Queen’s University

CISCI 351 Report

Kidney Cancer Imaging

Life Sciences

Avatar Team #11

Victor Zheng -20123494

Ryan Elliott - 05999293

Madi Polgar - 20055152

Emily Poon - 20062692

Derek Xu - 20114245

April 16, 2021

# Abstract

The high rate of human error in the medical diagnostic field leads to extreme financial and physical hardships for patients and healthcare systems. This is prevalent in cancer patients, specifically those with kidney cancer. Its common symptoms make this type of cancer hard to diagnose, and time is often wasted waiting for pathology and radiology results. The additional burden of human error results in death, disability, debt, minor health issues, and inconveniences. To solve this problem, the goal of this project was to improve the speed and accuracy of the diagnosis of clear cell renal cell carcinoma using the classification of pathology samples.

The dataset is from the National Institute’s Clinical Proteomic Tumor Analysis Consortium and contains 783 pathology images from 222 patients with clear cell renal cell carcinoma. The pathology images are 67% cancerous and 33% normal kidney tissue.

The images were scanned into SVS files then processed into 65-pixel tiles. Tiles from the center of each sample were fed through a tensor which was used to train the convolution neural network (CNN). The model contains a 2D convolution layer, followed by a 2D MaxPooling layer, and a batch normalization layer. The layers in the CNN model repeat three times then the images are vectorized and passed through a three-layered fully connected neural network. A cancerous image was classified using the vector [0,1] and a normal tissue image was assigned [1,0]. The testing accuracy of this method was 72.14%.

To improve the model’s accuracy in the future, the team would utilize multiple tiles from each image in training and testing data. Additionally, unsupervised learning would remove the need to label each tile. The model would be improved if it could extract features such as age, gender, tumor size, and tumor grade to use in conjunction with image classification.

Table of Contents

[Abstract i](#_Toc69491268)

[List of Figures ii](#_Toc69491269)

[List of Tables ii](#_Toc69491270)

[1.0 Introduction 1](#_Toc69491271)

[1.1 The Problem 1](#_Toc69491272)

[1.2 Motivation 1](#_Toc69491273)

[2.0 Related Work 1](#_Toc69491274)

[3.0 Dataset 2](#_Toc69491275)

[4.0 Methodology 2](#_Toc69491276)

[4.1 Data Processing 3](#_Toc69491277)

[4.2 Feature Extraction 3](#_Toc69491278)

[4.3 Modeling 4](#_Toc69491279)

[5.0 Experiments and Results 4](#_Toc69491280)

[5.1 Threats to Validity 4](#_Toc69491281)

[5.2 Possible Improvements 5](#_Toc69491282)

[6.0 Discussion 5](#_Toc69491283)

[7.0 Group Member Contributions 6](#_Toc69491284)

[8.0 Replication Package 6](#_Toc69491285)

[9.0 Conclusions 6](#_Toc69491286)

[9.1 Future Work 7](#_Toc69491287)

[References 8](#_Toc69491288)

# List of Figures

[Figure 1: The DeepPath framework takes the 20x magnification level of the SVS image and saves it as a 512x512 jpeg tile. Due to hardware constraints our model selects a 64-pixel box from the centre of the tile for evaluation by the CNN. The red box from the 20x tile is the region which is evaluated by the CNN. 3](#_Toc69491289)

# List of Tables

[Table 1: Group member contributions to the project. 6](#_Toc69491290)

# 1.0 Introduction

## 1.1 The Problem

Cancer is the leading cause of death in Canada where one in two Canadians will be diagnosed in their lifetime, and one in four of those individuals will pass away [1]. Initially, 12% of cancer patients are misdiagnosed which leads to complications in the future from incorrect or lack of treatment. One of the hardest cancers to diagnose is kidney cancer because symptoms include fatigue, back pain, and weight loss. It is often missed during annual physical exams because the kidneys are located far inside the body. Additionally, results of a pathology sample can take weeks to become available and once the patient receives these results, there is up to a 15% chance that there is an error in the diagnosis [2]. Diagnostic-related errors have cost hospitals like John Hopkins $38.8 billion in a 25-year span [3]. Not only does human error cost the patients and healthcare systems, but it also results in death, disability, inconveniences, and minor health issues. Overall, human error in diagnostics is a costly, deadly, and common problem that must be addressed.

## 1.2 Motivation

To combat the problem, the team’s goal was to use image classification techniques to accurately diagnose kidney cancer from medical images. These techniques included data preprocessing using TensorFlow as well as feature extraction and predictive modelling using a convolution neural network (CNN). The goal was to use the team’s knowledge of data analysis to decrease the human error present in diagnostics, provide faster and more accurate results to the patient, and reduce misdiagnoses in kidney cancer patients which would ideally decrease the overall cancer cases and resulting deaths in Canada. This would also financially benefit patients and healthcare systems by reducing malpractice lawsuits and treatment, either by avoiding a misdiagnosis or by catching the cancerous tumor early.

# 2.0 Related Work

A third-party analysis of the dataset was conducted by Crowds Cure Cancer in 2017, where it was used in conjunction with other datasets in The Cancer Imaging Archive to create labels and annotations for tumor images. The purpose of this project was to generate a large repository of publicly available cancer radiology images to aide with cancer prevention, diagnosis, monitoring, and treatment [4].

A related paper published on the topic of clear cell renal cell carcinoma and proteomic tumor analysis is the journal article published in October 2019, *Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma*. This article used tissue samples to determine the level of immune infiltration in the tumors and identified protein regulation problems with proteomic analysis. This was done through genome and RNA sequencing of tissue samples collected from 110 tumor samples. The authors concluded that there is a strong relation between high-grade tumors and immune infiltration [5].

The article *Classification of Tumor Samples from Expression Data Using Decision Trunks*, published in 2013, uses machine learning to classify prostate, bladder, breast, neuroblastoma, and lung tumor samples. The authors used decision trunks, SVM, kNN, and ROCC as predictive modeling techniques which had classification accuracies of 84%, 75%, 81%, and 83% respectively [6].

# 3.0 Dataset

The dataset used for this project was from the National Cancer Institute’s Clinical Proteomic Tumor Analysis Consortium - Clear Cell Renal Cell Carcinoma (CPTAC-CCRCC). The Clinical Proteomic Tumor Analysis Consortium (CPTAC) is a national effort to coordinate and accelerate the understanding of the cancer at a molecular basis through the application of large-scale proteome and genome analysis. The dataset on Clear Cell Renal Cell Carcinoma (CCRCC) contains radiology, clinical, proteomic, genomic and pathology data. For each patient there are radiology images with corresponding tissue slide images of the affected organ along with clinical data, genomic data and proteomic data [7].

The current version of the dataset used included 783 pathology images from 222 patients. Within the pathology data, 67% of the images are of tissue with cancerous tumors and 33% are of normal tissue. Of the 783 patients within the pathology data, 28% of the patients have radiology data with 61% of patients being male and 39% were female. The dataset also contained 94,500 radiology images from 63 patients. The type of radiology images in the dataset contained computed tomography, magnetic resonance, SR document, computed radiography, and digital radiography images. Of the 94,500 radiology images, 81% of them were computed tomography images. The atomical sites of the images included the abdomen, chest, liver and chest to pelvis. The genomic data provided us the gene sequences that looks at the patient’s disposition for the disease. The proteomic data provided the proteins within the patient’s body that act as biomarkers for the patient’s tendency to attract certain maladies. The clinical data contained the label of either tumor or normal, patient ID, the tumor site, focality and other variables such as age, gender, tumor stage and BMI. Within this dataset, we chose to use the pathology data for the project.

# 4.0 Methodology

The pathology samples consisted of 190GB of 782 slides of normal and cancerous kidney tissue. The samples were scanned at high resolution and magnification into a SVS files. The SVS images are not just a single scan at one magnification, instead, they consist of scans at multiple magnification levels. The opensource imaging library, OpenSlide, was used to bring the images into our workspace. Previous publications have developed tools for breaking the large SVS images into smaller image tiles for faster processing and evaluation. The DeepPath framework developed by Coudray *et al.* was used to process all the SVS images into smaller tiles.

## 4.1 Data Processing

The DeepPath framework requires users to point towards the directory of the SVS images, specify the desired number of pixels in the tiles, the tile overlap, number of threads to dedicate to processing the images, the image format, and the amount of background allowed in an image. Our model chose 512-pixel tiles with no overlap and 25% allowed background in a tile. This dataset was still too large for the CNN to evaluate the images on a 4GB Nvidia 980 GPU.

Images were passed to a CNN built from the Keras Python deep learning API as a tensor constructed with TensorFlow. Using 512x512 pixel images for the individual tiles remained too large for processing in the CNN so smaller tiles were selected to build the tensor. Tiles of 65x65 pixels were chosen for the final tensor chosen from the centre of all the tiles within the sample as shown in Figure 1 below. This tensor was used to train the CNN. Choosing a subregion of samples aids in the evaluation of spatially separated regions within the tissue. This is an important factor to consider as each sample of cancerous tissue is a collection of normal and malignant tissue. The average percent of tumor nuclei within a malignant tissue sample was 82% ± 11% STD. While we only evaluate 1.5% of a tissue surface area, each slide consists of more than 200 images per slide. We can reasonably assume that the images of cancerous tissue contain some number of cancerous cells no matter where the images were taken.

|  |
| --- |
| Figure 1: The DeepPath framework takes the 20x magnification level of the SVS image and saves it as a 512x512 jpeg tile. Due to hardware constraints our model selects a 64-pixel box from the centre of the tile for evaluation by the CNN. The red box from the 20x tile is the region which is evaluated by the CNN. |

## 4.2 Feature Extraction

The tensor is passed to a convolution neural network for feature learning. The model consists of a 2D convolution layer which feeds into a 2D MaxPooling layer and a batch normalization layer. These layers are repeated in triplicate before vectorizing the processed images and passing the images through a fully connected neural network with three layers. The final layer in the neural network classifies the images as malignant or normal tissue. Two vectors were assigned to the images to encode the classification, [1,0] for normal, and [0,1] for cancerous.

## 4.3 Modeling

A testing dataset was then created using five other images randomly sampled from the tissue. The five samples are not from the same subset as the training set. We posit that random sampling of regions not included in training is a suitable estimate of our model performance.

# 5.0 Experiments and Results

In our experiments, we constructed the neural networks in Keras, loaded the training and testing data, fit the model to the training data over five epochs, and finally evaluated the model based on the testing data. This method was used for all our experiments.

We evaluated our approaches based on % accuracy. This is a suitable measure of performance because the dataset is not heavily unbalanced (67% cancer positives and 33% cancer negatives). For a heavily unbalanced dataset (over 90% in one class), a different measure of performance would be required.

## 5.1 Threats to Validity

One major flaw with the approach to cropping the images is that cropping a cancerous slide could exclude the cancer nuclei, which would bias the predictor towards non-cancerous images even if they are initially labeled as cancerous. To minimize this effect, we took cropped tiles from the center of each image. We chose this approach because most of the cancer nucleus percentages were greater than 80%, which reduces the chance of a crop missing the tumor site entirely. In the notebook CISC351\_Kidney\_CNN\_10\_tiles\_All\_Slides.ipynb, the bounding box for the crop of the test data was in the center of the image, which resulted in a 72.14% testing accuracy compared to the 47.34% testing accuracy of the model in CISC351\_Kidney\_CNN.ipynb. The difference in accuracy is caused by the chosen cropping method. Since the tile in the first notebook was more central, the tile contained more relevant features, and thus was more reliable for the predictor.

Current methods of cancer nucleus detection use down-sampling to reduce the data to a workable size. The advantage of this method is that it preserves all features of a given sample. The disadvantage is that the features are preserved at a lower resolution.

## 5.2 Possible Improvements

We wanted to see if our method could improve on the existing methods by preserving higher-resolution samples than the existing method. Unfortunately, the tradeoff of higher resolution with less features was not worth it because our method’s 72.14% testing accuracy was below the current standard method’s testing accuracy of about 95%.

One possible next step that we considered was to use multiple tiles from each image in both training and testing and using the different tiles to vote on the result for the entire image. This would preserve more features than our current approach, while also preserving a higher resolution than the existing approaches. However, this type of model was beyond the skillset of the team.

Another approach we considered was unsupervised or semi-supervised learning. The advantage to this method is that our model could use the 64x64 crops without needing each individual tile to be labeled. This approach would overcome the problem of cropping out the cancer nuclei, since the model would cluster this tile in a different class than the other tiles in its image. Unfortunately, unsupervised learning is more computationally complex than the existing supervised method and was beyond the skillset of the team.

# 6.0 Discussion

The options that the group has researched and discussed were ultimately narrowed down to a convolutional neural network after an extensive process of iteration, experimentation, and research. When exploring novel options, the group approached a different way of pre-processing the image datasets. Current methods use down sampling to get their images into a suitable size for evaluation or train on a disproportionately low subset of the healthy cell tissue. Since the average number of tumor nuclei is 82% of the sample, to ease the exhaustive computing needs of the model, it is possible to input a smaller region of the tissue histology image without sacrificing substantive accuracy percentage from noise in image preprocessing. The most indicative features that the model needs to extract can be gathered from a significantly smaller sample image, thus whole, multi-channel images are redundant in information provided and are unnecessarily more time-consuming and disk space intensive to process.

The other potential avenue to curb performance issues was to host the model and data on the cloud. It would remove all local and physical disk space requirements and significantly improve the computation time for the model. This proved to be challenging as only a small portion of the team was familiar with cloud computing, and its specific platforms. Plus, there were concerns about the monetary cost of hosting with that volume of data that would exceed the free account limits.

Another route was unsupervised learning particularly through the invariant information clustering for unsupervised image classification. The main problem with this method was its poor performance on advanced datasets, which would translate directly into the complex and high dimensional problem posed in this paper. Some other methods were highly promising including edge mapping and adversarial learning, and autoencoders, but they were limited to detecting and segmenting the nucleus and had limited application to the goal of the project and additionally could not differentiate between cancerous cells and normal cells when segmenting.

# 7.0 Group Member Contributions

Table 1: Group member contributions to the project.

|  |  |
| --- | --- |
| **Group Member** | **Contributions** |
| Victor Zheng - 20123494 | Final Report- Discussion  Presentation- Discussion  Coding – Markdown, Repository Management |
| Ryan Elliott - 05999293 | Final Report – Methodology  Presentation - Methodology  Coding – Tiling of the SVS images, Implementation of the CNN |
| Madi Polgar - 20055152 | Final Report – Abstract, Introduction, Related Work (Formatting and Editing)  Presentation – Introduction to the problem, Motivation, Conclusion (Formatting and Editing) |
| Emily Poon - 20062692 | Final Report – Dataset and Conclusion  Presentation – Dataset |
| Derek Xu - 20114245 | Final Report – Experiments and Results  Presentation – Experiments and Results  Coding – Solution to the images too large problem, proposed solutions to avoid losing data when cropping |

# 8.0 Replication Package

The following URL is a link to the project’s GitHub repository:

<https://github.com/victorzheng1/351project>

# 9.0 Conclusions

Overall, the main goal for this model was to build a convolution neural network to diagnose kidney cancer in an accurate and efficient way from the given data. An incorrect medical diagnosis is a prevalent problem in the medical industry with 12% of cancer patients being initially misdiagnosed. Our model would ideally help decrease the malpractice incidents where patients are misdiagnosed, which in turn would help avoid health risks and any legal issues that come with a misdiagnosis.

Using our dataset from the National Cancer Institute’s Clinical Proteomic Tumor Analysis Consortium - Clear Cell Renal Cell Carcinoma, we chose to use the pathology data to predict if there was cancerous tissue present. To build our model, we scanned the images in the dataset into SVS files before processing them into 65-pixel tiles. Using the center of each image, the tile was processed through a tensor that was used to train the CNN. To build our neural network, we used a 2D convolutional layer, a 2D max pooling layer and a batch normalization layer. The different layers were repeated three times before the image was flattened into a column vector and passed through the three-layered fully connected neural network to be categorized as either cancerous [0,1] or normal [1,0].

Our model produced a testing accuracy of 72.14%, which reflects a realistic ability to diagnose cancer based on diagnostic images and tissue samples but is below the 95% testing accuracy of the current standard methods.

## 9.1 Future Work

To further improve on the application, the model should be able to output a 2D array of the cancer outline for prognosis by extracting cancer specific features e.g., size, growth patterns, grade, etc. and use other data such as genomics, and clinical to evaluate treatment options. The model should also be expanded further to be able to distinguish between different subtypes of CCRCC.

Another idea is to increase randomness for both the testing and training data, but to implement this we need access to a more powerful computer. With more access, we could use all of the images given in the dataset to test and train the data, as well as increase the number of tiles read. This would improve the accuracy of our model as all the tiles in an image be processed. Our current model was only able to see one tile from each image, so if an image contained cancerous cells but the tile selected only contains normal cells, then the image is still classified as normal tissue giving as a less accurate prediction.

# References

|  |  |
| --- | --- |
| [1] | "Kidney Cancer Statistics," Canadian Cancer Society, 2021. [Online]. Available: https://www.cancer.ca/en/cancer-information/cancer-type/kidney/statistics/?region=ab#:~:text=Incidence%20and%20mortality&text=In%202020%2C%20an%20estimated%3A,1%2C300%20will%20die%20from%20it.. [Accessed April 2021]. |
| [2] | J. Sinrich, "The Healthy," Trusted Media Brands, Inc., 08 April 2021. [Online]. Available: https://www.thehealthy.com/cancer/cancer-screening/. [Accessed 10 April 2021]. |
| [3] | The Johns Hopkins University, "Diagnostic Errors More Common, Costly, And Harmful Than Treatment Mistakes," Johns Hopkins Medicine, 23 April 2013. [Online]. Available: https://www.hopkinsmedicine.org/news/media/releases/diagnostic\_errors\_more\_common\_costly\_and\_harmful\_than\_treatment\_mistakes. [Accessed April 2021]. |
| [4] | "Crowds Cure Cancer," National Cancer Institute, 2017. [Online]. Available: https://www.crowds-cure.org/. [Accessed 2021]. |
| [5] | D. J. Clark, S. M. Dhanasekaran, F. Petralia and H. Zhang, "Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma," *National Library of Medicine: National Center for Biotechnology Information,* vol. 179, no. 4, pp. 964-983, 2019. |
| [6] | B. Ulfenborg, K. Klinga-Levan and B. Olsson, "Classification of Tumor Samples from Expression Data Using Decision Trunks," *Cancer Informatics,* no. 12, pp. 53-66, 2013. |
| [7] | "Office of Cancer Clinical Proteomics Research," 2021. [Online]. Available: https://proteomics.cancer.gov/programs/cptac. |
| [8] | P. S. O. T. S. N. N. M. S. D. F. A. L. M. N. R. A. T. Nicolas Coudray, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning.," *Nature Medicine,* 2018. |
| [9] | N. C. I. C. P. T. A. C. CPTAC, "Radiology Data from the Clinical Proteomic Tumor Analysis Consortium Clear Cell Renal Cell Carcinoma [CPTAC-CCRCC] collection [Data set].," 2018. |