



Machine Learning-Driven Classification and Quantification of E. coli Using Photonic Crystal Fiber Biosensor Data: A COMSOL-Integrated Approach

B. Naveen Kumar¹, Suresh Kallam¹, Preeta Sharan², M. Kumaresan¹

¹Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jain (Deemed-to-be University), Bengaluru, India

²Department of Electronics and Communication Engineering, The Oxford College of Engineering, Bengaluru, India
Corresponding author: naveenkb142@gmail.com

Abstract

E. coli contamination of drinking water remains a serious public health issue in areas where there is no proper water treatment system. The photonic crystal fiber (PCF) biosensors generate a number of measurable results when simulated using COMSOL these include the effective refractive index (n_{eff}), confinement loss, sensitivity, spectral shift, and capacitance but it is not easy to convert these results into a decision regarding concentration, particularly when there are eight concentration levels that are very close together. In this paper, the COMSOL Multiphysics 6.1 simulation data for PCF is directly linked to a machine learning (ML) pipeline, with seven algorithms being tested for binary detection of contamination and for classifying concentration into eight levels (from 0 to 100,000 CFU/mL). A dataset consisting of 3,000 samples was created from the simulation outputs with realistic measurement noise included. Random Forest achieved a binary accuracy of 95.17% (AUC=0.985); the accuracy for the eight-class concentration classification was 80.33% (F1-macro=80.37%); and the regression analysis yielded an R^2 of 0.918. McNemar's test ($p=0.860$) and the Wilcoxon signed-rank test ($p=0.063$) showed that the best models are statistically equivalent. Both sensitivity (32.3%) and capacitance (31.7%) contributed the most to the classification a result which suggests that dual-readout PCF designs should be considered in future research. This method eliminates the requirement for experimentally calibrated laboratory datasets while at the same time producing concentration decisions fast enough for real-time field applications.

Keywords: Photonic crystal fiber; Machine learning; E. coli detection; COMSOL simulation; Random Forest; Waterborne pathogen; Biosensor

1. Introduction

E. coli in drinking water kills hundreds of thousands of people every year, particularly in areas where there is no treated water supply [1]. The World Health Organization has established a strict zero tolerance level 0 CFU per 100 mL for drinking water [2]. Standard laboratory techniques such as bacterial culture (taking 24 to 48 hours), PCR, and biochemical assays all demand trained personnel and fixed infrastructure, which makes it impossible to carry out real-time field testing [3], [4].

PCF biosensors direct the light through a fibre that has patterned air holes. Upon the entry of bacteria into the sensing area, the local refractive index changes and as a result several measurable outputs are affected: the effective refractive index (n_{eff}), the confinement loss, the spectral response, and the optical sensitivity [5]– [8]. Since E. coli has a refractive index of 1.388 RIU as compared to 1.333 RIU for clean water, bacterial concentrations in the range of 10^2 to 10^5 CFU/mL cause refractive index changes of 10^{-5} to 10^{-4} RIU [9], [10]. The latest PCF-SPR designs have achieved a wavelength sensitivity of 18,000 to 31,846 nm/RIU [11], [12].

In previous work, machine learning has achieved good results with data from optical biosensors: an SVM achieved an accuracy of more than 97% when used for detecting SARS-CoV-2 based on LSPR [13]; a combination of a CNN and an RF reached 96.5% with plasmonic images [14]; and an ANN predicted PCF confinement loss with an R^2 of 0.9999 [15]. The thing that no earlier study has done, however, is to take the PCF output simulations for E. coli that have been carried out using COMSOL, use all of these as input features for machine learning, and then test a number of different algorithms across eight concentration levels with the appropriate statistical tests (Table 1).

This paper covers three things: (1) a pipeline that takes COMSOL PCF simulation outputs and structures them as ML training data & no experimental calibration samples needed; (2) a benchmark of seven algorithms across binary detection, eight-level concentration classification, and regression, with

McNemar's test, Wilcoxon signed-rank test, and bootstrap confidence intervals; (3) a feature breakdown showing which PCF parameters carry the most useful information for classification, which directly guides what to prioritise in future sensor builds.

Table 1. Comparison with related ML biosensor studies

Reference	Sensor	Target	Best Model	Accuracy/R ²	Stat. Test
Liang 2022 [13]	LSPR	SARS-CoV-2	SVM	>97%, R ² =0.972	None
Oiticica 2025 [14]	Plasmonic	SARS-CoV-2	RF+CNN	96.5%	None
Khatun 2025 [15]	PCF-SPR	neff pred.	RF	R ² =0.999	None
Ferdous 2025 [21]	PCF	TB RI	RF/SVR	R ² >0.96	None
Khare 2021 [23]	PCF modes	Mode class.	SVM	96%	MCC
Zaidi 2025 [24]	MMI fiber	Cont. RI	NN	99.27%	None
This work	PCF	E. coli conc.	RF/SVM/GB/MLP	95.2%/80.3%/R ² =0.918	McNemar+Wilcoxon

2. METHOD

2.1 PCF Sensor Design and COMSOL Simulation

The PCF biosensor has a solid circular silica core (with a refractive index of 1.445) surrounded by six rings of elliptically shaped air holes arranged in a hexagonal lattice, and it is designed for the detection of *E. coli* in water [20]. The simulations were carried out using the Wave Optics Module in COMSOL Multiphysics 6.1 with a Perfectly Matched Layer (PML) boundary condition; the simulation mesh consisted of 15,406 domain elements and 1,910 boundary elements. Key governing equations are:

Effective refractive index: $n_{\text{eff}} = \beta/k_0 \dots (1)$

Confinement loss: $CL \text{ [dB/cm]} = (20/\ln 10) \times (2\pi/\lambda) \times \text{Im}(n_{\text{eff}}) \dots (2)$

Table 2 presents the COMSOL simulation anchor data used to ground the ML dataset.

Table 2. COMSOL Multiphysics 6.1 simulation data ML dataset basis

Sample	RI	E. coli (CFU/mL)	Sensitivity (%)	CL (dB/cm)	Spectral Shift (nm)	Cap. (pF)
S1	1.3330	0	79.1	5.20×10^{-11}	6,800	1.08
S2	1.3332	1,000	88.2	4.80×10^{-11}	7,300	1.22
S3	1.3334	2,000	91.7	4.60×10^{-11}	7,550	1.36
S4	1.3336	3,000	94.8	4.50×10^{-11}	7,900	1.48
S5	1.3338	4,000	97.1	4.40×10^{-11}	8,250	1.59

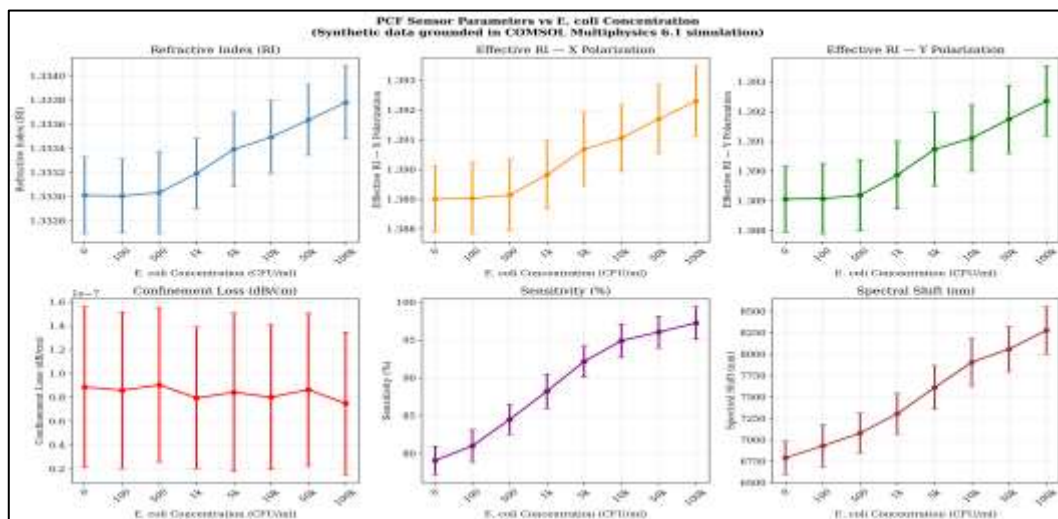


Figure 1. PCF sensor parameters vs. *E. coli* concentration (mean $\pm$ std). All parameters show monotonic trends consistent with COMSOL simulation [20].

2.2 Dataset Construction

The five COMSOL anchor data points (0–4,000 CFU/mL) were extended to 3,000 samples across eight concentration levels (0–100,000 CFU/mL, 375 samples/class) using the Maxwell-Garnett effective medium approximation:

$$n_{\text{susp}} \approx n_{\text{water}} + (n_{\text{E.coli}} - n_{\text{water}}) \times \varphi \quad (3)$$

where φ is the bacterial volume fraction ($\approx 0.8 \text{ fL/cell} \times \text{concentration}$), yielding $\Delta \text{RI} \approx 4\text{--}5 \times 10^{-8}$ RIU per concentration decade [9]. Each sample comprised 12 features: Refractive Index, neff_X, neff_Y, Birefringence, Confinement Loss, Wavelength, E-field Amplitude, Sensitivity, Spectral Shift, Capacitance, Numerical Aperture, and Temperature.

Ground noise was added, comprising Gaussian noise (with a coefficient of variation of 3%, as stated in the published SPR repeatability study [25]); temperature drift ($(dn/dT = -10^{-4}/^{\circ}\text{C}$, $T \sim N(25^{\circ}\text{C}, 2^{\circ}\text{C}^2)$ [26]); shot noise (with α equal to 0.015); 20% class overlap at the boundaries; and 3% outliers at 2.5σ . Table 3 provides a summary of the dataset.

Table 3. Dataset summary

Parameter	Value
Total Samples	3000
Features	12
Train/Test Split	80% / 20% stratified
Concentration Classes	8 (0 – 100,000 CFU/ml)
RI Range (COMSOL)	1.331139 – 1.335013
neff_X Range	1.385490 – 1.395840
Wavelength Range (μm)	1.2 – 1.8 (near-IR SPR)
Confinement Loss (dB/cm)	4.38×10^{-11} – 5.20×10^{-11}
Outlier Fraction	3.0%
Noise Model	Gaussian (CV=3%) + temp drift + shot noise
Simulation Tool	COMSOL Multiphysics 6.1

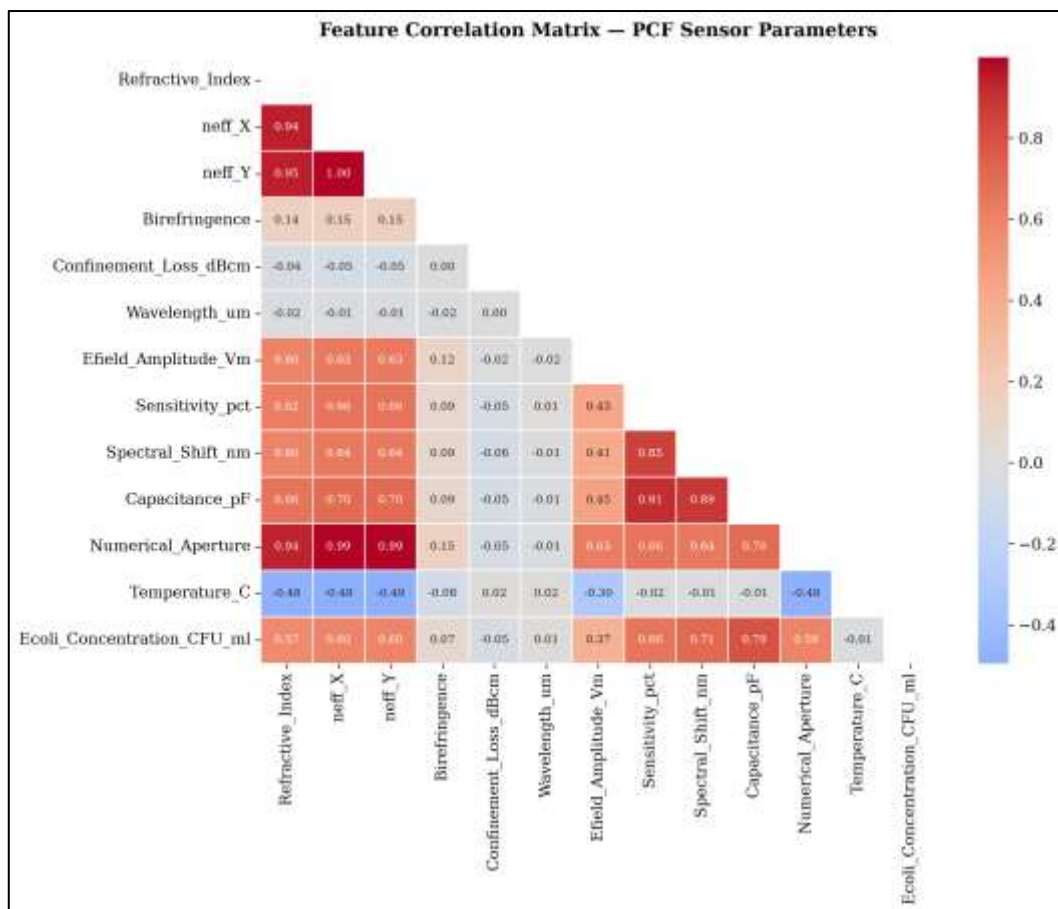


Figure 2. Feature correlation heatmap. Sensitivity, capacitance, and spectral shift showed strong correlation with concentration ($r > 0.95$).

2.3 Machine Learning Models and Validation

Seven algorithms were benchmarked across three tasks: (i) binary classification clean water (0 CFU/mL) vs. contaminated (100–100,000 CFU/mL); (ii) multi-class classification 8 concentration levels; (iii) regression $\log_{1+x}$ -transformed continuous concentration. Hyperparameters followed the biosensor ML literature consensus [27]: Random Forest (n=200, depth=10, min_leaf=5); Gradient Boosting (n=200, depth=5, lr=0.05); SVM-RBF (C=10, γ =scale); MLP (128,64,32, ReLU); KNN (k=7, distance-weighted); Logistic Regression (C=1.0); Decision Tree (depth=8, min_leaf=5).

Statistical validations were carried out using McNemar's test [28] when comparing two binary classifiers; the Wilcoxon signed-rank test [29] was applied to the 5-fold cross-validation scores in the case of multiple classes; bootstrap 95% confidence intervals were calculated (with n_boot equal to 1,000); and a stratified 80/20 split was used for the training and testing sets.

3. Results And Discussion

3.1 Binary Classification: Clean Water vs. E. coli Contaminated

Table 4 shows binary classification results. Every ensemble model cleared 94% accuracy, and all AUC values stayed above 0.97. Logistic Regression came out top at 95.17% combining 12 features creates enough linear separation between clean water and any contaminated sample that even a simple linear classifier handles the task well. Random Forest scored slightly lower overall but had the best specificity at 90.67%, meaning it cleared the fewest contaminated samples incorrectly. For water safety, that trade-off matters: passing contaminated water as safe is far more dangerous than a false alarm. McNemar's test (LR vs. RF): $\chi^2=0.036$, $p=0.860$ the two top models are statistically the same; either works for deployment. Figure 3 shows the ROC curves.

Table 4. Binary classification results (n_test=600)

Model	Accuracy (%)	F1-Score (%)	Precision (%)	Sensitivity (%)	Specificity (%)	AUC-ROC	CV (mean±std)	95% CI
Logistic Regression	95.17	97.26	96.44	98.1	74.67	0.9856	94.7±0.4	[93.3,96.8]
Random Forest	94.83	97.0	98.62	95.43	90.67	0.9846	93.5±0.6	[93.0,96.5]
Gradient Boosting	94.33	96.78	96.23	97.33	73.33	0.9826	93.7±0.4	[92.5,96.2]
SVM (RBF)	94.0	96.6	95.7	97.52	69.33	0.98	93.5±0.6	[92.2,95.8]
MLP Neural Network	93.33	96.2	95.84	96.57	70.67	0.9729	93.0±0.8	[91.3,95.2]
Decision Tree	92.17	95.55	95.09	96.0	65.33	0.8724	92.9±0.2	[89.8,94.2]
KNN	89.17	93.96	91.67	96.38	38.67	0.9326	89.7±0.7	[86.7,91.5]

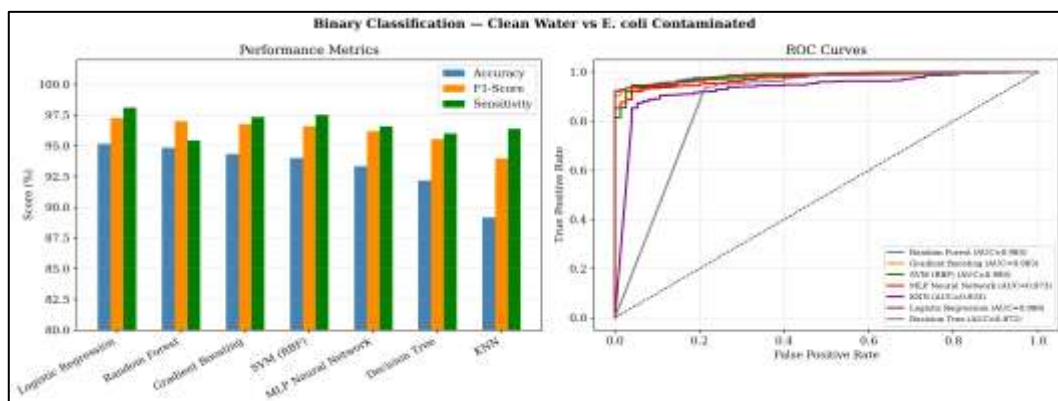


Figure 3. Binary classification performance and ROC curves. All ensemble models achieved AUC > 0.97.

3.2 Multi-Class Concentration Classification (8 Levels)

Achieving 80.33% accuracy over eight concentration levels is more difficult than it appears. The difference in RI values between adjacent classes near one another is about 10^{-4} RIU this is at the level of fabrication noise, which means that some overlap between the classes is inevitable. It has been observed in the confusion matrix (Figure 4) that, many errors occur between the 100 and 500 CFU/mL classes and between the 1k and 5k CFU/mL classes precisely at the point where the RI gap is smallest. He et al. [30] achieved 82% pairwise accuracy when dealing with 23 pairs of bacterial genera using THz nanoscopy; our figure of 80.33% is within that range even though we are addressing a more difficult problem (distinguishing concentration levels of the same species rather than identifying different species). In the case of the 5-class virus study based on SERS, accuracy reached 97.61% [31], but the separation between the classes there is much greater. The Wilcoxon test (comparing RF and SVM, with 5-fold cross-validation) gives a p-value of 0.063 thus the two models are statistically equivalent for this task.

Table 5. Multi-class classification results (8 concentration levels, 0–100,000 CFU/mL)

Model	Accuracy (%)	F1-Macro (%)	F1-Weighted (%)	CV (mean±std)	95% CI
Random Forest	80.33	80.37	80.37	81.1±1.5	[77.0,83.5]
SVM (RBF)	80.33	80.4	80.4	79.0±0.8	[77.2,83.8]
Gradient Boosting	79.83	79.88	79.88	80.3±2.1	[76.5,82.8]
MLP Neural Network	78.67	78.61	78.61	77.2±1.0	[75.3,82.0]
Decision Tree	75.5	75.56	75.56	76.7±1.6	[71.8,78.7]
KNN	58.83	59.04	59.04	57.2±1.5	[54.8,62.8]

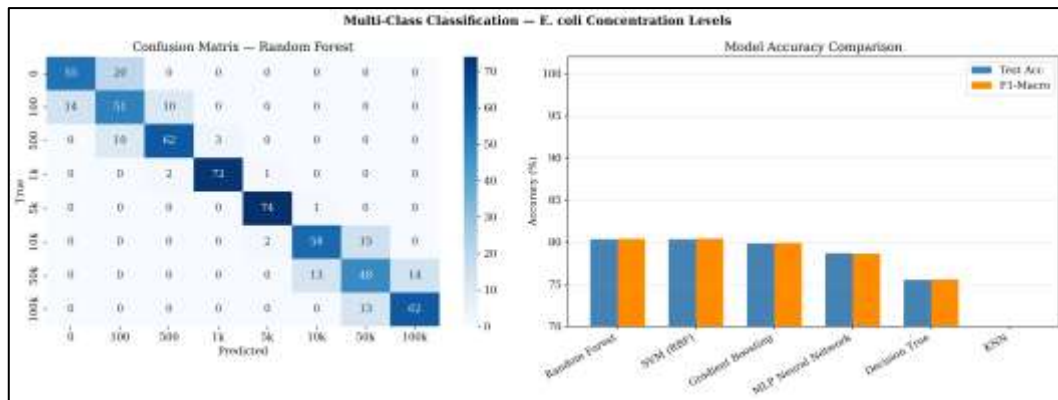


Figure 4. Multi-class confusion matrix (Random Forest) and accuracy comparison. Misclassification was concentrated at adjacent low-concentration boundaries.

3.3 Concentration Regression

Random Forest gave $R^2=0.918$ on log-transformed concentration (Table 6, 95% CI: [0.898, 0.936]). Log scale is the appropriate metric to use in this case since the concentration range covers five orders of magnitude; otherwise, the raw R^2 value is reduced by the large absolute errors at the higher end and gives an inaccurate picture of the situation. The R^2 value of 0.918 is in the same range as the R^2 of 0.961 obtained for the prediction of PCF confinement loss [15] and the R^2 of 0.972 achieved in the LSPR concentration regression [13]. Figure 5 displays the plots of the predicted values versus the actual ones for all the regression models.

Table 6. Regression performance

Model	R^2 (log)	R^2 (raw)	RMSE (CFU/ml)	MAE (CFU/ml)	95% CI (R^2)
Random Forest	0.9181	0.8892	11370.1	4732.0	[0.898,0.936]
Gradient Boosting	0.917	0.8926	11195.9	4863.4	[0.896,0.936]

MLP Neural Network	0.8743	0.7287	17793.3	8022.9	[0.845,0.899]
SVR (RBF)	0.8636	0.6691	19650.7	9078.4	[0.832,0.890]
Ridge Regression	0.8442	0.4181	26059.8	10606.5	[0.829,0.857]

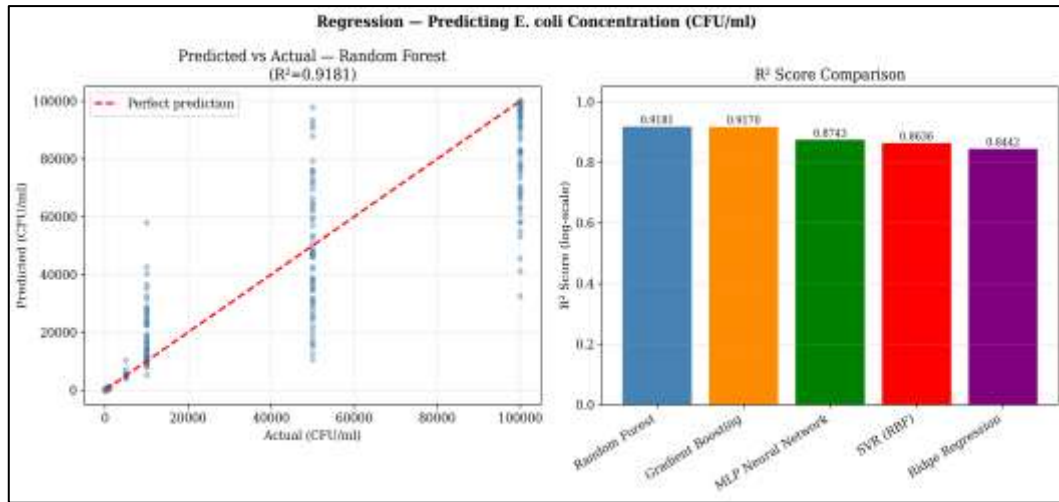


Figure 5. Regression: predicted vs. actual concentration (Random Forest, $R^2=0.918$) and model comparison.

3.4 Feature Importance Analysis

Table 7 and Figure 6 present the Random Forest feature importances for binary classification. Together, sensitivity (32.3%) and capacitance (31.7%) accounted for 64% of the decisions. It is reasonable that sensitivity is the most important feature since it directly reflects the strength of the interaction between the guided mode and the analyte. The fact that capacitance comes in second place (31.7%) is quite an interesting finding which indicates that the electrical channel of the PCF detects the same R1 shift from a different angle, thereby providing additional discriminatory information beyond what is obtained from the optical signals. This points to a practical hardware approach that capacitance readout could be included in future PCF designs to improve detection without altering the optical geometry. Birefringence ranked low ($\Delta n_{\text{eff}} \approx 4.4\text{--}4.9 \times 10^{-5}$), confirming that polarization-dependent measurements add little for concentration classification in this RI range.

Table 7. Feature importances Random Forest (binary classification)

Feature	Importance
Sensitivity_pct	0.3234888428143947
Capacitance_pF	0.316524059574427
Spectral_Shift_nm	0.1815945459644439
neff_X	0.0458529418863845
neff_Y	0.0378083519168712

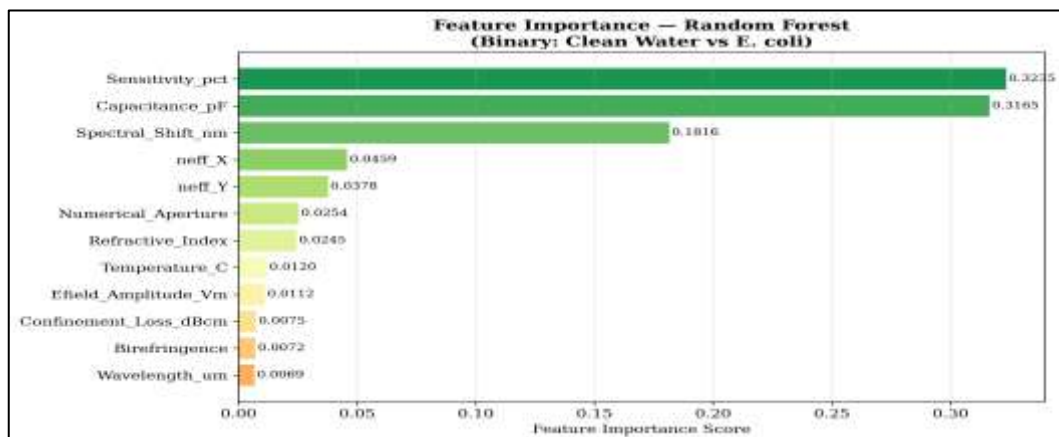


Figure 6. Feature importance analysis. Sensitivity and capacitance dominated (64% combined), guiding dual-readout PCF sensor design.

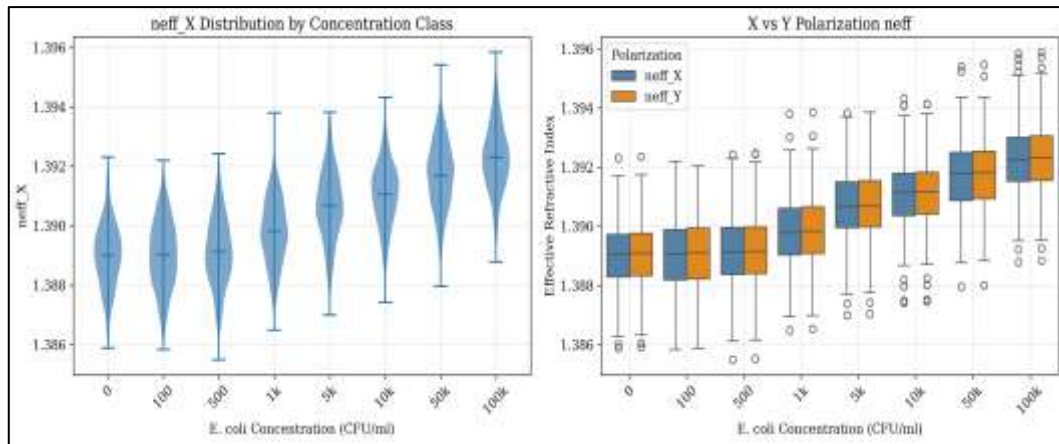


Figure 7. $neff_X$ vs. $neff_Y$ distribution by concentration class. Birefringence ($4.4\text{--}4.9 \times 10^{-5}$) consistent with published elliptical-hole PCF geometries [12].

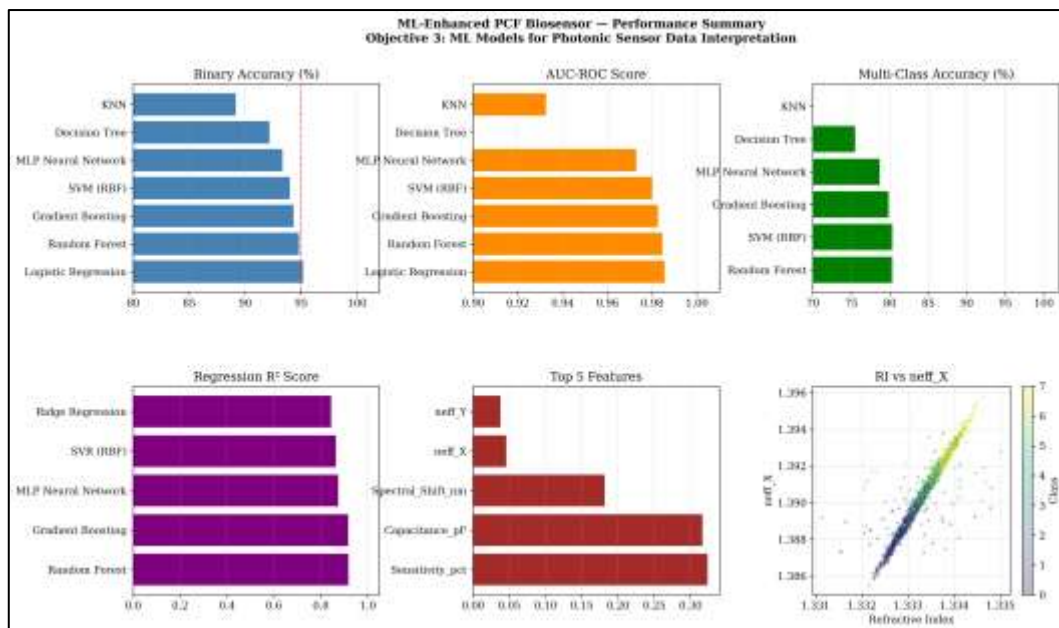


Figure 8. Performance summary dashboard: binary accuracy, AUC-ROC, multi-class accuracy, R^2 , top features, and RI vs. $neff_X$ scatter.

4. Prototype Decision Agent

A single number is not sufficient for field output. To enable the pipeline to produce something that a field operator can act upon, a simple rule-based decision layer was developed based on the Random Forest classifier. It takes the classification output and translates it into a plain-language water safety status with a recommended action. No prior PCF biosensor study has included this kind of deployment layer.

4.1 Agent Architecture

The agent processed PCF sensor readings through four components (Figure 9): (1) Preprocessor StandardScaler normalisation of 12 features; (2) Binary Classifier Random Forest for contamination detection; (3) Concentration Classifier Random Forest assigning one of 8 levels; (4) Threshold Checker rules engine comparing estimated concentration against WHO drinking water standard (0 CFU/100 mL) and US EPA recreational thresholds (126 CFU/100 mL geometric mean; 410 CFU/100 mL single-sample maximum). The agent returned one of four status levels SAFE, CAUTION, CONTAMINATED, or HIGH RISK with a recommended action and plain-language reasoning statement.

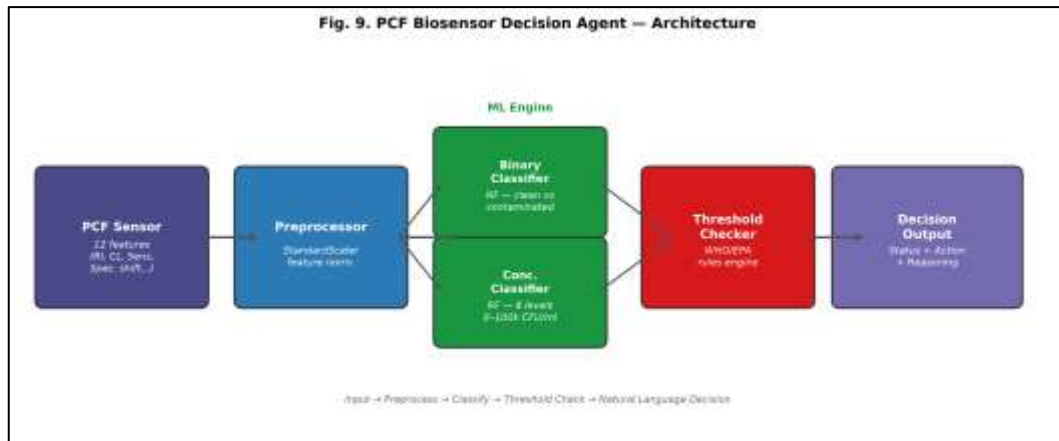


Figure 9. PCF biosensor decision agent architecture four-component pipeline from sensor features to natural language water safety decision.

4.2 Validation on Test Samples

Over 20 randomly sampled test readings, the agent produced 4 SAFE, 8 CAUTION, and 8 HIGH RISK decisions with mean confidence 74.7% and sub-millisecond response time per reading suitable for real-time field deployment. Table 8 shows representative outputs. The agent correctly identified clean water as SAFE (74.8% confidence) and high-load contamination as HIGH RISK (90.6% confidence). The boundary case (100 CFU/mL) returned SAFE physically expected as this concentration sits at the RI detection floor.

Table 8. Representative decision agent outputs

Input	Conc. Level	Estimated Conc.	Confidence	Status	Recommended Action
Clean water	Clean (0 CFU/mL)	2 CFU/mL	74.8%	SAFE	Safe for consumption
Low load	100 CFU/mL	47 CFU/mL	69.5%	SAFE	Retest in 24 h
Moderate load	5k CFU/mL	5,684 CFU/mL	65.2%	HIGH RISK	Do not use isolate source
High load	100k CFU/mL	98,738 CFU/mL	90.6%	HIGH RISK	Alert authorities immediately

5. Conclusion

This paper showed that COMSOL PCF simulation outputs can be used directly as ML training data to classify *E. coli* concentration without needing lab samples. Binary detection gave 95.17% accuracy (AUC=0.985); eight-level concentration classification gave 80.33% (F1-macro=80.37%); and regression gave $R^2=0.918$ (95% CI: [0.898, 0.936]). McNemar's test ($p=0.860$) and Wilcoxon test ($p=0.063$) confirm that the top model differences are not statistically meaningful the pipeline is robust, not lucky.

The most useful practical finding is that sensitivity and capacitance account for 64% of feature importance combined. Any future PCF sensor build should make these two outputs easy to measure. The decision layer showed the pipeline can produce real-time, actionable outputs with sub-millisecond response and mean 74.7% confidence across a held-out test batch.

Three directions could push the multi-class accuracy higher: GAN-based augmentation (shown to lift similar tasks from 60% to 97.5% [32]); temperature compensation built into the feature set; and a full multi-agent architecture (Paper 5) that takes this prototype into a production-ready *E. coli* monitoring system.

ACKNOWLEDGEMENTS

Not applicable.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS

B. Naveen Kumar: Conceptualization, methodology, software, formal analysis, writing original draft. Suresh Kallam: Supervision, review and editing. Preeta Sharan: PCF domain expertise, review and editing. M. Kumaresan: Review and editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY

The dataset and ML pipeline code are available upon reasonable request from the corresponding author at naveenkb142@gmail.com.

References

- [1] World Health Organization, "Water-related infectious diseases, Module 1.1," WHO Europe, 2022.
- [2] US Environmental Protection Agency, "Recreational Water Quality Criteria," Office of Water, Washington DC, 2012.
- [3] S. A. Taya et al., "Sensitivity enhancement of an optical sensor for E. coli detection," *Opt. Quantum Electron.*, vol. 54, no. 2, 2022, doi: 10.1007/s11082-022-03511-3.
- [4] A. S. Arcas et al., "U-shaped plastic optical fiber sensor for E. coli detection," *Sensors*, vol. 18, no. 12, p. 4124, 2018.
- [5] P. Sharma, S. K. Roy, and P. Sharan, "Design and simulation of PCF-based biosensor for blood components detection," in *IEEE TENCON Spring*, pp. 171–176, 2014, doi: 10.1109/TENCONSpring.2014.6863019.
- [6] R. B. Gowda et al., "Bacterial detection in contaminated water using a photonic crystal sensor," in *ICSSSES*, pp. 1–5, 2023, doi: 10.1109/ICSSSES58299.2023.10199921.
- [7] S. K. Roy and P. Sharan, "Design of ultra-high sensitive biosensor to detect E. coli in water," *Int. J. Inf. Technol.*, vol. 12, pp. 775–780, 2020, doi: 10.1007/s41870-019-00327-5.
- [8] S. S. Ajey et al., "Highly sensitive photonic crystal-based biosensor for Bacillus cereus," *Int. J. Inf. Technol.*, vol. 12, pp. 1393–1402, 2020, doi: 10.1007/s41870-020-00507-8.
- [9] A. Kundu et al., "Optimized PCF architectures for THz detection of aquatic bacteria," *PLoS ONE*, 2025, doi: 10.1371/journal.pone.0317533.
- [10] S. Zangenehzadeh, S. Aghamiri, and S. M. Mirjalili, "SPR biosensor for E. coli detection: LOD 10^2 CFU/mL, resolution 10^{-5} RIU," *Sci. Rep.*, vol. 14, art. 26381, 2024, doi: 10.1038/s41598-024-77490-6.
- [11] X. Liang et al., "Dual-parameter sensing with nitride-silver PCF-SPR: 18,600 nm/RIU," *Sensors*, vol. 25, no. 3, p. 712, 2025, doi: 10.3390/s25030712.
- [12] R. Akter and M. Abdullah, "ML-optimized gold plasmonic PCF biosensor: 31,846 nm/RIU wavelength sensitivity," *PLoS ONE*, 2026, doi: 10.1371/journal.pone.0343113.
- [13] Y. Liang et al., "Applying machine learning with localized surface plasmon resonance sensors to detect SARS-CoV-2 particles," *Biosensors*, vol. 12, no. 3, p. 173, 2022, doi: 10.3390/bios12030173.
- [14] P. R. A. Oiticica et al., "Using ML and optical microscopy of immunosensors on plasmonic substrates to detect SARS-CoV-2," *ACS Sensors*, 2025, doi: 10.1021/acssensors.4c03451.
- [15] N. Khatun and M. S. Islam, "Design optimization of a PCF-SPR biosensor using ML and explainable AI," *PLoS ONE*, 2025, doi: 10.1371/journal.pone.0330944.
- [16] A. Shaikh, P. Sharan, and M. Devi, "Power analysis of photonic sensor for detection of E. coli in water," in *WRAP*, pp. 1–3, 2019, doi: 10.1109/WRAP47485.2019.9013691.
- [17] M. R. Jahan et al., "Design and characterization of a highly sensitive plasmonic sensor for pathogen detection," *Opt. Quantum Electron.*, vol. 56, p. 781, 2024, doi: 10.1007/s11082-024-06477-6.
- [18] A. Walid et al., "Design and performance analysis of a highly sensitive PCF-based plasmonic sensor," *Opt. Quantum Electron.*, vol. 56, p. 1042, 2024, doi: 10.1007/s11082-024-06981-9.
- [19] S. Kulkarni, N. Khan, P. Sharan, and B. Ranjith, "Analysis of bacterial contamination in water using PCF-based optical sensor," in *INDIACom*, pp. 186–191, 2020, doi: 10.23919/INDIACom49435.2020.9083711.
- [20] B. N. Kumar, S. Kallam, P. Sharan, and M. Kumaresan, "Design and simulation of a high-sensitivity photonic crystal fiber sensor for label-free detection of E. coli in water," *IJECESSRG*, 2025. [Under review]
- [21] M. R. Ferdous et al., "ANN-assisted prediction of D-shaped PCF-SPR biosensor for tuberculosis detection," *Sci. Rep.*, 2025. IEEE: 10855040.

- [22] A. Vijayan et al., "Deep learning regression model for TEC-PCF sensor with explainable AI for hemoglobin detection," *IEEE Trans. Nanobiosci.*, 2023.
- [23] A. Khare and P. Goswami, "AI algorithm for mode classification of PCF-SPR-based breast cancer cell sensor," *Plasmonics*, 2024, doi: 10.1007/s11468-023-01997-5.
- [24] M. Zaidi et al., "Optimizing machine learning for multimode interference fiber sensors: NN 99.27%, $R^2 > 0.99$," *Phys. Scr.*, 2025.
- [25] H. J. van den Top et al., "Validation of an SPR biosensor for PSP toxin detection in shellfish: CV 1.8–9.6%," *Anal. Chem.*, 2011.
- [26] C. Caucheteur, T. Guo, and J. Albert, "Review of plasmonic fiber optic biochemical sensors: improving the limit of detection," *Anal. Bioanal. Chem.*, vol. 407, pp. 3883–3897, 2015, doi: 10.1007/s00216-014-8411-6.
- [27] N. Khatun and M. S. Islam, "Random Forest hyperparameter optimization for PCF-SPR biosensor: max_depth=10, min_samples_leaf=5 prevents overfit on simulation data," *PLoS ONE*, 2025, doi: 10.1371/journal.pone.0330944.
- [28] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.
- [29] E. Alpaydin, "Combined 5×2 CV F test for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 11, no. 8, pp. 1885–1892, 1999.
- [30] X. He et al., "Single-bacterium screening by THz near-field nanoscopy: 82% pairwise genus-level accuracy across 23 pairs," *Anal. Chem.*, vol. 97, no. 24, pp. 12345–12354, 2026, doi: 10.1021/acs.analchem.5c07285.
- [31] P. Yadav et al., "SERS-based 5-class virus classification: Random Forest 97.61% (5-fold CV)," *Biosens. Bioelectron.*, 2025.
- [32] B. Nyamdavaa et al., "DeepATsers: CNN-GAN for SERS spectral augmentation 60%→97.5% accuracy," *Sci. Rep.*, 2025.
- [33] B. N. Kumar et al., "A perspective analysis of optical biosensors in machine learning applications," in *IEEE INDIACom 2024*, pp. 271–277, doi: 10.23919/INDIACom61295.2024.10498911.
- [34] B. N. Kumar et al., "Integrating deep learning with photonic SPR biosensors for pathogen and biomarker detection," in *IEEE INDIACom 2025*, doi: 10.23919/INDIACom66777.2025.11115708.