Analysis of Morphological Features of Benign and Malignant Breast Cell Extracted From FNAC Microscopic Image Using the Pearsonian System of Curves

Nijara Rajbongshi, Kangkana Bora, Dilip C. Nath¹, Anup K. Das², Lipi B. Mahanta
Central Computational and Numerical Studies, Institute of Advanced Study in Science and Technology, Guwahati, Assam, India, ¹Department of Statistics, Gauhati University, Guwahati, Assam, India, ²Ayursundra Healthcare Pvt. Ltd., Guwahati, Assam, India

Abstract

Context: Cytological changes in terms of shape and size of nuclei are some of the common morphometric features to study breast cancer, which can be observed by careful screening of fine needle aspiration cytology (FNAC) images. Aims: This study attempts to categorize a collection of FNAC microscopic images into benign and malignant classes based on family of probability distribution using some morphometric features of cell nuclei. Materials and Methods: For this study, features namely area, perimeter, eccentricity, compactness, and circularity of cell nuclei were extracted from FNAC images of both benign and malignant samples using an image processing technique. All experiments were performed on a generated FNAC image database containing 564 malignant (cancerous) and 693 benign (noncancerous) cell level images. The five-set extracted features were reduced to three-set (area, perimeter, and circularity) based on the mean statistic. Finally, the data were fitted to the generalized Pearsonian system of frequency curve, so that the resulting distribution can be used as a statistical model. Pearsonian system is a family of distributions where kappa (k) is the selection criteria computed as functions of the first four central moments. Results and Conclusions: For the benign group, kappa (k) corresponding to area, perimeter, and circularity was −0.00004, 0.0000, and 0.04155 and for malignant group it was 1016942, 0.01464, and −0.3213, respectively. Thus, the family of distribution related to these features for the benign and malignant group were different, and therefore, characterization of their probability curve will also be different.

Keywords: Breast cell, FNAC, image processing, Pearsonian system

INTRODUCTION

Breast cancer is a state when cells in the breast begin to grow out of control and usually form a tumor that can often be seen on an X-ray or felt as a lump. The tumor is malignant (cancerous) if the cells can grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. Breast cancer occurs almost entirely among women, but men can get it, too.[1] The worldwide statistics in 2012 reveal that an age-specific rate of 25.16% of all new cases of female cancer patients (excluding nonmelanoma skin cancer) were diagnosed with breast cancer.[2] In India (where this study was carried out), the healthcare burden related to breast cancer is increasing every day. It is mentioned in the International Agency for Research on Cancer (WHO) report of India that the number of new cases of breast cancer increased from 115251 (22.2%) in the year 2008 to 144937 (27.0%) and that 17000 more deaths have occurred in the year 2012 compared to 2008.[3] Breast cancer is curable if it is detected at an early stage. At present, among cancer cases, more than 90% of women diagnosed with breast cancer at an earlier stage survive their disease for at least 5 years compared to approximately 15% women diagnosed with the most advanced stage of the disease.[4] Fine needle aspiration cytology (FNAC) is widely accepted as a reliable technique for the initial evaluation of palpable and nonpalpable (guided biopsy) breast lumps. The procedure is simple, safe, minimally invasive, and as sensitive

Address for correspondence: Dr. Lipi B. Mahanta, Central Computational and Numerical Studies, Institute of Advanced Study in Science and Technology, Guwahati - 781 035, Assam, India. E-mail: lipimahanta@yahoo.co.in

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

How to cite this article: Rajbongshi N, Bora K, Nath DC, Das AK, Mahanta LB. Analysis of morphological features of benign and malignant breast cell extracted from fnac microscopic image using the pearsonian system of curves. J Cytol 2018;35:99-104.
as a biopsy,\(^5\) where a sample of cells from breast lesion are collected and used for preparation of slides. Then, an experienced pathologist observes the slides under a microscope to determine malignancy. However, the entire process is time consuming, involves observer bias, and lacks quantitative evidence (as most observations are subjective in nature).\(^6\) In an FNAC image, changes in shape and size of nuclei play an important role in dysplasia detection. Pathologists consider these changes to be important for studying abnormality.

During the last several years different methods have been developed to classify the FNAC images into benign and malignant groups based on the different morphological features.\(^7\) Present research work is an attempt from a different viewpoint. It is known that the objective of curve fitting is to theoretically describe experimental data with a model (function or equation) and to determine the parameters associated with this model.\(^8\)

Of the two types of models, namely empirical and mechanistic, the ones of primary importance to us are the mechanistic models because they are specifically formulated to provide insight into a chemical, biological, or physical process considered to govern the event under investigation. Parameters derived from mechanistic models are quantitative estimates of real system properties. The selection of a specific statistical distribution as a model for describing the population behavior of a given variable is seldom a simple problem.\(^9\) One approach is to test different distributions (normal, lognormal, Weibull, etc.) and choose the one providing the best fit to the observed data and being the most parsimonious.\(^10\) The Pearsonian system of curves has been used for many types of applications till date. A set of models have been described\(^11\) that capture the stochastic nature of the volatility of rates, stocks, etc., and this family of distributions may prove to be one of the more important distributions. In the United States, the Log-Pearson III is the default distribution for flood frequency analysis.\(^12\) The Type III curve was found suitable for projection of insurance data in India.\(^13\) Andreev (2005) showed the applicability of the Pearsonian system using the example of randomly selected time series ranging from commodity markets to macroeconomic variable. Pizzutilo (2012) also used Pearsonian system of frequency curve to analyze the return distribution of the shares in all companies listed on the Italian stock exchange and found good output on it.\(^14\)

Our aim is to utilize the Pearsonian system of curves to identify the family of distribution for the morphological features which are used to quantify the dysplastic changes of nuclei. To our knowledge such an attempt has not been made in earlier studies. In the present study, cell nucleus is taken as the region of interest (ROI), as during the cancer detection stage, malignancy is always confirmed based on the different features of the cell nucleus. The morphometrics included in the present study are area, perimeter, eccentricity, compactness, and circularity.

This study has a twofold contribution to the existing literature. First, with respect to theory, it suggests a possible basis for the development of more realistic biostatistical models in the presence of deviations from normality of the distributions of cell nucleus. Second, with respect to practice, its implications are applicable for the design of diagnostic strategies for breast cancer management.

**Patients and Methods**

For this study FNAC images of breast lesion were generated at a healthcare centre of Eastern India. Experienced doctors collected the FNAC sample from 14 patients and prepared the slides following all hospital ethical protocols. The slides were identified and separated. Six patients were identified as benign cases and 8 were malignant cases. Next, 20 nonoverlapping images per slide were captured using Leica ICC50 HD microscope at 400× resolution using 24-bit color depth. These digitized images were reviewed by experienced and certified cytopathologists who selected the best images.

Finally, a pathologist also marked the ROI to be segmented. Subsequently, two repositories of cell level images were created comprising 693 normal (benign) samples and 564 abnormal (malignant) samples. Normal samples collected for this study mainly included fibroadenoma cases and abnormal samples included cancerous cases.

In the next phase, nuclei of the images were segmented using basic image processing techniques. Preprocessing, segmentation, postprocessing, feature extraction, and classification are some of the major steps of the image processing approach. In this approach, we concentrated up to the feature extraction level. Figure 1 displays the block diagram of the present segmentation and feature extraction technique.

The present image processing approach is completed in 4 steps. Step 1 comprised informative channel identification followed by two-level histogram equalization technique.\(^15\) This step can be termed as preprocessing phase. Channel identification is very important in image processing as information hidden in an image not reflected in one channel may be identified correctly in another channel. In doing so, 9 different channels using three different color models, viz. RGB, CMY, and HSI were studied properly. All segmentation experiments were tested on the 9 channels of the input image, and finally the best informative channel was identified. For rest of the work, cyan channel was finalized. The main objective of using the channel is that it does not contain much noise and some of the artefacts, namely the presence of RBC commonly seen in FNAC images, are reduced to an extent. Moreover, it is not easily affected by poor staining. This is followed by two-level histogram equalization to further reduce the noise as well as to highlight the regions where nuclei may be present.

In step 2, Otsu thresholding is performed which is one of the simplest thresholding techniques in literature.\(^16\) In this technique, the interclass variance between background and foreground pixels are monitored, and the gray level with the least variance is considered as the final threshold. For simplicity, we are concentrating on this basic segmentation.
approach which is easy to understand for all users. This step is followed by application of morphological operations namely erosion and dilation to properly distinguish the ROI by further reducing noise. This will produce output of an image containing only the nuclei. The final step is feature extraction where 5 shape features were extracted from the segmented nuclei. The description of shape features are as follows:

a. Area: Represents the number of pixels used to represent the object
b. Perimeter: Represents the number of pixels used to represent the boundary of the object
c. Eccentricity: The ratio of the minor axis to major axis of the object
d. Circularity: \[ \text{Circularity} = \frac{\text{Area}}{\text{New Area}} \text{ where } \text{New Area} = \frac{\text{Perimeter}^2}{4\pi} \]
e. Compactness: \[ \text{Compactness} = \frac{\text{Perimeter}^2}{2 \times \text{Area}} \]

These morphometric features can be useful in quantifying the dysplastic changes in cell nuclei. Area can give information about cell enlargement. Perimeter can be useful in studying the irregularity in size of nuclei. Similarly, eccentricity and circularity can help reveal dysmorphism and irregularity in nuclear membrane. Compactness helps study the polarity of the cell.

For statistical analysis, we first extracted the descriptive statistics measurement of the five features. We avoided those features where significant changes were not observed in the average values among the two groups, viz., benign and malignant. After selecting significant features, their extracted feature sets were fitted into Pearsonian system of frequency curve one by one to identify the specific type they followed.

Pearsonian system is a parametric family of distribution introduced by Karl Pearson (1895). It is system of curves used to model observations obtained from real life situations and can be effectively used to determine the properties of the observations based on which predictions can be made.

The value of the criterion \( \kappa \), which determines the properties, is calculated using the already developed FIPSYC algorithm accordingly, the value of the estimated frequency corresponding to the different observed frequency in our database were also calculated. Some adjustments were made in the comparisons of \( \kappa \) with values of zero. Although zero value is possible theoretically, it is not exactly so in numerical calculations, especially in computers. Hence, values “close to zero” were taken in the above mentioned algorithm for the purpose. To test the significance of goodness of fit of the types of curve estimated, defined by the \( \kappa \) value, Chi-square test is employed. The null hypothesis is that there is no difference in the values of the observed and estimated values of the extracted features. Ideally, if our distribution type is estimated correctly, the null hypothesis should be accepted. Theoretically, when \( P \) values are lesser than 0.05 (5% level of significance) or 0.01 (1% level of significance), the null hypothesis should be rejected. Hence, when \( P \) values were greater than 0.05 or 0.01 we accepted the null hypothesis at 5% and 1% level of significance, respectively.

**Results**

Implementations of image processing and feature extraction were carried out in MATLAB (The MathWorks, Inc., Natick, Massachusetts, United States) R2016a using Intel CORE i5 processor of 2.20 GHz and 4 GB RAM. All the statistical analyses were performed on the extracted features which comprised a dataset of 564 × 5 for malignant samples and 693 × 5 for benign samples. We performed the estimation of...
the parameter using algorithm developed in C language. And goodness fit test is done using Ms-Excel.

Figure 2a-c displays the sample database of FNAC images as well as the output images of the image processing algorithms adopted for this work.

From the segmented output of both benign and malignant samples five features viz. Area, Perimeter, Eccentricity, Compactness and Circularity of a cell nucleus were extracted. Different statistical values (namely mean, median, mode, standard deviation, range, skewness and kurtosis) for these datasets were calculated [Table 1].

After feature extraction, the features with no significant changes in the average value when it turns to malignant from benign lesions were excluded. Average values are highly different for the following features: area, perimeter, and circularity [Table 1]. Hence, only these three features were considered for further investigation. So set where

\[ f_j^{Original} = \{ \text{Area, Perimeter, Eccentricity, Compactness, Circularity} \} \]

\[ f_j^{Reduced} = \{ \text{Area, Perimeter, Circularity} \} \]

Generated reduced feature set of a cell nucleus are fitted into the generalized Pearsonian probability distribution system using the FIPSYC algorithm where the best fit types are automatically selected. The results and the related parameter values are depicted in Table 2.

The dataset for all the features in benign and malignant groups were divided into subintervals for analysis.

For the benign group, the dataset of the features area spread from 108 to 535. From the Table in Appendix II and Table 2 and Figure 3a, it can be observed that the “area” feature of a benign breast cell belongs to Type II family of probability distribution and the curve is symmetrical and bell shaped. For the malignant group, though the minimum and maximum values are 353 and 2405, respectively, it is observed that most datapoints range 500 to 1000. Hence, the probability curve is skewed and fits to Type I distribution [Figure 3b].

In the Chi-square goodness of fit test, area feature for both benign and malignant groups were found to have significant

### Table 1: Descriptive statistical measure of different morphological features (in pixel)

| Features   | Cell type | Mean  | Median | Mode  | Std. dev | Range  |
|------------|-----------|-------|--------|-------|----------|--------|
| Area       | Benign    | 310.32| 315    | 346   | 81.510   | 427    |
|            | Malignant | 986.26| 844    | 834   | 338.852  | 2052   |
| Perimeter  | Benign    | 65.66 | 67.94  | 67.30 | 9.93198  | 62.8700|
|            | Malignant | 119.48| 92.67  | 114.67| 31.23865 | 256.8945|
| Eccentricity| Benign    | 0.68  | 0.635  | 0.702 | 0.12572  | 0.7801 |
|            | Malignant | 0.62  | 0.237  | 0.620 | 0.14660  | 0.9060 |
| Compactness| Benign    | 14.18 | 13.46  | 13.79 | 1.89056  | 20.2499|
|            | Malignant | 16.66 | 14.11  | 14.83 | 5.91488  | 64.5858|
| Circularità| Benign    | 0.897 | 0.93   | 0.91  | 0.07345  | 0.4256 |
|            | Malignant | 0.111 | 0.10   | 0.11  | 0.02608  | 0.14787|

![Figure 2](image-url)
Regarding the “perimeter” feature, in the benign group the range was 36.38 to 99.25. The curve follows Type VII distribution, is symmetrical, and bell shaped [Figure 3c]. For the malignant group, the dataset spreads from 67.35 to 324.25 and belongs to Type IV distribution with skewed characteristics [Figure 3d]. In the chi square goodness of fit test, the \( P \) values corresponding to both (benign and malignant) perimeter features are significant at 1% level of significance (0.4429 and 0.1307, respectively).

Next, the dataset in the benign group for “circularity” feature lies between 0.39 and 1.04 and fits Type IV family of distribution, which is skewed and bell shaped [Figure 3e]. For malignant group, circularity ranges from 0.04 to 0.19 and is described by Type VII family of distribution. \( P \) value corresponding to Chi-square for goodness of fit for benign category was significant at 1% level of significance (0.7691). This feature for the malignant group is symmetrical in nature and was found to be mildly significant (i.e., at 10% level of significance) with a \( P \) value of 0.0526 [Figure 3f].

**Discussion**

In this paper, we discuss the family of probability distribution for three significant morphometric features of the breast cell, as observed from FNAC diagnostic procedure, which differentiates benign from malignant cases. The results reveal that the extreme values within the dataset for the selected features.
features are significantly different for both the benign and malignant groups. Further, the distributions described by the selected features are also different for each category within each group. In the benign group, area of a breast cell follows Type II probability distribution, perimeter follows Type VII, and circularity follows the Type IV. Furthermore, the area of a malignant cell is characterized by Type I family of distribution, perimeter by Type IV, and circularity by Type VII family of distribution. Visual observation from the plots of the probability curves also confirm the difference of characteristics for different morphological features between the two groups of breast cells. We hope that the findings of this work will be useful for the practitioners in various fields of theoretical and applied sciences and may serve as a computational basis for the computer scientist in the development of cancer diagnosis technique.

The limitation of the study is that the overlapping cells observed in a FNAC image are avoided due to computational complexity. Also, this work can be extended by taking the other significant features which are essential to detect a cancer cell in an FNAC image.

**Conclusion**

A lot of approaches are being studied for detection of cancer (malignancy or benign) from FNAC images. From this study it may be concluded that the statistical approach of using distribution as a model to represent the identifying features of a cell may successfully serve as a computational basis for the computer scientist in the development of cancer diagnosis technique.

**Financial support and sponsorship**

Nil.

**Conflicts of interest**

There are no conflicts of interest.

**References**

1. Cancer.org [Internet]. American Cancer Society, Inc.; c2016. Available from: http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-what-is-breast-cancer. [Last accessed on 2016 Sep 26].
2. Globocan.iarc.fr [Internet]. International Agency for Research on Cancer, Inc.; c2016. GLOBOCAN (2012), Population Fact Sheets. Available from: http://globocan.iarc.fr/Pages/fact_sheets_population.aspx. [Last accessed on 2016 Sep 26].
3. Breast Cancer India [Internet]. Statistics of Breast cancer in India, Global Comparison; c2015 by The Pink Initiative. Available from: http://www.breastcancerindia.net/statistics/stat_global.html. [Last accessed on 2016 Sep 26].
4. Cancerresearchuk.org [Internet]. Cancer Research UK, Inc.; Available from: http://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important. [Last accessed on 2016 Sep 26].
5. Nguansangiam S, Jedsapatarakul S, Tangjitgamol S. Accuracy of fine needle aspiration cytology from breast masses in Thailand. Asian Pacific J Cancer Prev 2009;10:623-6.
6. Kalhan S, Dubey S, Sharma S, Dudani S, Preeti DM. Significance of nuclear morphometry in cytological aspirates of breast masses. J Cytol 2010;27:16-21.
7. Chaiwun B, Settakorn J, Ya-In C, Wisedmongkol W, Rangdaeng S, Thomer P. Effectiveness of fine-needle aspiration cytology of breast: Analysis of 2,375 cases from northern Thailand. Diagn Cytopathol 2002;26:201-5.
8. Gray W, Mc Kee G T. Breast. In: Diagn Cytopathol. Churchill Livingstone. 3rd ed. USA: Elsevier Science Limited; 2003. pp. 235-79.
9. Saha M, Mukherjee R, Chakraborty C. Computer-aided diagnosis of breast cancer using cytological images: A systematic review. Tissue Cell 2016;48:461-74.
10. Intro Curve Fitting [Internet]. Cornell University c2016, http://collum.chem.cornell.edu/dbc6/Group_Resources.html. Available from: http://Collum.chem.cornell.edu/documents/Intro_Curve_Fitting.pdf. [Last accessed on 2016 Sep 26].
11. Lahcene B. On Pearson families of distributions and its applications. Afr J Math Comput Sci Res 2013:6:108-17.
12. Yuichi N. The PDF and CF of Pearson Type IV distributions and the ML estimation of the parameters. Stat Probabil Lett 1999;43:251-4.
13. Carpenter JA, Shenton LR. The Mills ratio and the probability integral for a Pearson Type IV Distribution. Biometrika 1964;52:119-28.
14. Lee J. Parameter estimation of the extended generalized Gaussian family distributions using maximum likelihood scheme. Info Technol J 2010;9:61-6.
15. Mahanta LB, Nath DC. The Pearsonian Type III curve and its potentials in projecting insurance data. IFRSA's Int Comput 2012;2:499-503.
16. Andrew A, Kanto A, Malo P. Simple Approach for Distribution Selection in the Pearson System. Helsinki School of Economics. Working Paper (W-388) 2005; ISBN 951-791-927-1. Available from: https://pdfs.semanticscholar.org/6262/ef804dd62fd5bc23191e8f9b994a84da1aaa.pdf. [Last accessed on 2016 Sep 16].
17. Pizzutilo F. Use of the Pearsonian System of frequency Curves for the Analysis of Stock Return Distributions: Evidence and implications for the Italian Market. Econ Bull 2012;32:272-28.
18. Tutorialspoint.com [Internet] Tutorials Point Simply easylearning, Inc.; c2016. Available from: http://www.tutorialspoint.com/dip/Histogram_Equalization.html. [Last accessed on 2016 Sep 27].
19. G Andrew. Lecture notes on Otsu Thresholding [Internet]. 2010. [about 1 p.]. Available from: http://www.labbookpages.co.uk/software/ImgProc/otsuThreshold.html. [Last accessed on 2016 Sep 16].
20. Elderton WP, Johnson NL. Pearson’s system of frequency-curves. In: Systems of Frequency Curves. 5th ed. Cambridge: Cambridge University Press; 1969. pp. 35-46.
21. Mahanta LB. On the Study of Pearsonian system of curves and development of computer algorithms [Internet]. 2004. [about 17 p.]. Available from: http://shodhganga.inflibnet.ac.in/handle/10603/69717?mode=simple. [Last accessed on 2016 Sep 26].