Classification of Cervical Cancer Cells using FTIR data

Erick Njoroge, Stephen R. Alty, *Member, IEEE*, Mahbub R. Gani, *Member, IEEE* and Maha Alkatib

***Abstract*— High false-negative rates of the Papanicolauo (so-called 'Pap') smear test and the shortage of colposcopists has led to the desire to find alternative non-expert (automated) approaches for accurately testing cervical smears for signs of cancer. Fourier-Transform Infra-Red (FTIR) spectroscopy has been shown to offer the potential for improving the accuracy (i.e. sensitivity and specificity) of these tests. This paper details the application of the machine learning methodology of Support Vector Machines (SVM) using FTIR data to enhance and improve upon the standard Pap test. A cohort of 53 subjects was used to test the veracity of both the Pap smear results and the FTIR based classifier. The Pap test achieved an overall classification of 43%, whereas our method achieved a rate of 80%.**

I. INTRODUCTION

Cervical cancer is the rapid uncontrolled growth of severely abnormal cells on the cervix. Scientific studies [1] point to the HPV (Human Papillomavirus) infection as a necessary prerequisite and the prime risk factor for the development of cervical cancer. Genital HPV serotype viruses are by far the most common sexually transmitted infections and it is estimated that 80% of sexually active humans have been infected with a strain of genital HPV at one time or another. At least 95% of cervical cancer cases result from high-risk HPV serotypes [2]. Hence, there is now considerable interest in vaccination against these particular forms of the virus. However, until their widespread use, physicians are still relying on the tried and tested Pap smear screening techniques. As a result, women are advised to have a Pap smear test annually after they become sexually active. Unfortunately, due to the high prevalence of the HPV virus, it is impossible to use it as an effective tool to test for cervical cancer though a negative result for the HPV virus can be reassuring.

Worldwide, cervical cancer is the second most common cancer among women after breast cancer [2]. It is also the third highest cause of death when compared to other cancers. In the USA, however, the statistics are different. Cervical cancer is only the eighth most common cancer among women, with incidence and mortality rate figures as low as half of the worldwide figures. This difference can be attributed in part to the success of effective and widespread screening using the Pap or smear test in the developed world. This makes effective screening a very important tool in the fight against cervical cancer.

A. Pap or smear test

The Pap test is the first step in cervical cancer screening. This test was developed by American Dr. Georgios Papanikolaous [3] in the 1940s. The test consists of a simple cervical swab to collect cell samples from a specific area on the cervix. These cells are then examined in a laboratory for abnormalities. Though abnormal results do not necessarily mean the patient has cervical cancer, they are an indication that changes may be taking place in the cervical cells and further action needs to be taken.

B. Colposcopy or colcoscopy

In most cases, a colposcopy is performed to further investigate abnormal results in Pap smear results. It is very accurate and is in fact considered 'the gold standard' test in the diagnosis of cervical cancer. This simply means that it is the most reliable test known to diagnose this condition. Hypothetically, it should have an accuracy of 100% but this is not always the case as no medical diagnostic technique is infallible. Colposcopy is a diagnostic procedure that utilizes a colposcope, which can be likened to a binocular microscope, to examine an illuminated and magnified view of the cervix, vulva and vagina. During this examination, the colposcopist distinguishes normal from abnormal cells and takes biopsies as required for further pathological examination.

The Pap smear is simple, effective and relatively cheap when compared to the colposcopy. On the other hand, it is not disease specific and has lower sensitivity and specificity. Though the colposcopy is highly accurate it is also resource intensive and time consuming. It also requires specially trained personnel to carry out the procedure (colposcopists). All these factors make it expensive and hence it is reserved only for follow up referrals based on abnormal or ambiguous Pap smear results. Moreover, it is invasive, may be painful and is definitely very uncomfortable for many women.

Therefore, there exists a demand for a highly accurate, non-specialist and relatively cheap method to improve the sensitivity of the Pap smear. This would be of great benefit to health screening professionals in general, but in particular benefit to the developing world where due to lack of screening, cervical cancer is the single largest cause of mortality in young women.

E. Njoroge, S. R. Alty and M. R. Gani are all with King's College London, Centre for Digital Signal Processing Research, Strand, London WC2R 2LS, UK.

M. Alkatib is with the Acute Gynaecology Unit, St. George's Hospital, Tooting, London SW17 0QT, UK.

Please address all correspondence to `steve.alty@kcl.ac.uk`

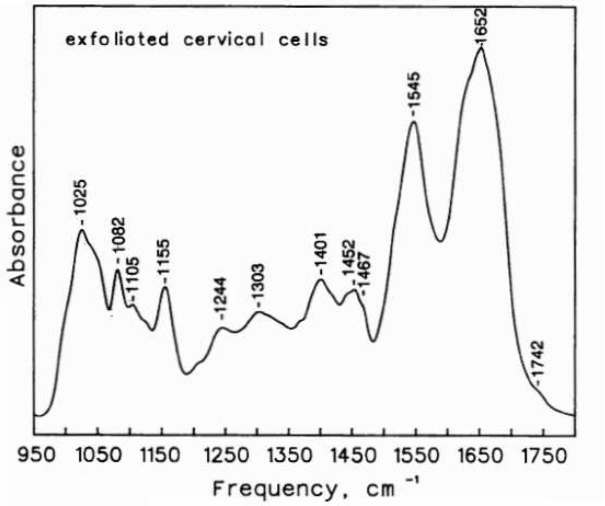

Fig. 1. FTIR spectrum of healthy cervical cells, from [4].

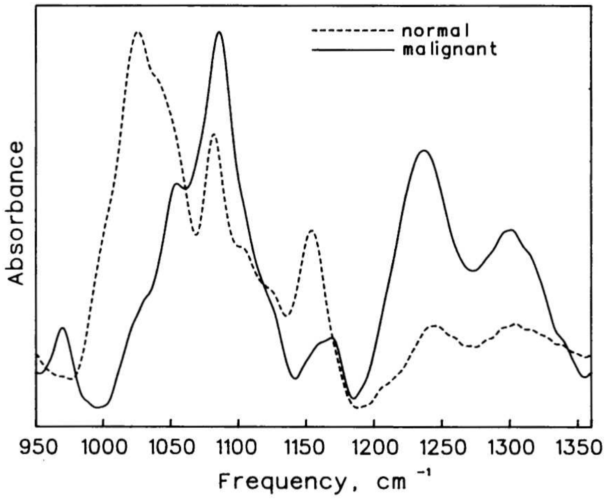

Fig. 2. FTIR spectra of healthy and malignant cells, from [4].

II. Background Theory

Recent medical research [4] has shown that infra-red spectra obtained from exfoliated cervical cells hold vital clues when it comes to diagnosing *dysplasia*, a distinct pre-malignant state, or full blown cancer. Cervical cells contain specific building blocks which result in certain absorption bands being prominent in the infra-red spectrum. These clues are contained in the changes observed in the dominant bands. Some of the bands that are of interest include those at:

- 1025cm^{-1} and 1047cm^{-1} which are mainly from the vibrational modes of -CH$_2$OH groups and the C-O stretching coupled with C-O bending of the C-OH groups of carbohydrates.
- 1082cm^{-1} which is due mainly to the symmetric phosphate (PO$_2^-$) stretching mode.
- 1244cm^{-1} which is due to the asymmetric phosphate (PO$_2^-$) stretching mode.

Fig. 1 shows the infra-red spectrum of a normal cervical cell sample showing the bands of interest. The infra-red spectra of malignant cervical samples display the differences outlined below:

- Significant changes in intensities of the bands at 1025cm^{-1}, 1047cm^{-1}, 1082cm^{-1}, 1155cm^{-1}, 1244cm^{-1} and 1303cm^{-1}.
- Significant shifts of the peaks normally appearing at 1082cm^{-1}, 1155cm^{-1} and 1244cm^{-1}.
- An additional band peaking at 970cm^{-1}.

The infra-red spectrum of a sample with dysplasia displayed similar features as the malignant samples but less significantly changed [4]:

- Intensity of the glycogen bands is intermediate between those of normal and malignant samples.
- The peak of 1082cm^{-1} band is not shifted.
- The center of gravity of the 1155cm^{-1} is shifted less than cervical cancer.
- There is still an additional band at 970cm^{-1} but it is less intense than that of cervical cancer.

A comparison of the FTIR response of a normal and malignant sample is shown in Fig. 2 clearly exhibiting the shift in peaks described.

III. Support Vector Machines

Support Vector Machines (SVMs) [5]–[7] have received a great deal of attention recently proving themselves to be very effective in a variety of pattern classification tasks. They have been applied to a number of problems ranging from hand-written character recognition, bioinformatics to automatic speech recognition (amongst many others) with a great deal of success. A brief summary of the mathematical theory of SVMs follows, for a complete and accessible treatment please see [7].

Consider a binary classification task with a set of linearly separable training samples

$$S = \left\{ \ (\mathbf{x}_1, y_1) \quad \cdots \quad (\mathbf{x}_m, y_m) \ \right\}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d$, i.e., $\mathbf{x}$ lies in a d-dimensional input space, and y_i is the class label such that $y_i \in \{-1, 1\}$. The label indicates the class to which the data belongs. A suitable discriminating function could then be defined as:

$$f(\mathbf{x}) = sgn(\langle \mathbf{w}, \mathbf{x} \rangle + b) \ . \tag{2}$$

Where vector $\mathbf{w}$ determines the orientation of a discriminant plane (or hyperplane), $\langle \mathbf{w}, \mathbf{x} \rangle$ is the inner product of the vectors, $\mathbf{w}$ and $\mathbf{x}$ and b is the *bias* or offset. Clearly, there are an infinite number of possible planes that could correctly classify the training data. Intuitively one would expect the choice of a line drawn through the "middle", between the two classes, to be a reasonable choice. This is because small perturbations of each data point would then not affect the resulting classification. This therefore implies that a good separating plane is one that is more general, in that it is also more likely to accurately classify a new set of, as yet unseen, test data. It is thus the object of an optimal classifier to find the best *generalizing hyperplane* that is equidistant or furthest from each set of points. The set of input vectors is said to be *optimally separated* by the hyperplane if they are separated without error and the distance between the closest vector and the hyperplane is maximal. This approach leads to the determination of just one hyperplane.

A. Soft-Margin Classifier

Typically, real-world data sets are in fact linearly inseparable in input space, this means that the maximum margin classifier approach is no longer valid and a new model must be introduced. This means that the constraints need to be relaxed somewhat to allow for the minimum amount of misclassification. Therefore the points that subsequently fall on the wrong side of the margin are considered to be errors. They are, as such, apportioned a lower influence (according to a preset *slack variable*) on the location of the hyperplane. In order to optimize the soft-margin classifier, we must try to maximize the margin whilst allowing the margin constraints to be violated according to the preset slack variable ξ_i. This leads to the minimization of: $\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$ subject to $y_i(\langle\mathbf{w},\mathbf{x}_i\rangle + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for $i = 1, \ldots, m$. The minimization of linear inequalities is typically solved by the application of Lagrangian duality theory [7]. Hence, forming the primal Lagrangian,

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\beta_i\xi_i - \sum_{i=1}^{m}\alpha_i\left[y_i(\langle\mathbf{w},\mathbf{x}_i\rangle + b) - 1 + \xi_i\right], \quad (3)$$

where α_i and β_i are independent *Lagrangian multipliers*. The *dual-form* can be found by setting each of the derivatives of the primal to zero thus, $\mathbf{w} = \sum_{i=1}^{m} y_i\alpha_i\mathbf{x}_i$ and $\sum_{i=1}^{m} y_i\alpha_i = 0$, then re-substituting into the primal thus,

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m} y_i y_j \alpha_i \alpha_j \langle\mathbf{x}_i, \mathbf{x}_j\rangle \ . \quad (4)$$

Interestingly, this is the same result as for the maximum margin classifier. The only difference is the constraint $\boldsymbol{\alpha} + \boldsymbol{\beta} = C$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta} \geq 0$, hence $0 \leq \boldsymbol{\alpha} \leq C$. This implies that the value C, sets an upper limit on the Lagrangian optimization variables α_i, this is sometimes referred to as the *box constraint*. The value of C offers a trade-off between accuracy of data fit and regularization, the optimum choice of C will depend on the underlying nature of the data and is usually determined by *cross-validation* (whereby the classifier is tested on a section of *unseen* data). These equations can be solved mathematically using Quadratic Programming (QP) algorithms. There are many online resources of such algorithms available for download, see website referred to in [7] for an up to date listing.

B. Kernel Functions

It is quite often the case with real-world data that not only is it linearly non-separable but it also exhibits an underlying non-linear characteristic nature. Kernel mappings offer an efficient solution by non-linearly projecting the data into a higher dimensional feature space to allow the successful separation of such cases. The key to the success of Kernel functions is that special types of mapping, that obey Mercer's Theorem, offer an *implicit* mapping into feature space. This means that the explicit mapping need not be known or calculated, rather the inner-product itself is sufficient to provide the mapping. This simplifies the computational burden dramatically and in combination with SVM's inherent generality largely mitigates the so-called "*curse of dimensionality*". Further, this means that the input feature inner-product can simply be substituted with the appropriate Kernel function to obtain the mapping whilst having no effect on the Lagrangian optimization theory. Hence, the relevant classifier function then becomes:

$$f(\mathbf{x}) = sgn\left[\sum_{i=1}^{nSVs} y_i\alpha_i K(\mathbf{x}_i, \mathbf{x}) + b\right] \quad (5)$$

and for regression

$$f(\mathbf{x}) = \sum_{i=1}^{nSVs}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + b \ , \quad (6)$$

where $nSVs$ denotes the number of support vectors, y_i are the labels, α_i and α_i^* are the Lagrangian multipliers, b the bias, $\mathbf{x}_i$ the *Support Vectors* previously identified through the training process, and $\mathbf{x}$ the test data vector. The use of Kernel functions transforms a simple linear classifier into a powerful and general non-linear classifier (or regressor). There are a number of different Kernel functions available [7], however, one of the most consistently useful is the *Gaussian Radial Basis Function* (RBF) Kernel, given by

$$K(\mathbf{x}_i, \mathbf{x}) = exp(-\|\mathbf{x}_i - \mathbf{x}\|^2/2\sigma^2) \ . \quad (7)$$

It was found that using this Kernel gave the best performance for both classification and regression results.

IV. RESULTS

A. Classification of Dyskaryosis

There are a number of different phases of cervical cancer defined by physicians and tested for by the screening process. *Dyskaryosis* is the term used to describe the abnormal cells taken at a cervical smear. They are observed microscopically. The smear may show mild, moderate or severe dyskaryosis. A borderline smear result suggests there may be some slightly abnormal cells on the cervix but there is some uncertainty. The cervix may be entirely normal but there are some cells on the smear which cause the clinicians some minor concerns. Most clinicians will ask for an HPV test on these smears. If the result is positive then a colposcopy is required. If the result is negative then colposcopy is not usually necessary. *Dysplasia* is the term given to abnormal cells seen on the biopsy. Dysplasia is divided into three grades, CIN1, CIN2 and CIN3 to describe the different levels of abnormality. Cervical intraepithelial neoplasia (CIN) is an abnormality confined to the epithelial layer and therefore not in itself cancerous. CIN1 is usually observed as it

	Pap Smear		
	Normal	**Border+CIN1**	**CIN2,3**
Colp. Normal	71.4%	14.3%	14.3%
Colp. Border+CIN1	59.0%	38.5%	2.6%
Colp. CIN2,3	28.6%	28.6%	42.9%

TABLE I

PAP CLASSIFICATION OF CERVICAL CELLS

	FTIR/SVM Classifier		
	Normal	**Border+CIN1**	**CIN2,3**
Colp. Normal	50%	50%	0%
Colp. Border+CIN1	7.1%	85.8%	7.1%
Colp. CIN2,3	0%	0%	100%

TABLE II

FTIR/SVM CLASSIFICATION OF CERVICAL CELLS

may well return to normal. CIN2 and 3 is regarded as an abnormality which may develop into cancer in some people if left untreated. Therefore most patients with CIN2 or 3 will require some form of treatment.

B. *Data preprocessing*

The data provided by St. George's hospital Tooting, London contained the FTIR spectra of 53 cervical cancer patients along with their Pap smear and colposcopy results. The FTIR spectra were normalized to a standard amplitude scale and certain bands according to section II were used as features for the support vector classifier and the colposcopy results as training targets. Given the fairly limited size of the cohort currently available the test classes were re-grouped to three classes from the original five. Hence, class 1 represents normal, class 2, borderline and CIN1 cases and class 3 represents CIN2 and CIN3 cases.

C. *Pap test results*

Table 1 shows the results for the comparison of the Pap smear results when compared to the colposcopy results. As can be seen, a perhaps surprising degree of mis-classification occurs. The overall weighted correct classification being around 44%. Although quite sensitive to normal cases, this drops off when differentiating between borderline, CIN1, CIN2 and 3 cases. This is why all Pap smear examinations that appear abnormal are referred for colposcopy or at least HPV testing.

D. *FTIR using SVM results*

The OSU SVM Maltab® toolbox [8] was used to train and test a Support Vector Classifier using the FTIR data. A Gaussian Radial Basis Function Kernel was used and after some experimentation a constraint factor of $C = 1000$ and $\sigma = 1.5$ was found to be optimum. 33 data sets were used to train the SVC and 20 to test it and Table 2 shows the results. The sensitivity to both borderline, CIN1, CIN2 and 3 classes is very good, significantly better than the Pap test. However, it can also be seen that half of the normal cases have been lumped into the borderline class. This result is less of a concern as it represents a tendency to false positive and notwithstanding the overall weighted classification when compared to the colposcopy is around 80%. In spite of the limitation of this initial study, the figures are clearly superior to those of the Pap smear and represent a very encouraging result.

V. CONCLUSIONS

This paper presents the application of a machine learning methodology to cervical cancer screening based on FTIR absorption data. These initial results show significant potential, giving higher sensitivity and specificity than the traditional Pap smear test. The method would require little expert knowledge to implement and could lead to a new technique for cheap and effective screening against cervical cancer. This would be particularly attractive as a replacement for the Pap smear and could be usefully deployed in developing countries where mortalities caused by this form of cancer are very high.

The results presented here represent an initial study based on a fairly small cohort. The authors intend to extend the work to include a larger number of patients and optimize the classification process to try to approach the accuracy of the colposcopy gold standard.

VI. ACKNOWLEDGMENTS

The authors would like to acknowledge the Acute Gynaecology Unit at St. George's Hospital in Tooting, London for providing the FTIR data used in this study.

REFERENCES

[1] R. J. Greenblatt, "Human papillomaviruses: Diseases, diagnosis, and a possible vaccine", Clinical Microbiology Newsletter, 27(18), pp. 139–145, 2005.

[2] C. M. Lowndes and O. N. Gill, "Cervical cancer, human papillomavirus, and vaccination", British Medical Journal, Vol. 331, pp. 915–916, 2005.

[3] G. N. Papanicolaou and H. F. Traut, "Diagnosis of Uterine Cancer by the Vaginal Smear", New York, The Commonwealth Fund, pp. 19–45, 1943.

[4] P. T. T. Wong, R. K. Wong, T. A. Caputo, T. A. Godwin and B. Rigas, "Infrared Spectroscopy of Exfoliated Human Cervical Cells: Evidence of Extensive Structural Changes During Carcinogenesis", Proceedings of the National Academy of Sciences, Vol 88, pp. 10988–10992, 1991.

[5] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[6] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, 2, pp. 121–167, 1998.

[7] N. Christianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, http://www.support-vector.net, 2000.

[8] J. Ma, Y. Zhao and S. Ahalt, *OSU SVM Classifier Matlab Toolbox (version 3.00)*, http://www.eleceng.ohio-state.edu/maj/osu_svm/, 2002.