**Running title: Early Detection of Malignant and Pre-malignant using DNA Image Cytometry**

Yao Liu1#, Jianying Li2#, Zheng Sun1\* and Xiaoxin Chen3\*

2 Frontier Bioinformatics Solution, Cary, NC 27519, USA

#These authors contributed equally

\*Corresponding Authors

**Main contribution and objectives:**

1. A novel way to process the image data -- ExGCRn (figure 2/3)
2. Model selection and assessment -- model\_assessment table (table 1/figure 4/5)
3. Render high-risk vs. low risk for OLK patient, and guide follow up schedule accordingly (figure 6)

**Abstract**

DNA ploidy status directly reflects the cellular neoplasm activity and the abnormal cell division can be detected when the aneusomy or aneuploidy is observed. Several methods have been developed to directly measure the DNA content and further convert to the ratio of G0/G1. This enables us to determine the ploidy status, which can serve as a reliable marker of cell proliferation, even before the clear histopathological sign is observed. Exfoliative cytology, which offers a simple and non-invasive procedure, has emerged as a prominent technology in early oral cancer diagnosis, though many technical hurdles largely limited this method from becoming an automated and robust clinical standard protocol. In this research, we explored a novel data analysis procedure which allows automatically processing the data and re-constructing informative new variables by integrating the expert guided parameters. We then leveraged the modern machine learning technique to efficiently utilize the newly constructed data to build statistical prediction models. Using resampling methods for pruning the model core parameters, we tested a series of models and a successful Support Vector Machine (SVM) model was finally determined. Our method showed high sensitivity (median > 0.98) and specificity (median > 0.99) obtained both during the training process and in predicting on a hold-off test data. Finally, we proposed a risk index metrics for the oral leukoplakis (OLK) diagnosis, clinically defined lesions likely leading to oral squamous cell carcinoma (OSCC). Such an index reflects the probability leading to OSCC predicted from our statistical model, and it will provide a valuable guide for the clinical professionals to develop a meaningful patient’s follow up plan.

**Introduction**

Oral and pharyngeal cancer, grouped together, rank sixth among most common cancer types in the world (ref). Oral cancer is one of the most common malignancies and contributes the majority of cancer morbidity and mortality worldwide (ref). Even with the sufficient surgery, radiotherapy and chemotherapy; unfortunately, the 5-year survival rate of patients with distant metastases at the time of the first diagnosis is only 19%. On the contrary, early diagnosis for operable tumors with localized stage renders a much higher survival rate (approximates 80%) (ref). Oral squamous cell carcinoma (OSCC), a common type of oral cancer, can develop from precancerous lesions, such as oral leukoplakia (OLK) (ref). Therefore, successfully predication of potential OSCC from the commonly observable OLK carries potential value in clinical practice; and it attracts great attention worldwide (ref). Deeper understanding the OLK clinical implication will provide meaningful guide for patient follow up plan, which almost ensures an early diagnosis of possible cancer before the distant metastasis occurs.

Histopathological diagnosis acts as the current golden criteria, but it is a much invasive method and could cause unnecessary trauma for the patients especially when they were diagnosed negative in the end. DNA ploidy status directly reflects the cellular neoplasm activity and the abnormal cell division can be detected when the aneusomy or aneuploidy is observed (ref). Several methods have been developed to directly measure the DNA content and further convert to the ratio of G0/G1. This primary measurement (commonly called DNA index, or D.I. value) can be converted to the equivalent assessment of the ploidy status, which can serve as a reliable marker of cell proliferation, even before the clear histopathological sign is observed. Exfoliative cytology is the currently accepted method for measuring the cellular DNA content (ref) worldwide; it proves a successful method in screening for epithelial dysplasia in situ or invasive carcinomas of the uteri cervix (ref). Due to the progress of Thin Cytologic Test (TCT) and Automatic Imaging Cytometer (AICM), the exfoliative cytology has also been used in diagnosing the oral cancer and premalignant diseases. Since the location of the carcinoma of oral mucosa is superficial, brushing the exfoliate cell can be done during common dental check up, Therefore, not only does it offer a safe and convenient practice, it also reduces the traumatic injury to the patients.

Although exfoliative cytology offers the greatest potential to be an effective method for early prediction of malignant or pre-malignant, many technical hurdles largely limited this method from becoming an automated and robust clinical standard protocol. Firstly, exfoliative cytology offers a quantitative DNA content measurement, it needs much human intervention to review series of files and look for aneuploidy peaks as well as count the number of cells with excessive increased DNA ploidy. It could be time consuming and sometimes the assessment results could be quite subjective. Secondly, exfoliate cells often consist of mixture of different populations and the results based on thousands of cells presents even harder situations when explaining and interpreting the findings. Lastly, owing to the unbalanced cell populations, statistical models, which have been proved successful in handling mixture of populations (ref), could fail in handling the exfoliative cytology data. The major difficulty has been that the useful signals often buried under the unbalanced amount of the non-informative data. In other words, it is hard to differentiate the signal from the noises.

Upon fully understanding the challenges inherited in this approach, our effort was focused on the data processing and cleaning. In the report, we proposed an Expert-guided data cleaning and reconstruction (**ExGCRn)**, in which we implemented a sequential process to strip out different cell populations while retaining the summary statistics and other useful parameters. Next, we reconstructed a new data set based on *a priori* defined parameters from a mixture of several “populations”. In the end, we defined a set of variables along the axis for initial DNA index values and empirically estimated the density under finite number of the intervals. With the newly constructed the dataset, we then leveraged the modern machine learning technique to build and evaluate a series of statistical prediction models. For each predication model, using resampling methods for pruning the model core parameters, we evaluated the model performance and finalized on best hyper-parameters. Among all the tested models and a successful Support Vector Machine (SVM) model was finally determined. Overall, our method showed high sensitivity (median > 0.98) and specificity (median > 0.99) obtained both during the training process and in predicting on a hold-off test data.

The aim of this study was to establish an analytical protocol and promote a clinical standard for clinical diagnosis of OLK patient, and improve the sensitivity and the specificity of DNA quantitative analysis for potential OSCC early diagnosis. To predict the progressing direction of a clinical defined the oral leukoplakia lesion, we tested our SVM model and found out that the prediction results scattered across the entire panel of “probability”, from almost normal to almost OSCC. It is not surprise that the result was aligned well with the common clinical diagnosis, but it also provides additional risk factor for those OLK patients. Finally, we proposed a risk index metrics for the oral leukoplakia (OLK) diagnosis. Such an index reflects the probability leading to OSCC predicted from our statistical model, and it will provide a valuable guide for the clinical professionals to develop a meaningful patient’s follow up plan.

**Method and materials**

**Patients and clinical specimens collection**

**Exfoliated cells preparation, DNA staining and imaging analysis (classifier)**

**Expert-guided data cleaning and reconstruction**

For each DNA imaging sample, all available DNA index (D.I.) values were exported from the “software name” [reference here]. With the unknown distribution of the random variable(s), kernel density smoothing function was applied to the mixtures of three potential cell populations (normal, mitotic, and aneuploidy). Gaussian kernel was chosen for the smoothing purpose. With the nonparametric estimation of the probability density function, a mathematical procedure [reference here] was applied to search for local peak(s), which was assumed to represent the mode for each cell population. Based on our biological theory, a few assumptions were introduced. It was assumed that the mean D.I. value for normal cell population was “1c”; for mitotics cell population was “2c”; and the cell population with abnormal dividing (aneuploidy population) was above “2.2c” [reference here].

From here a sequential parsing the D.I. value was implemented. (1) To successfully extract the normal cell population, we searched along the mixed density probability density functions (pdf) and located the peak fell around the vicinity of “1c”. Then, we only used the data on the left of the peak to estimate the mean and standard deviation for the first population. Using the background normal distribution as our assumption with the estimated mean and standard deviation, the normal population was extracted out of the dataset. (2) A similar procedure was applied to extract the mitotic cell population and in the end only data left (if any) belonged to the abnormally dividing population. In case, the dataset only contained normal cell population, no cleaning was applied.

The next step is to reconstruct the “useful” dataset for the statistical modeling. To do so, we need summary statistics (mean and standard deviation) from the first two populations and number of data points in each population. The newly constructed data would normally represent the mixtures of two or three population at a control ratio, which provides the basis for building the following prediction model. (1) If only the normal population was determined, a ratio at: 98:1.5:0.5 was used to reconstruct the final data (2) If both normal and mitotic population were determined, the actual ratio between the two families was used and together consisted 99.5% toward the total leaving unchanged 0.5% for the abnormal population (3) If all three populations were determined, the actual ratio between the two families was used and together consisted 90% toward the total leaving unchanged 10% for the abnormal population.

The last step is to convert the reconstructed D.I. values to sixteen measurements of hidden variables. Now, with the known mixture ratios of population(s) and both the summary statistics and raw data value, we expand the D.I. measurement value on the scale between 0 and 8 (for samples with D.I. values greater than 8, they were converted 8). A mixed density was further consulted and then discretized into 16 bins, the density for each bin was stored as the “measurement” for each hidden variable.

**Prediction performance metrics**

For a particular classifier, the confusion matrix can be summarized in various standard performance metrics

The Confusion Matrix

|  |  |  |
| --- | --- | --- |
| Total Samples (*TS*) | Actual Positives (*AP*) | Actual Negatives (*AN*) |
| Predicted Positives (*PP*) | True Positives (*TP*) | False Positives (*FP*) |
| Predicted Negatives (*PN*) | False Negatives (*FN*) | True Negatives (*TN*) |

Sensitivity is a measure of actual positives correctly identified as such and the specificity measures the proportion of negatives correctly identified:

*Sensitivity = TP / AP = TP / (TP + FN)*

*Specificity = TN / AN = TN / (TN + FP)*

There is a trade-off between sensitivity and specificity, making models difficult to compare on the basis of these performance metrics. In contrast, such measures as accuracy, the proportion of correct predictions, the ROC curve, and the Matthews Correlation Coefficient (MCC)74, 75, enable a single parameter comparison of performance of binary classification models. The MCC is defined as:

**Statistical prediction model performance evaluation and selection**

**Predication on clinical patient samples**

**Results**

**DNA staining and imaging analysis with classifier (Figure 1 a-c)**

**Data stripping and signal magnification (figure 1d, 2, 3)**

**Creating new variables from the processed data (figure 4)**

**Selecting SVM as predication model (figure 5)**

**Predication on OLK sample (figure 6)**

**Building clinical risk index metrics for OLK patients**

Discussion and conclusion

**Tables**

Table 1. **Parameter setting for ExGCRn process**

Table 2. **Prediction results on “olk” sample**

**Table 3.**

**Figures Legends**

Figure 1. **DNA Pap staining analysis. (A) (B) (C) (D)**

**Figure 2. Ex**pert **G**uided Data **C**leaning and **R**econstructio**n** (**ExGCRn**) work flow

**DNA Index (D.I.) values collected from a patient sample was processed and measurement for each arbitrarily created variables were produced**

**Figure 3. (A) Density plot of D.I. values from clinical samples**

**Figure 4. Boxplot of available variables of three types of clinical defined samples. OSCC: OLK:, and Normal**

**Figure 5. Model fitting assessment**