Intelligent Diagnostic System for Nuclei Structure Classification of Thyroid Cancerous and Non-Cancerous Tissues

Jamil Ahmed Chandio, M. Abdul Rehman Soomrani
Department of Computer Science
Sukkur Institute of Business Administration, Sukkur IBA
Sukkur, Pakistan

***Abstract*—Recently, image mining has opened new bottlenecks in the field of biomedical discoveries and machine leaning techniques have brought significant revolution in medical diagnosis. Especially, classification problem of human cancerous tissues would assume to be one of the really challenging problems since it requires very high optimized algorithms to select the appropriate features from histopathological images of well-differentiated thyroid cancers. For instance prediction of initial changes in neoplasm such as hidden patterns of nuclei overlapping sequences, variations in nuclei structures, distortion in chromatin distributions and identification of other micro-architectural behaviors would provide more meticulous assistance to doctors in early diagnosis of cancer. In-order to mitigate all above stated problems this paper proposes a novel methodology so called "Intelligent Diagnostic System for Nuclei Structural Classification of Thyroid Cancerous and Non-Cancerous Tissues" which classifies nuclei structures and cancerous behaviors from medical images by using proposed algorithm Auto_Tissue_Analysis. Overall methodology of approach is comprised of four layers. In first layer noise reduction techniques are used. In second layer feature selection techniques are used. In third layer decision model is constructed by using random forest (tree based) algorithm. Finally result visualization and performance evaluation is done by using confusion matrix, precision and recall measures. The overall classification accuracy is measured about 74% with 10-k fold cross validation.**

***Keywords*—Machine learning; decision support system; clustering; classification; cancer cells**

I. Introduction

Recently Image mining has become one of the well-established research area(s) of (ML) machine learning and (AI) artificial intelligence based techniques are vastly used in healthcare industry to facilitate doctors during the diagnostic and prognostic process of various kinds of malignant diseases such as breast cancer, lung cancer, thyroid cancer and so on. This paper addresses the classification problem of well-differentiated thyroid cancerous and non-cancerous cells of human tissues, which are carefully and systematically acquired from DICOM (Digital Communication in Medicine) images (Fig. 1). As per general observation, the misdiagnosis is one of the leading causes for rapid proliferation of cancer incidences among the world population and various related research approaches [1], [2] and [4] have been seen to reduce the chances of miss-diagnosis. All above stated approaches are providing very nice services to solve the classification problem of thyroid cancer. Since the prediction of initial changes in neoplasm, such as hidden patterns of nuclei overlapping sequences, variations in nuclei structures, distortion in chromatin distribution of human cells and other micro- architectural behaviours would provide more meticulous assistance to doctors in early diagnosis cancer. By considering above stated problems, this paper proposes a data preparation algorithm so called Auto_Tissue_Analysis which has to perform three task (1) NSA (Nuclei Seed Approximation) because gradient base techniques generate multiple seeds due to existence of high intensities at multiple places between single object. (2) TSA (Tissue Structure Approximation) function that helps to find out the minimum mean distances between the nuclei seed points on a distance matrix where particular arrangement of nuclei specifies a particular point based sequences to declare a thyroid cancer type. For example, anaplast cancers [Fig. 1(d)] are most aggressive cancers and consists upon the defused set of nuclei having varying distances at nuclei structural level and papillary carcinoma nuclei structures are likely to be found as finger like shapes, but it become more confused [Fig. 1(b)] when these features are presented with thyroid insular carcinoma. Every cancer is treated with different therapies, since it is very essential to know the cancer type for proper prognosis of cancer and restoration of human health. (3) NFST (Nuclei Feature Selection Tray) function avoids manual object cropping and it helps to selects the appropriate behaviours from every DICOM image. In-order to mitigate all above stated issues this paper offers a CAD (Computer Added Diagnostic) system so called "Intelligent Diagnostic System for Nuclei Structural Classification of Thyroid Cancerous and Non-Cancerous Tissues" to provide assistance to doctors by recommending a second systematic opinion. The system classifies cancerous and non-cancerous nuclei by considering a system generated decision variable used as sub-class label attribute in every observation. Since the system generated recommendations provides more precise assistance to doctors to understand the hidden behaviours of cytological material which may enable the doctors to address the all types of thyroid cancer as stated above.

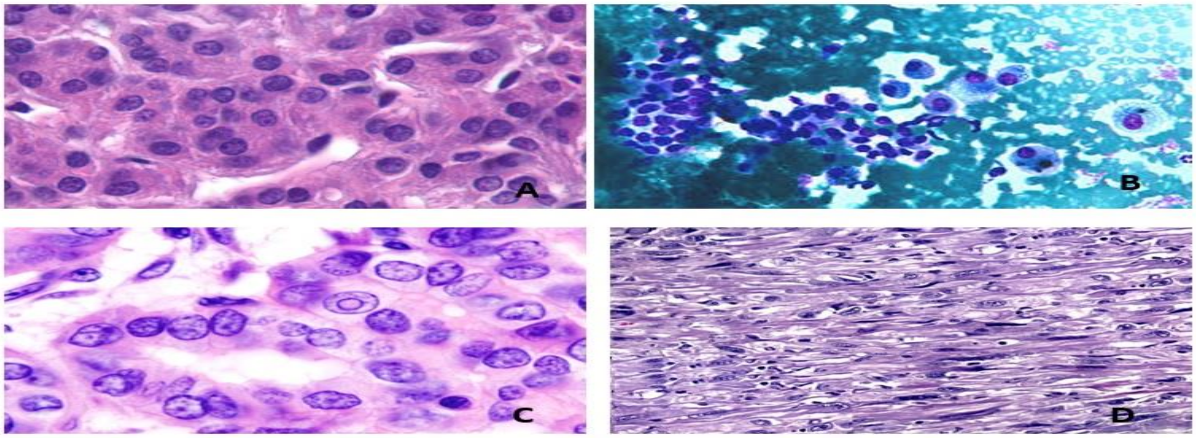


Fig. 1. Complex nuclei structures (a) Follicular (b) mixed papillary (c) papillary (d) anaplast cancers.

The methodology of proposed system comprises upon four layers and each layer is responsible to interact with each other. First two layers are responsible for data preparation. In first layer; noise reduction and classification of nuclei sequences are performed, in second layer auto-feature selection and auto-feature extraction are done, in third layer classification model have been constructed by using random forest tree based algorithm and in fourth layer the classification accuracy have been shown.

This paper is organized in several sections. Section one is used to describe introduction of this paper, related works are presented in section two, methodology is defined in section three, results are presented in section four and discussion &conclusion is discussed in section five whereas future work is presented in section six.

II. Literature Review

Classification of histopathological DICOM (Digital Communication in Medicine) images is of one the active research area(s) of ML (Machine Learning) and many approaches have been proposed to solve the classification problem of malignant diseases. This paper proposes an approach which is considered as productive modelling for thyroid disease classification in the domain of machine learning. Some of the related works have been presented as bellow.

A Comparison of different three ML (Machine Learning) algorithm [1] such like artificial neural network, decision tree and logistic regression were proposed to classify follicular thyroid cancer. This paper proposes an algorithm [Auto_Tissue_Analysis] to decide about the sub- class label attribute and used as decision variable and it is capable to identify all the types of thyroid cancer nuclei groves / structures, since DICOM structure identification at micro-architectural level would provide more refine results. Random forest tree base machine learning technique is used to classify cancerous and non-cancerous nuclei systematically with support of fully automated phenomena. A system [2] was proposed to classify cancerous thyroid disease tissues. Convolutional neural network based machine learning technique was used. In every DICOM image nuclei is key building block component and we approximate overlapped nuclei octo-edge distance value analysis compared with the equal to the size of neighbour set of nuclei. In most of approaches gradient base techniques have been used in segmentation stages, such kind of process would generate more than one centroids for each nuclei due to high intensity levels of pixels in nuclei regions. Proposed pre-processing algorithm selects the coordinates by considering size of each nuclei as per minimal set of edges between the set of neighbouring nuclei, if more than one edges are detected because of overlapping, regions are broken by measuring the historical data existing in the same DICOM image.

A comparison [3] of three machine learning techniques was proposed to classify the thyroid papillary cancer such as K-NN, Naive Bayesian and VPRS-CMR. In segmentation phase Otsu method is used and global level features were acquired. Micro-architectural level of nuclei detection would enhance the results and it provides more assistance to construct an unsupervised classification model for structure analysis.

A system [4] for thyroid cancer classification was presented and decision model was constructed by using support vector machine. As per their shown results malignant lesions lies between 0.97-80, P<.0001. In proposed pre-processing algorithm, the size of every nuclei is quantified with central location and compared with the nearer location by using unsupervised classification of nuclei structural states kwon as decision variable.

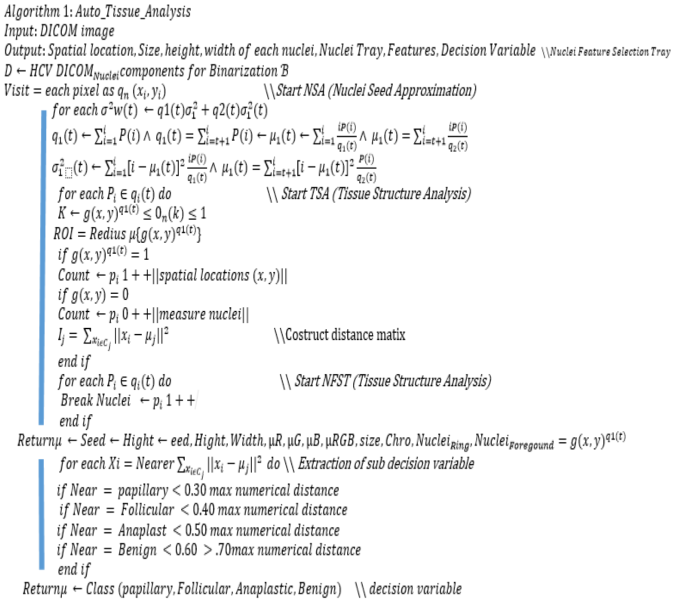
Using ultrasound images thyroid disease classification system [5] was proposed. Local binary patterns (LBP) base features were acquired through ROI (region of interest) feature selection was performed. There are certain limitations are associated with pixel base manual ROIs since the user may lose the important information of the content due human handling movements.

A system [6] was proposed for thyroid cancer and Otsu threshold was used to segment the nuclei from DICOM images. Using proposed method nuclei objects were detected by considering the nuclei rings and decision model is constructed for identification of set of nuclei grooves and use different levels of feature with the assistance of colour movements in DICOM image.

In an approach [7] six support vector machines were used to acquire ROI based features for thyroid disease using ultrasound images. All above stated approaches are providing very nice services but most of them consider thyroid cancer classification problem [8]-[13] at abstract levels, since thyroid disease classification may be more efficiently solved by considering the nuclei seed analysis, tissue structure analysis and nuclei feature selection tray, because structure identification needs special technique to understand nuclear grooves and arrangement of nuclei sequences, overlapping and other behaviours have significant importance in cancerous cell diagnosis. In literature thyroid follicular, medullary, papillary cancers have been classified. This paper proposes a novel technique which offers generalize algorithms for diagnosis of all thyroid cancers by proposing [Auto_Tissue_Analysis] algorithm which identifies all types of nuclei structures and other behaviours as stated above. Thus it is necessary to acquire the unsupervised classification by considering the nuclei regions / manual cropping with automated procedure, because manual cropping may select extra or less features from the regions DICOM [8]-[10] image. Secondly, follicular [14], [15], medullary, papillary sub cancer types have been classified in literature, whereas our proposed approach also classifies anaplastic and benign cancerous cells and tissues.

III. Methodology

The proposed approach is basically a predictive modelling in machine learning and deals with the classification problem of the histopathological cancerous cells and tissues that are taken from the DICOM images of FNAB [17], [18], and [19]. Using proposed methodology; thyroid papillary, follicular, anaplast and benign cancerous cells have been quantified. Due to the complexities of different nuclei staining materials (such as H, E, Biomarkers and so on) that are being used by cytologists to label the nuclei and to observe the heterogynous behaviours of nuclei such as different shapes, sizes, colour schemes, structures because micro architectural structures are observed always with dissimilar patterns. The methodology of proposed system (Fig. 2) is divided into four interconnected layers where each layer is assigned several inputs to interact with each other. In first layer; edges of nuclei are detected with support of noise reduction techniques and regions are formed, in second layer unsupervised classification of nuclei structures is performed by using seed analysis and auto-feature selection is done by NFST (Nuclei Feature Selection Tray). A decision variable S is derived from the nuclei sequences by using our proposed algorithm so called Auto_Tissue_Analysis [Algorithm 1]. In third layer classification model is constructed by using random forest tree based algorithm and in fourth layer the classification accuracy have been measured. Proposed pre-processing algorithm [Auto_Tissue_Analysis] is fully automated algorithm and it classifies the cancerous and non-cancerous nuclei derived from mimic, mix and confused pattern of DICOM images of FNAB. since these micro-architectural components produces lots of confusions related to nuclei structures, because nuclei sequences have significant importance to declare specific anatomical class of thyroid cancers such as well-differentiated, un-differentiated, poorly differentiated and benign as a global ontology of cancers.



```
Algorithm 1: Auto_Tissue_Analysis
Input: DICOM image
Output: Spatial location, Size, height, width of each nuclei, Nuclei Tray, Features, Decision Variable \\Nuclei Feature Selection Tray
$D \leftarrow$ HCV $DICOM_{Nuclei}$ components for Binarization $\mathcal{B}$
Visit = each pixel as $q_n (x_i, y_i)$          \\Start NSA (Nuclei Seed Approximation)
    for each $\sigma^2 w(t) \leftarrow q1(t)\sigma_1^2 + q2(t)\sigma_1^2(t)$
    $q_1(t) \leftarrow \sum_{i=1}^{i} P(i) \wedge q_1(t) = \sum_{i=t+1}^{i} P(i) \leftarrow \mu_1(t) \leftarrow \sum_{i=1}^{i} \frac{iP(i)}{q_1(t)} \wedge \mu_1(t) = \sum_{i=t+1}^{i} \frac{iP(i)}{q_2(t)}$
    $\sigma_{1}^2(t) \leftarrow \sum_{i=1}^{i} [i - \mu_1(t)]^2 \frac{iP(i)}{q_1(t)} \wedge \mu_1(t) = \sum_{i=t+1}^{i} [i - \mu_1(t)]^2 \frac{P(i)}{q_2(t)}$
     for each $P_i \in q_i(t)$ do          \\ Start TSA (Tissue Structure Analysis)
     $K \leftarrow g(x, y)^{q1(t)} \leq 0_n(k) \leq 1$
    ROI = Redius $\mu\{g(x, y)^{q1(t)}\}$
     if $g(x, y)^{q1(t)} = 1$
     Count $\leftarrow p_i\ 1 + +||$spatial locations $(x, y)||$
     if $g(x, y) = 0$
     Count $\leftarrow p_i\ 0 + +||$measure nuclei$||$
     $I_j = \sum_{x_{i \epsilon C_j}} ||x_i - \mu_j||^2$          \\Costruct distance matix
     end if
     for each $P_i \in q_i(t)$ do          \\ Start NFST (Tissue Structure Analysis)
      Break Nuclei $\leftarrow p_i\ 1 + +$
     end if
  Return$\mu$ $\leftarrow$ Seed $\leftarrow$ Hight $\leftarrow$ eed, Hight, Width, µR, µG, µB, µRGB, size, Chro, $Nuclei_{Ring}$, $Nuclei_{Foregound} = g(x, y)^{q1(t)}$
     for each Xi = Nearer $\sum_{x_{i \epsilon C_j}} ||x_i - \mu_j||^2$ do \\ Extraction of sub decision variable
    if Near = papillary < 0.30 max numerical distance
     if Near = Follicular < 0.40 max numerical distance
    if Near = Anaplast < 0.50 max numerical distance
    if Near = Benign < 0.60 > .70max numerical distance
    end if
  Return$\mu$ $\leftarrow$ Class (papillary, Follicular, Anaplastic, Benign)   \\ decision variable
```

Layer 1: Noise Reduction & formation of regions

For example well differentiated cancer class consists upon the follicular, papillary and other types of cancers. Papillary class can be further subdivided into papillary tall cell carcinoma, glass ground appearance of cells [Fig. 1(c)] and so on.

A. Noise Reduction

Set of Nuclei groves contain sufficient information to classify despite of having heterogynous behaviours and shapes. In literature lots of noise reduction techniques have been proposed. Some of them use image segmentation techniques (such as Watershed, Graph-cut, Super pixels and so on) meanwhile our technique is motivated from the graph-cut segmentation because colour information found in DICOM images is really important for doctors during the diagnostic process traditionally. For example in [Fig. 1(a-d)] lots of noise is persisting and nuclei sequences are heterogynous, especially; [Fig. 1(b)] is very difficult to interpret traditionally either it is belonging to papillary cancer class or other classes. Evidently; on applying of graph-cut segmentation algorithm, we observed that most significant object were detected as fore-ground and the remaining objects were removed as back-ground (Fig. 3). This phenomena is not only supporting to reduce the noise but also efficient to saved time and space complexities.

B. Grey Scale Conversion

For further image analysis [20], [21], DICOM images were converted into binary form, we use adaptive threshold algorithm as shown in [Fig. 4(a)] to covert the image into binary format. Since the Otsu method has several advantages in binary representation of medical image [22] because clustering base unsupervised classification of each pixel p= $\{x, y_1 \dots\dots x, y_n\}$ have to be represented as an object known The method follows the weighted pixels corresponding to either one or zero class as presented in (1) and (2).

$$\omega_0(t) = \sum_{i=0}^{t-1} p(i) \quad (1)$$

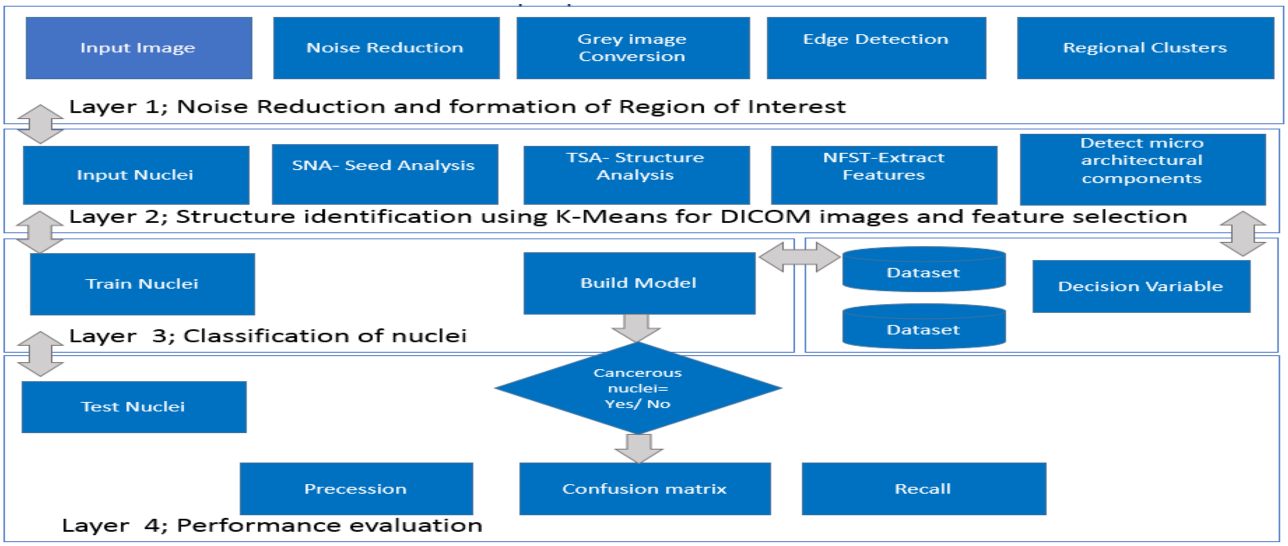

Input Image
Noise Reduction
Grey image Conversion
Edge Detection
Regional Clusters
Layer 1; Noise Reduction and formation of Region of Interest
Input Nuclei
SNA- Seed Analysis
TSA- Structure Analysis
NFST-Extract Features
Detect micro architectural components
Layer 2; Structure identification using K-Means for DICOM images and feature selection
Train Nuclei
Build Model
Dataset
Dataset
Decision Variable
Layer 3; Classification of nuclei
Test Nuclei
Cancerous nuclei= Yes/ No
Precession
Confusion matrix
Recall
Layer 4; Performance evaluation

Fig. 2. Intelligent diagnostic system for nuclei structural classification of thyroid cancerous and non-cancerous tissues workflow.

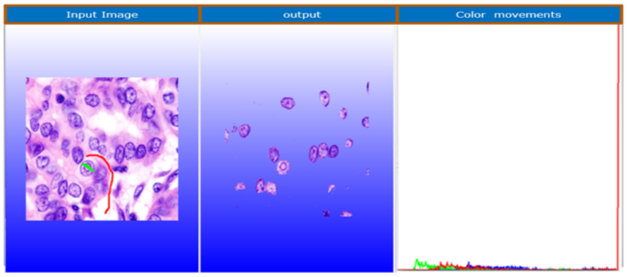


Fig. 3. Noise Reduction using graph-cut segmentation

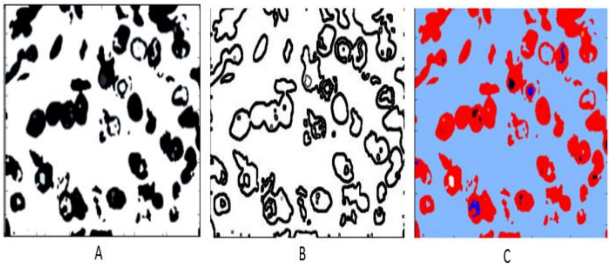


Fig. 4. (a) image banalization using adaptive threshold (b) Canny edge detection (c) Nuclei clusters

$$\omega_1(t) = \sum_{i=t}^{L-1} p(i) \tag{2}$$

The covariance of these corresponding pixel classes would be represented as per (3), where each label has significant impact upon the generated pixels.

$$(t) = \omega 0(t)\sigma_0^2(t) + \omega 1(t)\sigma_1^2(t) \tag{3}$$

So two regions may be constructed for classification of pixels and some regions would be appeared with the pixels having the value 1 whereas others may be represented with 0 quantity as shown in (4), (5) and (6)] where mean values are potted to visualize the wμ0+ wμ1 = μT and w0+w1=1

$$\mu_0(t) = \sum_{i=0}^{t-1} ip(i)/\omega_0 \tag{4}$$
$$\mu_1(t) = \sum_{i=0}^{L-1} ip(i)/\omega_1 \tag{5}$$

$$\mu_T = \sum_{i=0}^{L-1} ip(i) \tag{6}$$

Canny edge detection algorithm was used to find out the edges of all nuclei, since geometrical and morphological features have great importance in machine learning to let the classifier learn about the possible shapes of different objects as shown in [Fig. 4(b)] where each object is considered as region. Since the pixel base orientation is not enough to represent a medical component, therefore regions [Fig. 4(b)] are useful tool to apply further operations such as extracted centroid may be used for further operations as per Fig. 6.

Layer 2: Structure identification and feature selection

Appropriate Nuclei seed approximation is required to find out the spatial coordinates such as x and y but at pixel level the energy of the pixel may be found from 0 to 255. If we fix some of the ranges of DICOM images intensities we may find more the one point for each class, since our algorithm considers every nuclei as a region, detected by using canny edge detection algorithm in following way.

1) Nuclei Seed Approximation (NSA)

Proposed nuclei separation technique is inspired from the LBP (Linear Binary Patterns) at regional levels because there are several limitations associated with gradient base techniques because due to same intensity threshold may detect multiple cancroids for same nuclei such as DNA feature measuring algorithm. Let's consider a ring is created around the nuclei using canny edge detection so called region of interest and we have to acquire the radius of all regions [Fig. 5].

Above expression is plotted with eight points on the detected edges [23] region ROI= $\{ a_1, a_2, \dots \dots \dots \dots .. a_7 a_8\}$ where radius of each nuclei is the result of cross correlation and division of opposite points as per following equations [(7), (8) and (9)].

$$\sum_{a=8}^{a=0} g(x,y) \tag{7}$$

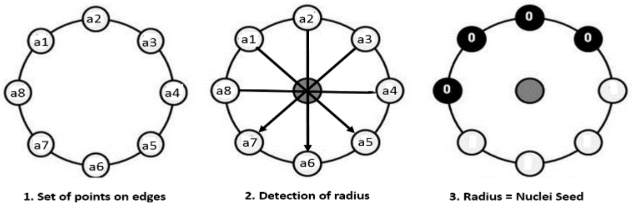


Fig. 5. Seeds extraction process.

$$ROI = Redius\ \{a_1, a_2, a_3, a_4, a_5, a_6, a_7 a_8\} \qquad (8)$$

$$\text{Where Radius}=\left.\begin{matrix}\mu\{a_1+a_5\}\\ \mu\{a_2+a_6\}\\ \mu\{a_3+a_7\}\\ \mu\{a_4+a_8\}\end{matrix}\right/ 8 \qquad (9)$$

The father analysis on these seeds is performed in (Section 3.4) to visualize the probabilities of different cancer sequences.

2) Tissue Structure Approximation (TSA)

DICOM distance matrix (Fig. 6) is very useful tool to measure the pixel coordinates where individual distances of each nuclei are considered by using dot product of spatial location x and y. By summarizing all the distances of nuclei and constructing a decision model to find out the nuclei sequences designated as S would assist to find the optimal sequences of nuclei. Since the optimal values of f(k) may be quantified which lies between $\mathrm{M} = \{\mathrm{m}_1 \dots\dots\dots..\mathrm{m}_\mathrm{n}\}$ distances of pixels of each object by considering edges of each nuclei. The ideal central location of each nuclei in DICOM dataset D is lying between the $\mathrm{x}^\mathrm{i}, \mathrm{y}^\mathrm{i}$ locations of$\mathrm{g(x)}$, which may be detected by considering the mean pixel values on gradient of every geometrical shape of nuclei. After acquiring the K values of each location the same may be separated by considering every nuclei as a partition. For example DICOM datasets with m^n number of nuclei could be considered in same group lies between locations 1….n but the central point is more significant to separate piece of medical image in an automated process. We may analyse k values of n objects designated as spatial location as decision variable and that may consider the clustering points by the adjustment of k clusters. Since the DICOM nuclei structures is an study of nuclei connected components could be used as decision variable S and central points would be assumed as following notation eq(12), where sk deviation of distances could achieved as global impact of clusters.

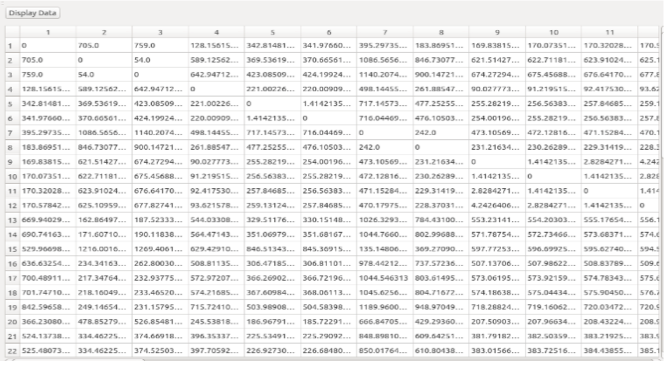
$$I_j = \sum_{x_{i\epsilon C_j}} ||x_i - \mu_j||^2 \qquad (10)$$

The global impact of all clusters' distortions is given by the quantity

$$S_k = \sum_{j=1}^{K} I_j \qquad (11)$$

The function of f (k) becomes optimal for DICOM structure as per following notation eq. (12).

$$f(k) = \begin{cases} 1 \qquad ifK =< .30 \\ \frac{S_K}{a_K S_{K-1}} ifS_{K-1}\ \neq 0, \forall K < .40 \\ 1 \qquad ifS_{K-1}\ = 0, \forall\ K < .50 \\ ifS_{K-1}\ = 0, \forall\ K < .60 \\ ifS_{K-1}\ = 0, \forall\ K < .50 > .70 \end{cases} \qquad (12)$$



	1	2	3	4	5	6	7	8	9	10	11	
1	0	705.0	759.0	128.15615...	342.81481...	341.97660...	395.29735...	183.86951...	169.83815...	170.07351...	170.32028...	170.5
2	705.0	0	54.0	589.12562...	369.53619...	370.66561...	1086.5656...	846.73077...	621.51427...	622.71181...	623.91024...	625.1
3	759.0	54.0	0	642.94712...	423.08509...	424.19924...	1140.2074...	900.14721...	674.27294...	675.45688...	676.64170...	677.8
4	128.15615...	589.12562...	642.94712...	0	221.00226...	220.00909...	498.14455...	261.88547...	90.027773...	91.219515...	92.417530...	93.62
5	342.81481...	369.53619...	423.08509...	221.00226...	0	1.4142135...	717.14573...	477.25255...	255.28219...	256.56383...	257.84685...	259.1
6	341.97660...	370.66561...	424.19924...	220.00909...	1.4142135...	0	716.04469...	476.10503...	254.00196...	255.28219...	256.56383...	257.8
7	395.29735...	1086.5656...	1140.2074...	498.14455...	717.14573...	716.04469...	0	242.0	473.10569...	472.12816...	471.15284...	470.1
8	183.86951...	846.73077...	900.14721...	261.88547...	477.25255...	476.10503...	242.0	0	231.21634...	230.26289...	229.31419...	228.3
9	169.83815...	621.51427...	674.27294...	90.027773...	255.28219...	254.00196...	473.10569...	231.21634...	0	1.4142135...	2.8284271...	4.242
10	170.07351...	622.71181...	675.45688...	91.219515...	256.56383...	255.28219...	472.12816...	230.26289...	1.4142135...	0	1.4142135...	2.828
11	170.32028...	623.91024...	676.64170...	92.417530...	257.84685...	256.56383...	471.15284...	229.31419...	2.8284271...	1.4142135...	0	1.414
12	170.57842...	625.10959...	677.82741...	93.621578...	259.13124...	257.84685...	470.17975...	228.37031...	4.2426406...	2.8284271...	1.4142135...	0
13	669.94029...	162.86497...	187.52333...	544.03308...	329.51176...	330.15148...	1026.3293...	784.43100...	553.23141...	554.20303...	555.17654...	556.1
14	690.74163...	171.60710...	190.11838...	564.47143...	351.06979...	351.68167...	1044.7660...	802.99688...	571.78754...	572.73466...	573.68371...	574.6
15	529.96698...	1216.0016...	1269.4061...	629.42910...	846.51343...	845.36915...	135.14806...	369.27090...	597.77253...	596.69925...	595.62740...	594.5
16	636.63254...	234.34163...	262.80030...	508.81135...	306.47185...	306.81101...	978.44212...	737.57236...	507.13706...	507.98622...	508.83789...	509.6
17	700.48911...	217.34764...	232.93775...	572.97207...	366.26902...	366.72196...	1044.546313	803.61495...	573.06195...	573.92159...	574.78343...	575.6
18	701.74710...	218.16049...	233.46520...	574.21685...	367.60984...	368.06113...	1045.6256...	804.71672...	574.18638...	575.04434...	575.90450...	576.7
19	842.59658...	249.14654...	231.15795...	715.72410...	503.98908...	504.58398...	1189.9600...	948.97049...	718.28824...	719.16062...	720.03472...	720.9
20	366.23080...	478.85279...	526.85481...	245.53818...	186.96791...	185.72291...	666.84705...	429.29360...	207.50903...	207.96634...	208.43224...	208.9
21	524.13738...	334.46225...	374.66918...	396.35337...	225.53491...	225.29092...	848.89810...	609.64251...	381.79182...	382.50359...	383.21925...	383.9
22	525.48073...	334.46225...	374.52503...	397.70592...	226.92730...	226.68480...	850.01764...	610.80438...	383.01566...	383.72516...	384.43855...	385.1

Fig. 6. Distance matrix by using x, y location of nuclei.

In the above equation $\mathrm{N_d}$ are the number of attributes in decision matrix where $\mathrm{a_k}$ is the weighted distance points such as papillary structures with mean distances of <0.30, follicular with <0.40, anaplastic cancers <0.40 and benign with <0.50 >0.70. Above 0.70 values were eliminated due to no significant set of object could be considered for partition as a DICOM structure class. In results section [Fig. 8] is shown to represent the distance matrix and [Fig. 9] is visualized to show the optimum percentage of split for each DICOM decision variable.

3) Nuclei Feature Selection Tray (NFST) algorithm

Manual selection of ROI (region of interest) and cropping image into sub-images as feature selection is labours work and there are number of limitations are associated with manual cropping / ROIs because user may select reduce set of features or human handling may not crop properly. Thus there are fare chances for the loss of useful information. NFST (Nuclei Feature Selection Tray) algorithm avoids manual cropping and ROI (region of interest) of nuclei selection, thus lots of laborious work of feature selection is saved in terms of time and complexity. Besides we have presented results of our approach from the perspective of feature selection in Fig. 6 so called Nuclei Feature Selection Tray (NFST) algorithm. Let's consider every DICOM image is consisting upon the several number of nuclei where $\mathrm{DICOM} = \{D^1, D^2, \dots\dots\dots\dots..\mathrm{D^n}\}$ on feature vector space. The spatial coordinates at this stage are really challenging to separate because overlapping of nuclei does not allow finding individual nuclei from a set of nuclei. At this stage we recorded sizes of each object and approximated the overlapped nuclei size by constructing spatial boundaries between the nuclei to count them as closed individual object. After acquiring the individual set of nuclei as objects we recorded colour movements with RGB mean values. Due to in-availability of such datasets for DICOM images of FNAB, we prepared our own datasets for training and testing purposes. The datasets comprise upon RGB movements [24]-[26] of nuclei, sizes, shapes and sub-class (SC) decision derived from (Section 3.3), which provides second opinion to doctors that what is the structure of particular DICOM image.

Therefore central point p of nuclei in a DICOM image is unique location containing x, y may be represented by using distance matrix (Fig. 6) where distance of each nuclei may be measured by using the [(13) and (14)] where height (h) and width (w) may be cropped with by considering the central point known as seed.

$$h,w = \sum_{a(y)=n}^{a(x)=1} + \mu h/w \qquad (13)$$

$$seed(\mu x, y) = \{Sum(a(x,y)_{5......8}^{1......4}) \qquad (14)$$

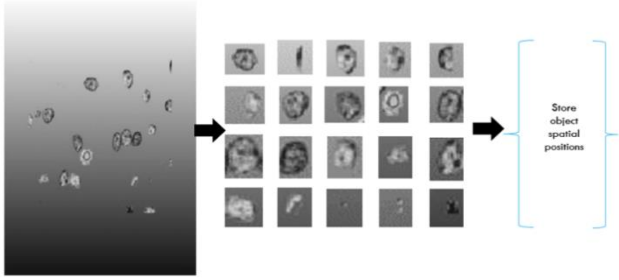


Fig. 7. Auto Nuclei Feature Selection Tray (NFST) for diagnosis.

Layer 3: Classification of nuclei

A wide variety of machine learning algorithms is available since the feature engineering needs special contributions to select the related observations from the medical images. This research uses random forest machine learning algorithm for classification of cancerous disease because tree base algorithms have been found with higher rank relationships and deep decision advantages.

Let's suppose medical image dataset D contains a number of significant class instances designated as $D = \{(\mathrm{X_i}, \mathrm{Y_i})\}\ _{\mathrm{i=1}}^{\mathrm{n}}$ to construct a decision model by using the aggregated measured variables fitted values during the training phase by considering the $\mathrm{J_{th}}$ feature in terms of deep decision trees based upon the aggregated permutation functions since the out of bag errors have been used to represent overall trees.

$$\mathrm{D} = \{(\mathrm{X_i}, \mathrm{Y_i})\}\ _{\mathrm{i=1}}^{\mathrm{n}} \qquad (15)$$

$$\hat{y} = \sum_{i=1}^{n} W(x_i, \acute{x}) y_i \qquad (16)$$

In (16) and (17) $\hat{y}$ variable is used to represent the cancerous and non-cancerous disease where x is the collection of formulized weighted and associated variables designated as W. the rest of function have assigned several number of class label attributes to classify by considering the highest level of tree base ranks.

$$\hat{y} = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} W_j(x_i, \acute{x}) y_i = \sum_{i=1}^{n} \left(\frac{1}{m} \sum_{j=1}^{m} W_j(x_i, \acute{x}) \right) y_i \qquad (17)$$

Layer 4: Performance Evaluation

In performance evaluation we use confusion matrix (Table 1). The classifier was trained by parsing total number of 602 observations, we selected 60% for training and 40% for testing purposes, where (Table 2) 218 instances were classified as true positive out of 301 and 813 instances were classifies as true positive out of 857 instances. The precision (18) and recall (19) was recorded for cancerous classes 83.2%, 73.2% and 91.00%, 94.90% was measured for non-cancerous classes.

$$Precision = \frac{NumberofTrePositives}{NumberofTruePositives + FalsePositives} \qquad (18)$$

$$Recall = Sensitivity = \frac{TruePositive}{FalsePositive} \qquad (19)$$

$$Sepecify = \frac{TrueNegetive}{TruePositive + FalsePositive} \qquad (20)$$

IV. Results

Fig. 7 is plotted to present the results of fully automated object separation process where each nuclei is fully cropped without human interaction. In Fig. 8, Column 4 is describing the presentation of distance matrix quantities in a leaner set of nuclei sequences, where each nearer colour have been plotted in blue spectrum to form the identity matrix whereas deviated values on an identity matrix are presented with different colour scales ranging from 120-540 colour schemes where most blue colour is object value. Pre-process method have been illustrated in Fig. 8 where column 1 is showing input image and column 2 demonstrates TSA estimation where each most probable centre of normal and abnormal nuclei is detected. Further pre-processing of initial guess about the spatial locations have been depicted in column 3 where group of nuclei / nuclei grooves have been captured with the assistance of NSA algorithm. Column 4 presents the results of extracted nuclei for measuring the nuclei grooves and sequences to determine the class label of particular image. Four type of object threshold have been observed deeply by the proposed algorithm to determine the structure of nuclei sequences such as Papillary class <0.30 with k-distances and finger like structure, Follicular class <0.40 with k-distances having random distances, Anaplast class <0.50 with k-distances with random vector spaces and values of <0.60 having k-distances shows un-known class in Dataset D.

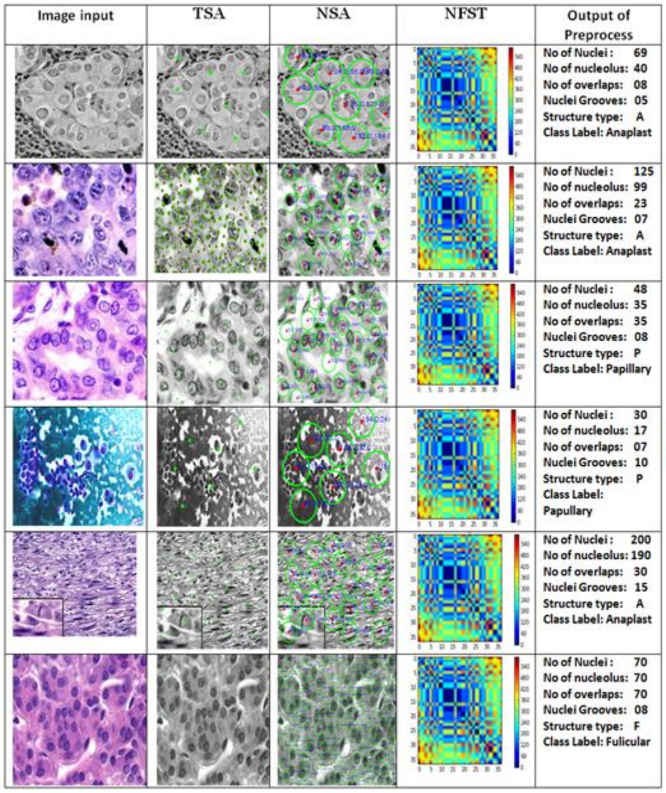


Fig. 8. Pre-process results, Column 1 shows input image, column 2 represents TSA estimation, Column 3 demonstrates group of nuclei / nuclei grooves, column 4 shows distance matrix estimation developed to determine the structure of nuclei and column 5 is represents the summary of finding with class label attribute.

TABLE. I. CONFUSION MATRIX

	Papillary	Follicular	Anaplast	Benign
Papillary	158	20	25	16
Follicular	27	145	30	17
Anaplast	25	15	258	26
Benign	24	23	24	215

TABLE. II. CONFUSION MATRIX

	CANCEROUS NUCLEI =YES	CANCEROUS NUCLEI = NO
CANCEROUS NUCLEI= YES	217	70
CANCEROUS NUCLEI = NO	202	559
	OVERALL ESTIMATED ACCURACY = 74%	

TABLE. III. OVERALL PERFORMANCE OF PROPOSED METHODOLOGY

	Raw DICOM images	No of Extracted nuclei	No of classified Nuclei	No of miss-classified Nuclei	Precession	Recall
Papillary	20	219	158	61	67.52%	72.14%
Follicular	20	219	145	74	71.42%	66.21%
Anaplast	20	324	258	66	76.55%	79.63%
Benign	20	286	215	71	87.46%	75.15%
			Overall estimated accuracy = 74%			

TABLE. IV. COMPARISON OF OUR SYSTEM WITH LITERATURE

Approaches	Image Type	Cancer Type	Technique	Accuracy
1	Ultrasound Image	Follicular	SVM	97.50 %
2	FNAC Images	Follicular	SVM	91.00 %
3	FNAC Images	Medullary	Decision Tree	98.00%
4	FNAB Images	Papillary	SVM	93.30%
5	FNAC Images	Papillary	SVM	95.00%
Our Proposed approach	FNAB Images	All Thyroid Cancers	Random forest	74.00%

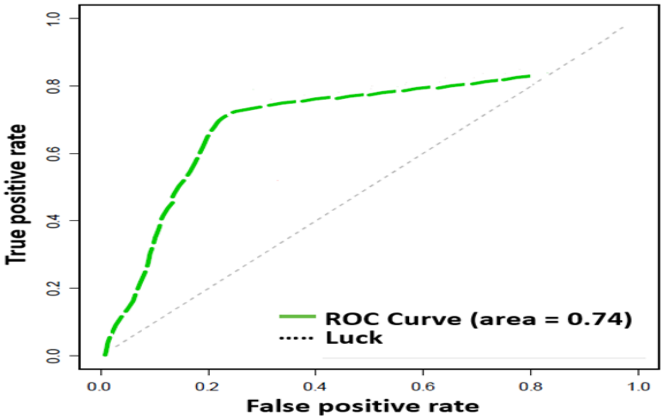


Fig. 9. Estimated ROC Curve for proposed system.

Confusion matrix (Table 1) show that 158 observations have been classified for papillary class label attribute and 145 instance have been classified for follicular class whereas 258 instance have been classified for anaplast class label attribute and 215 observation have been classified for benign class label attribute. The malignant and non-malignant observations have been presented in (Table 2) where 217 number of nuclei observation have been classified as cancerous cells yes and 559 instances of cancerous cells No have been classified by the classifier. The precision measure for each class such as papillary, follicular, anaplast and benign were recorded respectively 67.52%, 71.42, 76.55% and 87.46% whereas recall measure was approximated, respectively 72.14%, 66.21%, 89.63% and 75.15%. ROC Curve have been shown in Fig. 9 for overall system performance and measured classification accuracy (Table 3) of our system is about 74% with 10-k fold cross validation. Table 4 is presented for comparison with literature which shows that our proposed approach classifies all the classes of well-differentiated thyroid cancers. Overall performance of the pre-process method have been described in Fig. 8 which show decision variable after examining the number of detected nuclei, number of nucleolus, number of overlaps, number of nuclei grooves / sequences determined by the algorithm and type of sequence as P for papillary, F for follicular, A for anaplast and B for benign class label attribute.

V. Discussion and Conclusion

Due to the heterogeneous and complex nature of micro-architectural components of histopathological DICOM (Digital Communication in Medicine) images, automated nuclei structure identification is one of the significant problems. Since follicular, medullary, papillary classification approaches are reported in literature and automated segmentation with classification of thyroid disease structure is yet not reported, thus it is direly needed to propose a system to detect the nuclei groves by considering micro-architectural component analysis as a decision output or sub-class label attribute (such as well differentiated, poorly differentiated and benign cancers) followed by constructing a classification model for thyroid cancer variants. This paper proposes an automated computer based decision support system as second opinion to doctors which may enrich the assistance during the diagnostic process of cancer. Finally reproducibility of results would assume to be one of the convinced advantages to save the precious time and finance of the patients. This paper proposes a novel methodology for nuclei structure identification which selects the most significant DICOM (Digital Communication in Medicine) image behaviours from thyroid papillary, follicular carcinoma and anaplast cancers by using our algorithm Auto_Tissue_Analysis which is combination of our three proposed jobs (1) NSA (Nuclei Seed Approximation), (2) TSA (Tissue Structure Approximation) and NFST (Nuclei Feature Selection Tray). Over all methodology of our approach is comprising upon four layers, In first layer; noise reduction is by grey scale segmentation, edges of nuclei are detected, regions are transformed and nuclei seeds are extracted with respect to edges of every nuclei, in second layer unsupervised classification of tissue structures are performed by using the seed analysis and auto-feature selection is done. A decision variable is derived from the nuclei sequences which generates sub-class label attribute for every cancerous and non-cancerous class. In third layer we construct the decision model by using random forest (tree based) algorithm to extract the decision variable dependencies. Finally result visualization and performance evaluation is conducted by using confusion matrix, precision and recall measures respectively. The overall classification accuracy is measured about 74% with 10-k fold cross validation.

VI. Future Work

There are several types of histopathological DICOM (Digital Communication in Medicine) images each have its own properties and behaviours. This research work addresses the classification problem of thyroid cancer. As a future work; the different real-world datasets will be trained and tested for various types of histopathological images of cancers such as classification of lung cancer, breast cancer, brain cancer, blood cancer, skin cancers and others. In skin cancer especially zero derma pigmentosa will be one of the important dimension because of heterogynous images may contain the same intensity of nuclei representation.

References

[1] Pourahmad, S., Azad, M., Paydar, S., & Reza, H. Prediction of malignancy in suspected thyroid tumour patients by three different methods of classification in data mining. In First International Conference on advanced information technologies and applications: pp.1-8, 2012.

[2] Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M. and Chang, May. Deep learning of feature representation with multiple instance learning for medical image analysis. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, pp. 1626-1630, 2014.

[3] Jothi, J.A.A. and Rajam, V.M.A. Effective segmentation and classification of thyroid histopathology images. Applied Soft Computing, 2016.

[4] Iakovidis, D. K., Keramidas, E. G., & Maroulis, D. Fusion of fuzzy statistical distributions for classification of thyroid ultrasound patterns. Journal of Artificial Intelligence in Medicine 50 (2010): pp.33–41, 2010.

[5] Ding, J., Cheng, H., Ning, C., Huang, J., & Zhang, Y. (2011). Quantitative measurement for thyroid cancer characterization based on elastography. Journal of Ultrasound in Medicine, 30(9): pp.1259-1266, 2011.

[6] Bell A.A, Kaftan, J.N, Schneider, T. E. Imaging and Image Processing for Early Cancer Diagnosis on Cytopathological Microscopy Images towards Fully Automatic AgNOR Detection, WrithArchin science journal, 2006.

[7] Chuan, Y. Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound image. Journal of Pattern Recognition 43 (2010): pp.3494–3506, 2010.

[8] Dimitris G., Panagiota S., Stavros T., Giannis K., Nikos D., George N., Dionisis C. Unsupervised segmentation of fine needle aspiration nuclei images of thyroid cancer using a support vector machine clustering methodology, 1st IC-SCCE Athens, pp. 8-10 September, 2004.

[9] Suzuki, H., Saita, S., Kubo, M., Kawata, Y., Niki, N., Nishitani, H., & Moriyama, N. An automated distinction of DICOM images for lung cancer CAD system. In SPIE Medical Imaging (). International Society for Optics and Photonics, pp. 72640Z-72640Z, 2009.

[10] Glotsos, D, Tsantis, S. Kybic J, Daskalakis, A. Pattern recognition based segmentation versus wavelet maxima chain edge representation for nuclei detectionin microscopy images of thyroid nodules. Euromedica medical center, Department of Medical Imaging, Athens, Greece 2013.

[11] Gopinath, B, Shanthi, N. Support Vector Machine Based Diagnostic System for Thyroid Cancer using Statistical Texture Features, Asian Pacific J Cancer Prev, 14 (1), pp. 97-102, 2013.

[12] Gopinath, B, Shanthi, N. Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images, Australas Phys EngSci Med pp.36:219–230, 2013.

[13] Stavros T. Improving diagnostic accuracy in the classification of thyroid cancer by combining quantitative information extracted from both ultrasound and cytological images. 1st IC-SCCE-Athens, 8-10 September, 2004.

[14] Smith, Russell B., et al. "Preoperative FDG-PET imaging to assess the malignant potential of follicular neoplasms of the thyroid." Otolaryngology-Head and Neck Surgery Vol 138(1), pp. 101-106, 2008

[15] Gupta, Nidhi, et al. "Development and Validation of a Pediatric Endocrine Knowledge Assessment Questionnaire: Impact of ac Pediatric Endocrine Knowledge Assessment Questionnaire Intervention Study." Journal of clinical research in pediatric endocrinology Vol-8 (4), pp. 411-416, 2016.

[16] Han J., Kamber M., and Pei P. "Data Mining: Concepts and Techniques". 3rd (ed.) the Morgan Kaufmann Series in Data Management Systems, 2011.

[17] Rafeal, C. G., Richard, E. W. "Digital Image Processing", 3rd (ed). Prentice Hall. 2007.

[18] Hussein, Mohamed, Amitabh Varshney, and Larry Davis. "On implementing graph cuts on cuda." First Workshop on General Purpose Processing on Graphics Processing Units. Vol. 2007. 2007.

[19] Boykov, Y., &Funka-Lea, G. Graph cuts and efficient ND image segmentation. International journal of computer vision, 70(2), pp.109-131, 2006.

[20] Malik, J., Belongie, S., Leung, T., & Shi, J. Contour and texture analysis for image segmentation. In Perceptual Organization for artificial vision systems. Springer US, pp. 139-172, 2000.

[21] Al-Kofahi, Y., Lassoued, W., Lee, W., &Roysam, B. Improved automatic detection and segmentation of cell nuclei in histopathology images. Biomedical Engineering, IEEE Transactions on, 57(4), pp. 841-852, 2010.

[22] Liu, X., Huo, Z., & Zhang, J. Automated segmentation of breast lesions in ultrasound images. In Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the IEEE pp. 7433-7435, 2006.

[23] Al-Kofahi, Y., Lassoued, W., Lee, W., &Roysam, B. Improved automatic detection and segmentation of cell nuclei in histopathology images. Biomedical Engineering, IEEE Transactions on, 57(4), pp.841-852, 2010.

[24] Xu, N., Ahuja, N., & Bansal, R. Object segmentation using graph cuts based active contours. Computer Vision and Image Understanding, 107(3), pp.210-224, 2007.

[25] Freedman, D., & Zhang, T. Interactive graph cut based segmentation with shape priors. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 755-762). IEEE, 2005.

[26] Moukaddam, H., Pollak, J. and Haims, A., 2009. MRI characteristics and classification of peripheral vascular malformations and tumors. Skeletal Radiology, 6(38), pp..535-547, 2009.