Breast Tissue Characterization Based on Generalised Extreme Value Distribution of Ultrasound Radio-Frequency Data

Jaafar Raad Baqer¹, Aws Alazawi², Abbas Fadhal Humadi ³
¹,²,³ Electrical Engineering Technical College, Middle Technical University, Baghdad, Iraq
¹ eng.jaafarraad@gmail.com, ²aws_basil@mtu.edu.iq, ³drabbas@ mtu.edu.iq

Abstract. In this study, an ultrasound Radio-Frequency (RF) data of healthy and tumour breast regions are considered for tissue characterisation. The main aim of this study is to differentiate the consistent statistical distribution of backscattered RF data with the objective of performing semi-automatic segmentation based on statistics. The differentiation considers Gamma, Generalised Extreme Value (GEV), Log-normal, and Rayleigh distributions. The accuracy of the statistical parameters is measured based on the criteria of both the Kolmogorov-Smirnov Test (KS) and Mean Square Error (MSE) goodness of fit. Results show that there is a possibility of using the parameters of Rayleigh, Gamma, and GEV for different healthy and tumour tissue regions, where GEV yields the best goodness of fit test and its parameters have a good potential to be exploited for further study for segmentation purposes.

1. Introduction

Breast cancer is one of the most common diseases that cause death to women [1]. It has evolved from a normal breast and formed into a lump that feels unlike normal breast mass and is usually detected in this stage through fingertips sensing. The latest world health organisation report illustrates that breast cancer is the most cancer type that occurs to women worldwide, whereas it reaches 24.2% of total cancer cases during 2018 [2]. Although therapy techniques of breast cancer treatment were publically known, the need for further diagnostic safe modalities is in demand. One of the safest imaging modalities is the ultrasound [3]; it is commonly used for monitoring the growth of tumours.

The main drawback of ultrasound is the low image visibility that occurs due to the speckle phenomena. Speckles are considered as a diffuse backscatter reflection that is generated by an interaction of ultrasound waves with the scatterers involved in a point spread function, which is in turn depends on the density and type of tissue scanned. From that end, it is necessary to get a benefit from data information by obtaining accurate identification of the tumours.

In the literature, an investigation of using the distribution function of RF data showed that a Rayleigh distribution could be used to characterise the normal breast, liver, or heart tissues, and also considered in their findings a 1.91 Signal-to-Noise Ratio (SNR) for healthy tissue [4]. Shankar et al. showed that K-distribution was a good fit to identify breast tissue, whether the tissue is healthy or not, by finding multiple parameters to the tissue, like mean value, SNR, and standard deviation, the type of tissue determined based on the value of the parameters [5].

Further studies were considered a differentiation breast tumour using Nakagami distribution and sort them into the Breast Imaging Reporting and Data System classification (BI-RADS) using multiple parameters [6], [7]. BI-RAD system refers to Breast Imaging-Reporting and Data System. It is a system used by radiologists after interpreting a mammogram and used in the MRI and ultrasound. Byra and Hanna used homodyne K-distribution and Nakagami for tumour segmentation [8], [9].
Recently, distributions were considered Gamma, Weibull, Normal, Log-normal [10], where the four distributions were used to differentiate heart tissue and blood regions and the empirical model was evaluated using Rao–Rabson goodness of fit test combined with misclassification test, where Gamma distribution was considered as the best fit in their study. Others were considered the heart tissue and blood pool segmentation [11], investigation of Rayleigh, Rician, K-distribution, Nakagami, and compound distributions were conducted in their study, where statistical analysis was done using Kolmogorov-Smirnov test as a goodness of fit, the result shows K-distribution is the best for both heart tissue and blood pool.

In this study, four distribution functions of Rayleigh, Gamma, Log-normal, and GEV are investigated by considering minimum error using both KS and MSE as a goodness of fit methods, where the reliability of the distribution parameters are determined for a forthcoming study that concerning semi-auto segmentation to differentiate breast tumour regions.

2. Material

The RF data utilised in this paper were published by the department of ultrasound, institute of the fundamental technological research polish academy of sciences [12]. Image data acquired using L14-5/38 linear array transducer operating at 10MHz centre frequency then digitised with 40 MHz sampling frequency and 28Hz frame rate. The type of tumour was histologically assessed by core needle biopsy and classified into the BI-RAD system [13].

The database of RF data has converted to envelop the data by converting the RF data rf(t) to analytic data (an(t)) using Hilbert transformation, then analytic data converted to an enveloped data env(t) by using absolute operation as shown in Figure 1.

![Figure 1. Radiofrequency data conversion into envelope](image)

3. Methods

Four common statistical distributions of Rayleigh, Gamma, Log-normal, and GEV are investigated for backscattered RF data of breast tumours. The healthy regions of interest were selected manually, while tumour regions were marked by radiologist [12]. Two goodness of fit techniques are considered to determine which of the four distributions is consistent with the enveloped-data distribution in terms of least error. The enveloped data set represented by 100 breast tumour tissues, where the regions of interest were determined by a radiologist.
A. Rayleigh Distribution

Rayleigh is a special case of Weibull distribution that is formed when the signal takes multiple routes until it reaches the receiver. Rayleigh distribution was widely used to characterise the tissue in ultrasound imaging [14]. In ultrasound imaging, Rayleigh distribution is generated when the number of uniform cross-section scatterers within a point spread function is very high (fully developed speckles) [15]. The feature that makes it widely used is the simplicity of mathematical operations and it has only one parameter. The probability density function of Rayleigh represented by [16]:

\[ f(x) = \frac{x}{b^2} e^{\frac{-x^2}{2b^2}} \]  

where \( b \) is denoted as a scale parameter.

B. Gamma Distribution

Gamma distribution is commonly used to anticipate the time needed for the operation to be done. Segmentation of heart regions of tissue and blood was performed based on the Gamma distribution [17], where it has two parameters of shape denoted by \( \alpha \) and scale denoted by \( \beta \), the probability density function of Gamma represented by the mathematical formula as [18]:

\[ f(x) = \frac{1}{b^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{b}} \]  

When \( \alpha \) is large, the Gamma distribution is closely approximate a normal distribution.

C. Generalized Extreme Value Distribution

GEV distribution is usually used for modelling the smallest or the largest value among a large set of Independent Identically Distributed (IID) random values representing measurements or observations. The unique characteristic of this distribution is represented by three types. These types are combined in a single form, which allows a variety of continuous shape to be generated based on appropriate distribution. GEV distribution is more flexible to the change of tissue shape and density. In medical applications, GEV was recently used in Intra-cardiac echocardiography (ICE) [19]. The probability density function of GEV is represented as [20]:

\[ f(x) = \left( \frac{1}{\alpha} \right) \exp \left( -\left( 1 + k \frac{x-\mu}{\alpha} \right)^{-1/k} \right) \left( 1 + k \frac{x-\mu}{\alpha} \right)^{-1-1/k} \]  

where \( \mu \) is a location parameter, \( k \) is a shape parameter and \( \alpha \) is a scale parameter.

D. Log-normal Distribution

This distribution is also called the Galton distribution, and can only be applied to positive data as log(x) that cannot have a negative value. It deals with a random variable that has a normal distribution to generate Log-normal distributed data. The probability density function of Log-normal represented as [21]:

\[ f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp \left\{ \frac{-(\ln x - \mu)^2}{2\sigma^2} \right\} \]  

\( \mu \): mean of logarithmic values  
\( \sigma \): standard deviation of logarithmic values
4. Analysis

The above mentioned four distributions have to be evaluated in terms of their consistency to the enveloped-data distribution for the normal and tumour regions based on the goodness of fit techniques. According to the rule of thumb, if the data are congruent with the probability density function that would be considered as an indication of a good fit. On the contrary, if the data do not coincide with the probability density function that would be considered a trivial fit. Both KS and MSE are considered to measure how the four distributions are fit to the enveloped-data distributions of both normal and tumour regions, where the mathematical representation of both fitting techniques are as follows.

A. Kolmogorov-Smirnov Test (KS)

KS is used to determine the distance between two empirical distribution functions or between sample and reference Cumulative Distribution Function (CDF). The KS test is sensitive to the position and shape of the two samples makes it popular to use when statistical analysis is required particularly in medical ultrasound. The test statistic is the maximum absolute difference between the empirical CDF calculated from x and the hypothesised CDF [22]:

\[ KS = \max(|\hat{F}(x) - G(x)|) \]  \hspace{1cm} (5)

Where \( \hat{F}(x) \) is the empirical CDF and \( G(x) \) is the CDF of the hypothesised distribution, and when using the KS between two functions, the test uses the maximum absolute difference between the CDF of the distributions of the two data vectors. The test statistic equation is:

\[ KS = \max(|\hat{F}_1(x) - \hat{F}_2(x)|) \]  \hspace{1cm} (6)

B. Mean Square Error (MSE)

The MSE measures the average square error between the actual value and the expected value. MSE always has a positive outcome According to the mathematical formula [21]:

\[ \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \]  \hspace{1cm} (7)

where \( Y \) is a vector of the recorded results, and \( \hat{Y} \) is a vector of the predicted results.

5. Results and Discussion

The RF data of the region of interest is considered to determine which distribution model is the best fit with the objective of using their parameters in the forthcoming research. A visual investigation has been considered first to evaluate the consistency of the four distributions representing the data. Figure 2(a-f) shows these four distributions for three regions of interest: healthy tissue, benign tumour, and malignant tumour, and how they fit the enveloped-data distribution. In Figure 2, both GEV and Gamma distributions are in good agreement with the data distribution of the three tissue regions compared to the other two distributions of Rayleigh and Log-normal. It has also noticed that the scale of data distribution for the healthy tissue region has a widespread compared to the tumor regions (up to 6mV for healthy and up to 2mV for tumour regions) as shown in Figure 2(b, d, f). The reason behind the variation in the scale of healthy and tumour regions is due to the acoustic properties of tissue under imaging, where the scatterer density within the point spread function in tumour regions is lower than in the healthy region. In addition, the scatterers are in uniform repartition in health regions compared to non-uniform repartition in the tumour region [23]. In a point spread function, scatterers of high density and uniform repartitions attain black spots in the sonography image that reveals low intensity.

From that end, visual distinguish of tissue regions can be obtained based on B-mode image, where tumour area can be marked, yet it cannot characterise tumour type, and the intensity of the backscattered signal depends on
the position and the orientation of the tissue with respect to the ultrasound probe leading to a high error. Therefore statistical-based characterisation has to be considered as a reliable technique.

Figure 2. Four distribution functions fit into the envelope data, healthy ROI, (b) density function of healthy ROI, (c) benign ROI, (d) density function for benign ROI, (e) malignant ROI, (f) density function for malignant ROI.

Table 1 illustrates the results for the goodness of fit techniques that have been calculated using the above equations 6 and 7. It has been noticed that GEV has the minimum error in both techniques of MSE and KS for healthy and tumour tissue regions. From that end, GEV is regarded as an optimum distribution that describes breast tissue concerning the lowest error. As for Gamma and Log-normal, the goodness of fit yields tolerable results to be used in the prospect of different tissue regions either individually or combined with other distribution types.
Table 1. Goodness of fit of both healthy and tumor tissues regions of interest

| Distribution | Healthy tissue | Tumor tissue |
|--------------|----------------|--------------|
| MSE          | KS             | MSE          | KS             |
| Rayleigh     | 3.36 e-09      | 0.503        | 2.37 e-09      | 0.088          |
| Gamma        | 7.60 e-10      | 0.505        | 1.84 e-10      | 0.063          |
| GEV          | 5.19 e-10      | 0.171        | 1.05 e-10      | 0.029          |
| Log-normal   | 2.87 e-09      | 0.530        | 4.88 e-10      | 0.038          |

Furthermore, it has to test the reliability of distribution parameters and know how it changed based on the tissue region. Table 2 illustrates the behaviour of the parameters for the four distributions, where the Rayleigh scale exhibits great differences in value between healthy and tumour tissue, which makes it common to be used for tissue region discrimination. For Gamma parameters, the scale parameter exhibits a large change between healthy and tumour regions that can be exploited to differentiate tissue regions, while its shape parameter does not display any perceptible change for tissue regions.

For GEV parameters, both scale and location parameters are significantly changed as a function of tissue regions, where they can be considered toward tissue region differentiation. While GEV’s shape parameter displays a large variance in magnitude, with a range of values in both the negative and positive regions, it is noticed that the shape parameter tends to be positive when the ROI has tissue diversity and negative values in ROI with homogeneous tissues.

Finally, the Log-normal parameters do not change which leads to the inability to utilise the parameters for breast tissue characterisation.

Table 2. Mean value of the estimated parameters for both healthy and tumour tissue regions of interest

| Distribution | Parameters | Healthy tissue | Tumour tissue |
|--------------|-----------|----------------|---------------|
|              | Scale     | Scale          | Scale         |
| Rayleigh     | 2.53 e+03 | 703.58         |               |
| Gamma        | 2.32      | 1330           | 2.04          |
| GEV          | 0.15      | 1410           | 2100          |
|              | 0.22      | 334            | 480           |
| Log-normal   | 7.65      | 0.75           | 6.28          |
|              | 0.78      |                |               |

6. Conclusions

Statistical analysis has been considered to characterise the ultrasound radiofrequency data of breast healthy and tumour tissue regions. Consistent distribution has been determined, which has been considered as in good agreement with enveloped-data distribution according to both MSE and KS goodness of fit techniques. Four distributions of Rayleigh, Gamma, GEV, and Log-normal are conducted in the determination, where investigation shows how the parameters of the four distributions are influenced by tissue regions. Results show that GEV distribution is consistent with the data set in terms of minimum error compared to Rayleigh, Gamma, and Log-normal distributions. The finding is GEV distribution can be utilised to differentiate between healthy and tumour regions by considering their statistical parameters. Unlike the context of using backscattered echo to identify tissue regions, where scatterers orientation subtle to the changes in statistics of phase, may influence by imaging frequency-dependent attenuation, dynamic focusing, and diffraction. Further investigation can also be recommended, for GEV’s shape parameter to explore its behaviour.
References

[1] "IARC/WHO, GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012, Globocan 1–6, 2012".

[2] WHO report on cancer: setting priorities, investing wisely and providing care for all. Geneva: World Health Organization; 2020.

[3] AIUM, "Overview of Therapeutic Ultrasound Applications and Safety Considerations," Ultrasound Medicine, vol. 31, p. 623–634, 2012.

[4] R. F. Wagner, M. F. Insana, and D. G. Brown, "Statistical properties of radio-frequency and envelope detected signals with applications to medical ultrasound," J. Opt. Soc. Am. A, vol. 4, pp. 910-922, May 1987.

[5] P.M. Shankar, J.M. Reid, H. Ortega, C.W. Piccoli, B.B. Goldberg, "Use of Non-Rayleigh Statistics for the Identification of Tumors in Ultrasound B-Scans of the Breast," IEEE Transactions on Medical Imaging, vol. 12, no. 4, pp. 687 - 692, Dec 1993.

[6] Andrzej Nowicki, Hanna Piotrzkowska-Wróblewska, Jerzy Litniewski, Michal Byra, Barbara Gambin, Eleonora Kruglenko, Katarzyna Dobruch-Sobczak, "Differentiation of normal tissue and tissue lesions using statistical properties of backscattered ultrasound in breast," in IEEE International Ultrasonics Symposium (IUS), Taipei, Taiwan, 21-24 Oct. 2015.

[7] Po-Hsiang Tsui, Chih-Kuang Yeh, Yin-Yin Liao, Chien-Cheng Chang, Wen-Hung Kuo, King-Jen Chang, Chuang-Nien Chen, "Ultrasonic Nakagami Imaging: A Strategy to Visualize the Scattering Properties of Benign and Malignant Breast Tumors," Ultrasound in Medicine & Biology, vol. 36, no. 2, pp. 209-217, February 2010.

[8] Byra, Michal, et al. "Classification of breast lesions using segmented quantitative ultrasound maps of homodyned K distribution parameters." Medical physics 43.10 (2016), pp. 5561-5569.

[9] Piotrzkowska-Wróblewska, Hanna, et al. "Breast carcinoma tissues characterization using statistics of ultrasonic backscatter" 7th Forum Acusticum, 2014.

[10] Tao, Zhong, Hemant D. Tagare, and James D. Beaty. "Evaluation of four probability distribution models for speckle in clinical cardiac ultrasound images." IEEE transactions on medical imaging 25.11 (2006), pp. 1483-1491.

[11] Bernard, Olivier, "Segmentation in echocardiographic imaging using parametric level set model driven by the statistics of the radiofrequency signal," Diss. Ph. D. thesis, INSA-Lyon, 2006, p39-54.

[12] Hanna Piotrzkowska-Wróblewska, Katarzyna Dobruch-Sobczak, Michal Byra, Andrzej Nowicki, "Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions," Medical Physics, vol. 44, no. 11, pp. 6105-6109, 2017.

[13] Mendelson E, Bohm-Velez M, Berg W., "ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System," American College of Radiology, Reston, VA, 2013.

[14] Deng, Yinhui, Yuanyuan Wang, and Yuzhong Shen, "Speckle reduction of ultrasound images based on Rayleigh-trimmed anisotropic diffusion filter.," Pattern Recognition Letters, vol. 32, no. 13, pp. 1516-1525, 1 October 2011.

[15] Weng, Li, John M. Reid, P. Mohana Shankar, and Kawan Soetanto, "Ultrasound speckle analysis based on the K distribution," The Journal of the Acoustical Society of America, vol. 89, no. 6, pp. 2992-2995, 1991.

[16] Papoulis, Athanasios, and S. Unnikrishna Pillai, Probability, random variables, and stochastic processes., Tata McGraw-Hill Education, 2002.

[17] Z. Tao, H.D. Tagare, J.D. Beaty, "Evaluation of Four Probability Distribution Models for Speckle in Clinical Cardiac Ultrasound Images," IEEE Transactions on Medical Imaging, vol. 25, no. 11, pp. 1483 - 1491, 2006.

[18] Hahn, Gerald J., and Samuel S. Shapiro." Statistical Models in Engineering" Wiley Classics Library. New York: Wiley, 1994.

[19] Rettmann, Maryam E., et al. "Segmentation of left atrial intracardiac ultrasound images for image guided
cardiac ablation therapy." Medical Imaging 2013: Image-Guided Procedures, Robotic Interventions, and Modeling. Vol. 8671. International Society for Opt.

[20] Muraleedharan, G., C. Guedes Soares, and Claudia Lucas. "Characteristic and moment generating functions of generalised extreme value distribution (GEV)." Sea Level Rise, Coastal Engineering, Shorelines and Tides. Nova, 2011, pp. 269-276.

[21] Evans, M., N. Hastings, and B. Peacock. Statistical Distributions. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, p. 131.

[22] D’Agostino, R. B., and M. A. Stephens. "Goodness-of-fit techniques. New York: Marcel A. 1986.

[23] O. Bernard, "Segmentation in echocardiographic imaging using parametric level set model driving by the statistics of the radiofrequency signal," Diss. Ph. D. thesis INSA-Lyon, pp. p16-20, 2006.