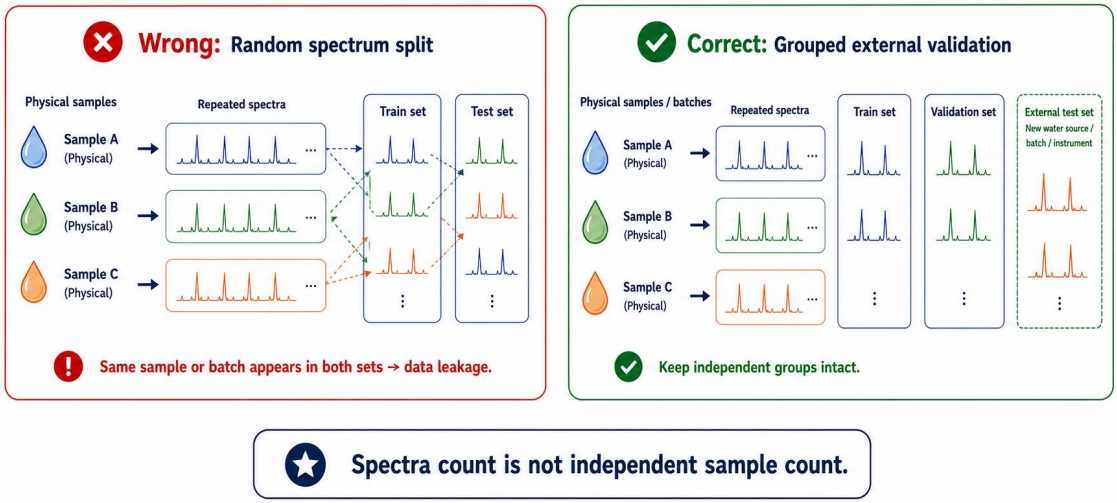

图 5 SERS 光谱错误随机划分与样品级分组验证对比

模型测试应保证配样和基底相互独立，并进一步考察实验日期、基底批次和水样来源变化造成的性能差异。对于明显偏离训练分布的样品，允许输出“不确定”或“未知”比强制归入已知类别更符合实际检测需求。

真实基质偏移可能远大于算法之间的性能差异。Huang等[20]的废水研究表明，复杂基质能够显著恶化SERS检测性能，说明非目标物竞争吸附和基底选择性仍可能成为限制因素。

Wei等[37]利用去离子水中的As^{3+} 响应数据预训练一维CNN，再以少量加标二级处理污水数据进行迁移微调，最终用于未经加标的原始污水，并以92%的准确率判断As^{3+} 水平是否低于WHO推荐限值。该结果表明迁移学习可减少模型适应新基质所需的数据，但其SERS信号来自大肠杆菌受重金属胁迫后的代谢响应，属于间接传感，不能等同于污染物分子的直接SERS指纹检测。

跨仪器和跨实验室稳定性是另一类外部泛化问题。Fornasaro等[34]通过15个实验室参与的多仪器SERS研究直接考察了跨实验室重现性。目前在同一环境污染物智能SERS体系中同时完成跨水样、跨基底和跨仪器验证的研究仍较少。

6.3 AI模型解释性与分析可信度

热图、注意力权重或变量重要性主要反映统计关联，不能单独作为化学解释。模型关注区域还需对应稳定的目标物谱带，并结合振动归属和干扰实验判断；若模型主要利用背景、批次或仪器噪声完成预测，其实际价值仍有限。

Tan等[9]将RF特征重要性与实验光谱和DFT振动分析结合，Yuan等[27]使用SHAP比较CNN关注区域与抗生素特征峰的关系，均提供了将模型归因与光谱信息结合的思路。但这类统计归因仍需通过光谱归属和独立对照实验验证。

6.4 智能便携环境检测

便携式拉曼仪器与柔性、纸基SERS基底为环境现场筛查提供了硬件基础。Li等[26]已在农田水、河水和鱼塘水加标实验中验证深度学习辅助多种有机磷农药识别；Huang等[38]进一步采用SERS与多层感知机在多种实际水样基质中开展5种磺胺类抗生素及其混合物识别，经数据增强后识别准确率达到93%，并以LC-MS验证实际样品结果。

实际部署仍需保留波数校准、信噪比和异常信号等质量控制，并将超出适用域或预测不确定的样品转入实验室确证。便携SERS-AI更适合快速筛查，而LC-MS/MS、GC-MS等方法承担确证；现场应用还需考虑检测时间、耗材成本和基底稳定性。相关现场筛查、实验室确证及未来发展路线如图6所示。

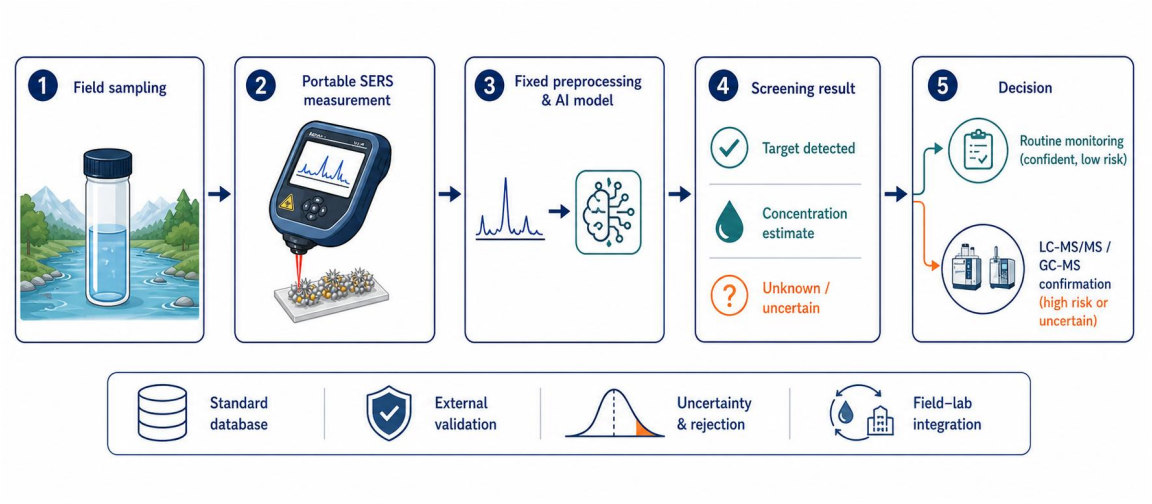

图 6 智能 SERS 现场筛查、实验室确证及未来发展路线

7 结论

SERS依托金属纳米结构产生的局域场增强，可在痕量污染物检测中获得具有分子特征的振动光谱；智能算法则主要用于复杂谱图的信息提取和结果判读。已有研究表明，传统机器学习和部分深度学习方法可用于污染物分类、浓度预测及混合体系分析，但其实际性能仍受样品、基底及测量条件影响，不能仅以模型内部准确率评价检测能力。Tan等[9]基于纯组分SERS光谱构建两阶段迁移框架，表明纯组分中获得的分子特征可在一定条件下用于未参与训练的混合体系分析，为减少混合样本建库需求提供了方法思路。但该研究仍基于受控肉碱体系，其“未知”属于预定义候选库内的组成未知，面向真实环境污染物的适用性仍需进一步验证。

面向实际环境监测，仍需进一步加强真实水样和变化实验条件下的验证，并提高模型对未知样品的识别能力。随着数据和验证体系逐步完善，智能算法辅助SERS有望进一步拓展其在环境水体痕量污染物检测中的应用。

致谢

本工作得到了江苏省大学生创新创业训练计划项目（S202610320059）的支持。

数据可用性

数据将根据请求提供。

利益冲突声明

作者声明他们没有已知的可能影响本文所报告工作的竞争性财务利益或个人关系。

参考文献

[1] SCHWARZENBACH R P, EGLI T, HOFSTETTER T B, et al. Global water pollution and human health[J]. Annual Review of Environment and Resources, 2010, 35: 109-136.

[2] MOJIRI A, ZHOU J L, OHASHI A, et al. Comprehensive review of polycyclic aromatic hydrocarbons in water sources, their effects and treatments[J]. Science of the Total Environment, 2019, 696: 133971.

[3] PŁOTKA-WASYLKA J, RUTKOWSKA M, OWCZAREK K, et al. Extraction with environmentally friendly solvents[J]. TrAC Trends in Analytical Chemistry, 2017, 91: 12-25.

[4] LANGER J, JIMENEZ DE ABERASTURI D, AIZPURUA J, et al. Present and future of surface-enhanced Raman scattering[J]. ACS Nano, 2020, 14(1): 28-117.

[5] CIALLA D, MÄRZ A, BÖHME R, et al. Surface-enhanced Raman spectroscopy (SERS): Progress and trends[J]. Analytical and Bioanalytical Chemistry, 2012, 403(1): 27-54.

[6] SCHLÜCKER S. Surface-enhanced Raman spectroscopy: Concepts and chemical applications[J]. Angewandte Chemie International Edition, 2014, 53(19): 4756-4795.

[7] QI Y, HU D, JIANG Y, et al. Recent progresses in machine learning assisted Raman spectroscopy[J]. Advanced Optical Materials, 2023, 11(14): 2203104.

[8] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521: 436-444.

[9] TAN E X, CHEN J R T, PANG D W C, et al.

Transfer learning-assisted SERS: Predicting molecular identity and concentration in mixtures using pure compound spectra[J]. Angewandte Chemie International Edition, 2025, 64(36): e202508717.

[10] SMITH E, DENT G. Modern Raman spectroscopy: A practical approach[M]. Chichester: John Wiley & Sons, 2005.

[11] FLEISCHMANN M, HENDRA P J, MCQUILLAN A J. Raman spectra of pyridine adsorbed at a silver electrode[J]. Chemical Physics Letters, 1974, 26(2): 163-166.

[12] MOSKOVITS M. Surface-enhanced spectroscopy[J]. Reviews of Modern Physics, 1985, 57(3): 783-826.

[13] LOMBARDI J R, BIRKE R L. A unified approach to surface-enhanced Raman spectroscopy[J]. The Journal of Physical Chemistry C, 2008, 112(14): 5605-5617.

[14] LUSSIER F, THIBAULT V, CHARRON B, et al. Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering[J]. TrAC Trends in Analytical Chemistry, 2020, 124: 115796.

[15] WENG S, CHEN S, LI M, et al. Quantitative analysis of thiram based on SERS and PLSR combined with wavenumber selection[J]. Analytical Methods, 2014, 6(1): 242-247.

[16] WOLD S, SJÖSTRÖM M, ERIKSSON L. PLS-regression: A basic tool of chemometrics[J]. Chemometrics and Intelligent Laboratory Systems, 2001, 58(2): 109-130.

[17] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer, 1995.

[18] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[19] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[20] HUANG Y H, WEI H, SANTIAGO P J, et al. Sensing antibiotics in wastewater using surface-enhanced Raman scattering[J]. Environmental Science & Technology, 2023, 57(12): 4880-4891.

[21] WENG S, ZHU W, DONG R, et al. Rapid detection of pesticide residues in paddy water using surface-enhanced Raman spectroscopy[J]. Sensors, 2019, 19(3): 506.

[22] GUO M, LI M, FU H, et al. Quantitative analysis of polycyclic aromatic hydrocarbons (PAHs) in water by surface-enhanced Raman spectroscopy (SERS) combined with Random Forest[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2023, 287: 122057.

[23] SAHIN F, CELIK N, CAMDAL A, et al. Machine learning-assisted pesticide detection on a flexible surface-enhanced Raman scattering substrate prepared by silver nanoparticles[J]. ACS Applied Nano Materials, 2022, 5(9): 13112-13122.

[24] LIU J, OSADCHY M, ASHTON L, et al. Deep convolutional neural networks for Raman spectrum recognition: A unified solution[J]. Analyst, 2017, 142(21): 4067-4074.

[25] WENG S, YUAN H, ZHANG X, et al. Deep learning networks for the recognition and quantitation of surface-enhanced Raman spectroscopy[J]. Analyst, 2020, 145(14): 4827-4835.

[26] LI R, WANG Z, ZHANG Z, et al. Deep learning-based multicapturer SERS platform on plasmonic nanocube metasurfaces for multiplex detection of organophosphorus pesticides in environmental water[J]. Analytical Chemistry, 2022, 94(46): 16006-16014.

[27] YUAN Q, YAO L F, TANG J W, et al. Rapid discrimination and ratio quantification of mixed antibiotics in aqueous solution through integrative analysis of SERS spectra via CNN combined with NN-EN model[J]. Journal of Advanced Research, 2025, 69: 61-74.

[28] YUAN Q, WEN X R, LIU W, et al. Simultaneous detection and quantification of ciprofloxacin, doxycycline, and levofloxacin in municipal lake water via deep learning analysis of complex Raman spectra[J]. Environmental Technology & Innovation, 2025, 37: 103987.

[29] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[30] ZHOU J, WU Z, JIN C, et al. Machine learning assisted dual-functional nanophotonic sensor for organic pollutant detection and degradation in water[J]. npj Clean Water, 2024, 7: 3.

[31] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30. Long Beach: Curran Associates, 2017: 5998-6008.

[32] WANG Z, YANG Z, SONG X, et al. Raman spectrum model transfer method based on Cycle-GAN[J]. Spectrochimica Acta Part A: Molecular

and Biomolecular Spectroscopy, 2024, 304: 123416.
[33] BELL S E J, CHARRON G, CORTES E, et al. Towards reliable and quantitative surface-enhanced Raman scattering (SERS): From key parameters to good analytical practice[J]. Angewandte Chemie International Edition, 2020, 59(14): 5454-5462.
[34] FORNASARO S, ALSAMAD F, BAIA M, et al. Surface enhanced Raman spectroscopy for quantitative analysis: Results of a large-scale European multi-instrument interlaboratory study[J]. Analytical Chemistry, 2020, 92(5): 4053-4064.
[35] BOCKLITZ T, WALTER A, HARTMANN K, et al. How to pre-process Raman spectra for reliable and stable models?[J]. Analytica Chimica Acta, 2011, 704(1-2): 47-56.
[36] GUO S, POPP J, BOCKLITZ T. Chemometric analysis in Raman spectroscopy from experimental design to machine learning-based modeling[J]. Nature Protocols, 2021, 16: 5426-5459.
[37] WEI H, HUANG Y, SANTIAGO P J, et al. Decoding the metabolic response of Escherichia coli for sensing trace heavy metals in water[J]. Proceedings of the National Academy of Sciences of the United States of America, 2023, 120(7): e2210061120.
[38] HUANG Z, LIANG W, LEI Y, et al. Homogeneous multi-antibiotics residual identification in various actual water via SERS spectra multilayer perceptron algorithm combined with Gaussian kernel density estimation data augmentation[J]. Analytica Chimica Acta, 2026, 1383: 344896.

Research Progress on Intelligent Algorithm-Assisted Surface-Enhanced Raman Spectroscopy for Environmental Trace Pollutant Detection

Zhou Tianyu, You Chunxia*

School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou 221116

Abstract: Trace pollutants in environmental waters are often present at low levels and in complex mixtures, while matrix interference can further complicate the interpretation of SERS spectra. This paper reviews the use of intelligent algorithms for SERS-based environmental analysis, focusing on pollutant identification, concentration estimation, and multicomponent spectral analysis. Machine-learning methods such as PCA, SVM, RF, and PLSR, together with CNN-based deep learning, have been used to extract information from full spectra and improve qualitative and quantitative analysis. More recent studies have also explored models trained on spectra of individual compounds and then applied to mixture analysis, offering a possible way to reduce the experimental burden of preparing large numbers of mixed samples. Despite these advances, matrix effects, variations among SERS substrates and instruments, data leakage, and limited generalization still restrict practical use. Improving data standardization, external validation, unknown-class rejection, and the connection between portable screening and laboratory confirmation will be important for developing more reliable intelligent SERS methods for environmental detection.

Keywords: Surface-enhanced Raman spectroscopy; Environmental trace pollutants; Machine learning; Deep learning; Transfer learning