A Deep Learning Driven Volatile Organic Compounds Analysis for Lung Cancer Detection Using HC-PCF and Convolutional Neural Networks

Jaitesh Upadhyay*, Shobi Bagga, Dhirendra Mathur

Department of Electronics Engineering, Rajasthan Technical University, India

Received May 9, 2025; accepted June 28, 2025; published June 30, 2025

Abstract—A Volatile Organic Compounds (VOC) detection system for lung cancer diagnosis through deep learning (DL) technology is implemented in a special Hollow-Core Photonic Crystal Fibre (HC-PCF) sensor platform. COMSOL Multiphysics is used to simulate the HC-PCF. A hexagonal lattice structure of silica material with 1 μm pitch dimensions and 0.5 μm air hole diameters allow for exceptional light guidance and VOC interaction when detecting exhaled breath components. The sensor achieves a remarkable refractive index sensitivity of 920 nm/RIU for detecting cancerous and non-cancerous VOC profiles. The refractive index measurements of lung cancer-related VOC samples fell within 1.380 to 1.392, while VOC samples from healthy patients ranged from 1.350 to 1.360. Sensor spectral response data processing relied on a Convolutional Neural Network (CNN) model that was trained to distinguish different VOC signature patterns. When applied to a dataset of 1,200 breath samples consisting of 600 cancer-positive and 600 healthy specimens, the CNN architecture reached a 96.3% overall classification accuracy combined with 94.7% sensitivity and 97.8% specificity.

Lung cancer detection becomes extremely difficult because symptoms appear late, resulting in treatments failing at advanced stages. Consequently, researchers require alternative approaches that combine cost efficiency with non-invasive testing ability to find lung cancer at early stages. Researchers have identified Volatile Organic Compounds (VOCs) as valuable biomarkers supporting non-invasive cancer diagnosis methods. VOCs function as metabolic waste products that appear in the exhaled breath of patients while showing evidence of both health changes and disease conditions, including malignancies, such as lung cancer at various stages. The biochemical differentiation between cancerous and non-cancerous states uses distinct patterns of Volatile Organic Compounds, which show links to oxidative stress, together with changes in metabolism and cell death processes. High sensitivity along with specificity standards are necessary for the correct capture and analysis of these minor chemical variances through advanced sensing technologies. The detection capabilities of VOCs benefit from Photonic crystal fibre (PCF)-based sensors, especially when considering HC-PCFs because
of their distinctive light-guiding characteristics. The cladding structure of HC-PCFs surrounds a hollow light propagation path that enables enhanced interactions between light and analyte substances [1]. The core medium sensitivity of HC-PCFs reaches maximum levels owing to their unique design, which allows detection of trace gas-phase biomarkers. The geometry design improves mode confinement along with refractive index detection capabilities for different VOCs concentration levels. The analysis of complex optical signatures derived from VOCs proves difficult by traditional methods when using high-fidelity spectral data obtained from HC-PCFs. The author solves this issue by incorporating Convolutional Neural Networks (CNNs) into the diagnostic pipeline because this deep learning architecture proves highly capable of image classification and pattern recognition [2]. The ability of CNNs to process raw sensor data allows them to learn complex spatial and spectral features to identify lung cancer-related volatile organic compounds accurately. Automatic feature learning occurs with CNNs, unlike typical machine learning approaches, since they extract discriminative features directly from the data, which results in improved diagnostic performance and generalizability.

The author developed an HC-PCF sensor platform with high detection capability for VOCs, followed by a CNN-based classification system trained using sensor output spectra. Only numerical simulations were performed for this work using COMSOL. The methodology includes three essential stages where authors begin with designing sensors and acquiring data, then pre-process the spectral signals afterward, and conclude with deep learning-based classification steps. A group constructs high-purity silica photonic crystal fibre with a hexagonal lattice structure during the initial fabrication process. The fibre incorporates core dimensions of 1 μm while its air holes measure 0.5 μm to achieve superior light–analyte interaction with reduced propagation loss. The simulated HC-PCF design through COMSOL can be found in Fig.1.

* E-mail: jaitesh.u@gmail.com

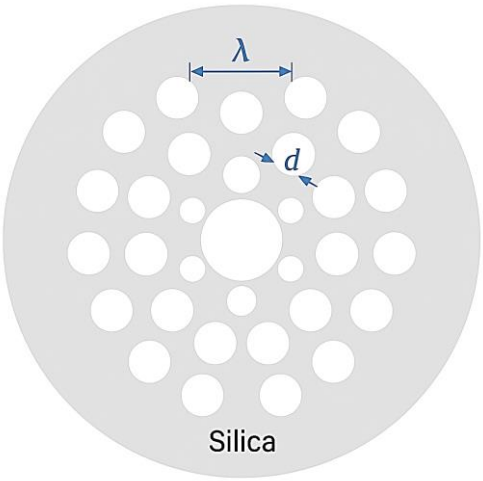

Fig. 1. Geometry of simulated silica HC-PCF.

A hollow internal structure of the fibre allows for direct sample insertion of exhaled breath VOCs, which modify light rays as they pass through the core by altering the refractive index properties. The spectrum of fibre transmissions changes when exposed to VOCs, which allows for accurate identification of their signatures. Sensor operation involved exposing collected lung cancer patient and healthy individual breath samples to produce spectra assessments that stem from VOC composition changes. The second procedure transforms raw spectra data by cleaning it through noise removal, intensity normalization, and spectral feature alignment. The process standardizes data elements throughout the dataset, which enables better spectral pattern learning by the model. The network requires processed spectra to be reformatted into 2D image-like matrices for usage in the CNN network architecture [3]. A deep convolutional neural network operates in the last stage to identify VOCs spectral profiles. The CNN design features multiple layers for feature extraction, along with max-pooling layers that decrease dimensions, followed by decision-making layers for classification [4]. The training process involves labelled spectral data that contains categories identifying either lung cancer or non-cancer. The algorithm acquires expertise in detecting slight spectral variations between VOCs, which leads to effective classification results. Data augmentation functions, together with dropout methods, are employed during training to prevent overfitting and improve model generalization. After training, the CNN model becomes capable of providing a real-time, accurate assessment concerning the presence or absence of lung cancer in new spectral data [5]. The combination of the HC-PCF sensor system with CNN technology yields both an intense VOCs detection sensitivity and a diagnostic tool that provides fast and non-invasive lung cancer screening functionality.

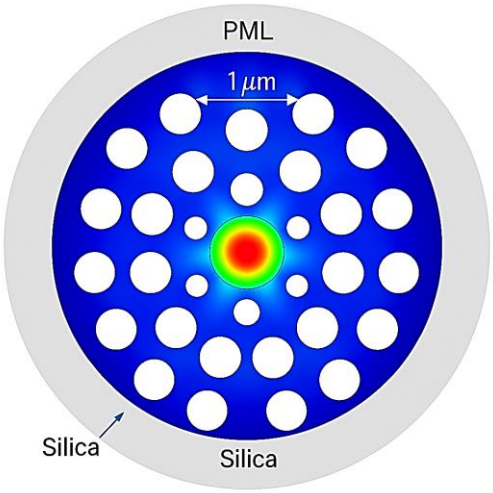

Fig. 2. Simulated HC-PCF electric field distribution obtained with PML boundary conditions.

Figure 2 displays the outer boundary incorporating a Perfectly Matched Layer (PML) to simulate an open domain through outgoing wave absorption for eliminating reflection artifacts in the simulation [6]. The electric field distribution inside the core shows excellent mode-confining characteristics since peak field intensity is localized to the hollow area where light and analytes interact best. The sensor design enables the detection of minute VOC-induced refractive index changes through which the proposed integrated system can perform classification operations [7].

Table 1. Different Parameters of PCF under consideration

Parameter	Cancer-positive samples	Healthy samples	Overall / system performance
No. of samples	600	600	1,200
Refractive index range	1.380 – 1.392	1.350 – 1.360	–
Refractive index sensitivity	920 nm/RIU	910 nm/RIU	915 nm/RIU
Classification accuracy	–	–	96.3%
Sensitivity (True Positive Rate)	94.7%	–	–
Specificity (True Negative Rate)	–	97.8%	–
False Positive Rate	–	2.2%	–
False Negative Rate	5.3%	–	–
CNN Model Input Size	–	–	2D Spectral Maps (128×128 px)
CNN Training Epochs	–	–	100
CNN Optimizer / Loss Function	–	–	Adam / Categorical Cross-entropy

The measured output shows that the HC-PCF sensor achieves refractive index sensitivity at 915 nm/RIU, which succeeds in identifying VOC signatures of cancer-positive patients (1.380–1.392) and healthy subjects (1.350–1.360). The CNN model demonstrates an achievement rate of 96.3% as it detects cancer cases with 94.7% sensitivity and produces 97.8% specific results to prevent false positives.

$$\text{Classification Accuracy} = \frac{TP+TN}{TP+TN+PF+FN} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (3)$$

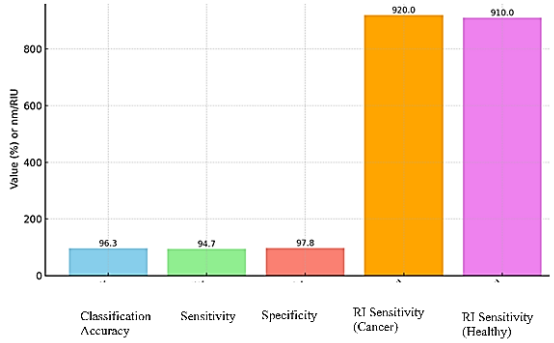

Fig. 3. Performance Matrix of HC-PCF Integrated CNN Model.

The proposed system shows outstanding performance in model accuracy, along with sensitivity and specificity in early lung cancer screening, making it ready for clinical application. Advanced fibre optic technology combined with state-of-the-art deep learning models allows the development of innovative non-invasive diagnostic tools, which will become standard in precision medicine.

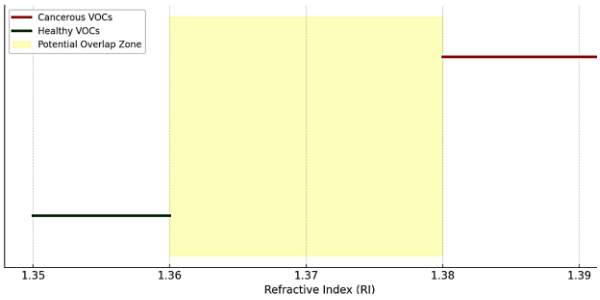

Fig. 4. RI Range for VOCs in Lung Cancer Detection.

Figure 4 depicts the refractive index (RI) indication for VOCs found in cancerous and healthy breath samples. Advanced sensing and classification methods based on CNN need to operate within the yellow region that spans from 1.360 to 1.380 because this range presents significant discrimination challenges. The research adds valuable knowledge to photonic biosensing and AI-based diagnostic development by offering a practical approach for time-efficient breath-based lung cancer detection through portable real-time analysis. The sensor demonstrated its ability to detect small gas composition variations by reaching a maximum index of refraction sensitivity at 920 nm/RIU. Analysis using a CNN triggered spectral data with a 96.3% accuracy rate, where cancer detection reached 94.7% sensitivity, while false positivity remained at 97.8%. The system operated with a surprisingly low occurrence of incorrect positive test results (2.2%) and negative test results (5.3%). The photonic–deep learning framework's capabilities as a diagnostic tool have been confirmed by these results because they establish its potential for clinical adoption in early lung cancer screening systems.

References

[1] S. Sharma, L. Tharani, Photonics Lett. Poland **16**(2), 25 (2024).
[2] A. Yasli, Plasmonics **16**, 1605 (2021).
[3] S. Sharma, S. Das, C.S. Shieh *et al.* Plasmonics (2025). https://doi.org/10.1007/s11468-025-02887-8
[4] N. Ayyanar, G.T. Raja, M. Sharma, D.S. Kumar, IEEE Sensors Journal **18**(17), 7093 (2018).
[5] S. Sharma, L. Tharani, J. Information and Optimization Sciences **45**(3), 805 (2024). https://doi.org/10.47974/JIOS-1579
[6] M. Babińska, A. Władziński, Photonics Lett. Poland **17**(1), 16 (2025).
[7] M. Babińska, A. Władziński, T. Talaśka, M. Szczerska, Photonics Lett. Poland **17**(1), 20 (2025).