Staging of clear cell renal cell carcinoma using random forest and support vector machine

D Talaat¹, F Zada and R Kadry
Computer Department, Faculty of Engineering and Technology,
Arab Academy for Science and Technology, Abou Keer, Alexandria, Egypt

Abstract. Kidney cancer is one of the deadliest types of cancer affecting the human body. It’s regarded as the seventh most common type of cancer affecting men and the ninth affecting women. Early diagnosis of kidney cancer can improve the survival rates for many patients. Clear cell renal cell carcinoma (ccRCC) accounts for 90% of renal cancers. Although the exact cause of the kidney cancer is still unknown, early diagnosis can help patients get the proper treatment at the proper time. In this paper, a novel semi-automated model is proposed for early detection and staging of clear cell renal cell carcinoma. The proposed model consists of three phases: segmentation, feature extraction, and classification. The first phase is image segmentation phase where images were masked to segment the kidney lobes. Then the masked images were fed into watershed algorithm to extract tumor from the kidney. The second phase is feature extraction phase where gray level co-occurrence matrix (GLCM) method was integrated with normal statistical method to extract the feature vectors from the segmented images. The last phase is the classification phase where the resulted feature vectors were introduced to random forest (RF) and support vector machine (SVM) classifiers. Experiments have been carried out to validate the effectiveness of the proposed model using TCGA-KRRC dataset which contains 228 CT scans of ccRCC patients where 150 scans were used for learning and 78 for validation. The proposed model showed an outstanding improvement of 15.12% for accuracy from the previous work.

Keywords: Detection, classification, feature extraction, segmentation, ccRCC, watershed, GLCM, random forest, support vector machine.

1. Introduction
The kidney is one of the most vital organs in the human body. It plays the most important role in detoxification of body toxins. The average human body has two kidneys located in the lower abdomen, one on each side of the spin. They are bean-shaped and each one has the size of a fist. A healthy kidney filters around half a cup of blood every minute, removing wastes and extra water in the form of urine. This filtration process balances fluids all over the body, controlling blood pressure as well as assisting in the production of red blood cells. Kidneys are also responsible for removing acids produced by the cell to balance the percentages of water, salts, and minerals—such as sodium, calcium, phosphorus, and potassium—in the blood [1].

¹ Corresponding author: doha.talaat.12@gmail.com
One of the most dangerous diseases that can infect the kidney is kidney cancer (renal cancer). Renal cell carcinoma (RCC) is the most common type of kidney cancers affecting the adults. Though causes of renal cell carcinoma (RCC) are yet to be clearly discovered, there are several risk factors that increase the susceptibility of becoming an RCC patient. These factors include: first, age, as the risk of becoming an RCC patient increases with older age. Second, smoking, as smoking is an unhealthy habit that destroy the human body gradually. Third, obesity, as people who tend to be obese are more likely to be infected with RCC. Fourth, blood pressure, as high blood pressure or (hypertension) increases the risk of kidney cancer. Fifth, kidney failure treatment or long-term dialysis, as people who suffer from chronic kidney failure have a greater risk of developing kidney cancer. Sixth, certain inherited syndromes, such as – Hippel-Lindau disease, or Birt-Hogg-Dube syndrome, may increase the risk of kidney cancer. Seventh, genetic history as people who have a strong family history of RCC tend to have a greater risk of renal cancer. Eighth, exposure to certain chemicals such as – cadmium or specific herbicides increases the risk of kidney cancer. These factors may increase the risk of renal cancer but not necessarily prove that their presence would lead to RCC, as some people may have none of them yet still develop RCC [2], [3]. Medical experts discovered that kidney cancer begins with DNA mutations in some kidney cells. These mutations tend to divide and grow rapidly until they become tumors. These tumors can extend beyond the kidney, breaking off and spreading to distant parts of the body (metastasize) [3]. As a matter of fact, early detection of the tumor helps control the disease and thus increases survival rate. Renal cell carcinoma (RCC) shows about 90% of the renal malignant tumors, also it is among the top 10 of the most malignant tumors [4]. 2% to 3% of the adult cancers are regarded as RCC [5]. Renal cell carcinoma is shown as the seventh most common cancer in men and the ninth in women [6]. For the last 75 years, the occurrence of RCC increase by 2% every year. The death-rate of renal cancer accounts for 25% which is considered the highest rate compared to the rest of the urological cancers [7]. Renal cell carcinoma is subtyped into 75% clear cell renal cell carcinoma (ccRCC), 7% to 15% papillary renal cell carcinoma (pRCC), and 5% chromophobe renal cell carcinoma (chRCC) [8]. According to the subtypes percentages of the RCC, clear cell renal cell carcinoma represents the most common subtype. As RCC is insurgent and cannot be cured by using chemotherapy or radiation therapy, so the surgical removal of the tumor is the best choice of treatment. Computed tomography (CT) scan is the best way to detect RCC and its stages. CT is characterized by its fast acquisition time. Therefore, it is considered as the golden choice for the renal tumor recognition [8]. Knowing the cancer stage helps find the suitable way of treatment.

In this paper, a novel semi-automated model is proposed for early detection and staging of clear cell renal cell carcinoma (ccRCC). The model aims to classify tumor stage from TCGA-KRIC dataset of CT scans for patients with ccRCC, assisting in early detection, diagnosis, and treatment of this disease [9].

2. Literature Survey

Different systems, to detect kidney tumors have been modeled and implemented by many researchers. Most of the models are in medical or clinical form. P. Hallscheidt et al. [10] used multidetector computed tomography (CT) and magnetic resonance imaging (MRI) in staging and estimating renal cell carcinomas with caval thrombus. The dataset acquired consisted of 18 CT scans and 17 MRI scans to be evaluated and correlated with surgical and histopathological staging. The results provided an accuracy of 75%, sensitivity of 93%, and specificity of 80%. A. El-Hefnawy et al. [11] used the data of 693 patients to assess the accuracy of multidetector computed tomography (MDCT) in the staging of RCC. All the MDCT scans of the patients were performed using 4-slice MDCT scanner to produce proper results. Radiological data produced from MDCT scanner was contrasted with surgical and histopathological results. The results provided an overall accuracy of 78%, sensitivity of 95%, and specificity of 80%. Y. Liu et al. [12] acquired the data of 312 patients with RCC to measure the accuracy of 64-slice MDCT staging technique. Sensitivity and specificity are measured according to: perinephric fat invasion, tumor thrombosis, invasion of the adrenal gland, involvement of lymph nodes, and distant metastasis. Results provided an accuracy of 75.48%, sensitivity of 32.26%, and
specificity of 85.87%. Most of the previous work depended on medical or clinical techniques but the proposed model mainly depends on the machine learning and computer vision algorithms in detecting and staging the clear cell renal cell carcinoma.

3. Proposed Model
The system is proposed for ccRCC detection, classification, and staging is shown in figure 1. The approach contains three phases; the first phase is segmentation in which the CT scans were masked to produce masked images containing only the two kidneys then the watershed segmentation algorithm is applied to separate the tumor from the infected kidney. The second phase is the feature extraction where the segmented images were given to Gray Level Co-occurrence Matrix (GLCM) and normal statistical methods were introduced to extract feature vectors from the segmented images to be given to the next phase. The third phase is the classification one, the resulted feature vectors were fed to Random Forest (RF) and Support Vector Machine (SVM) classifiers. These classifiers classified the CT scans into 4 classes with respect to the stage of the ccRCC: stage 1, stage 2, stage 3, and stage 4. In the first stage, the tumor size ranges from 4 cm to 7 cm and is located only inside the kidney, and it does not reach the lymph nodes surrounding the kidney or any other organ. In the second stage, the size of the tumor is larger than the 7 cm but it is still inside the kidney only, it does not extend to the outer lymph nodes or the other organs. In the third stage, the tumor can be of any size and grow outside the kidney, but it does not spread beyond the Gerota’s Fascia. It is growing to a major vein (the renal vein or the vena cava). It can reach to the lymph nodes but not to the other organs. Finally, in the fourth stage, the tumor is growing to the Gerota’s Fascia and beyond and may reach to the adrenal gland above the kidney and it may grow to the surrounding lymph nodes but not to the other organs [13]. Figure 2 illustrates the kidney structure [14].

3.1. Dataset
The dataset used in this proposed model is TCGA-KRIC and it consists of 228 CT scans for patients with ccRCC [9]. The images are in DICOM format. The dataset consists of four groups of patients; the
first group contains 118 patients with ccRCC in stage 1 (between 4 cm and 7 cm and only inside the kidney), the second group has 20 patients in stage 2 (greater than 7 cm and only located in the kidney), the third group consists of 55 patients in stage 3 (can be of any size and inside the kidney or spread to the Gerota’s Fascia), and the fourth group contains 35 patients in stage 4 (spread to the Gerota’s Fascia and beyond). Figure 3 shows examples of the four stages.

![Figure 3](image3.png)

**Figure 3.** Example of ccRCC four stages: (a) stage 1 (between 4 cm and 7 cm and only inside the kidney), (b) stage 2 (greater than 7 cm and only located in the kidney), (c) stage 3 (can be of any size and inside the kidney or spread to the Gerota’s Fascia), (d) stage 4 (spread to the Gerota’s Fascia and beyond).

### 3.2. Segmentation

Segmentation is the process of extracting the region of interest from the background and isolate it from the surrounding objects and structures which are in the images. CT segmentation includes the region of interest separation from the background and other objects in order to be studied [15]. In the proposed model, firstly, the binary mask was used to segment the kidney lobes. Then the segmented images of the kidneys were introduced to watershed algorithm to segment the tumor.

#### 3.2.1. Binary Mask

A binary mask image is created to segment the two kidneys from the CT scans. It is applied on the original image to remove all the other objects except the kidneys. The pixels that are created in the shape of the kidney are white (have the value 1) and the rest pixels of the background are black (have the value 0). The mask is multiplied by the original image so that, the masked image – containing the two kidneys only – is created. This mask differs from one image to another as the kidney size and shape are not the same for all the images [15]. Figure 4 shows the segmented image produced by applying the binary mask.

![Figure 4](image4.png)

**Figure 4.** The two kidneys segmented by binary mask: (a) original image, (b) binary mask, (c) segmented image.

#### 3.2.2. Watershed Segmentation

One of the most difficult image processing operations is the separation of touching objects in CT images. In order to solve this problem, watershed segmentation (also called watershed transform) is used. Watershed segmentation is a mathematical morphological operating tool that detects all the different shapes in an image and can split any of them up [16]. It is
one of the best techniques used to assemble the pixels of the same intensities together, where the high intensity pixels are light and the low intensity pixels are dark, so it finds the catchment basins and watershed ridge lines in the image [17]. Therefore, the watershed transform is a better algorithm used to segment the tumor from the masked images of the kidneys. Figure 5 shows the tumor separation using the watershed segmentation.

![Figure 5. Tumor Segmentation: (a) segmented image by the binary mask, (b) detecting the tumor, (c) tumor segmentation.](image)

3.3. Feature Extraction
The second phase in the proposed model is the feature extraction phase. Feature extraction is used to increase the efficiency and the accuracy of the classifier, as it keeps the most relevant features of the segmented images in feature vectors and discards the rest of features. These feature vectors are then fed into the classifiers categorizing the dataset into four classes with respect to the tumor stage. There are lots of feature extraction techniques. In this model, gray level co-occurrence matrix was integrated with normal statistical method to produce the feature vectors of all the segmented images.

3.3.1. Gray Level Co-occurrence Matrix (GLCM). Gray Level Co-occurrence Matrix (GLCM) is the most commonly used algorithm for the analysis and classification of the medical images [18]. GLCM technique is used to mark the proportional position of two pixels relative to each other. This method is formed by calculating the number of each pixel pair appearance at a definite distance. GLCM method can produce about 22 features, in this model, only 5 features are the most effective and helpful in the classification phase: auto-correlation, contrast, correlation, homogeneity, and energy [19].

3.3.2. Normal Statistical Method. Normal statistical method is the method where area, mean, variance, and standard deviation features where calculated. This method was integrated with the GLCM algorithm to help the classifiers to give more accurate and efficient results [19].

Auto-correlation, contrast, correlation, homogeneity, energy, area, mean, variance, and standard deviation are the chosen features because they are the most effective features that help in separating the ccRCC four stages easily in the classification phase leading to more accurate results.

3.4. Classification
Classification is the last phase in the model which is the most important phase as its main function is to classify the dataset into different classes. After applying the feature extraction phase, the features extracted by GLCM and the features calculated by the normal statistical method are introduced to the classification phase. This model is used to classify the dataset into 4 classes according to the stage of the tumor; stage 1-class, stage 2-class, stage 3-class, and stage 4-class. In this paper, to check the effectiveness of the used process, the model is experimented on the Random Forest (RF) and the Support Vector Machine (SVM) classifiers.

3.4.1. Random Forest Algorithm (RF). One of the simplest and easiest ways of classification in machine learning is Random Forest algorithm. Random forest is characterized as a supervised
ensemble classification technique. It is also characterized by its robustness and simplicity while classifying large datasets. In other words, it merges more than one technique of different or same type of classification algorithms such as Support Vector Machine, Naïve Bayes, and decision trees [20]. Random Forest can be used for classification as well regression. It is an easy to use algorithm and is regarded as a promising classifier. RF operates by creating decision trees based on randomly selected data samples and by means of voting it gets a prediction from each tree to select the best solution. The large number of decision trees participating in the process makes RF robust and highly accurate, although this make RF a slower classifier when compared to other classifiers [21]. Experiments on the given dataset using RF as a classifier provided great results in terms of accuracy, sensitivity, and specificity, yet SVM provided better results.

3.4.2. Support Vector Machine Algorithm (SVM). Support Vector Machine (SVM) is a supervised and relatively simple machine learning algorithm. It is mainly used for linear and non-linear classification or regression problems [22]. SVM performs classification by constructing a hyper-plane that maximize the probabilistic multiclass classifier as it takes the input data and predicts which class is an appropriate one. Thus, SVM is characterized by its good performance in many fields like the pattern, text recognition, etc. SVM proved to achieve significantly higher accuracy in the classification of images when compared to traditional query schemes, after just three to four rounds of relevance feedbacks. That’s why it was adopted as one of the classifiers in the proposed model producing promising results. SVM can be interpreted as an extension of the perceptron, as it belongs to a family of generalized linear classifiers. SVM is also known as maximum margin classifier as it tends to minimize the empirical classification error and maximize the geometric margin. In the proposed model, experiments proved that SVM was more efficient than Random Forest classifier.

4. Results and Discussion
The proposed model was able to classify CT scans in the dataset according to ccRCC stage. The classification phase classifies images according to the cancer stage. ccRCC has 4 stages. In the first stage, the tumor is enclosed inside the kidney and its size ranges from 4 cm to 7 cm. In the second stage, tumor is also enclosed inside the kidney but its size is usually greater than 7 cm. In the third stage, the tumor can be of any size and not necessarily be enclosed inside the kidney as it may extend to the Gerota’s Fascia. In the fourth stage, the tumor grows extending even beyond the Gerota’s Fascia. The dataset contains 228 CT scans that were divided into two groups of images. The first group consisted of 150 images used for learning the classifiers (77 images from stage 1, 13 images from stage 2, 36 images from stage 3, and 23 images from stage 4), while the second group consisted of 78 images used for validation and testing (41 images from stage 1, 7 images from stage 2, 19 images from stage 3, and 12 images from stage 4). In order to evaluate the performance of the proposed approach, the following performance indicators were calculated: accuracy, sensitivity, and specificity. They were calculated using equations (1), (2), and (3) respectively, where TP is true positive (the result where the classifier predicts the positive class accurately), TN is true negative (the result where the classifier predicts the negative class accurately), TP is false positive (the result where the classifier predicts the positive class inaccurately), and FN is false negative (the result where the classifier predicts the negative class inaccurately).

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP}
\]
The CT scans were classified using SVM classifier producing the overall results of accuracy 99.12%, sensitivity 98.78%, and specificity 99.41% compared to Random Forest classifier which produced overall results of accuracy 98.25%, sensitivity 98.43%, and specificity 99.16%. Experiments on the given dataset showed that when comparing the results of SVM classifier to Random Forest classifier, SVM yielded overall better results.

As shown in table 1 SVM is more accurate than RF in staging the ccRCC. Table 2 illustrates the proposed model result compared to the other systems.

| Classifier | Stage 1 Specificity | Stage 2 Specificity | Stage 3 Specificity | Stage 4 Specificity | Stage 1 Sensitivity | Stage 2 Sensitivity | Stage 3 Sensitivity | Stage 4 Sensitivity | Total Accuracy |
|------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|-----------------|
| SVM        | 97.62%              | 100%                | 100%                | 100%                | 99.12%              | 100%                | 99.12%              | 100%                | 99.12%          |
| RF         | 97.62%              | 99.03%              | 100%                | 98.61%              | 95.12%              | 100%                | 98.25%              | 100%                | 98.25%          |

Table 1. Comparison of the accuracy of the proposed model on different classifiers.

| Paper                  | Sensitivity | Specificity | Accuracy |
|------------------------|-------------|-------------|----------|
| P. Hallscheidt et al, 2005 [10] | 93%         | 80%         | 75%      |
| A. El-Hefnawy et al, 2011 [11] | 95%         | 80%         | 78%      |
| Y. Liu et al, 2012 [12]    | 32.26%      | 85.87%      | 75.48%   |
| P. Alongi et al, 2015 [6]  | 74%         | 80%         | 84%      |
| Proposed Model          | 98.78%      | 99.41%      | 99.12%   |

Table 2. Comparison of the proposed model with other previous researches.

5. Conclusion
In this paper, a novel semi-automated model for early detection and staging of clear cell renal cell carcinoma has been introduced. Performance analysis shows that, the proposed semi-automated model provided better results in terms of sensitivity, specificity, and accuracy, in comparison with previous clinical models. Binary mask was used to segment kidney lobes from the CT scans. Watershed algorithm was then applied to the masked images to segment the tumor. GLCM was integrated with normal statistical method to extract the most important features from segmented images. Choosing the right classification algorithm was as important as choosing proper feature extraction technique, that’s why the resulted feature vectors were fed into Random Forest (RF) and Support Vector Machine (SVM) classifiers, leading to an increase in accuracy, sensitivity, and specificity with 15.12%, 3.78%, and 13.54% respectively. Experiments were carried on TCGA-KRIC dataset which includes 228 CT scans. Further experiments can be performed on MRI scans using the proposed model to enhance performance and increase efficiency.

6. References
[1] Information H, Disease K, Work Y, Work Y, Center T and Health N 2019 Your Kidneys & How They Work | NIDDK National Institute of Diabetes and Digestive and Kidney Diseases
[2] Anon 2019 Understanding Kidney Cancer WebMD
8

[3] Anon 2019 Kidney cancer - Symptoms and causes Mayo Clinic
[4] Wang C, Hu J, Lu M, Gu H, Zhou X, Chen X, Zen K, Zhang C, Zhang T, Ge J, Wang J and Zhang C 2015 A panel of five serum miRNAs as a potential diagnostic tool for early-stage renal cell carcinoma Scientific Reports 5 1-8
[5] Low G, Huang G, Fu W, Moloo Z and Girgis S 2016 Review of renal cell carcinoma and its common subtypes in radiology World Journal of Radiology 8 484-500
[6] Alongi P, Picchio M, Zattoni F, Spallino M, Gianolli L, Saladini G and Evangelista L 2015 Recurrent renal cell carcinoma: clinical and prognostic value of FDG PET/CT European Journal of Nuclear Medicine and Molecular Imaging 43 464-73
[7] Sacco E, Pinto F, Totaro A, D’Addessi A, Racicoppi M, Gulino G, Volpe A, Marangi F, D’Agostino D and Bassi P 2011 Imaging of renal cell carcinoma: state of the art and recent advances Urologia Internationalis 86 125-39
[8] Oostenbrugge T, Fütterer J and Mulders P 2018 Diagnostic imaging for solid renal tumors: a pictorial review Kidney Cancer 2 79-93
[9] Anon 2019 TCGA-KIRC - The cancer imaging archive (TCIA) public access – cancer imaging archive wiki Wiki.cancerimagingarchive.net
[10] Hallscheidt P, Fink C, Haferkamp A, Bock M, Luburic A, Zuna I, Noeldge G and Kauffmann G 2005 Preoperative staging of renal cell carcinoma with inferior vena cava thrombus using multidetector CT and MRI Journal of Computer Assisted Tomography 29 64-8
[11] El Hefnawy A, Mosbah A, Shaaban A, El Diasty T, Hassan M and Ibrahim E 2011 UP-01.153 Accuracy of multi-detector computed tomography (MDCT) in staging of renal cell carcinoma (RCC): analysis of risk factors for mis-staging and its impact on surgical intervention Urology 78 887-91
[12] Liu Y, Song T, Huang Z, Zhang S and Li Y 2012 The accuracy of multidetector computed tomography for preoperative staging of renal cell carcinoma International braz j urol 38 627-36
[13] Anon 2019 Kidney Cancer Stages | Renal Cell Carcinoma Staging Cancer.org
[14] Samaha D 2019 - TNM Staging of Renal cell carcinoma (illustrated). Simple-med.blogspot.com
[15] Ali A, Hassan E and Nazar A 2016 Analysis and classification of kidney images using watershed segmentation and texture properties International Journal of Computer Science and Mobile Computing 5 592-604
[16] El Allaoui A 2012 Medical image segmentation by markercontrolled watershed and mathematical morphology The International journal of Multimedia & Its Applications 4 1-9
[17] Mustaqeem A, Javed A and Fatima T 2012 An Efficient brain tumor detection algorithm using watershed & thresholding based segmentation International Journal of Image, Graphics and Signal Processing 4 34-9
[18] Shah B, Sawla C, Bhanushali S and Bhogale P 2017 Kidney tumor segmentation and classification on abdominal CT scans International Journal of Computer Applications 164 1-5
[19] Victor A and Ghalib M 2017 A hybrid segmentation approach for detection and classification of skin cancer Biomedical Research 28 6947-54
[20] Raczk E and Zagajewski B 2017 Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images European Journal of Remote Sensing 50 144-54
[21] Sonobe R, Tani H, Wang X, Kobayashi N and Shimamura H 2014 Random forest classification of crop type using multi-temporal TerraSAR-X dual-polarimetric data Remote Sensing Letters 5 157-64
[22] Sinha P and Sinha P 2015 Comparative study of chronic kidney disease prediction using KNN and SVM International Journal of Engineering Research and 4 608-12