Machine learning for predictions of cervical cancer identification – preliminary investigation based on refractive index

Michał Kruczkowski (✉ michal.kruczkowski@utp.edu.pl)

Bydgoszcz University of Science and Technology

Anna Drabik-Kruczkowska

Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń

Anna Marciniak

Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń

Martyna Tarczevska

Bydgoszcz University of Science and Technology

Monika Kosowska

Bydgoszcz University of Science and Technology

Małgorzata Szczerska

Gdańsk University of Technology

Research Article

Keywords: cervical, cancer, doctors, measurement, early, diagnosis

Posted Date: October 1st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-948525/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Machine learning for predictions of cervical cancer identification – preliminary investigation based on refractive index

Michał Kruczkowski^{1,*}, Anna Drabik-Kruczkowska², Anna Marciniak^{3,1}, Martyna Tarczewska¹, Monika Kosowska^{1,*}, Małgorzata Szczerska^{4,*}

¹ Department of Telecommunication, Informatics and Electronics, Bydgoszcz University of Science and Technology, Al. prof. S. Kaliskiego 7, 85-796 Bydgoszcz, Poland

² Department of Obstetrics, Gynaecology and Oncology, Faculty of Medicine, Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, 85-094 Bydgoszcz, Poland

³ Department of Forensic Medicine, Department of Molecular and Forensic Genetics, Faculty of Medicine, Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, 85-094 Bydgoszcz, Poland

⁴ Department of Metrology and Optoelectronics, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, 11/12 Narutowicza Street, 80-233 Gdańsk, Poland

* Correspondence: michal.kruczkowski@utp.edu.pl, monika.kosowska@utp.edu.pl, malszcze@pg.edu.pl

Abstract

Cervical cancer is one of the most commonly appearing cancers, which early diagnosis is of greatest importance. Unfortunately, many diagnoses are based on subjective opinions of doctors – to date, there is no general measurement method with a calibrated standard. The problem can be solved with the measurement system being a fusion of an optoelectronic sensor and machine learning algorithm to provide reliable assistance for doctors in the early diagnosis stage of cervical cancer. We demonstrate the preliminary research on cervical cancer assessment utilizing optical sensor and prediction algorithm. Since each matter is characterized by refractive index, measuring its value and detecting changes give information about the state of the tissue. The optical measurements provided datasets for training and validating the analyzing software. We present data preprocessing, machine learning results utilizing three algorithms (Random Forest, eXtreme Gradient Boosting, Naïve Bayes) and assessment of their performance for classification of tissue as healthy or sick. All of them provided high values (>89%) of the measures describing them. Our solution allows for rapid sample measurement and automatic classification of the results constituting a potential support tool for doctors.

1. Introduction

Cervical cancer is one of the most common cancers worldwide ^{1,2}. Every year around the world cervical cancer is diagnosed in about half a million women, including about 2,5 thousand in Poland ³. The incidence and mortality of cervical cancer have been dramatically reduced by screening programs ⁴. However, in many cases diagnoses are still dependent on the doctor's subjective interpretation, creating a strong need for solutions supporting them ⁵.

Particularly important in the diagnosis of cervical cancer is the precise determination of the depth of neoplastic lesions, which is of clinical importance during procedures such as LEEP (Loop Electrosurgical Excision Procedure), cold knife conization, laser ablation, cervical brachytherapy in order to correctly define the margin of pathological changes. Commonly used measurement methods such as colposcopy, visual inspection with acetic acid and Lugol's iodine are limited by the subjective judgment of the examiner and the lack of reliable measurement calibration standards ⁶. The imprecise definition of the type of neoplastic lesion may lead to far-reaching consequences such as extended diagnosis time, high treatment costs, patient's exposure to unnecessary procedures, and in extreme cases even the patient's death .

The majority cause of the premalignant changes in the cervical epithelium is associated with the infection of HPV - Human Papilloma Virus ^{7,8}. Although there are approximately 100 types of HPV virus, only several create a high risk. The typical microscopic recognized view of the HPV infection cells is Koilocytosis. The Atypical cell with „halo” and with enlarged, irregular nucleus, which suggested an intensive mitosis process. The solution to this issue may be the fusion of the most dynamically developing technologies: optical sensing and machine learning techniques ⁹⁻¹¹. With a fast, reliable and non-destructive optical method, we can investigate the biological sample and then analyze the acquired data with dedicated software ^{12,13}, allowing for auto-identification of neoplastic cervical lesions which will be invaluable support for doctors at the stage of initial diagnosis ^{14,15}. The identification will be based on refractive index values of measured tissues.

The refractive index is one of the most important physical properties characterizing materials. In case of biological tissues, it is highly correlated with the morphological features including the cell density and the nuclear-cytoplasm ratio. Based on cervical cancer's state of the art ¹⁶, refractive index of normal cells and cancerous cells are different, hence refractive index changes constitute a basis for relatively easy differentiation between the normal and cancerous cells ¹⁷. Table 1 presents typical refractive index values obtained for both healthy and sick cervical cells.

Table 1. Associated refractive index values of cervical cells at different neoplastic progression.

Cell type	Basal	Midzone	Superficial
Normal	1,387	1,372	1,414
Cancer	1,426	1,404	1,431

In this study, we propose a method of preliminary cervical cancer identification based on a prediction algorithm, taught on data obtained from low-coherence measurements of certified refractive index liquids. We have measured and analyzed samples within the range of actual refractive index values for healthy cervical tissues and neoplastic lesions. The acquisition and preparation of datasets, machine learning process and results of the investigation are described. To date, no research applying machine learning for peculiar analysis of low-coherence data obtained for various refractive indices was reported. Our approach allows for fast and reliable analysis of such data and their classification, which is the starting point for the development of a system able of the initial identification of neoplastic cervical lesions. This can be a helpful tool for doctors greatly impacting and improving the effectiveness of early cervical cancer diagnosis.

2. Methodology

The classification of cervical intraepithelial neoplasia (CIN) is based on a histological evaluation that differentiates three advancement stages: CIN1, CIN2, CIN3¹⁸. The grade of dysplasia is the proportion of cervical changes in the epithelium. CIN1 has a low potential for progression to malignancy. CIN1 is confined to the basal one-third of the epithelium. CIN2 has more marked nuclear abnormalities than CIN1. The dysplastic cellular is observed to the lower of two-thirds of the epithelium. The CIN3 occurs if the atypical cells are found in all layers of the epithelium. The characteristic features are a low potential for malignancy and a high potential for regression. The L-SIL (Low-grade Squamous Intraepithelial Lesion) corresponds histologically to CIN1. The H-SIL (High-Grade Squamous Intraepithelial Lesion, CIN2 and CIN3) has a higher potential for progression and lower potential for regression.

The main goal of the cervical cancer identification method is to detect neoplastic lesions according to the designed methodology as shown in Fig. 1.

Figure 1. Methodology workflow.

The proposed methodology includes four relevant modules: low coherence interferometric measurements, data preprocessing (row mapping, filtering), training of supervised machine learning model and testing the built predictive model..

On the basis of a literature study, the assignment of individual samples with known refractive index values to two classes (healthy or cancer) was defined^{19,20}. The predictive capabilities of selected supervised machine learning algorithms were built and analyzed to select the optimal classification model. Moreover, the proposed method was tested on the basis of completely new test datasets that were not involved in the training process. It should be noted that the cancer is diagnosed when the basal membrane is invaded due to differences in treatment. However, the evaluation of the refractive index should be correlated with the identification of the basal layer. Therefore an essential element of the elaborated method is sensitivity to the Fabry-Perot interferometer length changes. This parameter corresponds to the depth of the cervical epithelium of the measurement sample that determines the grade of dysplasia.

2.1 Dataset acquisition

The optical determination of refractive indices of the investigated liquids was performed in a Fabry-Perot interferometer. The measurement setup was built in a reflective configuration using fiber-optic technology. The components of the system were a superluminescence diode (SLD-1550-13-, FiberLabs Inc., Fujimino, Japan), an optical spectrum analyzer (Ando AQ6319, Yokohama, Japan), a 2x1 optical coupler (Lightel, Renton, Washington, USA) and a micromechanical stand. The light source operated at the central wavelength of 1550 ± 20 nm with a spectral width of 35 nm. The Fabry-Perot resonance cavity was formed by the polished fiber end-face and a silver mirror^{21,22}.

The light from the light source was guided through the fiber to the cavity. Partial reflections occurred at the two boundaries: fiber end-face/medium and medium/silver mirror. The reflected light beams interfered giving a signal recorded by the optical spectrum analyzer. The phase shift between interfering beams is dependent on their optical path difference (which is influenced by the geometrical path length and refractive index of the medium) according to the following formula:

$$\phi = \frac{4\pi nL}{\lambda} \quad (1)$$

where ϕ – phase shift, n – refractive index, L – geometrical path length, λ – wavelength.

In our investigation the geometrical path length difference (the width of the resonance cavity) was constant throughout the whole measurement process, hence the refractive index change was the only variable impacting the acquired signal²³.

For precise measurements of the refractive index of liquids, we used the Certified Refractive Index Liquids by Cargille®(Cargille Labs, Cedar Grove, USA). The investigated liquids were characterized by refractive indices in the range of 1.3-1.5 with a step of 0.01. The choice of this measurement range was based on the values of refractive indices of healthy and diseased tissues^{24–26}. The range was extended to include inter-individual differences and assure a larger dataset for algorithm learning. This way, the results obtained by the proposed method can be directly translated into biological tissues. In this article, we refer to each oil using the label value (measured for 589.3 nm in 25°C) for clarity. However, the data analysis takes into account the nominal values given in datasheets for the wavelength equal to 1550 nm, as the source used in the experiment²⁷.

The highest signal contrast of $V = 0.9956$ was obtained for the cavity length equal to 280 μm . The reference signal was acquired to control the intact cavity setting. Next, 30 μL of the liquid sample with a known refractive index was introduced into the cavity. The optical spectra were recorded and the cavity was cleaned. The whole procedure was then repeated for all liquids (a total of 10 spectra for each sample).

2.2 Dataset preparation

Interferograms obtained in accordance with the adopted methodology were the basis for further analyzes. 210 interferograms were taken for analysis, each data consists of two columns representing the wavelength and the optical power of the signal. The representative signal is shown in Fig. 2.

Figure 2. Sample interferogram.

The first step in the analysis was mapping the measurement data to the feature vector – this way we obtained a dataset adapted to the training supervised learning model. The mapping process was based on 18 features extraction as described in Table 2. The interferogram was filtered with a threshold that represents a percentage of the global maximum. A part of signal rejected from the analyses – by multiplication of global maximum and threshold noises were eliminated. A noisy part of the signal was eliminated by the multiplication of a global maximum and a threshold value.

Table 2 Selected features description.

Feature	Description
Number of local maxima	Calculated with use of find_peaks function from scipy.signal module; saved as a list of all local maximums
Global maxima	Maximum value from local maximum list
Threshold	A variable used to filter amplitude to smooth the signal
Amplitude normalization factor	Used to rescale experimental plot compared to the simulation one, due to different ranges of amplitude; factor was calculated as shown in Equation 2.
Local maxima distance – average	Average wavelength distance between local maxima
Local maxima distance – maximum	Maximum wavelength distance between local maxima
Local maxima distance – minimum	Minimum wavelength distance between local maxima
Local maxima distance – median	Median wavelength distance between local maxima
Dissimilarity measure	Dissimilarity between simulated and experimental interferogram (integral calculated with the use of Simpson rule as shown in Equation 3)
Chart axial shift	Used to shift the experimental plot so that the global maximums of both plots are on the same wavelength
Roots Mean Squared Error (RMSE)	Used to determine the difference between the simulation plot and the experimental data

Cavity length	Value read from the configuration of the measuring set
Minimum wavelength	Minimum value for wavelength parameter
Maximum wavelength	Maximum value for wavelength parameter
Amplitude	Difference between maximum and minimum y value, where y is representative of amplitude column from input data
λ_0	Wavelength for maximum amplitude
λ_0 for simulation signal	Wavelength for maximum amplitude (for base signal)
Target variable	1 - cancer, 0 – healthy

Each row in the dataset represents one sample and consists of 18 columns. Each column is representative of one from selected features. The target variable was assigned based on refractive index value: refractive indices between 1.30 and 1.38 were assigned as ‘healthy’ tissue while refractive indices between 1.39 and 1.50 were labeled as ‘sick’ tissue. The dataset was balanced, consisting of 43% of healthy samples and 57% of sick representatives. Prepared dataset allowed to build a machine learning model based on selected supervised learning algorithm.

The following formulas were introduced into preprocessing procedure in order to estimate the distortion of the measurement interferogram with relation to the theoretical interferogram. Factor f is responsible for the fit of the theoretical signal amplitude to the measured interferogram as shown in equation 2.

$$f = \frac{ssmax}{global\ max} \quad (2)$$

Where: global max – global maximum for measured signal, ssmax – maximum value for simulation signal. Finally, the distortion level D of the measurement interferogram is calculated numerically using the surface area below the interferometric signal as shown in equation 3.

$$D = \left| \frac{1 - (area_exp \times f)}{area_sym} \right| \quad (3)$$

Where: area_sym – integral under the curve of the simulation plot, area_exp – integral under the curve of the plot of experimental data, f – factor.

Before the model training process began, the k stratified fold cross-validation method was used to divide the data into the validation and training dataset, where k equals 3 (Figure 3). The value of k was chosen experimentally from odd numbers set in the range from 3 to 9, due to the fact that each of the considered values of k, quite similar cross-validation results were obtained. On the other hand, the smaller the km value, the shorter time of obtaining cross-validation results. Cross-validation is a resampling procedure, which is used to evaluate machine learning

models on a limited data sample. Its main goal is to randomly divide data into a given number of sets on which the machine learning model is later tested.

Figure 3: Graphical representation of Stratified 3-Fold Cross Validation on a prepared dataset.

2.3 Machine learning

Referring to reported research where similar analytical problems were solved^{28–33}, three algorithms were selected for further analysis: Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Naïve Bayes (NB). It should be noted that the use of well-known algorithms in the combination with the novel methodology of data preprocessing³⁴ and enrichment is an unprecedented approach in the analysis and prediction of optical properties of measured substances. For each algorithm, optimal parameters were selected experimentally.

Random Forest^{35,36} and eXtreme Gradient Boosting^{37,38} classifiers utilize ensembles of classifications are receiving increased interest. Ensemble learning algorithms use the same base classifier to produce repeated multiple classifications of the same data or use a combination of different base classifiers to generate multiple classifications of the same data or to target different subsets of the data³⁹. The collection of multiple classifiers of the same data are combined using a rule-based approach (such as maximum voting, product, sum or Bayesian rule) or based on an iterative error minimization technique by reducing the weights for the correctly classified samples (e.g. boosting). Ensemble learning techniques have higher accuracy than other machine learning algorithms because the group of classifiers performs more accurately than any single classifier, and utilizes the strengths of the individual group of classifiers while the classifier weaknesses are circumvented. Whereas Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions between the features⁴⁰. They are among the simplest Bayesian network models but coupled with kernel density estimation, they can achieve higher accuracy levels⁴¹.

The first algorithm we tested was RF, where the following parameters were selected: `n_estimators` – 100, `criterion` – gini, `min_samples_split` – 2, `min_samples_leaf` – 1. To test the possibility of improving the RF results, an XGBoost algorithm was used and the following parameters were selected: `booster` – gbtrees, `learning_rate` – 0.3, `min_split_loss` – 0, `max_depth` – 6 and `sampling_method` – uniform. As a part of the application of a different approach to classification, an NB algorithm was used. Following parameters were selected: `priors` – None, `var_smoothing` – 1e-9.

3. Results

Since the presented problem can be treated as binary classification, confusion matrices^{42,43} were used to evaluate and compare the ML-based methods. Four measures were defined as follows:

TP – true positives – cancer tissue classified as cancer;

FP – false positives – healthy tissue classified as cancer;

FN – false negatives – cancer tissue classified as healthy;

TN – true negatives – healthy tissue classified as healthy.

A graphical representation of these measures is presented in Figure 4.

		Predicted	
		Negative (N) -	Positive (P) +
Actual	Negative -	True Negatives (TN)	False Positives (FP) Type I error
	Positive +	False Negatives (FN) Type II error	True Positives (TP)

Figure 4: A graphical representation of evaluation measures: True Positives, False Positives, False Negatives, True Negatives.

In order to reliably evaluate the predictive ability of the model, we introduce the following classifier evaluation metrics: Accuracy (equation 4), Precision (equation 5), Recall (equation 6) and F1-score (equation 7).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

207

208
$$F1 = \frac{2(Precision \cdot Recall)}{Precision + Recall} \quad (7)$$

209 The use of these metrics provides us information not only about the accuracy of the classification but
 210 especially important properties like sensitivity and specificity and the model's insensitivity to overfitting and
 211 underfitting. All measures used in those equations were mentioned above (TP, FP, FN and TN). The obtained results
 212 are presented in Table 3.

213 Table 3 Classification results

Classifier	Fold	Accuracy	Precision	Recall	F1-score
Random Forest	1	0.97	0.97	0.98	0.97
	2	1.00	1.00	1.00	1.00
	3	1.00	1.00	1.00	1.00
	Validation:	0.91	0.91	0.92	0.92
XGBoost	1	1.00	1.00	1.00	1.00
	2	1.00	1.00	1.00	1.00
	3	1.00	1.00	1.00	1.00
	Validation:	0.89	0.90	0.90	0.89
Naïve Bayes	1	0.96	0.96	0.96	0.96
	2	0.95	0.95	0.95	0.95
	3	0.97	0.97	0.97	0.97
	Validation:	0.92	0.93	0.93	0.92

214 It can be seen that all classifiers give results with accuracy above 95%, precision above 95%, recall above
 215 95% and F1-score above 95% for training datasets. On validation, the results were as follows: accuracy above 89%,
 216 precision above 90%, recall above 90% and F1-score above 89%. Thus, the most promising results on training were
 217 obtained with XGBoost (Accuracy equals 100%, Precision equals 100%, Recall equals 100%, F1-score equals
 218 100%). However, XGBoost did not accomplish the best results on the validation set, where the best results were
 219 obtained for Naïve Bayes (Accuracy equals 92%, Precision equals 93%, Recall equals 93%, F1-score equals 92%).
 220 The obtained results are presented also as confusion matrices in Figure 5.

Figure 5: Confusion matrices for selected algorithms: A1: Random Forest test dataset fold 1, A2: Random Forest validation dataset, B1: XGBoost test dataset fold 1t, B2: XGBoost validation dataset, C1: Naïve Bayes test dataset fold 1, C2: Naïve Bayes validation dataset.

Additionally, to extend the model evaluation, the learning time from the training data and making predictions was measured for each algorithm. The results are presented in Table 4.

Table 4 Average time for training and prediction for chosen algorithms.

Algorithm	Training	Prediction
Random Forest	212 ms	15.5 ms
XGBoost	21 ms	2.08 ms
Naïve Bayes	7.54 ms	1.81 ms

It can be noted that the Naïve Bayes method not only gives the best results for the validation test, but also is the fastest regarding the training and prediction phases.

4. Conclusions

In this study, we presented a novel approach to the analysis of data acquired by a low-coherence interferometer. The optical sensor is able to detect changes in the refractive index of samples, including the biological range of values. Hence, it can be used for measurements and initial assessment of the neoplastic cervical lesions stage. The data obtained for test liquids were acquired with a Fabry-Perot interferometer and then applied in the machine learning algorithm. Interferograms representing the optical properties of measured substances in conjunction with meta-data from the measurements are transformed into multidimensional datasets. A number of heuristics have been defined on the basis of which these datasets are constructed, taking into account their use in predictive modeling. A particularly important stage in the machine learning process was the development of an original approach to the initial processing and enrichment of data sets. Part of data was used to train the algorithm, and the other served for validation of its proper operation. The proposed solution allows for the identification and classification of healthy and sick tissues. The tested classifiers were characterized by high accuracy above 95%, precision above 95%, recall above 95% and F1-score above 95% for training datasets, and for validation accuracy above 89%, precision above 90%, recall above 90% and F1-score above 89%. The method we reported can be of great assistance for doctors in early cervical cancer diagnosis.

Acknowledgments

This work has been supported by the DS funds of Faculty of Telecommunication, Informatics and Electronics, Bydgoszcz University of Science and Technology and the DS funds of Faculty of Electronics, Telecommunications and Informatics of Gdańsk University of Technology.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Zhang, X., Zeng, Q., Cai, W. & Ruan, W. Trends of cervical cancer at global, regional, and national level: data from the Global Burden of Disease study 2019. *BMC Public Health* **21**, 894 (2021).
2. Zhang, S., Xu, H., Zhang, L. & Qiao, Y. Cervical cancer: Epidemiology, risk factors and screening. *Chin J Cancer Res* **32**, 720–728 (2020).
3. Pikala, M., Burzyńska, M. & Maniecka-Bryła, I. Years of Life Lost Due to Cervical Cancer in Poland in 2000 to 2015. *Int J Environ Res Public Health* **16**, 1545 (2019).
4. Nowakowski, A. *et al.* The implementation of an organised cervical screening programme in Poland: an analysis of the adherence to European guidelines. *BMC Cancer* **15**, 279 (2015).

5. Conceição, T., Braga, C., Rosado, L. & Vasconcelos, M. J. M. A Review of Computational Methods for Cervical Cells Segmentation and Abnormality Classification. *International Journal of Molecular Sciences* **20**, 5114 (2019).
6. Duesing, N. *et al.* Assessment of cervical intraepithelial neoplasia (CIN) with colposcopic biopsy and efficacy of loop electrosurgical excision procedure (LEEP). *Arch Gynecol Obstet* **286**, 1549–1554 (2012).
7. Zhang, J., Cheng, K. & Wang, Z. Prevalence and distribution of human papillomavirus genotypes in cervical intraepithelial neoplasia in China: a meta-analysis. *Arch Gynecol Obstet* **302**, 1329–1337 (2020).
8. Sitarz, K. *et al.* HPV Infection Significantly Accelerates Glycogen Metabolism in Cervical Cells with Large Nuclei: Raman Microscopic Study with Subcellular Resolution. *International Journal of Molecular Sciences* **21**, 2667 (2020).
9. Decaro, C., Montanari, G. B., Bianconi, M. & Bellanca, G. Prediction of hematocrit through imbalanced dataset of blood spectra. *Healthcare Technology Letters* **8**, 37–44 (2021).
10. Venkat, S. *et al.* Machine Learning based SpO2 Computation Using Reflectance Pulse Oximetry. *Annu Int Conf IEEE Eng Med Biol Soc* **2019**, 482–485 (2019).
11. Hornung, R. *et al.* Quantitative near-infrared spectroscopy of cervical dysplasia in vivo. *Human Reproduction* **14**, 2908–2916 (1999).
12. Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell* **5**, 221–232 (2016).
13. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).
14. Chang, W. *et al.* A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. *Diagnostics* **9**, 178 (2019).
15. Mustafa, N. & Li, J.-P. Medical data classification scheme based on hybridized SMOTE technique (HST) and Rough Set technique (RST). in *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* 49–55 (2017). doi:10.1109/ICCCBDA.2017.7951883.
16. Giannios, P. *et al.* Visible to near-infrared refractive properties of freshly-excised human-liver tissues: marking hepatic malignancies. *Sci Rep* **6**, 27910 (2016).

17. Sharma, V. & Kalyani, V. L. Nano-cavity coupled waveguide photonic crystal based biosensor detection of cervical cancer using nucleus and cytoplasm. in *2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC)* 1–5 (2017). doi:10.1109/ICOMICON.2017.8279111.
18. Bruno, M. T., Cassaro, N., Bica, F. & Boemi, S. Progression of CIN1/LSIL HPV Persistent of the Cervix: Actual Progression or CIN3 Coexistence. *Infectious Diseases in Obstetrics and Gynecology* **2021**, e6627531 (2021).
19. Panda, A. & Puspa Devi, P. Photonic crystal biosensor for refractive index based cancerous cell detection. *Optical Fiber Technology* **54**, 102123 (2020).
20. Parvin, T., Ahmed, K., Alatwi, A. M. & Rashed, A. N. Z. Differential optical absorption spectroscopy-based refractive index sensor for cancer cell detection. *Optical Review* **28**, 134–143 (2021).
21. Kosowska, M. *et al.* Microscale diamond protection for a ZnO coated fiber optic sensor. *Sci Rep* **10**, 19141 (2020).
22. Kosowska, M. *et al.* Incorporation of nitrogen in diamond films – A new way of tuning parameters for optical passive elements. *Diamond and Related Materials* **111**, 108221 (2021).
23. Karpienko, K., Wróbel, M. S. & Jędrzejewska-Szczerska, M. Determination of refractive index dispersion using fiber-optic low-coherence Fabry–Perot interferometer: implementation and validation. *OE* **53**, 077103 (2014).
24. Jabin, Md. A. *et al.* Surface Plasmon Resonance Based Titanium Coated Biosensor for Cancer Cell Detection. *IEEE Photonics Journal* **11**, 1–10 (2019).
25. Giannios, P. *et al.* Complex refractive index of normal and malignant human colorectal tissue in the visible and near-infrared. *J Biophotonics* **10**, 303–310 (2017).
26. Lin, X., Wan, N., Weng, L. & Zhou, Y. Light scattering from normal and cervical cancer cells. *Appl. Opt., AO* **56**, 3608–3614 (2017).
27. Labs, C. Available Refractive Indices, SDS & Datasheets – Cargille Labs. <https://www.cargille.com/available-refractive-indices-sds-datasheets/>.
28. Ma, B. *et al.* Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Comput Biol Med* **121**, 103761 (2020).
29. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* **13**, 8–17 (2015).

- 320 30. Chai, H. *et al.* Integrating multi-omics data through deep learning for accurate cancer prognosis prediction.
321 *Comput Biol Med* **134**, 104481 (2021).
- 322 31. Deng, F. *et al.* Predict multicategory causes of death in lung cancer patients using clinicopathologic factors.
323 *Computers in Biology and Medicine* **129**, 104161 (2021).
- 324 32. Chen, M., Hao, Y., Hwang, K., Wang, L. & Wang, L. Disease Prediction by Machine Learning Over Big Data
325 From Healthcare Communities. *IEEE Access* **5**, 8869–8879 (2017).
- 326 33. Dubey, V. *et al.* Low coherence quantitative phase microscopy with machine learning model and Raman
327 spectroscopy for the study of breast cancer cells and their classification. *Appl. Opt., AO* **58**, A112–A119 (2019).
- 328 34. García, S., Luengo, J. & Herrera, F. Tutorial on practical tips of the most influential data preprocessing
329 algorithms in data mining. *Knowledge-Based Systems* **98**, 1–29 (2016).
- 330 35. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
- 331 36. Qi, Y. Random Forest for Bioinformatics. in *Ensemble Machine Learning: Methods and Applications* (eds.
332 Zhang, C. & Ma, Y.) 307–323 (Springer US, 2012). doi:10.1007/978-1-4419-9326-7_11.
- 333 37. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM*
334 *SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for
335 Computing Machinery, 2016). doi:10.1145/2939672.2939785.
- 336 38. Wang, C., Deng, C. & Wang, S. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-
337 imbalanced classification with XGBoost. *Pattern Recognition Letters* **136**, 190–197 (2020).
- 338 39. Biau, G. & Scornet, E. A random forest guided tour. *TEST* **25**, 197–227 (2016).
- 339 40. Yang, F.-J. An Implementation of Naive Bayes Classifier. in *2018 International Conference on Computational*
340 *Science and Computational Intelligence (CSCI)* 301–306 (2018). doi:10.1109/CSCI46756.2018.00065.
- 341 41. Chandrasekar, P. & Qian, K. The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier.
342 in *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)* vol. 2 618–619
343 (2016).
- 344 42. Batarseh, F. A. & Yang, R. *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and*
345 *Knowledge Engineering*. (Academic Press, 2020).
- 346 43. Al-Jabery, K., Obafemi-Ajayi, T., Olbricht, G. & Wunsch, D. *Computational Learning Approaches to Data*
347 *Analytics in Biomedical Applications*. (Academic Press, 2019).