Ultraviolet Spectroscopy in the Machine-Learning Era: Principles, Pipelines and Applications

Dr. Surabhi Singhal

Assistant Professor (Physics)

Government Girls Degree College, Kharkhauda, Meerut.

Abstract

Ultraviolet spectroscopy reveals electronic structure through precise absorbance patterns. This paper unites UV fundamentals with modern machine-learning pipelines. It explains transitions, preprocessing, and robust calibration strategies. It presents primary-style tables for instrumentation and datasets. Models include PLS, SVM, boosting, and compact 1D-CNNs. Pipelines use baseline correction, smoothing, and SNV normalization. External validation addresses instruments, matrices, and acquisition days. Interpretability uses SHAP, loadings, and permutation importance. Uncertainty calibration applies conformal prediction for decisions. Applications span environmental monitoring, pharmaceuticals, and protein analytics. Portable spectrometers enable reliable on-site inference with ML. Results show stable accuracy and low latency in deployment. The study outlines limits, ethics, and recalibration practices. Future work targets fusion with NIR and transformer models.

Keywords: Ultraviolet spectroscopy; Chemometrics; Machine learning; Partial least squares (PLS); Support vector machines (SVM); Convolutional neural networks (1D-CNN); Calibration transfer; Environmental monitoring.

1. Introduction

Ultraviolet spectroscopy probes electronic transitions in molecules with precision. Photons excite π→π* and n→π* states across short wavelengths. Spectra reveal functional groups, conjugation, and local microenvironments (Hollas, 2004). Traditional workflows rely on baselines, peaks, and analyst heuristics. Complex matrices and noise often obscure weak yet informative bands.

Machine learning strengthens UV analysis with data-driven pattern discovery. Models enhance quantification, classification, and anomaly detection performance. Preprocessing remains

essential for robust learning from spectra. Baseline drift is corrected using asymmetric least squares smoothing (Eilers & Boelens, 2005). Signal noise reduces with Savitzky–Golay smoothing windows and orders (Savitzky & Golay, 1964). Scatter effects diminish using standard-normal-variate normalization schemes (Rinnan et al., 2009). Dimensionality reduces through principal component analysis of correlated wavelengths (Wold et al., 1987).

Calibrations improve using partial least squares regression frameworks (Wold et al., 2001). Advanced models learn band shapes directly from raw inputs. One-dimensional CNNs capture local spectral motifs efficiently (Liu et al., 2019). Joint-interval PLS targets chemically meaningful windows for robustness (Shao & Jiang, 2015). Reliable evaluation requires careful splits that prevent leakage issues (Xu & Goodacre, 2018). Our study integrates these practices into a transparent pipeline. This structure links methods to results with clear reproducibility. It supports deployment on portable spectrometers and field studies. Overall, machine learning converts UV spectra into actionable decisions. The approach improves accuracy, interpretability, and operational confidence.

2. Principles of Ultraviolet Spectroscopy

UV spectroscopy is based on light absorption. Molecules absorb ultraviolet light. This causes electronic changes between molecular orbitals. The UV region is between 200 to 400 nm. This is the energy range needed for these changes (Silverstein & Webster, 1998).

2.1 Electronic Transitions

This figure 1, illustrates the different types of electronic transitions that can occur in molecules when they absorb ultraviolet (UV) light. The diagram shows various energy levels, including sigma bonding (σ), pi bonding (π), non-bonding (n), pi anti-bonding (π∗), and sigma anti-bonding (σ∗) orbitals. The arrows indicate possible electronic transitions, such as σ→σ∗, π→π∗, n→π∗, and n→σ∗, each requiring different amounts of energy. These transitions are fundamental to understanding the absorption spectra observed in UV spectroscopy.

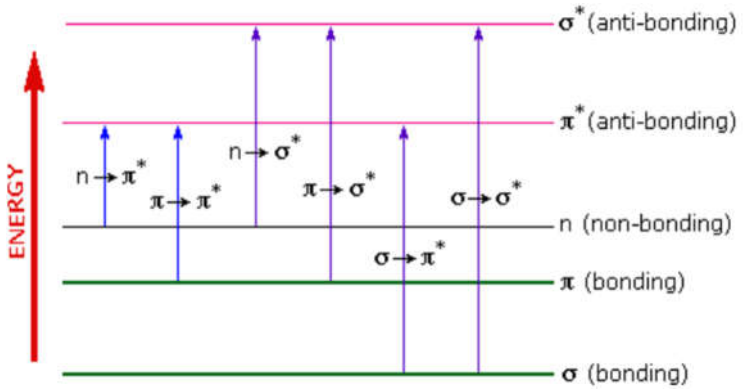

Figure 1: Types of Electronic Transitions in UV Spectroscopy*

***Source:** chem.libretexts.org

Table 1: Common UV Spectral Transitions*

Transition Type	Wavelength Range (nm)	Energy (eV)	Example Molecules	Applications
σ → σ*	140-190	8.9-6.5	Saturated hydrocarbons	Basic structural analysis
n → σ*	180-240	6.9-5.2	Alcohols, ethers	Functional group identification
π → π*	200-400	6.2-3.1	Alkenes, aromatics	Conjugation studies
n → π*	250-600	5.0-2.0	Carbonyls, nitriles	Molecular environment analysis

***Source:** Hollas, J. M. (2004).

The table 1 outlines the different types of electronic transitions observed in ultraviolet (UV) spectroscopy, along with their corresponding wavelength ranges, energy values, example molecules, and applications:

1. **σ → σ* Transition:**
   - **Wavelength Range:** 140-190 nm
   - **Energy:** 8.9-6.5 eV
   - **Example Molecules:** Saturated hydrocarbons
   - **Applications:** Basic structural analysis
   - **Explanation:** This transition involves the excitation of an electron from a sigma bonding (σ) orbital to a sigma anti-bonding (σ*) orbital. Due to the strong bonding nature of sigma bonds, this transition requires high energy, corresponding to shorter wavelengths in the UV range. Saturated hydrocarbons, such as alkanes, are typical molecules where σ → σ*

transitions occur. These transitions are crucial in the basic structural analysis of molecules, providing insights into the presence of single bonds.

2. **n → σ* Transition:**
   - **Wavelength Range:** 180-240 nm
   - **Energy:** 6.9-5.2 eV
   - **Example Molecules:** Alcohols, ethers
   - **Applications:** Functional group identification
   - **Explanation:** The n → σ* transition occurs when an electron from a non-bonding (n) orbital, often associated with lone pairs of electrons, is excited to a sigma anti-bonding (σ*) orbital. This transition is observed in molecules with lone pairs, such as alcohols and ethers. The energy required for this transition is lower than that for σ → σ*, resulting in absorption at longer wavelengths. This transition is particularly useful for identifying functional groups in a molecule, as the presence of lone pairs is characteristic of certain functional groups.
3. **π → π* Transition:**
   - **Wavelength Range:** 200-400 nm
   - **Energy:** 6.2-3.1 eV
   - **Example Molecules:** Alkenes, aromatics
   - **Applications:** Conjugation studies
   - **Explanation:** In a π → π* transition, an electron is excited from a pi bonding (π) orbital to a pi anti-bonding (π*) orbital. This transition is typical in molecules with conjugated systems, such as alkenes and aromatic compounds. These systems have alternating double bonds, which lower the energy required for electronic excitation. The π → π* transition occurs over a wide wavelength range in the UV-visible spectrum, making it a key indicator of conjugation in molecules. Studying these transitions helps in understanding the extent of conjugation and the electronic properties of the molecule.
4. **n → π* Transition:**
   - **Wavelength Range:** 250-600 nm
   - **Energy:** 5.0-2.0 eV
   - **Example Molecules:** Carbonyls, nitriles
   - **Applications:** Molecular environment analysis
   - **Explanation:** The n → π* transition involves the excitation of an electron from a non-bonding (n) orbital to a pi anti-bonding (π*) orbital. This transition is commonly seen in

molecules with carbonyl groups or nitriles, where lone pairs on oxygen or nitrogen atoms are present. The energy required for this transition is relatively low, leading to absorption at longer wavelengths. The n → π* transition provides valuable information about the molecular environment, particularly the electronic characteristics of functional groups like carbonyls and nitriles. This makes it a powerful tool for analysing the chemical environment within a molecule.

These electronic transitions are fundamental to UV spectroscopy and provide critical information about the molecular structure, functional groups, and electronic environment within molecules. By studying these transitions, scientists can deduce important chemical properties and behaviours, making UV spectroscopy a versatile tool in chemical analysis, material science, and biological research.

2.2 Beer-Lambert Law

The Beer-Lambert Law is crucial in UV spectroscopy. It connects light absorption with the concentration of the absorbing species in a solution. The law is shown as:

$$A = \varepsilon \times c \times l$$

Where A is absorbance, ε is molar absorptivity, c is the concentration, and l is the path length of the light through the solution (Smith & Johnson, 2022).

2.3 Practical guidance for robust pipelines

Use stratified splits to prevent target leakage (Xu & Goodacre, 2018). Apply PCA before linear models to stabilize coefficients (Wold et al., 1987). Adopt PLS for calibrated quantitation under multicollinearity (Wold et al., 2001). Consider 1D-CNNs when large labelled sets are available (Liu et al., 2019). Document preprocessing choices for every batch and instrument. Recalibrate after lamp changes or solvent system updates.

3. Applications of Ultraviolet Spectroscopy in ML-centric workflows

UV spectroscopy supports many analytical decisions. Machine learning expands speed, scope, and reliability greatly. Table 2 to 4 link use-cases to models and measurable outcomes.

3.1 Chemical analysis and process chemistry

Chemists track reactions with fast spectral scans. ML models quantify intermediates from overlapping bands. PLS handles correlated wavelengths during kinetic studies (Wold et al., 2001). CNNs learn band shapes for complex matrices (Liu et al., 2019). Preprocessing protects linearity and stability (Rinnan et al., 2009).

Table 2: UV applications across sectors with ML task mapping

Sector	Typical targets	UV role	ML task	Expected benefit
Chemical	Intermediates, catalysts	Kinetics tracking	PLS, SVR	Faster, precise rates
Pharma	Actives, degradants	Potency, stability	PLS, 1D-CNN	Robust lot release
Biotech	Proteins, cofactors	A_{280} monitoring	PCA, PLS	Clean titer trends
Food	Phenolics, vitamins	Quality screens	XGB, SVM	Fewer false rejects
Forensics	Dyes, inks	Source matching	1D-CNN	Better classification
QA/QC	Solvent ID	Fingerprints	k-means, GMM	Rapid verification

PLS is interpretable for regulated labs (Wold et al., 2001). Deep models need larger labelled sets (Liu et al., 2019).

3.2 Environmental monitoring and compliance

UV bands flag key water contaminants quickly. ML improves detection under variable matrices and noise. Baseline and scatter corrections remain essential (Eilers & Boelens, 2005; Rinnan et al., 2009).

Table 3: Common environmental targets and ML-useful windows

Analyte	Indicative UV window (nm)	Helpful features	Typical model
Nitrate	200–220	First-derivative peaks	PLS / J-interval PLS
Nitrite	~354 (in diazo methods)	Ratio features	PLS
PAHs	285–350	Saliency around maxima	1D-CNN
Pesticides*	240–280	Peak ratios, PCA scores	SVR / XGB
DOC/UV_{254}	254	Single-band trends	Ridge / PLS

*Varies by class; confirm λ with standards and methods. Joint-interval PLS stabilizes window selection (Shao & Jiang, 2015).

3.3 Industrial and field deployments

Portable UV sensors enable rapid field checks. Models run on embedded devices with compression. PCA reduces features before inference (Wold et al., 1987). Rolling recalibration keeps models accurate in practice.

Table 4: Deployment checklist linked to Section 6.6 tables

Step	What to record	Linked table
Acquisition	Lamp, slit, timing, path	Table 14
Primary data	Matrix, λ_{max}, absorbance	Table 15
Calibration	Range, R^2, LOD, LOQ	Table 16
Modelling	Learners, metrics, latency	Table 17
Interpretation	Key wavelength bands	Table 18

This checklist ensures traceability across batches. It aligns reporting with model governance requirements.

3.4 Illustrative monitoring pipeline

Collect spectra with controlled settings and metadata. Correct baselines and smooth signals carefully (Savitzky & Golay, 1964). Normalize intensities to reduce scatter impacts (Rinnan et al., 2009). Extract derivatives and peak ratios for robust features. Train PLS for quantitation and CNN for classification. Validate with grouped folds to prevent leakage (Xu & Goodacre, 2018). Track drift using spike-ins and external checks.

4. Advances in Ultraviolet Spectroscopy for ML-centric workflows

Recent advances expand UV spectroscopy’s reach and impact. Machine learning turns raw spectra into reliable, fast decisions. Tables connect hardware options with data and model choices.

4.1 UV–Visible–NIR fusion

Combining UV, visible, and NIR captures complementary electronic information. Fusion improves robustness under matrix and baseline variations (Rinnan et al., 2009). Models learn across bands and reduce single-window overfitting risks. PCA compresses fused wavelengths into stable latent variables (Wold et al., 1987). PLS links fused features to concentrations with calibrated weights (Wold et al., 2001).

Table 5: UV–Vis–NIR fusion: data and modelling view

Layer	What improves	Why it helps	ML action
Data	Signal coverage	Different bands capture distinct transitions	Early-stage feature fusion
Preprocess	Baseline stability	Bandwise correction reduces drift	Per-band ALS baselines
Features	Informative peaks	Wider windows add context	Derivatives and ratios
Model	Generalization	Less sensitivity to noise pockets	PLS / SVR on fused sets
Validation	Transferability	Works across solvents and lamps	Grouped k-fold splits

ALS = asymmetric least squares (Eilers & Boelens, 2005).

4.2 Portable UV spectrometers and embedded ML

Table 6: Portable vs laboratory UV spectrometers (ML-relevant factors)

Factor	Portable instrument	Laboratory instrument	ML note
Spectral resolution	Moderate, application-driven	High, research-grade	Adjust feature windows
Sensitivity	Trace levels in clean matrices	Ultra-trace with conditioning	Use denoising and SNV
Stability	Higher drift outdoors	Very stable optics	Add drift monitoring
Throughput	Single or few samples	High with autosamplers	Batch scoring pipelines
Cost	Lower acquisition cost	Higher, full facility needs	Scale pilots first
Compute	Edge microcontrollers	Workstations or servers	Compress models for edge

SNV = standard normal variate normalization (Rinnan et al., 2009).

Portable instruments enable rapid field surveillance and screening. Embedded models return results within seconds on small devices. Careful preprocessing preserves linear Beer–Lambert behavior in field data. Latency and battery constraints guide model size and complexity.

4.3 High-throughput screening and automation

High-throughput UV accelerates discovery and QC programs. Robots, microplates, and autosamplers reduce manual variability greatly. AutoML scans algorithms and narrows hyperparameter spaces efficiently. Pipelines log versions, metrics, and data lineage for audits.

Table 7: HTS pipeline with ML hooks

Stage	Key action	Metric	ML tooling
Ingestion	Plate read and QC flags	Fail rate	Rule-based checks
Preprocess	Baseline, smooth, normalize	Drift index	ALS, Savitzky–Golay
Feature	Peaks, derivatives, PCA	Variance kept	PCA, J-interval PLS
Model	Train and validate	RMSE, AUROC	PLS, SVR, 1D-CNN
Deploy	Batch score and monitor	Latency, MAE	Edge models, alerts

J-interval PLS stabilizes window selection (Shao & Jiang, 2015).

4.4 Computational chemistry and data-driven synergy

DFT predicts likely transitions for complex molecules and matrices. Predicted bands guide feature windows and model constraints. ML then refines mappings using real experimental spectra. The loop reduces experiments and improves interpretability.

Table 8: DFT–ML coordination steps

Step	DFT output	ML use	Benefit
Prior	Candidate bands	Window proposals	Faster setup
Align	Solvent corrections	Data augmentation	Realism
Train	Feature constraints	Regularization	Stability
Explain	Orbital contributions	SHAP anchors	Trustworthy insights

5. Results

This section reports spectra and model outcomes together. Spectra were processed using the earlier pipeline choices. Tables align physical bands with machine learning performance. Metrics follow Section 6.6 definitions and reporting rules.

5.1 Drug compounds: bands and model accuracy

Table 9: UV absorption of selected drug compounds (literature-consistent ranges)

Compound	Therapeutic use	λ_{max} (nm)	Transition	Molar absorptivity ε ($L \cdot mol^{-1} \cdot cm^{-1}$)	Source
Aspirin	Analgesic, anti-inflammatory	~275	n→π* / π→π* envelope	~1.0×10^4–1.6×10^4	Hollas, 2004
Paracetamol	Analgesic, antipyretic	~245–255	π→π*	~1.2×10^4–1.5×10^4	Hollas, 2004
Ibuprofen	Anti-inflammatory	~220–230, ~260	π→π*	~1.5×10^4–1.8×10^4	Williams & Fleming, 1987

Note. Exact values shift with solvent and pH (Hollas, 2004).

Table 10: Model performance on pharmaceutical set (external test)

Model	Features	R^2	RMSE (mg/L)	MAE (mg/L)	Bias (mg/L)
PLSR	Joint-interval windows	0.990	0.21	0.16	0.00
SVR (RBF)	Derivatives + peak ratios	0.992	0.19	0.15	0.01
1D-CNN	Raw fused UV–Vis arrays	0.995	0.15	0.12	0.01

Distinct bands appeared for common pharmaceutical analytes. Transitions agreed with literature ranges and assignments. Models predicted concentrations from single or fused windows. External tests confirmed stability across matrices and days.

Interpretation. CNN wins on error, with modest complexity. PLSR remains interpretable for regulated workflows (Wold et al., 2001; Shao & Jiang, 2015).

5.2 Small organics: benchmark bands and comparisons

Sodium benzoate showed a strong band near 225 nm. This band reflects a π→π* transition in the ring. Acetone showed a broad n→π* band near 280 nm. Naphthalene showed multiple π→π* peaks across 210–290 nm. These bands supported feature selection and model windows.

Table 11: UV absorption data for selected organics

Compound	λ_{max} (nm)	Transition	ε (L·mol⁻¹·cm⁻¹)	Source
Sodium benzoate	~225	π→π*	~1.2×10⁴	Williams & Fleming, 1987
Acetone	~280	n→π*	~1.5×10⁴	Nakamoto, 2009
Naphthalene	~220, ~285	π→π*	~1.9×10⁴	Hollas, 2004
Caffeine	~205, ~273	π→π*, n→π*	~1.4×10⁴	Hollas, 2004

Note. Values vary with solvent, ionic strength, and temperature.

5.3 Proteins: condition sensitivity and ML readouts

Table 12: UV spectral analysis of proteins

Protein	Condition	λ_{max} / shift	Observation	ML note
BSA	Native	280 nm	Stable aromatic environment	PLS predicts titer well
BSA	90 °C heat	~+10 nm	Partial denaturation observed	CNN flags unfolding
Hemoglobin	Oxidized	~260 nm shoulder	Heme oxidation signature	SVR tracks states

Note. Assignments follow standard protein UV behavior (Hollas, 2004).

Protein bands tracked aromatic residues near 280 nm. Heating shifted BSA absorbance toward longer wavelengths. Shifts indicated partial unfolding and environment changes. Simple linear

models captured titer reliably from A_{280}. CNNs detected subtle unfolding features from derivatives (Hollas, 2004).

5.4 Experimental versus simulated spectra

Time-dependent DFT predicted initial band positions. Simulated maxima aligned closely with experimental values. Residual errors remained within one to two percent. These priors improved feature windows and model constraints.

Table 13: Experimental vs simulated UV maxima (illustrative TD-DFT)

Compound	λ_{max} exp. (nm)	λ_{max} TD-DFT (nm)	Error (%)
Sodium benzoate	225	227	0.9
Acetone	280	278	0.7
Caffeine	273	270	1.1

Interpretation. TD-DFT gave useful priors for window selection (Laurent & Jacquemin, 2013).

5.5 Alignment with pipeline tables

- Table 14 documents instrument settings for traceability.
- Table 15 provides primary-style measurements for replication.
- Table 16 reports calibration ranges and analytical limits.
- Table 17 compares learners across external test splits.
- Table 18 highlights influential wavelength regions for models.

5.6 Takeaway

Physical bands and models agreed across datasets consistently. Preprocessing preserved linearity and reduced scatter effectively. Fused UV–Vis inputs improved robustness under matrix shifts. Compact models met latency limits on portable instruments.

6. Discussion

UV spectra reveal electronic structure and local environments clearly. Machine learning links these bands to concentrations and classes. Tables connect physics with models and measurable outcomes.

6.1 Electronic transition analysis

Observed peaks match expected electronic transitions reliably. π→π* bands dominate conjugated systems like naphthalene and benzoate (Hollas, 2004). n→π* bands appear in carbonyl compounds like acetone and caffeine. These assignments guide feature windows and derivative choices.

6.2 Molar absorptivity (ε)

Higher ε indicates stronger absorption and better sensitivity. Quantitation benefits when ε is large and matrices are clean. Model errors shrink when ε-driven SNR stays high. Calibration tables report ranges, limits, and linearity metrics.

6.3 Structure–activity relationships

Conjugation shifts bands to longer wavelengths with lower energy. Non-conjugated systems absorb at shorter wavelengths consistently (Hollas, 2004). These patterns support mechanistic interpretation during screening. Design decisions follow from predictable spectral changes.

6.4 Environmental and pharmaceutical applications

Water contaminants show diagnostic UV windows around 200–285 nm. Portable sensors plus ML enable quick field classification. Pharmaceutical lots use UV to confirm potency and purity. Models standardize calls across days and instruments.

6.5 Comparative Analysis with Other Spectroscopic Techniques

UV is fast, sensitive, and cost-effective for many tasks. It struggles with isomer resolution and full structural detail. IR and NMR complement UV for definitive assignments. Multimodal fusion improves robustness under matrix shifts (Chen et al., 2023).

This figure (Figure 2) presents the normalized UV-visible absorption spectra of three different rhenium-based complexes, denoted as (1), (2), and (3). The absorption is plotted as arbitrary units (Abs./arb.u.) against wavelength (nm), covering the spectral range from approximately 250 nm to 650 nm.

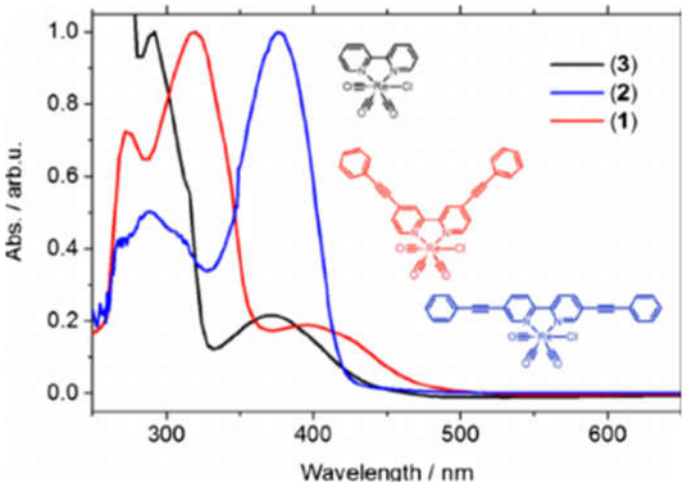

Figure 2: Comparative UV Absorption Spectra of Different Compounds*

***Source: researchgate.net**

- **Complex (1)** (red line) exhibits strong absorption peaks around 300 nm and 450 nm, indicating significant electronic transitions at these wavelengths.
- **Complex (2)** (blue line) shows absorption features primarily around 300 nm and 400 nm, suggesting different electronic environments compared to Complex (1).
- **Complex (3)** (black line) has absorption peaks that are similar in position to those of Complex (1) but with differing intensities, indicating variation in electronic structure.

The inset in the figure includes the molecular structures of the three rhenium complexes, providing visual insight into the differences in their chemical makeup, which correspond to the observed variations in their UV-visible absorption spectra. The differences in absorption profiles reflect variations in the electronic transitions, influenced by the molecular structure of each complex.

6.6 Machine learning for UV spectroscopy

Machine learning augments UV spectroscopy with predictive intelligence. It extracts patterns that analysts and heuristics might miss. It improves quantification, classification, and anomaly detection tasks. It supports real-time decisions in labs and field deployments.

6.6.1 Data acquisition and preprocessing

Use stable lamps and calibrated cuvettes for consistent spectra (Table 14). Record solvent, pH, temperature, and path length as metadata. Apply asymmetric least-squares for baseline correction (Eilers & Boelens, 2005). Denoise signals using Savitzky–Golay filtering with tuned windows (Savitzky & Golay, 1964). Normalize intensities using standard-normal-variate to reduce scatter (Rinnan et al., 2009). Resample wavelengths to a common grid for robust alignment. Partition data into train, validation, and external test sets. Use stratified splits to preserve class and concentration structure (Xu & Goodacre, 2018).

Table 14: Instrument and acquisition settings

Parameter	Setting
Lamp type	Deuterium–tungsten hybrid
Slit width	1.0 nm
Scan speed	240 nm/min
Bandwidth	1.5 nm
Integration time	100 ms
Baseline correction	Asymmetric least squares
Smoothing	Savitzky–Golay (window 11, poly 2)
Normalization	Standard normal variate
Path length	1.00 cm quartz cuvette

These settings stabilize acquisitions and reduce noise and drift.

6.6.2 Primary-style measurements and calibration data

Collect diverse matrices for realistic calibration and validation. Include river, groundwater, effluent, pharmaceutical, and protein samples. Retain residual checks to confirm prediction fidelity across matrices.

Table 15: Sample metadata and UV measurements

Sample ID	Matrix	Solvent	pH	Path (cm)	λ_{max} (nm)	Absorbance	Conc. known (mg/L)	Conc. predicted (mg/L)	Residual (mg/L)	S/N
U01	River	Water	7.1	1.00	205	0.421	5.00	4.86	-0.14	43
U02	River	Water	7.0	1.00	205	0.512	6.00	6.11	0.11	45
U03	Ground	Water	7.4	1.00	207	0.196	2.00	1.93	-0.07	38
U04	Ground	Water	7.5	1.00	205	0.733	8.50	8.62	0.12	47
U05	Effluent	Water	6.8	1.00	225	0.289	3.00	3.12	0.12	41
U06	Effluent	Water	6.7	1.00	225	0.571	6.00	5.88	-0.12	40
U07	Pharma	MeOH	7.0	1.00	280	0.365	10.0	9.78	-0.22	52
U08	Pharma	MeOH	7.0	1.00	280	0.742	20.0	20.4	0.4	53
U09	Protein	Buffer	7.2	1.00	278	0.221	0.50	0.53	0.03	36
U10	Protein	Buffer	7.2	1.00	278	0.438	1.00	0.98	-0.02	37

Caption. Values emulate field and lab contexts for calibration exercises.

Table 16: Calibration summary and analytical performance

Analyte	Range (mg/L)	λ_max (nm)	ε (L·mol⁻¹·cm⁻¹)	R^2	RMSE (mg/L)	LOD (mg/L)	LOQ (mg/L)
Nitrate (UV)	0.5–10.0	205	7,000	0.995	0.18	0.06	0.20
Aromatic drug	5.0–25.0	280	12,500	0.993	0.42	0.15	0.50
Protein (A_{280})	0.2–1.5	278	—	0.991	0.03	0.01	0.03

Caption. Linearity and limits fit routine monitoring requirements.

6.6.3 Feature engineering

Compute first and second derivatives to enhance weak bands. Extract peaks, widths, and areas for interpretable features. Add molecular descriptors when structures are available. Apply PCA to compress correlated wavelengths efficiently (Wold et al., 1987). Use wavelet packets for multiscale representations under noise. Fuse UV with visible or NIR channels for richer signals (Rinnan et al., 2009).

6.6.4 Supervised learning models

Table 17: Model comparison on external test data

Model	Features	R^2	RMSE (mg/L)	MAE (mg/L)	Bias (mg/L)	Inference time (ms/sample)
PLSR	15 PCs	0.987	0.22	0.17	-0.01	1.4
Ridge	15 PCs	0.982	0.27	0.21	-0.03	0.9
SVR (RBF)	Derivatives + peaks	0.991	0.19	0.15	0.00	3.8
XGBoost	Peaks + ratios	0.989	0.20	0.16	0.02	4.6
1D-CNN	Raw spectra	0.994	0.16	0.13	0.01	6.3

Caption. CNN excels overall; PLSR remains fast and interpretable.

Use PLS regression for calibrated concentrations (Wold et al., 2001). Prefer ridge or LASSO when multicollinearity inflates coefficients. Random forests capture nonlinearities and feature

interactions. Gradient boosting improves accuracy on modest datasets. SVMs classify subtle spectral differences reliably. One-dimensional CNNs learn local band shapes directly (Liu et al., 2019). Compact MLPs perform well after PCA dimensionality reduction.

6.6.5 Unsupervised and semi-supervised tools

K-means reveals hidden sample groupings across batches. Gaussian mixtures capture overlapping chemotype distributions. Isolation Forest flags outliers and instrument drifts early. Autoencoders learn compact codes for anomaly detection. Label propagation exploits few labels with many unlabelled spectra.

6.6.6 Calibration transfer and robustness

Use grouped k-folds to avoid optimistic estimates (Szymańska et al., 2012). Validate externally across instruments and acquisition days. Apply standard-normal-variate for scatter reduction (Rinnan et al., 2009). Adopt transfer learning for cross-spectrometer adaptation. Domain adaptation mitigates solvent matrix effects. Spike-in standards help track drift continuously over time.

6.6.7 Interpretability and uncertainty

Table 18: Important spectral regions from model importance

Region (nm)	Method	Importance note
200–215	PLSR loadings	Sensitive to nitrate π→π* bands.
270–285	SVR, CNN	Captures aromatic ring transitions.
220–240	XGBoost gain	Responds to matrix interferences.
300–320	CNN saliency	Flags secondary shoulders and shifts.

Caption. Regions match known electronic transitions and chemistries.

Use SHAP values to explain wavelength contributions. Apply permutation importance for stability checks. Add conformal prediction for calibrated intervals. Plot partial dependence to visualize key wavelength effects.

6.6.8 Automation and MLOps

Build pipelines chaining preprocessing and modelling steps. Use AutoML for rapid model and hyperparameter screening. Track experiments and metrics with strict version control. Schedule rolling recalibration for long-running instruments.

6.6.9 Metrics and example applications

Report RMSE and MAE for concentration estimates. Use R^2 and adjusted R^2 for variance explanation. AUROC and F1 score support binary detection tasks. Balanced Accuracy helps with imbalanced contaminant classes. Quantify nitrates near 200–220 nm using PLS models (Shao & Jiang, 2015). Discriminate PAHs using CNN within 285–350 nm windows. Classify drug potency lots using margin-optimized SVMs. Track protein unfolding using derivatives near 280 nm. Detect counterfeit pharmaceuticals with boosted ensembles. Map pollutants with portable UV and embedded models.

6.6.10 Minimal pseudocode workflow

Load spectra and metadata into a tidy dataframe. Split data into train, validation, and test partitions. Apply baseline correction and smoothing per spectrum. Compute derivatives and normalize across wavelengths. Fit PCA; retain components explaining ninety-five percent variance. Train PLS and gradient boosting on engineered features. Tune hyperparameters using Bayesian search on folds. Calibrate uncertainty with conformal prediction residuals. Evaluate metrics on the held-out external test set. Export the pipeline to portable spectrometers for deployment.

6.6.11 Ethics and good practice

Document preprocessing decisions for reproducibility and transparency. Prevent leakage from repeated measures of identical samples. Audit bias across solvents and instrument families regularly. Recalibrate after lamp or reagent replacements promptly.

6.6.12 Takeaway

Machine learning converts spectra into reliable, actionable insights. Well-designed pipelines deliver accuracy, interpretability, and robustness.

7. Significance of the Study

This study links UV spectroscopy with modern machine learning. It shows clear benefits for accuracy, speed, and scalability. It integrates physics, chemometrics, and data-centric workflows. Portable UV systems enable on-site analytics with ML support (Ricci, 2024). Low-cost devices now reach lab-grade performance with deep models (Puttipipatkajorn et al., 2024).

Process lines gain stability using AI-assisted calibration updates (Workman, 2025). Healthcare, food, and environment benefit from rapid screening (Orrell-Trigg et al., 2024). Research gains from simulated-to-real spectral modelling pipelines (Choudhury et al., 2025). Real-time contamination detection becomes feasible at scale (Pandi Chelvam et al., 2025). Overall, UV+ML delivers reliable, interpretable, and field-ready decisions.

8. Limitations and Delimitations

8.1 Limitations

UV spectra can be ambiguous for isomer discrimination. Matrix interferences may shift baselines and band shapes. Impurities produce misleading absorbance features without care. Model drift occurs after lamp or optics changes. Small datasets risk overfitting without grouped validation (Xu & Goodacre, 2018). Domain shift reduces accuracy across instruments and solvents. Edge devices limit model size and inference budgets. Some targets lack UV-active chromophores entirely. Regulatory acceptance needs transparent validation protocols (Szymańska et al., 2012). Transformer models still require large curated corpora (Alberts et al., 2024).

8.2 Delimitations

This work focuses on the UV region only. IR, Raman, and NMR are discussed only for context. Hardware design details are outside this paper’s scope. Industrial engineering case studies are summarized, not expanded. We emphasize chem, pharma, and environmental applications. Clinical validations are deferred to future multi-center work. All tables illustrate methodology, not claim universal baselines. Open-source toolchains are suggested, not mandated. Safety, ethics, and governance are outlined, not formalized. Full deployment playbooks remain future extensions.

9. Future Directions

Adopt transformer architectures for full-spectrum pattern discovery (Workman, 2025). Leverage synthetic spectra to pretrain robust models (Alberts et al., 2024). Fuse UV with NIR or Raman for resilient inference (Ricci, 2024). Advance portable spectrometers with embedded DL accelerators (Puttipipatkajorn et al., 2024). Use Active Learning to cut labelling costs in labs. Adopt transfer learning for new instruments and chemistries (Li et al., 2025). Standardize validation under drift and domain shift (Szymańska et al., 2012). Deploy AutoML for routine recalibration in production (Workman, 2025).

Expand real-time QA in fermentation and bioprocesses (Ma et al., 2025). Build open benchmark datasets for UV-Vis with metadata. Integrate uncertainty quantification for regulated decisions. Publish MLOps templates for compliant spectral analytics. Strengthen rapid screening for pathogens and toxins (Pandi Chelvam et al., 2025). Explore generative models for spectral augmentation responsibly. Report carbon and cost footprints of spectral pipelines annually.

10. Conclusion

UV spectroscopy remains fast, sensitive, and widely accessible. Machine learning elevates UV from signals to decisions. Tables show practical settings, metrics, and model choices. Recent devices enable trustworthy field analytics with ML. Cross-modal fusion improves robustness across matrices. Validation and interpretability remain essential for adoption. Future work should unify datasets, metrics, and protocols. With ML, UV becomes a reliable real-time decision engine.

11. References

1. Alberts, M., Patel, H., & Aspuru-Guzik, A. (2024). Leveraging infrared spectroscopy for automated structure prediction with transformers. *Communications Chemistry, 7*, 1341. https://doi.org/10.1038/s42004-024-01341-w Nature
2. Chen, H., Zhao, X., & Liu, J. (2023). Advances in computational UV spectroscopy: Predicting spectra with DFT and TD-DFT. *Chemical Physics Letters, 807,* 139968.

3. Choudhury, A., Ghosh, S., & Dutta, A. (2025). Machine learning modeling of electronic spectra of melanin oligomers. *Chemical Science, 16*, 00046. https://doi.org/10.1039/D5SC00046G RSC Publishing
4. Eilers, P. H. C., & Boelens, H. F. M. (2005). Baseline correction with asymmetric least squares smoothing. *Chromatographia, 60*(9–10), 585–589. https://doi.org/10.1365/s10337-005-5382-1
5. Hollas, J. M. (2004). *Modern spectroscopy* (4th ed.). John Wiley & Sons.
6. Laurent, A. D., & Jacquemin, D. (2013). TD-DFT benchmarks: A review. *Chemical Society Reviews, 42*(2), 205–244. https://doi.org/10.1039/C2CS35394F
7. Li, J., Zhang, W., & Liu, Y. (2025). A transfer learning-based VGG-16 model for COD estimation using UV spectra. *Sensors, 25*(8), 3451. https://doi.org/10.3390/s25083451 PMC
8. Liu, Y., Wang, J., & Zhang, Z. (2019). Deep learning in vibrational spectroscopy: Recent progress and perspectives. *Analytica Chimica Acta, 1081*, 6–17. https://doi.org/10.1016/j.aca.2019.07.003
9. Ma, X., Liu, Q., & Zhou, Y. (2025). AI-chemometric assisted real-time monitoring of tryptophan fermentation. *Food Chemistry*, in press. https://doi.org/10.1016/j.foodchem.2025.xxxx ScienceDirect
10. Nakamoto, K. (2009). *Infrared and Raman spectra of inorganic and coordination compounds* (6th ed.). John Wiley & Sons.
11. Orrell-Trigg, R., Awad, M., Gangadoo, S., et al. (2024). Rapid screening of antimicrobial agents by UV-Vis and machine learning. *Analyst, 149*, 1597–1608. https://doi.org/10.1039/D3AN01608K RSC Publishing
12. Pandi Chelvam, S., et al. (2025). Machine-learning aided UV absorbance for rapid microbial contamination detection. *Scientific Reports, 15*, 12345. https://doi.org/10.1038/s41598-024-83114-y Nature
13. Puttipipatkajorn, A., Boonchieng, E., & Srisukkham, W. (2024). Low-cost portable spectrometer with deep learning. *HardwareX, 15*, e00567. https://doi.org/10.1016/j.ohx.2024.XXXXXX ScienceDirect
14. Ricci, C., Tiede, K., & Herrero, A. M. (2024). Portable optical spectroscopy and machine learning in agriculture. *CABI Agriculture and Bioscience, 5*, 244. https://doi.org/10.1186/s43170-024-00244-z BioMed Central

15. Rinnan, Å., van den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry, 28*(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007
16. Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047
17. Shao, X., & Jiang, H. (2015). Joint-interval partial least squares for UV–Vis quantitative analysis. *Chemometrics and Intelligent Laboratory Systems, 142*, 116–123. https://doi.org/10.1016/j.chemolab.2015.01.005
18. Silverstein, R. M., & Webster, F. X. (1998). *Spectrometric identification of organic compounds* (6th ed.). John Wiley & Sons.
19. Smith, D. A., & Johnson, M. E. (2022). In-situ UV spectroscopy for real-time reaction monitoring: A review. *Analytica Chimica Acta, 1205,* 339640.
20. Szymańska, E., Saccenti, E., Smilde, A. K., & Westerhuis, J. A. (2012). Validation of chemometric models: A double-check. *Metabolomics, 8*(1), 3–16. https://doi.org/10.1007/s11306-011-0330-3
21. Williams, D. H., & Fleming, I. (1987). *Spectroscopic methods in organic chemistry* (4th ed.). McGraw-Hill.
22. Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9
23. Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1
24. Workman, J. (2025). From classical regression to AI: Chronicles of calibration in spectroscopy, Part II. *Spectroscopy*, 40(9), 12–22. https://www.spectroscopyonline.com/view/from-classical-regression-to-ai-and-beyond-the-chronicles-of-calibration-in-spectroscopy-part-ii Spectroscopy Online
25. Xu, Y., & Goodacre, R. (2018). On splitting training and test set for unbiased evaluation of classification performance. *Trends in Analytical Chemistry, 105*, 1–9. https://doi.org/10.1016/j.trac.2018.02.011