A data-driven modelling based approach to evaluating prognostic value of Electrical Impedance Spectroscopy for cervical cancer diagnosis

Ping Li*, Peter E. Highfield, Zi-Qiang Lang*, Darren Kell****

**Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield,*
S1 3JD, UK (e-mail: p.li@sheffield.ac.uk; z.lang@sheffield.ac.uk)
***Zilico Ltd, Rutherford House, Pencroft Way, Manchester,*
M15 6SZ, UK (e-mail:peter.highfield@zilico.co.uk; darren.kell@zilico.co.uk)

Abstract: Electrical impedance spectroscopy (EIS) has been used as an adjunct to colposcopy for cervical cancer diagnosis for many years. The study presented in this paper was a longitudinal EIS data analysis where women with a negative colposcopy were followed up to three years and their initial EIS readings were analyzed to see if it was possible to predict the women who subsequently developed cervical cancer. A data-driven modelling approach was proposed to extract features from EIS readings and cross validation techniques were then used to choose the best classification model constructed from the selected features to separate the group of women who developed cervical cancer from those who didn't within the follow-up years. The developed method was applied to analyze a real EIS data set and the results showed that EIS does offer prognostic information on the risk of cervical cancer development over three follow-up years. The method developed is of long-term benefit for EIS–based cervical cancer diagnosis and, in conjunction with standard colposcopy, there is potential with the developed method to provide more effective and efficient patient management strategies for clinic practice.

Keywords: Electrical impedance spectroscopy (EIS), spectrum curve fitting, multivariate analysis of variance (MANOVA), logistic regression, cross validation, cervical cancer, bio impedance, Cole equation

1. INTRODUCTION

Electrical Impedance Spectroscopy (EIS) is a method of measuring the electrical impedance of a substance as a function of the frequency of an applied electrical current. The research on cervical cancer, or more specifically, high-grade cervical intraepithelial neoplasia (HG-CIN) detection using EIS, had been carried out for many years (see e.g. Brown et al 2000, 2005; Abdul et al 2006) and the EIS measurement device ZedScan™ (see Fig 1) has been developed for real-time diagnostics (see Tidy et al 2013). Cervical epithelium is a highly structured and stratified tissue that exhibits changes as it progresses from normal epithelium to high-grade CIN. These changes are associated with the losses in the layer of flattened epithelial cells close to the surface of the cervix, and the increases both in the nuclear cytoplasmic ratio and in the extra cellular space. All of these changes caused by the disease will eventually lead to a change in the impedance compared with that for cells from a normal cervix. As a result, the impedance spectra generated will alter markedly as the cells transform into CIN, enabling healthy tissue to be differentiated from abnormal and precancerous lesions.

EIS has currently been used as an adjunct to colposcopy for HG-CIN detection. The impedances are measured with ZedScan™ at 14 frequencies, logarithmically spaced between 76Hz and 625 kHz. The template matching method has been used for analysing the 14-frequency EIS spectra measured from a maximum of 12 reading sites around the cervix for diagnosis (see Tidy et al 2013, 2018), where the measured spectra are compared with the 'template' spectra generated from the 3-D finite element models of the normal and abnormal cervical tissues and matching between the measured spectra and templates is made using the least squares method, finally the results from matching are then used to generate a probability index for the detection of HG-CIN.

Fig. 1. The ZedScan™ *handset for making the EIS measurements used in this paper. The handset is shown placed on the base*

Complementary to colposcopy, EIS provides an objective scientifically-proven method to differentiate between normal, pre-cancerous and cancerous tissues. It plays an important role in improving performance of diagnosis as shown in previous studies (see Tidy et al 2013, 2018). In the study presented in this paper, instead of HG-CIN detection, our attention will be focused on evaluation of the prognostic value of EIS for cervical cancer diagnosis. To this end, a longitudinal EIS data analysis was carried out where the EIS data to be analysed were from the women who underwent a colposcopy (with EIS as an adjunct) examination and had been diagnosed as HG-CIN negative initially. Those women were followed up and, some of them were then found to develop HG-CIN over the subsequent three years. The objective of this longitudinal EIS data analysis is to evaluate the prognostic value of EIS so as to see if it is possible to identify women who are likely to develop HG-CIN within three follow-up years based on the EIS measurement taken at their initial colposcopy.

The EIS readings used in this study were taken from 569 women. Of these, 35 women were found to develop HG-CIN within three follow-up years and 534 women were not. In the rest of this paper, the entire population was divided into two groups, with one group including all women who had developed HG-CIN within three follow-up years and another group including women who had not. Analysing these EIS data for evaluating prognostic value of EIS is then formulated as a classification problem. Due to the limited size of available longitudinal EIS data, the features used for classification need to be identified using domain knowledge. As such, a model-based feature extraction method is proposed to derive features from measured EIS data for classification in Section 2. This is followed by feature and classification model selection using multivariate analysis of variance (MANOVA) and cross validation techniques in Section 3. The results of data analysis using the developed method are presented in Section 4 with discussion in Section 5 and conclusions in Section 6.

2. MODEL-BASED FEATURE EXTRACTION

The basic idea behind the EIS-based template match method for HG-CIN detection as mentioned above is to identify the difference in spectrum shapes between diseased and non-diseased tissues. Similar ideas were used in this study to evaluate the prognostic value of the EIS for diagnosis. However, instead of directly comparing the measured spectra with the template spectra to generate features for diagnosis, a model-based spectrum curve fitting approach was used. Specifically, we try to fit a model to the measured spectrum, and then derive the required features from the fitted model parameters.

2.1 Model-based bio-impedance spectrum curve fitting

The commonly used model for biological tissue impedance is the Cole-Cole equation of the following form (Cole and Cole 1941, Brown et al 2000):

$$Z(f) = R_\infty + \frac{R_0 - R_\infty}{1 + (\frac{jf}{f_c})^{1-\alpha}} \qquad (1)$$

This is an equivalent model that is commonly used by researchers in the field to describe the relationship between the measured tissue impedance Z and frequency f. In equation (1), R_0 and R_∞ are the resistances at zero and infinite frequency respectively. f is the frequency of excitation current used in measurement and 14 logarithmically spaced frequencies (with $f_1 = 76$Hz and $f_{14} = 625$kHz) are used in measurement. f_c is often referred to as the characteristic frequency and α is a constant $(0 \le \alpha \le 1)$. These four model parameters are associated with the tissue structure and properties under investigation and need to be determined with the measured EIS data.

Equation (1) is a nonlinear complex model and spectrum curve fitting for determination of the model parameters and can be formulated as a complex nonlinear optimization problem. This can be solved using the trust-region-reflective algorithm (Coleman and Li 1996), subject to the bounds determined with the measured EIS spectra. Fig. 2 below shows some typical results of Cole model-based EIS fitting.

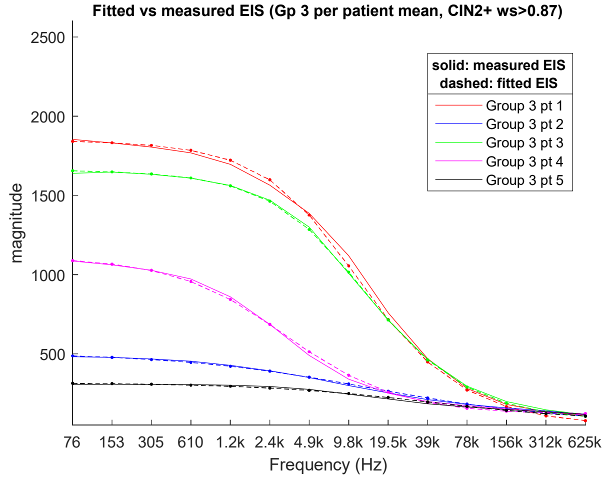

Fig. 2. Comparison between measured and model fitted EIS.

2.2 Feature extraction from fitted model

The structure of biological tissue is complex and the impedance change with frequency will depend upon many factors, such as cellular arrangement (layering of cells), extracellular space, cell size, conductivity of extracellular fluid, thickness of cell membrane, electrical properties of cell membrane and so on. When cervical epithelium progresses from normal epithelium to high-grade CIN, the tissue properties mentioned above will also be altered which are reflected in the changes in the measured EIS spectra, hence allowing EIS to be used for disease detection (Moqadam et al 2018). Ultimately, these changes will lead to changes in the four estimated parameters R_0, R_∞, f_c and α of the Cole model (1), this enable us to derive features from four estimated Cole model parameters for classification so as to evaluate the prognostic value of EIS.

A commonly used interpretation (Brown et al 2020) of the four Cole model parameters for tissue structure is that the inverse of extracellular volume determines R_0, the inverse of the total volume determines R_∞, cell sizes determine f_c, which is the centre of the dispersion, and α is determined by the inhomogeneity of the cells within the dispersion. The conductivity of the intracellular and extracellular spaces will also affect both R_0 and R_∞. When used as an adjunct to colposcopy, EIS spectral measurements are made at up to 12 reading sites around cervix (minimum number of sites is 8) of individual women. As the lesion can either be large to cover many sites or be small just for few /or even a single site around cervix, two types of features were derived from Cole model parameter estimates. The first type of features consist of the four Cole model parameter estimates $(\bar{R}_\infty, \bar{R}_0, \bar{f}_c, \bar{\alpha})$ of the mean spectrum over all reading sites of an individual women which aims to provide information for detecting large lesion that cover many reading sites. The second types of features consist of the four maximum differences of Cole model parameter estimates over all (up to 12) reading sites around the cervix of an individual woman defined as follows:

$$\begin{aligned} \Delta R_\infty &= \max_i R_\infty^i - \min_i R_\infty^i \\ \Delta R_0 &= \max_i R_0^i - \min_i R_0^i \\ \Delta f_c &= \max_i f_c^i - \min_i f_c^i \qquad (i = 1,2,\cdots,12) \qquad (2) \\ \Delta \alpha &= \max_i \alpha^i - \min_i \alpha^i \end{aligned}$$

The features defined by (2) can be viewed as a measure of spatial inhomogeneity of the tissue around cervix and are expected to provide information for detecting small lesions presented in few or just a single reading site. The rationale behind this is that, if there are no lesions around cervix, EIS taken at all sites will have approximately the same shape, thus similar Cole model parameter estimates are expected when performing spectrum curve fitting and the differences defined by equation (2) will be small. However, if a lesion does exist and only presents in a few or a single site, the EIS taken at these sites will significantly differ from those taken at sites where no lesions were present. Hence, the differences defined in equation (2) will be large. To sum up, using both Cole model parameter estimates associated with the mean spectrum and the differences defined by equation (2) (i.e. $\bar{R}_\infty, \bar{R}_0, \bar{f}_c, \bar{\alpha}, \Delta R_\infty, \Delta R_0, \Delta f_c, \Delta\alpha$) from individual women as features will allow both large and small lesions to be detected.

3. FEATURE SELECTION AND CROSS-VALIDATION FOR CLASSIFICATION MODEL DETERMINATION

The evaluation of prognostic value of EIS for cervical cancer diagnosis can be viewed as a problem of detecting early signs in the EIS taken at the initial colposcopy which is caused by the incipient change in tissue structure as neoplasia develops and this is formulated as a classification problem in this paper. Specifically, the feature/predictor vector defined as:

$$\boldsymbol{x} = [\bar{R}_\infty, \bar{R}_0, \bar{f}_c, \bar{\alpha}, \Delta R_\infty, \Delta R_0, \Delta f_c, \Delta\alpha]^T \qquad (3)$$

derived from the fitted Cole model in last section will be used to solve this classification problem.

3.1 Multivariate analysis of variance for evaluating impact of neoplasia on derived features

The complexity of any classifier depends on the number of input dimensions (i.e. the number of features to be used). In the last section, eight handcrafted features have been derived from the EIS measurements. However, the effect of neoplasia on the four Cole model parameters, hence the features derived, is complex and some of these features may be redundant or not informative. Statistically, it is often more attractive to estimate a simpler model with non-informative features being removed as this usually leads to a reduced estimation variance and improved robustness in prediction, and also prevents over fitting for the given data set of limited size. From a practical point of view, a simpler model may also be more interpretable. To this end, a univariate hypothesis test was used for evaluating the capability of each individual feature collected in $\boldsymbol{x}$ to separate two groups. Specifically, the Wilcoxon rank sum test (see e.g. Gibbons and Chakraborti 2011) was applied to each individual feature and the two smallest p-values obtained were 0.14404 ($\bar{\alpha}$) and 0.17515 ($\Delta\alpha$). This indicated that the change caused by neoplasia in any single feature collected in $\boldsymbol{x}$ alone was not statistically significant (at the traditional 5% significance level) to allow a separation of two groups and more features may need to be used to solve the problem. This motivated the use of multivariate analysis of variance (MANOVA) (Rencher 2002) for evaluating the capability of the various combinations of the derived features to separate two groups and the results from analysis are summarized in Table 1, which enable us to identify the most informative feature combination for classification.

Table 1. *p*-values from MANOVA

Feature combinations	*p*-values	Feature combinations	*p*-values
$\bar{\alpha}, \Delta\alpha$	0.016774	$\bar{f}_c, \Delta\alpha$	0.028598
$\bar{\alpha}, \Delta R_0$	0.023126	$\bar{R}_0, \bar{\alpha}$	0.029462
$\bar{f}_c, \bar{\alpha}$	0.025591	$\bar{R}_\infty, \bar{\alpha}$	0.029579
$\bar{\alpha}, \Delta R_\infty$	0.027387	$\bar{f}_c, \bar{\alpha}, \Delta\alpha$	0.031422
$\bar{\alpha}, \Delta f_c$	0.027491	$\bar{R}_0, \bar{\alpha}, \Delta\alpha$	0.033538

Table 1 shows the results for comparing the multivariate means of the different combination of features from the two groups (i.e. HG-CIN developed *vs* no HG-CIN developed within three follow-up years). Columns 1 and 3 in Table 1 specify the feature combinations to be compared and columns 2 and 4 show the corresponding p-values for testing whether the specified combinations have the same means (i.e. the corresponding mean vectors lie in a space of dimension 0). The smaller the p-value, the more confident for us to believe that the corresponding feature combination from two groups has different means. Hence, these p-values can be used as the indices to measure the capability of the corresponding feature combinations to separate two groups. Extensive multivariate analysis of variance had been carried out and Table 1 lists the ten feature combinations with the smallest p-values among all possible combinations of eight features. From Table 1 and for the given data set, it appears that using more features does not necessarily increase capability to separate two groups and the most informative features for separating two groups are associated with the inhomogeneity of the cells within the tissue and around cervix (i.e. $\bar{\alpha}$ and $\Delta\alpha$). These results provide useful information for selecting features to build classifier.

3.2 Stratified cross validation for classification model selection

The problem to be solved in this study can be viewed as a binary classification problem, specifically the goal is to differentiate women who are likely to develop HG-CIN within three follow-up years from those who are not with the selected features derived in the previous section. Once the features to be used for classification are determined, classification models can be trained and the final model for our problem will be determined by cross-validation. However, a major challenge for this study is the limited size of available EIS data set, in particular, the number of women who developed HG-CIN in the whole population is very small (35 of 569). Hence simple partitioning of the data into two (i.e. training and test) sets for building and validating model may not work as this is likely to result in substantially different class distributions between the training and test sets and even no HG-CIN sample at all in some sets. To overcome this difficulty, k-fold cross validation (see e.g. Kuhn and Johnson 2013) with stratified random sampling was applied to the available data set. Specifically, the original EIS data in each group was randomly partitioned into 5 equal sized subsamples respectively. This ensured that each fold contains roughly the same proportions of the two types of classes as in the original population (in this case, each fold will contain 7 HG-CIN data points) and the 5-fold cross validation procedure was then used to choose the best classification model to be used.

Logistic regression, a widely used classification method in medical/clinical data analysis (Christodoulou et al 2019) for disease diagnosis, was selected in this study to perform classification for evaluating the prognostic value of EIS due to its simplicity and interpretability. Logistic regression is concerned with direct modelling the posterior probability $P(C_1|\boldsymbol{x})$ that an instance belongs to a particular class or group C_1 (e.g. women likely to develop HG-CIN within follow-up years) given the observed feature vector $\boldsymbol{x}$. In logistic regression, this posterior probability $P(C_1|\boldsymbol{x})$ is model with the logistic function defined below (James et al 2017):

$$P(C_1|\boldsymbol{x}) = \frac{1}{1+e^{-a(\boldsymbol{x})}} \qquad (4)$$

where $a(\boldsymbol{x})$, in the basic form, is a linear function of $\boldsymbol{x}$ defined as:

$$a(\boldsymbol{x}) = \boldsymbol{\beta}^T \begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix} \qquad (5)$$

and the regression coefficient vector $\boldsymbol{\beta}$ (with up to 9 elements i.e. $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \beta_2 \cdots \beta_8]^T$ in this study) will be estimated from the training data. Classification using the above linear logistic regression model will result in a linear decision boundary (hyperplane $a(\boldsymbol{x}) = [1 \ \boldsymbol{x}]\boldsymbol{\beta} = 0$) which does not have enough flexibility for classifying the data that is not linearly separable. However, it can easily be extended to obtain a non-linear decision boundary by using e.g. polynomial functions of the predictors. In general, $a(\boldsymbol{x})$ can be expressed as:

$$a(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i \varphi_i(\boldsymbol{x}) \qquad (6)$$

where $\varphi_i(\boldsymbol{x})$ $(i = 1, \cdots, k)$ are some known (e.g. polynomial) functions of $\boldsymbol{x}$. In such a case, $a(\boldsymbol{x})$ is still linear-in-the-parameters and can actually be viewed as the linear logistic regression model in terms of new features/or predictors $\varphi_i(\boldsymbol{x})$ $(i = 1, \cdots, k)$. The 5-fold cross validation with stratified random sampling discussed previously can then be employed to evaluate performance of various classification models and determine the best model to be used for evaluation of prognostic value of the EIS. As can be seen, the logistic regression model defined by equations (4), (5) and (6) is computationally simple. The posterior probability $P(C_1|\boldsymbol{x})$ is expressed as an explicit function of the features, hence has good interpretability. This allows us to get a better idea about the relationship between the increased risk of developing HG-CIN and the changes in cervix tissue structure.

4. RESULTS

Following the discussion in the last section, the area under the receiver operating characteristic (ROC) curve (abbreviated as AUC), a commonly used index for measuring the performance of classifier (Fawcett 2006), together with the stratified 5-fold cross validation procedure discussed previously, were used in this study for evaluating the performance of various logistic regression models so as to determine the final model to be used for evaluating the prognostic value of the EIS and the results were summarized in Table 2 and Table 3 below. Columns 1 and 3 of these tables specify the feature combinations used for building the logistic regression models and columns 2 and 4 show the corresponding mean AUC values from 100 repeated 5-fold cross validation runs. A different partitioning of the dataset into 5 folds was implemented (via random permutation of data points in two groups respectively) for each run.

Table 2. Mean AUC values from 100 5-fold cross validation runs with linear logistic regression models

Feature combinations	Mean AUC	Feature combinations	Mean AUC
$\bar{\alpha}, \Delta\alpha$	0.58702	$\bar{R}_0, \alpha, \Delta\alpha$	0.57225
$\bar{\alpha}, \Delta R_\infty$	0.57765	$\bar{f}_c, \alpha, \Delta\alpha$	0.57163
$\bar{f}_c, \bar{\alpha}$	0.57448	$\bar{\alpha}, \Delta R_0$	0.57153
$\bar{f}_c, \Delta\alpha$	0.57443	$\bar{\alpha}, \Delta f_c$	0.56862
$\bar{f}_c, \Delta R_\infty, \Delta\alpha$	0.57358	$\bar{R}_0, \bar{\alpha}$	0.56779

Table 3. Mean AUC values from 100 5-fold cross validation runs with nonlinear logistic regression models

Feature combinations	Mean AUC	Feature combinations	Mean AUC
$\bar{\alpha}^2, \Delta\alpha^2$	0.61029	$\bar{f}_c, \bar{\alpha}^2, \Delta\alpha^2$	0.59105
$\Delta\alpha, \bar{\alpha}^2$	0.59922	$\bar{R}_0^2, \alpha^2, \Delta\alpha^2$	0.58986
$\bar{\alpha}, \Delta\alpha^2$	0.59893	$\Delta R_\infty, \bar{\alpha}^2, \Delta\alpha^2$	0.58946
$\bar{\alpha}^2, \bar{\alpha} \cdot \Delta\alpha, \Delta\alpha^2$	0.59461	$\Delta R_\infty^2, \bar{\alpha}^2, \Delta\alpha^2$	0.58909
$\alpha, \Delta\alpha, \Delta\alpha^2$	0.59391	$\bar{\alpha}^2, \Delta f_c, \Delta\alpha^2$	0.58853

Table 2 shows the ten linear combinations of features for building logistic regression models that have the largest mean AUC values among all possible linear combinations of 8 features defined in (3). As can be seen from Table 2, including more features in the linear logistic regression model does not necessarily improve the classification performance and the best linear logistic regression model (in terms of mean AUC value) is constructed with $\bar{\alpha}$ and $\Delta\alpha$. This is in agreement with the results obtained from the multivariate analysis of variance carried out in last section.

Table 3 shows the ten nonlinear combinations of features for building the logistic regression models that have the largest mean AUC values among all possible nonlinear combinations of (up to the second order polynomial) 8 features. Similarly, it can be seen from Table 3, using more features/or polynomial terms in the logistic regression model does not necessarily improve the classification performance and the best nonlinear logistic regression model (in terms of mean AUC value) is constructed with the polynomial terms $\bar{\alpha}^2$ and $\Delta\alpha^2$, hence it has an ellipse decision boundary. Fig. 3 below is the 2-D histogram of the feature data $\bar{\alpha}$ and $\Delta\alpha$. It can be observed that the $\bar{\alpha}$-$\Delta\alpha$ data points from women who did not develop HG-CIN within follow-up years tend to be concentrated in an area relatively close to the origin; whereas the data points from women who did develop HG-CIN within follow-up years tend to be distributed over the outskirts of this area away from the origin which means that those women tend to have large $\bar{\alpha}$ or/and $\Delta\alpha$ values.

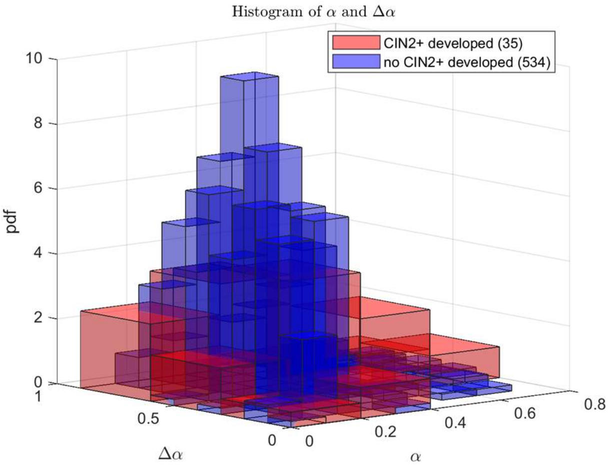

Fig. 3. 2-D histogram of $\bar{\alpha}$-$\Delta\alpha$ data points from two groups

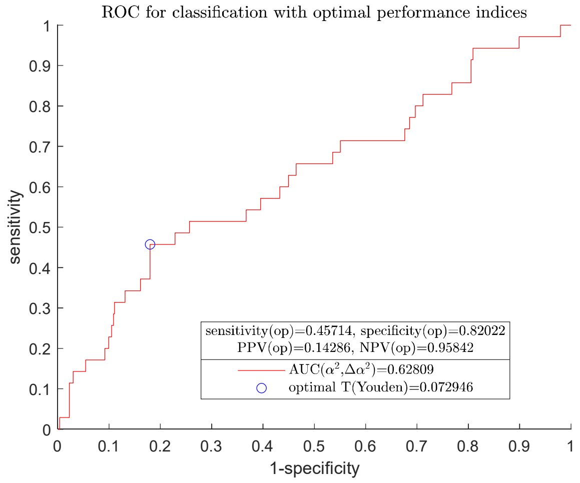

Fig. 4. An ROC curve of final model for separating two groups with OOP and the associated performance indices.

Once the "winning" model structure (in this study, the model constructed with the polynomial terms $\bar{\alpha}^2$ and $\Delta\alpha^2$) was determined, we could then train this model with the whole date set to finalize our classification model and determine the optimal operating point (OOP). The OOP was chosen in this study such that Youden index (Schisterman et al 2005) *J*=sensitivity+specificity-1 was maximized. This could readily be obtained from the ROC curve of the final model and the results are shown in Fig.4.

5. DISCUSSION

The research carried out in this paper is the continuation of the study presented in Brown et al 2020. All the women in the study had a negative outcome at their initial colposcopy and then followed up for three years. The main objective of the research was to see if we were able to identify any increased risk of HG-CIN developing over the follow-up years based on the EIS readings taken at the initial colposcopy so as to evaluate the prognostic value of EIS readings. Ultimately, this will provide useful information for determining the clinical implication and usage of these EIS readings. In the previous study (Brown et al 2020), two single features derived from mean spectra of individual women, i.e. the impedance at 152Hz and the slope of the EIS spectra between frequencies 1.22 and 2.44kHz (used as a proxy for α), were respectively used to build a classifier for separating the two groups. The classification performance for the given data set of these two classifiers were compared with that of the new logistic regression classifier developed in this paper and the results were summarized in Table 4.

Table 4. Classification performance comparison between the new classifier developed and the previous classifiers

Classifier	AUC	Sensitivity	Specificity
Logistic regression ($\bar{\alpha}^2$, $\Delta\alpha^2$)	0.628	45.714%	82.022%
Impedance at 152Hz	0.621	38.7%	83.4%
Slope (between 1.22 and 2.44kHz) as α	0.596	45.2%	70.1%

In Table 4, the sensitivity and specificity were calculated at the OOP determined from ROC curves of the corresponding classifiers. It can be seen from Table 4, that the logistic regression classifier developed in this paper can achieve balanced sensitivity and specificity and has, overall, a better performance in comparison with the previous classifiers.

An added advantage of the new classifier is that the posterior probability of a woman developing HG-CIN over follow-up years, given the measured EIS, is expressed as an explicit function of the selected features. This may help interpretability of the resulting classification model and may provide us with valuable information to understand the reasons why the changes we observed in the EIS spectra were associated with an increased risk of HG-CIN in some women. Specifically, the study presented in the paper has shown that the probability of developing HG-CIN is related to the handcrafted features $\bar{\alpha}^2$ and $\Delta\alpha^2$ which, in turn, are associated with inhomogeneity of the cells within the tissue and the spatial inhomogeneity of the tissue around the cervix. This could be helpful in explaining the evolution of the CIN from histopathological perspective.

As can be seen from Table 4, the classifier can achieve a relatively higher specificity in comparison with sensitivity. This means that, unlike the most diagnostic classifiers which are focused on detecting the high risk patients, it appears that the classifier developed here is good at picking out the low risk patients and the negative predictive value (NPV) for the given EIS data set with logistic regression classifier is 95.842%. This should have some clinical implications and could be used for developing a patient management strategy. Currently if there is any question of HG CIN risk, patients are called back within 6-12 months. If, however, we can confidently classify a group as low risk, those patients then need not come back for 3-5 years, which could save in terms of resources.

6. CONCLUSIONS

A data driven modelling-based approach has been proposed to evaluate the prognostic value of EIS for HG-CIN diagnosis, where the handcrafted features are derived from Cole parameters estimated from EIS measurements. These features are then used to construct the classification model to distinguish two groups so as to determine the prognostic value of EIS. The key novelty of the proposed approach is to introduce the maximum differences of the Cole model parameter estimates over all reading sites around the cervix as features, in addition to the Cole parameter estimates of the mean spectra. These maximum differences can be viewed as a measure for the spatial inhomogeneity of tissue around cervix and allow the earlier signs caused by small incipient lesion to be detected. These earlier signs could be smoothed out by averaging or covered by diversity of conditions between individual patients, hence may be difficult to be detected using features derived from the mean spectra alone.

The results obtained in this study suggests that the increased risk of developing HG-CIN is associated with the increase in inhomogeneity of the cells within tissue and the spatial inhomogeneity of the tissue around cervix. This is consistent with the speculation postulated in Brown et al 2020. The weakness of the classification model developed in this study is mainly from the small data set, in particular, a very small portion of high risk patients within the available population. Nevertheless, a sensible separation in the EIS data from the patient groups has been achieved. This shows that EIS does contain prognostic information on evolving cervical neoplasia and can provide the basis for development of a practical patient management strategy for clinic use.

ACKNOWLEDGEMENT

The authors would like to thank Innovate UK Knowledge Transfer Partnerships for supporting this work through grant KTP-11443.

REFERENCES

Abdul, S., Brown, B.H., Milnes, P., and Tidy, J.A. (2006). The Use of electrical impedance spectroscopy in the detection of cervical intraepithelial neoplasia. *Int.J. Gynecol Cancer,* Vol. (16), 1823-1832.

Brown, B.H., Tidy, J., Boston, K., Blackett, A.D., Smallwood, R.H., and Sharp, F. (2000). The relationship between tissue structure and imposed electrical current flow in cervical neoplasia. *Lancet.* 355, 892-895.

Brown, B.H., Milnes, P., Abdul, S., and Tidy, J.A. (2005). Detection of cervical intraepithelial neoplasia using impedance spectroscopy – prospective study. *BJOG,* Vol. (112), 802-806.

Brown, B.H., Highfield, P.E., and Tidy, J. (2020). Prognostic value of Electrical Impedance Spectroscopy (EIS) when used as an adjunct to Colposcopy – a longitudinal study. *Journal of Electrical Bioimpedance,* Vol. (11), 81-86.

Christodoulou, E., Ma, J., Gollins, G.S., Steyerberg, E.W., Verbakel, J.Y., and Calster, B.V. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology,* Vol. (110), 12-22.

Cole, K.S., and Cole, R.H. (1941). Dispersion and absorption in dielectrics. *J. Chem. Phys,* 9341-9351. https://doi.org/10.1063/1.1750906

Coleman, T.F., and Li, Y. (1996). An Interior Trust Region Approach for Nonlinear Minimization Subject to Bounds. *SIAM J. Optimization,* Vol.6, No. 2, 418-445.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* Vol.27, 861-874.

Gibbons, J.D., and Chakraborti, S. (2011). *Nonparametric Statistical Inference*, 5th Edition, Chapter 8. Chapman & Hall/CRC, Boca Raton, FL 33487-2742, USA.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017). *An Introduction to Statistical Learning*. Springer Science, New York, USA.

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling*, Chapter 4. Springer Science, New York, USA.

Moqadam, S.M., Grewal, P.K., Haeri, Z., Ingledew, P.A., Kohli, K., and Golnaraghi, F. (2018). Cancer detection based on electrical impedance spectroscopy: A clinical study. *Journal of Electrical Bioimpedance,* Vol. (9), 17-23.

Rencher, A.C. (2002). *Methods of Multivariate Analysis,* 2nd Edition. John Wiley & Sons, Inc., New York, USA.

Schisterman, E.F., Perkins, N.J., Liu, A., and Bondell, H. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples *Epidemiology,* Vol. (16), 73-81.

Tidy, J.A., Brown, B.H., Healey, T.J., Daayana, S., Martin, M., Prendiville, W., and Kitchener, H.C. (2013). Accuracy of detection of high-grade cervical intraepithelial neoplasia using electrical impedance spectroscopy with colposcopy. *BJOG,* Vol. (120), 400-410.

Tidy, J.A., Brown, B.H., Lyon, R.E., Healey, T.J., and Palmer, J.E. (2018). Are colposcopy and electrical impedance spectroscopy complementary when used to detect high-grade cervical neoplasia? *European Journal of Gynaecological Oncology,* Vol. (39), 70-75.