

**PREDICTING VON HIPPEL LINDAU (VHL),
POLYBROMO-1 (PBRM1) MUTATIONS AND STAGES OF
CLEAR CELL RENAL CELL CARCINOMA FROM
COMPUTED TOMOGRAPHY IMAGES BY MACHINE
LEARNING**

by

Harika Beste Ökmen

B.S., in Biomedical Engineering, Yeditepe University, 2015

Submitted to the Institute of Biomedical Engineering
in partial fulfillment of the requirements
for the degree of
Master of Science
in
Biomedical Engineering

Boğaziçi University

2019

**PREDICTING VON HIPPEL LINDAU (VHL),
POLYBROMO-1 (PBRM1) MUTATIONS AND STAGES OF
CLEAR CELL RENAL CELL CARCINOMA FROM
COMPUTED TOMOGRAPHY IMAGES BY MACHINE
LEARNING**

APPROVED BY:

Assoc. Prof. Dr. Albert Güveniř
(Thesis Advisor)

Prof. Dr. Cengizhan Öztürk

Dr. Mehmet Kocatürk

DATE OF APPROVAL: 02.08.2019

ACKNOWLEDGMENTS

First, I would like to extend my thanks to my thesis advisor Assoc. Prof. Dr. Albert Güveniř for his guidance during my master research.

I am very grateful to our radiologist Dr. Hadi Uysal for his endless patience, helps and supports. Also, I express my gratitude to my classmates and teachers at Boğaziçi University for inspiration to me while handling my research.

I would like to thank my patient friends Aysu Öznaneci, Havva Araz and Tuğçe Hatipođlu. They always supported and motivated me.

Finally, I would like to state my deep gratitude to my family: my dear mother Hamiyet Ökmen, and my dear father Dr. Recep Kemal Ökmen for their love, patience, and belief to me. They shared their precious experiences with me and were very supportive and understanding during my academic career.

ACADEMIC ETHICS AND INTEGRITY STATEMENT

I, Harika Beste Ökmen , hereby certify that I am aware of the Academic Ethics and Integrity Policy issued by the Council of Higher Education (YÖK). I fully acknowledge all the consequences due to its violation by plagiarism or any other way.

Name :

Signature:

Date:

ABSTRACT

PREDICTING VON HIPPEL LINDAU (VHL), POLYBROMO-1 (PBRM1) MUTATIONS AND STAGES OF CLEAR CELL RENAL CELL CARCINOMA FROM COMPUTED TOMOGRAPHY IMAGES BY MACHINE LEARNING

RCC is the most prevalent renal malignancy and ccRCC is the most common subtype of RCC. It is reported that the prognosis has a strong association with VHL alteration. It is also reported that PBRM1 gene, second most common alteration in ccRCC, has a critical role in ccRCC progression and great potential to identify ccRCC. Moreover, available treatment opportunities are mostly related to stage information. The treatment options are limited in stage 3 and 4. Studies of ccRCC indicate that there is a correlation between cancer CT imaging features and gene expression (radiogenomics). We hypothesized that from quantitative 2D CT images via one slice with the biggest tumor, both VHL and PBRM1 mutations and stages can be predicted with accuracy using ML algorithms. TCGA-KIRC data were collected and tumors were segmented by an expert radiologist. Classification was done by using CL and ANN on MATLAB. Our results showed that Fine Gaussian SVM model is able to predict VHL and NON-VHL data with 68.6%, k-NN with Random Subspace model is able to predict PBRM1 and NON-PBRM1 with 84.9% ,and ANN predicted stages with 91.90% accuracies. From this study, it appears that ML-based quantitative 2D CT analysis using one slice for each patient is a feasible and potential method for predicting the status of VHL and PBRM1 mutations and stages for ccRCC patients.

Keywords: Renal Cell Carcinoma (RCC) , Clear cell Renal Cell Carcinoma (ccRCC), Von Hippel Lindau (VHL), Polybromo-1 (PBRM1), Computed Tomography (CT), Machine Learning (ML), Artificial Neural Network (ANN), Classification Learner (CL).

ÖZET

BİLGİSAYARLI TOMOGRAFİ GÖRÜNTÜLERİNDEN MAKİNE ÖĞRENMESİ İLE BERRAK HÜCRELİ BÖBREK KARSİNOMUN VON HIPPEL LINDAU (VHL) VE POLYBROMO-1 (PBRM1) MUTASYONLARININ VE EVRELERİNİN TAHMİN EDİLMESİ

BHK en fazla görülen böbrek kanseridir ve BHBK, en sık görülen BHK alt tipidir. BHBK çalışmaları kanserli BT görüntüleri ile gen mutasyonları arasında bir korelasyon olduğunu göstermektedir (radyogenomik). Ayrıca, prognozun VHL mutasyonunu ile güçlü bir ilişkisi olduğu rapor edilmiştir. PBRM1 geni, BHBK’de en yaygın ikinci mutasyondur ve BHBK’yi tanımlamak için büyük bir potansiyele ve ilerlemede kritik bir role sahiptir. Dahası, mevcut tedavi olanakları çoğunlukla evre bilgileriyle ilgilidir. Tedavi seçenekleri evre 3 ve 4’te sınırlıdır. Bu nedenle erken tanı hastalar için önemlidir. Bir axial slayt üzerinden en büyük tumore göre nicel 2D BT görüntüleriyle, VHL ve PBRM1 mutasyonları ile evrenin doğrulukla tahmin edilebileceğini varsayıldı. TCGA-KIRC datası kullanıldı ve ilgili bölge uzman bir radyolog tarafından çizildi. MATLAB’da hem CL hem de ANN kullanılarak sınıflandırma yapıldı. Alınan sonuçlar, CL’de Fine Gaussian SVM modelinin VHL ve NON-VHL verilerini 68.6%, k-NN with Random Subspace modelinin PBRM1 ve NON-PBRM1’i % 84.9 dereceleriyle doğru tahmin edebildiğini gösterirken, ANN modelinin evreyi % 91.9 doğru tahmin edebildiğini gösterdi. Bu çalışmadan, ML ile tek slayt bazlı nicel 2D BT doku analizinin, BHBK’li hastalarda VHL ve PBRM1 mutasyonlarını ve evreyi tahmin etmek için uygun ve potansiyelli bir yöntem olduğu anlaşılmaktadır.

Anahtar Sözcükler: Böbrek Hücresi Kanseri (BHK), Berrak Hücreli Böbrek Kanseri (BHBK), Von Hippel Lindau (VHL), Polybromo-1 (PBRM1), Bilgisayarlı Tomografi (BT), Makine Öğrenmesi (ML), Classification Learner (CL), Yapay Sinir Ağı (ANN).

TABLE OF CONTENTS

ACKNOWLEDGMENTS	iii
ACADEMIC ETHICS AND INTEGRITY STATEMENT	iv
ABSTRACT	v
ÖZET	vi
LIST OF FIGURES	ix
LIST OF TABLES	xi
LIST OF SYMBOLS	xii
LIST OF ABBREVIATIONS	xiii
1. INTRODUCTION	1
1.1 Staging Of RCC	1
1.2 A Suggested Solution - Radiogenomics	2
1.3 Effect Of VHL On ccRCC	3
1.4 Effect Of PBRM1 On ccRCC	4
1.5 Some Previous Machine Learning Approaches And Radiogenomics	5
1.6 Plan of Thesis	7
2. METHODS	9
2.1 Data	9
2.2 Experimental / Computational Method	9
2.3 Classification Tools	15
3. RESULTS	20
3.1 Classification Learner	20
3.1.1 For VHL Mutation Status	20
3.1.2 For PBRM1 Mutation Status	21
3.1.3 For Stage Status	22
3.2 Artificial Neural Network	24
3.2.1 For Stage Status	24
4. DISCUSSION	26
4.1 Radiogenomics Feature Extraction	26

4.2	Data Preprocessing	27
4.3	Extracting And Selecting Radiogenomics Features	27
4.4	Classification Process And Results	28
4.4.1	For Classification Learner	29
4.4.2	For ANN	30
4.5	Discussion Of The Results	30
5.	CONCLUSION	33
	APPENDIX A. HYPERPARAMETERS FOR CLASSIFICATION MODELS .	34
A.1	Hyperparameter Options For Classification Learner	34
A.2	Hyperparameter Options For ANN	34
	APPENDIX B. RADIGENOMICS FEATURES	35
B.1	ImageJ	35
B.2	MIPAV	35
B.3	LifeX	39
	REFERENCES	41

LIST OF FIGURES

Figure 1.1	Followed Path For This Work.	7
Figure 2.1	Patients With ccRCC. (a) The tumor which had PBRM1 mutation were shown in right kidney, was indicated by yellow line. (b) The tumor which had PBRM1 mutation were shown in left kidney, was indicated by yellow line. (c) The tumor which had VHL mutation were shown in right kidney, was indicated by yellow line. (d) The tumor which did not have VHL mutation were shown in left kidney, was indicated by yellow line.	10
Figure 2.2	Feature Weight/Feature Index For VHL Mutation Status Data.	13
Figure 2.3	Feature Weight/Feature Index For PBRM1 Mutation Status Data.	13
Figure 2.4	Feature Weight/Feature Index For Stage Status Data.	13
Figure 2.5	An Illustration Of A Neural Network Architecture.	18
Figure 2.6	Illustration Of ANN Process.	19
Figure 3.1	Confusion Matrix Of Classification Learner For VHL Mutation Status.	20
Figure 3.2	Number Of Observations For Classification Learner-VHL Mutation Status.	21
Figure 3.3	Confusion Matrix Of Classification Learner for PBRM1 Mutation Status.	22
Figure 3.4	Number Of Observations For Classification Learner-PBRM1 Mutation Status.	22
Figure 3.5	Confusion Matrix Of CL For Stage Status. A= Stage 1, B= Stage 2, C= Stage 3 and D= Stage 4.	23
Figure 3.6	Number Of Observations For Classification Learner - Stage Status. A= Stage 1, B= Stage 2, C= Stage 3 and D= Stage 4.	23

Figure 3.7	Confusion Matrices For Stage Status. (a) shows confusion matrix for training, (b) shows confusion matrix for validation, and (c) shows confusion matrix for testing.	24
Figure 3.8	Neural Network Training Performance-for Stage Status.	24
Figure 3.9	Minimizing Cross-Entropies (CE) And Percent Errors For Training, Validation Testing For Stage Status.	25
Figure 3.10	Illustration Of ANN For Stage Status.	25
Figure B.1	List Of Extracted Features From ImageJ.	36
Figure B.2	List Of Extracted GLCM Texture Features From ImageJ.	37
Figure B.3	List Of Extracted Features From MIPAV.	37
Figure B.4	List Of Extracted Features From MIPAV.	38
Figure B.5	Table For Advance Options For Radiomic Features In LifeX.	39

LIST OF TABLES

Table 1.1	Some of The Previous Works Including ML And Radiogenomics.	6
Table 2.1	Table For The Number Of Patients According To Gender, Mutation and Stage Information.	10
Table 2.2	Number Of The Data Before And After SMOTE Process.	12
Table 2.3	Features And Their P-Values For Classification Of VHL Mutation Status. *-Features From ImageJ, **-Features From LifeX, ***-Features From MIPAV.	16
Table 2.4	Features And Their P-Values For Classification Of PBRM1 Mutation Status. *-Features From ImageJ, **-Features From LifeX, ***-Features From MIPAV.	16
Table 2.5	Feature Names For Classification Of Stage Status. *-Features From ImageJ, **-Features From LifeX, ***-Features From MIPAV.	17
Table 3.1	Accuracies of models to predict mutations and stage status.	20
Table 3.2	Tables For Results And Hyperparameters Of Random Subspace And KNN Templates For PBRM1 Mutation Status.	21
Table 3.3	Tables For Results And Hyperparameters Of Random Subspace And KNN Templates For Stage Status.	22
Table A.1	Table For Hyperparameters Of Each Mutation Status And Stage.	34
Table A.2	Table For Hyperparameters for The ANN Model.	34

LIST OF SYMBOLS

<i>VHL</i>	In Tables, the data which contains VHL mutations
<i>NON – VHL</i>	In Tables, the data which does not contain VHL mutations
<i>PBRM1</i>	In Tables, the data which contains PBRM1 mutations
<i>NON – PBRM1</i>	In Tables, the data which does not contain PBRM1 mutations
%E	Percent Errors
CE	Minimizing Cross-Entropies
HIF	Hypoxia Inducible Factor
HIF- 1α	Hypoxia - inducible factor 1- alpha
HIF- 2α	Hypoxia - inducible factor 2- alpha
BX	BX is the coordinates of the upper left corner of the rectangle
XM	Center of mass-for x coordinate all pixels in the selection.
YM	Center of mass-for y coordinate all pixels in the selection.
Circ.	Circularity
MaxThr	Maximum Threshold
GLZLM	Grey-Level Zone Length Matrix
GLRLM	Grey-Level Run Length Matrix
NGLDM	Neighborhood Grey-Level Different Matrix
GLCM	Grey Level Co-occurrence Matrix
GLRLM_SRHGE	Short-Run High Gray-level Emphasis
GLZLM_SZE	Short-Zone Emphasis
Std.Dev	Standard deviation
GLZLM_SZHGE	Short-Zone Emphasis
GLRLM_LRLGE	Long-Run Low Gray-level Emphasis
GLRLM_LRE	Long-Run Emphasis
GLZLM_ZP	Zone Percentage

LIST OF ABBREVIATIONS

RCC	Renal Cell Carcinoma
ccRCC	Clear Cell Renal Cell Carcinoma
VHL	Von Hippel Lindau
PBRM1	Polybromo-1
ROI	Region Of Interest
ANN	Artificial Neural Network
CL	Classification Learner
TCIA	The Cancer Imaging Archive
TCGA	The Cancer Genome Atlas
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
VEGF	Vascular Endothelial Growth Factor
BAP1	BRCA1 Associated Protein-1
SETD2	Set Domain Containing 2
KDM5C	Lysine Demethylase 5C
BAF	BRG1-Associated Factor
ROI	Region Of Interest
VOI	Volume Of Interest
FSCNCA	Feature Selection Using Neighborhood Component Analysis
NCA	Neighborhood Component Analysis
SVM	Support Vector Machine
KNN	(K Nearest Neighborhood
SMOTE	Synthetic Minority Oversampling Technique
MIPAV	Medical Image Processing, Analysis, and Visualization
WHO	World Health Organization

1. INTRODUCTION

Renal Cell Carcinoma (RCC) is the most common renal malignancy and at least 3.7 % of new cancer cases are represented by this disease. Just in The United States, RCC accounted for an estimated 61.560 new patients and 14.080 deaths in 2015. It is known that certain genes which have mutations can activate intracellular molecular pathways. These specific pathways lead to an increased risk of specific histological subtypes of RCC. This knowledge has assisted us for better understanding the pathogenesis of RCC. RCC has been divided into subtypes related to genetic structure and mutation status. According to the WHO, there are eight major subtypes of adult-onset RCC and the most common one is Clear cell Renal Cell Carcinoma (ccRCC, 75%). Approximately 20% of patients have metastatic disease at presentation. More than half of the patients develop metastases after the initial diagnosis. Recently, the comprehension of the genetic base of RCC has prominently improved for researching the disease and using novel anticancer agents targeting specific intracellular pathways [1].

1.1 Staging Of RCC

The therapeutic approach to ccRCC is related directly to the stage or degree of tumor spreading. For example, Cabozantinib is a drug which targets the VHL pathway. The drug was approved in 2016 for patients with metastatic RCC whose tumors did not respond to their first treatment or whose cancer had returned [2]. Staging by TNM procedure hich including tumor information, node information and metastasis information was defined by The American Joint Committee on Cancer (AJCC) [3]. The size of the Stage 1 tumor is less than 7 centimeters or smaller. Both stage 1 and 2, the tumor is found in the kidney only. For Stage 2, the size of the tumor is greater than 7 centimeters. The tumor begins to spread in stage 3. It could be any size but could spread to nearby lymph nodes, the blood vessels in or near the kidney, structures of urine collection, or fatty tissue around the kidney. Metastasis could occur in stage

4. More than 50% of patients with early stages of RCC are cured, but the outcome for stage 4 of the disease is poor [4].

Since the standard treatment for kidney cancer related with the stage of cancer, if the patient is eligible for surgery and the tumor is small enough, partial nephrectomy may be a viable option. The kidney is spared with this operation, but the tumor and some of the surrounding tissue are removed. In more advanced cases, a full nephrectomy in which an entire affected kidney is removed may be necessary. In some cases, where surgery is not an option, cryoablation - the freezing of cancer cells may offer a solution if the tumor is solid and in a contained area. Targeted therapies are drugs that target specific receptors or molecules along the cancer cell growth pathways. These could make slower, or halt cancer growth if cancer has spread. However, it is a fact that chances of survival are much better when the disease is diagnosed before it has spread [4]. Therefore, early diagnosis remains a significant point for the patients [1].

Recently, experiences have shown that simple size measurements are inadequate. Even sometimes these measurements may cause to misleading for accurate response assessment with targeted therapies. More various features are required [1]. In addition, the current decision methods for detecting ccRCC involves an invasive procedure, namely a biopsy [4] which has technical drawbacks because of the manual extraction of the tumor. The biopsy always has the possibility of destroying the patient emotionally, damaging healthy tissue or causing the tumor to spread.

1.2 A Suggested Solution - Radiogenomics

Since these cancer patients race against time, the way of precise diagnosis and proper treatments based on the patients individually have become a requirement to acquire effective treatment results. Precision medicine represents a field that works on this specific problem [5].

Radiogenomics is one of the branches of precision medicine. It gives an opportu-

nity to make a connection between medical imaging and molecular biology and genetics area to generate new biomarkers. The first and also the most quoted study about relationship between medical CT imaging and mutation status of ccRCC belongs to Karlo et al [6]. The aim of the study was investigating associations between computed tomographic (CT) imaging features of ccRCC and mutations in VHL, PBRM1, SETD2, KDM5C, or BAP1 genes. According to the study, VHL mutation was significantly associated with well-defined tumor margins, enhancement of nodular tumor, and gross appearance of intratumoral vascularity. Solid ccRCC's genotype differed significantly from the genotype of multicystic ccRCC. It was also reported that VHL and PBRM1 mutations were more common among solid ccRCC [6]. The result of the study gave us valuable information about mutation types that we need to look at for ccRCC. Shinagare et al. also studied on radiogenomics for ccRCC. According to the study, VHL and PBRM1 were the most common mutation [7].

However, radiogenomics still remains an evolving area because it is a field related to many varied and developing fields of disciplines [8]. Moreover, standardization, overfitting, consistency of feature determination among readers are still problems with radiogenomics studies [7].

1.3 Effect Of VHL On ccRCC

The Von Hippel Lindau gene (VHL) is located on chromosome 3p25. A 213 amino acid the tumor suppressor protein which plays a key role in the regulation of the hypoxia response pathway is encoded by this gene. In low oxygen conditions, this pathway is vital for tumor survival [9]. The potential role of VHL mutation as a prognostic and predictive marker for RCC was also reported in [9]. The VHL gene codes the VHL protein. Under its standard functional state and under normoxic conditions, hypoxia-inducible transcription factors (e.g. HIF- α , HIF-2 α) are targeted by VHL complex for ubiquitin-mediated proteolysis. HIF regulates a host of significant downstream targets like the vascular endothelial growth factor (VEGF) which promotes angiogenesis, platelet-derived growth factor, and erythropoietin. Due to the fact that the VHL com-

plex cannot bind HIF- α for degradation, HIF- α accumulates under hypoxic conditions [10]. Therefore, RCC shows a proneness to be a vascular tumor with high expression of VEGF. It tends to be VEGF receptor, PDGF receptor, and basic fibroblast growth factor (bFGF). High VEGF expression is related to tumor aggressiveness and resulting in poor survival for RCC [11].

In addition, carbonic anhydrase IX maintains a significant role in pH regulation in cancer cells. It allows these cells for adaptation to the negative conditions of the tumor microenvironment. There is, furthermore, a study [12] indicated that low CAIX expression is associated with the absence of VHL alteration and aggressive tumor features. A significant prognosis in patients with ccRCC was reported.

Moreover, with the process of developing molecularly targeted therapies, the therapeutic landscape for RCC has changed in the last few years. There are four FDA approved drugs which are sorafenib, sunitinib, pazopanib, and axitinib for VEGF receptor inhibitors. And one drug for anti-VEGF monoclonal antibody which is Bevacizumab [10] to treat RCC. Plus, Cabozantinib is one of the tyrosine kinase inhibitors (TKI), which blocks VEGF. Considering the importance of VEGF, determining the situation of VHL mutation for RCC patients could help to lead the treatment options.

1.4 Effect Of PBRM1 On ccRCC

Polybromo-1 (PBRM1) gene is the second most common mutation and is seen 40% of these patients. The mutation encodes protein BRG1-associated factor (BAF) 180 [13]. It is a crucial distinct component of polybromo BAF SWI/SNF chromatin remodeling complex which is macromolecular types of machinery. It uses ATP to mobilize nucleosome. As a result, it affects the critical cellular process by regulating cell-cycle changes, metabolism, and DNA repair [14].

Although there is not gene-related FDA approved drug yet, there are studies indicating this gene is valuable because it has an impact on survival. One of the studies

about PBRM1 indicated that in patients with RCC, decreasing in PBRM1 expression is linked with a poor prognosis and advanced clinicopathological features [15]. Another study with RCC stage-4 patients was indicated that this gene could have potential as a prognostic marker for advanced RCC [16]. Moreover, other studies indicate PBRM1 mutation status has great potential to identify ccRCC [17], [18]. It was reported that PBRM1-mutated and BAP-mutated tumors exhibit different biology [18]. In addition to its noticeable effects on disease progression and being an indicator of the disease, if the specific role of PBRM1 in chromatin modification is taken into considering, this may positively affect to new treatment strategies in the future [19].

1.5 Some Previous Machine Learning Approaches And Radiogenomics

Some of studies which used machine learning methods and radiogenomics approaches were shown in Table 1.1. Using ANN with Matlab toolbox for classification was reported in [20]. The study used data source as National Instrument Hardware. Another study which was about predicting PBRM1 gene mutation for ccRCC was reported in [21]. They used The Cancer Genome Atlas as a data source with preoperative CT (CECT) images corticomedullary phase contrast-enhanced and artificial neural network (ANN) algorithm and a random forest (RF) algorithm for prediction. The study was conducted with a total of 45 ccRCC patients. Another study about to predict specific gene mutations of ccRCC using with their own model was reported in [22]. They used 57 ccRCC patients from two independent cohorts and a multi-classifier multi-objective predictive model for prediction. However the main disadvantage of these two studies was the low number of patients and not having enough of a large independent test dataset.

Using Matlab Classification Learner application to generate a model was reported in [23] and the study used the dataset which was based from the Trust-Hub. Besides supervised machine learning techniques with segmentations, a deep learning

method for classification of gene mutation was also reported in [24]. The study used TCIA database as data source and they compared their convolutional neural network model with a random forest model.

Table 1.1
Some of The Previous Works Including ML And Radiogenomics.

Name of The study	Year	Author	Data Source	Information About Data	Classification Model and software
Hand Motion Recognition From Single Channel Surface EMG Using Wavelet & Artificial Neural Network	2015	S.M. Manea et al.	Single channel data is acquired from National Instrument Hardware	three different hand motions of four healthy person.	ANN with MATLAB neural network toolbox.
Radiogenomics in Clear Cell Renal Cell Carcinoma: Machine Learning–Based High-Dimensional Quantitative CT Texture Analysis in Predicting PBRM1 Mutation Status	2019	Kocak B. Et al.	The Cancer Genome Atlas–Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) database - preoperative CT (CECT) images corticomedullary phase contrast-enhanced	45 patients. With PBRM1 mutation=16, w/o PBRM1 mutation=29.	an artificial neural network (ANN) algorithm and a random forest (RF) algorithm.
Hardware Trojan Identification Using Machine Learning-based Classification	2017	Noor et al.	The dataset was based from the Trust-Hub.	12 group of HTs	Decision Tree (DT), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) by using MATLAB Classification Learner Apps.
Noninvasive Determination of Gene Mutations in Clear Cell Renal Cell Carcinoma Using Multiple Instance Decisions Aggregated CNN	2018	Hussain et al.	CT datasets from the TCIA database	99VHL mutations, 51 PBRM1 mutations, 19 SETD2 mutations , 17 BAP1 mutations	CNN and RF
Reliable Gene Mutation Prediction in Clear Cell Renal Cell Carcinoma through Multi-classifier Multi-objective Radiogenomics Model	2018	Chen et al.	From an institutional review board-approved, Health Insurance Portability and Accountability Act-compliant (HIPAA) and TCIA	57 ccRCC patients from two independent cohorts	A proposed new multi-classifier multi-objective predictive model

Both machine learning and radiogenomics areas are constantly evolving and new methods are emerging to the fields. Therefore, different pathways like choosing multiple slice of ROI could become a shortcoming. Also, using deep learning with medical images may have disadvantages for nowadays since deep learning needs lots of images for obtaining solid results [25]. Besides deep learning, using random forest technique may lead to some defects like instability, especially when there is noisy and / or unbalanced data [26]. Moreover, the consistency of feature determination among readers seems like another essential point to be needed attention.

1.6 Plan of Thesis

Our goal was to develop an alternative solution to these problems in the future by using a machine learning (ML) algorithm with quantitative CT texture analysis that can be implemented to detect gene mutations and stages of ccRCC to increase the speed of diagnosis. In addition, detecting specific mutations and stages via an intelligent system may be a time-saver and affect the treatment options.

We hypothesized that with quantitative 2D CT images, VHL and PBRM1 mutations and stage status of ccRCC can be predicted with precision and accuracy. Deep learning methods require more data for training [25]. Considering the drawbacks of deep learning, we aimed to use artificial neural networks (ANNs) to predict stages of ccRCC from 2D CT images. The results were compared, and it is novel for the area of radiogenomics on ccRCC. Various ML models have been studied. The flowchart of this study is shown in Figure 1.1.

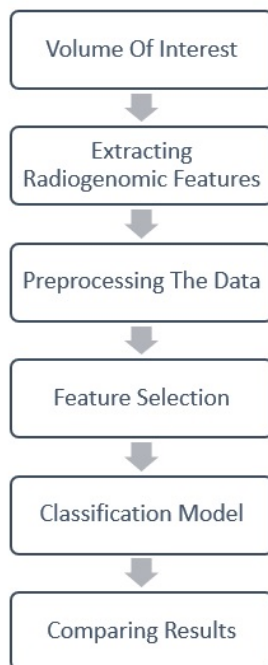


Figure 1.1 Followed Path For This Work.

The workflow of radiogenomics includes three fundamental components: data measurements (which include gathering clinical data of histopathological characteriza-

tion and imaging results), data preprocessing (quality control and preprocessing steps) and the analysis process (subsets of features) [8].

Our approach was based on machine learning (ML) techniques. Features such as shape textural properties and related to the relationships between image voxels were utilized to develop learning models. Then, these models were used for identifying tumor characteristics to properly classify specific gene mutation status previously mentioned and stages. The selected features for producing correlations to stage level or specific gene mutations were used as inputs for classification models. Once the training step was complete, it was the time to see if the model was any good by tuning the adjustable parameters according to the results.

2. METHODS

2.1 Data

There is an ongoing project called Cancer Genome Atlas (TCGA). The collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) has provided comprehensive maps of key genomic changes in 33 identified cancer types, including ccRCC [27],[28]. TCGA-KIRC data set which has the disease type Adenomas and Adenocarcinomas of Kidney was used for this study. "The results <published or shown> here are in whole or part based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>".

2.2 Experimental / Computational Method

Our aim was to develop an algorithm which will improve the accuracy of early diagnosis of ccRCC in the future. To achieve the aim, first, we needed preparing the data for the learning process. Preprocessing the data which were used as an input was a crucial step to improve the efficiency of neural network training.

How much data and type of data required were decided by taking into consideration scanning the data of ccRCC. The data to be studied were selected by an expert radiologist. Image quality, the effect of noise in the image, image modality (since the data included both CT and MR modalities) and nephrectomy criteria were taken into consideration during data selection. Accordingly, the study was performed with 191 patients data. We gathered and labeled the axial CT images to have training data for the network. The numbers of patients which were linked to the gender information, mutation status and stage information were shown in Table 2.1.

After data acquisition process, the tissue of interest was carefully delineated

Table 2.1

Table For The Number Of Patients According To Gender, Mutation and Stage Information.

Total Number Of Patient	191
Female Patients	69
Male Patients	122
Patients With VHL Mutation	81
Patients With PBRM1 Mutations	63
Stage 1 Patients	92
Stage 2 Patients	19
Stage 3 Patients	50
Stage 4 Patients	30

by the radiologist with over 10 years old experiences using ImageJ [29] software. The slices were considered to include the largest tumor area for each patient obtained from TCGA-KIRC data. Sample ROI's are shown in Figure 2.1.

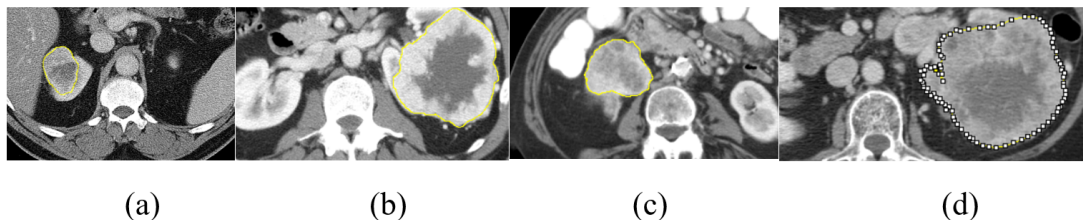


Figure 2.1 Patients With ccRCC. (a) The tumor which had PBRM1 mutation were shown in right kidney, was indicated by yellow line. (b) The tumor which had PBRM1 mutation were shown in left kidney, was indicated by yellow line. (c) The tumor which had VHL mutation were shown in right kidney, was indicated by yellow line. (d) The tumor which did not have VHL mutation were shown in left kidney, was indicated by yellow line.

After the region of interests (ROIs) was drawn, Radiogenomics features were extracted from the ROI. These included gray level patterns, inter-voxel relationships, shape and texture features. In this step, 136 radiographic features were generated and a high dimensional feature matrix was created.

In ImageJ software [29], the Analyze section was used for feature extraction. In addition, the Texture Analyzer plugin was used for GLCM-Texture features [30]. In MIPAV software [31], features were extracted from images by operating Statistic

Generator section. Selecting VOI was used as an input for Statistic generator. In LifeX [32], Texture section was operated for feature extraction. Radiogenomic features include first-order features (shape and histogram), GLRLM, GLZLM, GLCM, and NGLDM. Visuals of corresponding sections of software and all extracted features were shown in APPENDIX B.

After feature extraction from the software, a high dimensional matrix was generated, and the presence of missing data and non-quantitative data was checked by using the MATLAB code which is below:

[R,TF] = rmmissing(_) R represented the data with missing entries removed and TF represented removed entry indicator. With this step, 114 out of 136 features were left.

After the checking missing features and non-numeric features process, the data were scaled to the standard deviation for normalization.

N = normalize (A). This algorithm returns the vectorwise z-score of the data in A with center 0 and Std.Dev 1. In our cases, A is the matrix which includes quantitative features and this algorithm was used for operating on each column of data separately for normalization.

Unbalanced data were [33] handled for balancing by using the ADASYN algorithm which represents an extended version of SMOTE. It was used to decrease class imbalance by synthesizing minority class examples. The purpose of the algorithm was to produce more examples around the boundary between the two classes than within the minority class. Synthetic examples were generated by using linear interpolation between the majority class data and related minority class data. The related MATLAB code is given below:

function [out_featuresSyn, out_labelsSyn] = ADASYN [in_features, in_labels, in_beta, in_kDensity, in_kSMOTE, in_featuresAreNormalized]

The explanations of each section could be found in [33]. The default numbers were stated in [33] for `in_beta`, `in_kDensity`, `in_kSMOTE` and `in_featuresAreNormalized`. The number of data after the algorithm applied for the unbalanced problem are shown in Table 2.2.

Table 2.2
Number Of The Data Before And After SMOTE Process.

CASE	BEFORE SMOTE	AFTER SMOTE
VHL	81	97
NON-VHL	110	110
PBRM1	63	131
NON-PBRM1	128	128
STAGE I	92	92
STAGE II	19	92
STAGE III	50	90
STAGE IV	30	96

In ML and statistical analysis, reducing the dimension of features is described as an essential procedure since a combination of many features and a limited number of observations is useless for producing the desired learning result. This situation may result in the learning algorithm to overfit. By reducing the number of features, more storage space can be used, and computation time can be saved. Feature selection was used for this study [34].

After radiogenomics features were calculated, feature selection algorithms were required to find out whether features can separate specific gene mutation and stage status. Two sample T-test statistical analysis was selected as the first step of feature selection to classify the mutation status ($p < .05$) [35]. The related MATLAB code is

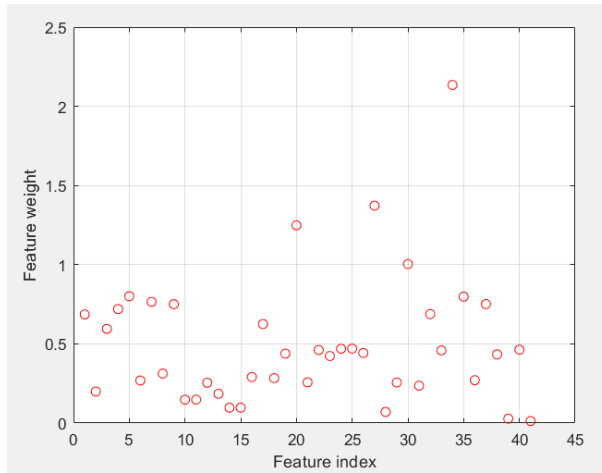


Figure 2.2 Feature Weight/Feature Index For VHL Mutation Status Data.

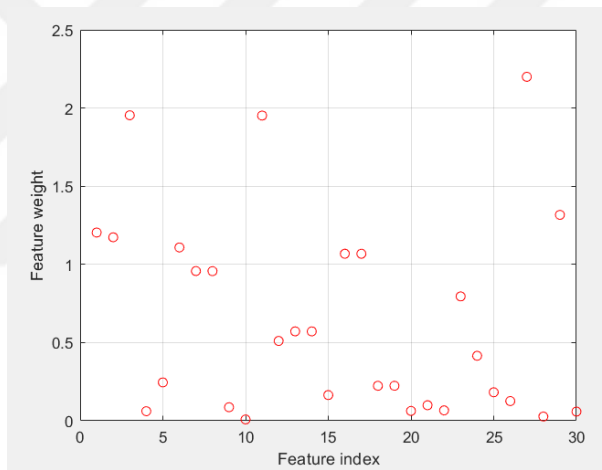


Figure 2.3 Feature Weight/Feature Index For PBRM1 Mutation Status Data.

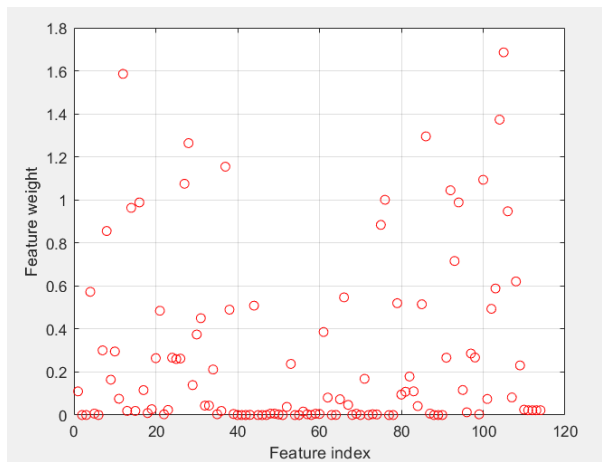


Figure 2.4 Feature Weight/Feature Index For Stage Status Data.

given below:

```
dataTrain1 = obs(grp2idx(mutType)==1,:);
```

```
dataTrain2 = obs(grp2idx(mutType)==2,:);
```

```
[h,p,ci,stat] = ttest2 (dataTrain1,dataTrain2,'Vartype','unequal');
```

obs represented each column of data which means features and it was a numeric matrix **mutType** was cell array type of data for mutation status and represented one column. **grp2idx** was used to create an index vector from the grouping variable. **p** represented p-value for each column.

The t-test was applied to each feature, and the p-values were compared for each feature to understand how effective the features for separating the groups. For VHL gene data, there were 41 features with a p-value less than 0.05 and for PBRM1 gene data, there were 30 features.

Feature selection using neighborhood component analysis (FSCNCA) was the second step to define the final features for mutation classification. Also, it was the first step for reducing the dimension of the feature matrix and defining final features for stage status. The working principle of FSCNCA is learning the weights of feature by implementing a diagonal adaptation of NCA with regularization [36]. The related MATLAB code is given below:

```
mdl = fscnca(X,y,'Solver','sgd','Verbose',1);
```

X represented the data which had the desired p values and y represented cell array of the data which indicated the type of the data. For example for VHL mutation, cell array was like [VHL, VHL, VHL, NON, VHL, NON, VHL...]. For stage data, X represented the whole feature matrix and y represented cell array of the data which indicated the type of stage. For plotting the selected features, the related MATLAB

code which was used is given below:

```
figure()  
  
plot mdl.FeatureWeights,'ro')  
  
grid on  
  
xlabel('Feature index')  
  
ylabel('Feature weight')
```

In Figure 2.3, 2.4 and 2.5, feature weights/feature index graphs are shown for VHL mutation, PBRM1 mutation, and stage status. The weights of features which were irrelevant close to zero. The weights which had > 0.5 were considered for creating subsets of features. For VHL gene data, the number of features obtained using this method was 10. The name of the features and their p-values are shown in Table 2.3. For PBRM1 gene data, the number of features obtained using this method was 9. The name of the features and their p-values are shown in Table 2.4. And for stage data, the number of the features obtained using this method was 10. The name of features and their p-values are shown in Table 2.5.

2.3 Classification Tools

After the feature selection process for each specific gene mutation and stages, classifications were performed by using MATLAB classification learner (CL) [37]. Also ANN on MATLAB was used for stage classification [38] to train a model and see the accuracy.

The selected features were used as inputs for the classification models to produce correlations to the status of gene mutations and stages. The confusion matrix was used

Table 2.3

Features And Their P-Values For Classification Of VHL Mutation Status. *-Features From ImageJ, **-Features From LifeX, ***-Features From MIPAV.

Feature Name	P-Value
Circularity***	0.019296
GLZLM_SZHGE**	0.011092
GLRLM_LRLGE**	0.005773
GLZLM_ZP**	0.007101
Major*	0.019428
Area*	0.019152
Height*	0.027418
Feret*	0.016682
stdValue**	0.014597
GLRLM_LRE**	0.006249

Table 2.4

Features And Their P-Values For Classification Of PBRM1 Mutation Status. *-Features From ImageJ, **-Features From LifeX, ***-Features From MIPAV.

Feature Name	p-value
Circularity***	0.009752378
YM*	0.002295351
minValue**	0.010231968
Std Dev of Intensity***	0.017907719
FeretAngle*	0.008218469
PARAMS_YSpatialResampling**	0.013848689
PARAMS_XSpatialResampling**	0.013848689
MaxThr*	0.033478727
GLZLM_SZE**	0.020053672

Table 2.5
Feature Names For Classification Of Stage Status. *-Features From ImageJ, **-Features From LifeX, ***-Features From MIPAV.

Feature Names
Mean Curvature***
Number of Indentations Curvature***
Circ.*
Fractal Dimension Box Count***
GLZLM_SZHGE**
Eccentricity***
XM*
BX*
GLRLM_SRHGE**
Median Intensity***

to check for comparing the results visually. It yielded the percentages of correct and incorrect classifications for both classification methods. Correct classifications were represented as the green squares on the matrix diagonal.

Classification learner (CL) is an application which is available on MATLAB R2019a to solve classification problems. It was used to classifying data in terms of supervised machine learning by using varied classifiers. Supervised learning was a learning type which learns the mapping function from the pre-known input-output data. The pre-known data were labeled and set for being a response variable. The extracted features were set as observations. Validation selection was set to 5-cross-validation. Models were compared by using the accuracy of the trained models. The model type was k-NN with Random Subspace for PBRM1 mutation and stage cases. The main principle of ensemble methods was based on combining the results from weak learners and turning them a high-quality ensemble model by ensemble classifiers. k-NN with Random Subspace used a typical KNN model to produce an ensemble model. KNN is a good way for prediction accuracy in low dimension data. It categorizes query points based on their distance to points or in this case, neighbors to classify new points. The model type which yielded the highest accuracy was Fine Gaussian SVM for VHL mutation status. SVM is a type of supervised classifiers which is known as a

kernel-based algorithm. The working principle is based on finding the best hyperplane, the largest margin between the two classes. The hyperplane separates one class' data points from the other classes [39]. After the initial results, hyperparameters of classifiers were optimized in CL. The hyperparameters which were used for CL are shown in APPENDIX A.

Lastly, the ANN method was utilized to classify patients related to stage information. ANN **nnstart** was a graphical user interface (GUI) on R/MATLAB2019a. A window was opened and pattern recognition and classification option was chosen. The desired number of hidden neurons could be decided by hand in it. Also, the percentage of training, testing and validating data could be decided to create a model and get results. For performing the identifying and classifying the data, neural networks were especially well suited. An input layer which was the selected features in our case, hidden layers, and an output layer which represented four stage status in our case, were included in ANN. The layers were interconnected via nodes. The output of the previous layer was implemented by each layer as its input (Figure 2.5).

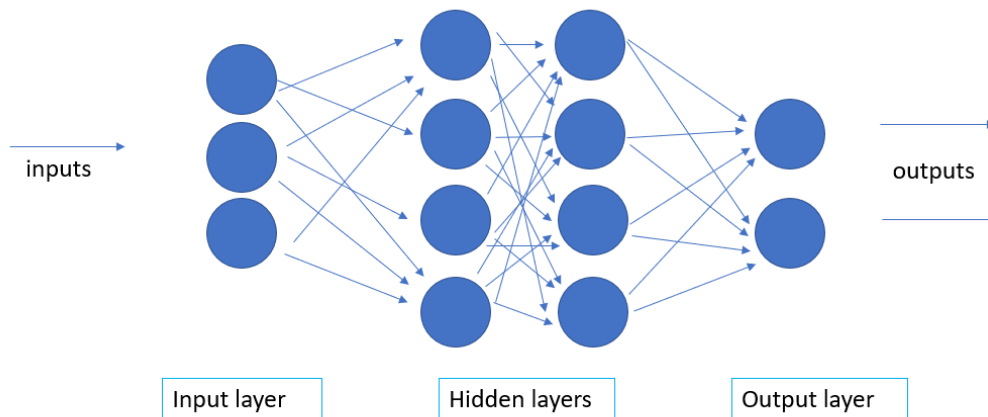


Figure 2.5 An Illustration Of A Neural Network Architecture.

80% of the data were randomly selected for training. 10% of the data for validation and 10% of the data for testing were randomly selected. The number of hidden neurons was shown in APPENDIX A, Table A.2. An illustration of the neural network as shown in Figure 2.6.



Figure 2.6 Illustration Of ANN Process.

3. RESULTS

Our results showed that using classification learner, Fine Gaussian SVM model can correctly predict VHL and NON-VHL data with 68.6% overall accuracy, and k-NN with Random Subspace model can correctly predict PBRM1 and NON-PBRM1 data with 84.9% overall accuracy. ANN model can correctly predict stages with 91.90% accuracy respectively (Table 3.1).

Table 3.1
Accuracies of models to predict mutations and stage status.

Case	Model	Accuracies
VHL	Fine Gaussian SVM	68.6%
PBRM1	k-NN with Random Subspace	84.9%
STAGE	k-NN with Random Subspace	85.4%
STAGE	ANN	91.9%

3.1 Classification Learner

3.1.1 For VHL Mutation Status

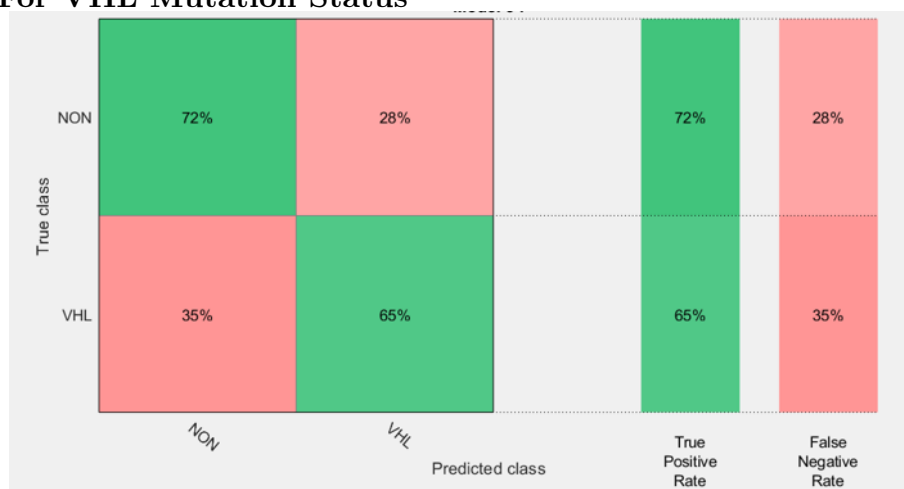


Figure 3.1 Confusion Matrix Of Classification Learner For VHL Mutation Status.

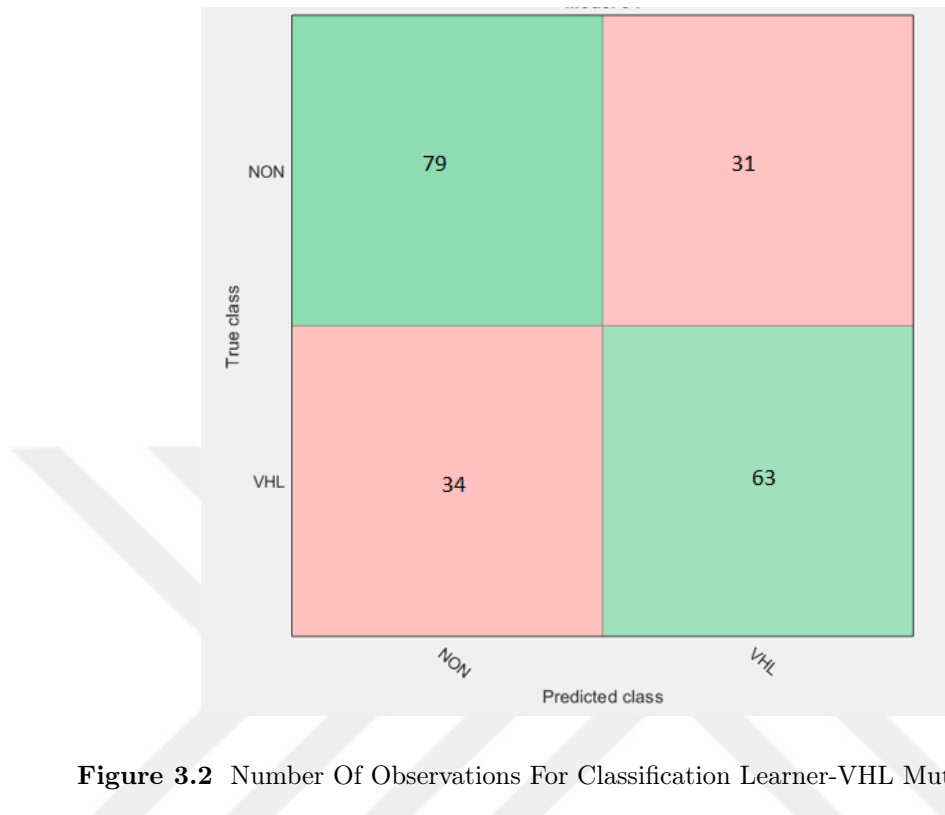


Figure 3.2 Number Of Observations For Classification Learner-VHL Mutation Status.

3.1.2 For PBRM1 Mutation Status

Table 3.2

Tables For Results And Hyperparameters Of Random Subspace And KNN Templates For PBRM1 Mutation Status.

Num.Of.Neighbor	Accuracy	Cross-Validation	Distance Metric	Distance Weight	NumLearningCycles	Subspace Dimension	Weighted KNN Accuracy
k=1	83.80%	5	Euclidean	Equal	30	6	77.80%
k=3	84.90%	5	Euclidean	Equal	33	6	78.40%
k=4	83.80%	5	Euclidean	Equal	33	5	79.20%
k=5	83.10%	5	Euclidean	Equal	30	5	78.00%

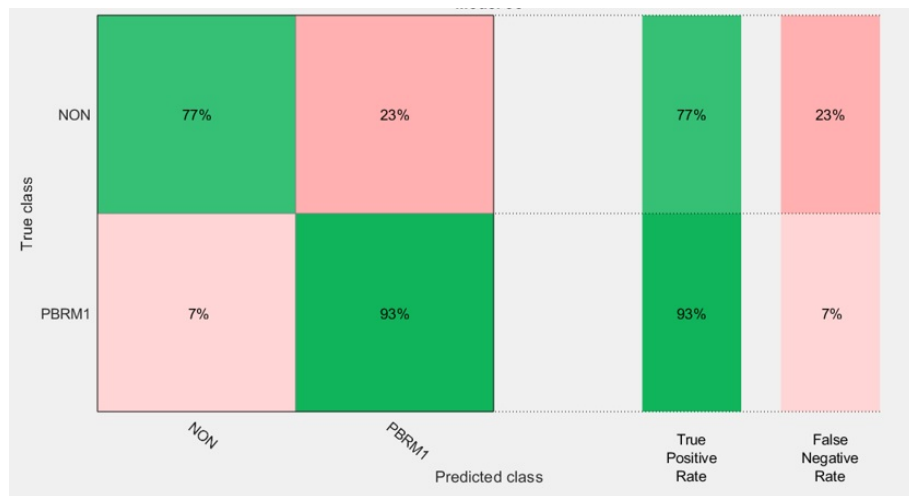


Figure 3.3 Confusion Matrix Of Classification Learner for PBRM1 Mutation Status.

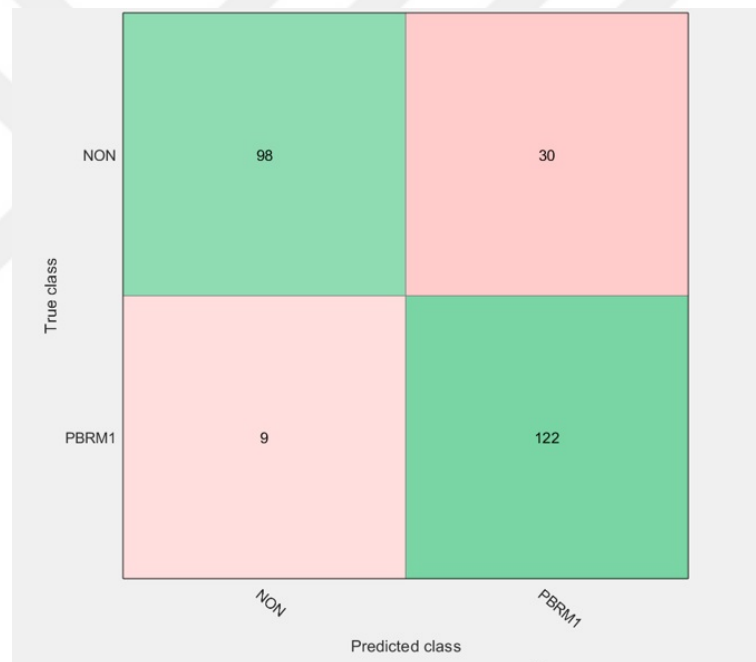


Figure 3.4 Number Of Observations For Classification Learner-PBRM1 Mutation Status.

3.1.3 For Stage Status

Table 3.3

Tables For Results And Hyperparameters Of Random Subspace And KNN Templates For Stage Status.

Num.Of.Neighbor	Accuracy	Cross-Validation	Distance Metric	Distance Weight	NumLearningCycles	Subspace Dimension	Weighted KNN Accuracy
k=1	83.80%	5	Euclidean	Equal	30	6	81.30%
k=3	86.80%	5	Euclidean	Equal	36	5	78.60%

True class	A	78%	7%	11%	4%	78%	22%
	B		96%	3%	1%	96%	4%
	C	9%	6%	74%	11%	74%	26%
	D	2%	3%	2%	93%	93%	7%
		A	B	C	D	True Positive Rate	False Negative Rate
		Predicted class					

Figure 3.5 Confusion Matrix Of CL For Stage Status. A= Stage 1, B= Stage 2, C= Stage 3 and D= Stage 4.

True class	A	72	6	10	4
	B		88	3	1
	C	8	5	67	10
	D	2	3	2	89
		A	B	C	D
		Predicted class			

Figure 3.6 Number Of Observations For Classification Learner - Stage Status. A= Stage 1, B= Stage 2, C= Stage 3 and D= Stage 4.

3.2 Artificial Neural Network

3.2.1 For Stage Status

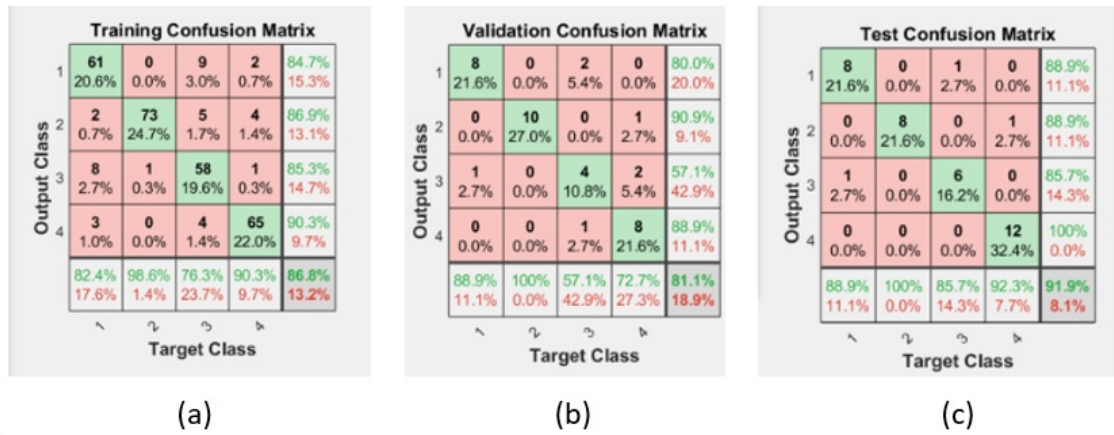


Figure 3.7 Confusion Matrices For Stage Status. (a) shows confusion matrix for training, (b) shows confusion matrix for validation, and (c) shows confusion matrix for testing.

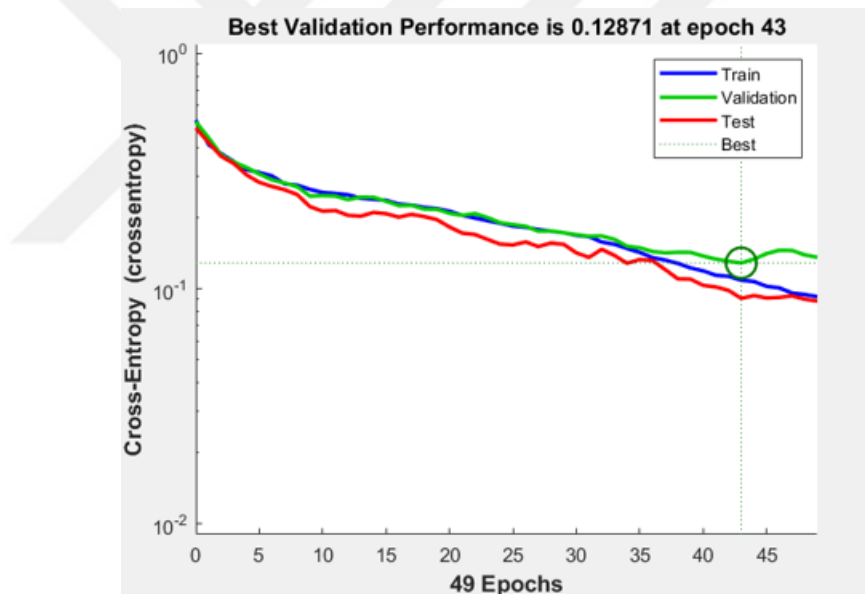


Figure 3.8 Neural Network Training Performance-for Stage Status.






	 Samples	 CE	%E icon" data-bbox="688 261 703 276"/> %E
 Training:	296	7.87279e-1	13.17567e-0
 Validation:	37	3.20567e-0	18.91891e-0
 Testing:	37	3.18817e-0	8.10810e-0

Figure 3.9 Minimizing Cross-Entropies (CE) And Percent Errors For Training, Validation Testing For Stage Status.



Figure 3.10 Illustration Of ANN For Stage Status.

4. DISCUSSION

The aim of this project was to develop an algorithm which may aid for the accuracy of early diagnosis of ccRCC in the future by using features of radiogenomics and ML. We examined a population which has been taken TCGA [27],[28]. The data was examined by both authors and a radiologist who has over 10 years of experience. This examination provided to us the number of images which include only CT images and the data without renal nephrectomy.

To begin with, the data were extracted by its quality, absence of kidney and presence of MR images. With 191 patients, ImageJ [29] toolbox was employed to acquire ROI according to the biggest tumor on each axial slice of CT images.

4.1 Radiogenomics Feature Extraction

When ImageJ was used, the analyze menu was used to measuring, calculating and displaying for area statistics based on ROI [29]. Texture Analyzer plugin was additionally used for acquiring GLCM texture features [30]. In APPENDIX B, the extracted features using ImageJ were listed.

When MIPAV was used, the Statistics Generator section was used for feature extraction which includes statistical and morphological features. The Statistics Generator automatically saved the results in a tab-delimited file [31]. In APPENDIX B, the extracted features using MIPAV were listed.

When LifeX was used, first-order features and texture features were calculated using Texture Feature Extraction section [32]. The results were stored as an Excel file named by default patientName-patientID-Texture-time.xls. In APPENDIX B, the extracted features using LifeX were listed.

114 out of 136 features were quantitative and related to 2D images. The results of these quantitative features were properly registered with the name of ALL-q-features.xlsx as an Excel file and as a result, we had a high dimensional matrix.

4.2 Data Preprocessing

Data preprocessing step had importance to acquire robust results. After the data were scaled to the standard deviation for normalization, we had unbalanced data problem since the observation distribution of both mutation status and stages were unequal. Unbalanced data may lead to misclassifying, and this could bring about severe consequences for training and accuracy [34]. To solve the imbalance problem, ADASYN algorithm [33] was used to solve class imbalance by synthesizing minority class examples. For example, to produce synthetic Stage 2 data, Stage 1 data was used as a majority class. The results of this step were shown in Table 2.2.

4.3 Extracting And Selecting Radiogenomics Features

After the data preprocessing steps, we had to reduce the dimension of our matrix to acquire robust results. Having a high dimensional matrix may lead the learning algorithm to overfit and cause noise. Besides helping overfitting and noise problems, reducing features can also aid to save storage for a CPU and computation time for training. Therefore, feature selection algorithms were required to accelerate the process and determine the subsets of features which have any specific gene information and stage status in common. The first step for reducing the size of the matrix and feature selection was applying the T-test statistical analysis ($p < .05$) to [35] both mutational status data. For each mutation status, features whose p-values were smaller than 0.05 were extracted and saved.

FSCNCA was applied as the second step of feature selection to define the final

features for classification of gene mutation status [36]. Plus, it was the first step for generating a subset of features for stage classification. Feature weight/Feature index plot for VHL mutation status data was shown in Figure 2.2. Feature weight/Feature index plot for PBRM1 mutation status data was shown in Figure 2.3. Feature weight/Feature index plot for stage status data was shown in Figure 2.4. The weights of the irrelevant features were close to zero. Feature weights equal, near or higher than 0.5 were considered. 10 features for VHL mutations, 9 features for PBRM1 mutations and 10 features for classifying stages were used. Final features for VHL status were shown in table 2.3, final features for PBRM1 status were shown in Table 2.4 and final features for stage status were shown in table 2.5 and these final features were used as inputs for classification of each mutation status and stages. Before the classification process, histograms of each feature were also examined on Excel program for quality control.

4.4 Classification Process And Results

To train models and obtaining accuracy, Classification Learner App [37] and ANN [38] on MATLAB R2019a [40] were used. These methods produced the results which gave eligible correlations to the status of gene mutations and stages. The confusion matrix was used to check for comparing the results. Correct classifications were shown as green squares on the matrix diagonal.

First, classification learner (CL) was used to classifying data according to both mutation status and stages. Its working principle was based on supervised machine learning with many classifiers. 5-cross-validation was selected for CL. Then, ANN was used for predicting stage status. ANN was a computing model whose layered structures act like the networks of the neurons in the brain and it was used for stage classification too. The hidden neurons were connected via nodes with both inputs and outputs sections. Labeled data were set as a response variable and extracted features were set as predictors. Since there are no fix criteria for the ratio of training validation and testing data; the data were set randomly %80, %10 and %10 for training, validation and testing respectively. The hyperparameters which were used for both CL and ANN

were shown in APPENDIX A. Moreover, when classifying PBRM1 mutation status, 4 number of neighbor(k) were examined to obtain the best accuracy. According to this examination, the best accuracy for PBRM1 mutation status was obtained from k=4. The hyperparameters for both ensemble method and nearest neighbor method and the results of these methods were shown in Table 3.2.

4.4.1 For Classification Learner

1-VHL MUTATION STATUS:

Fine Gaussian SVM was the model type which yielded the highest accuracy for VHL mutation status. The best hyperplane was found by SVM. Figure 3.1 shows the confusion matrix of CL for VHL mutation status. VHL data were estimated 65% accuracy and NON-VHL data was estimated 72% accuracy. Figure 3.2 shows the number of observations for VHL status on CL. Overall accuracy was 68.6% and training time was 0.78476 sec. Prediction speed was ~12000 obs/sec.

2- PBRM1 MUTATIONAL STATUS:

The model type which yielded the highest accuracy was k-NN with Random Subspace for PBRM1 mutation status. k-NN with Random Subspace was one of the ensemble classifiers in CL. An ensemble classifier combines the results of many weak learners and creates one high-quality ensemble model. The confusion matrix of CL for PBRM1 mutation status including true positive rate and false negative rate was shown in Figure 3.3. PBRM1 data were estimated 93% accuracy and NON-PBRM1 data was estimated 77% accuracy. Figure 3.4 shows the number of observations for PBRM1 status on CL. Overall accuracy was 84.90% and training time was 1.957 sec. Prediction speed was ~830 obs/sec.

3- STAGE STATUS:

The model type which yielded the highest accuracy was k-NN with Random Subspace for stage status. Figure 3.5 shows the confusion matrix of CL for stage status. Stage1 data were estimated 78% accuracy, Stage2 data was estimated 96% accuracy, Stage3 data was estimated 74% accuracy and Stage4 data was estimated 93% accuracy. Figure 3.6 shows the number of observations for stage status on CL. Overall accuracy was 85.40% and training time was 3.3917 sec. Prediction speed was ~ 850 obs/sec.

4.4.2 For ANN

Confusion Matrix for training, validation, and testing for stage status were shown in Figure 3.15. The test result was 91.9% accuracy. Figure 3.16 showed Neural Network Training Performance-for stage Status. Best validation performance was 0.12871 at epoch 43. Minimizing Cross-Entropies (CE) and Percent Errors for training, validation, and testing for stage status were shown in Figure 3.17. CE was $3.18817e-1$ and %E was $8.10810 e-0$. Illustration of ANN-for stage status was shown in 3.18.

4.5 Discussion Of The Results

This was the first study with using k-NN random subspace. It was observed that classification learner (CL) k-NN with Random Subspace gave higher accuracy for status of PBRM1 mutation and stages. Also, using the ANN method improved the results for stage status. We consider that the size of the dataset may have caused the situation. When comparing the size of our data after handling unbalanced data, the least difference occurred for VHL mutation data. When comparing the numbers, the number of PBRM1 mutation data after the process was way smaller than stage status data.

There was no such study related study with staging status of ccRCC in the literature during this study. ANN and k-NN random subspace methods were used

for classification. With consideration of disadvantages of number of neighbor equals 1 (most likely to overfit), we also examined one more different number of neighbor for both ensemble method and KNN method. Also, we did the same process for PBRM1 mutation status. The hyperparameters and the results of the models for stage status were shown in table 3.3. $k=3$ was the most eligible model for prediction of stage status of ccRCC.

After stage status, the most precise accuracy was observed when examining PBRM1 mutation status. The hyperparameters and the results of the models for PBRM1 mutation status were shown in table 3.2. When the results were examined, $k=3$ was the most eligible for the accuracy. Our results showed the patients could be classified according to their PBRM1 mutation status with 84.9% accuracy. Figure 3.3 showed that the confusion matrix of predicted classes. 90% true positive rate for patients with PBRM1 mutations and 77% true positive rate with patients who do not have PBRM1 mutations were observed.

And when we looked at VHL mutation status, Fine Gaussian SVM model did not provide sufficient accuracy. It cannot be said that Fine Gaussian SVM model gave a random prediction, but also the results were not adequate when comparing them with results of PBRM1 mutation status and stage status. Our results showed the patients could be classified according to their VHL mutation status with 68.6% accuracy. Figure 3.1 showed that the confusion matrix of predicted classes. 65% true positive rate for patients with VHL mutations and 72% true positive rate with patients who do not have VHL mutations were observed.

While k -NN with Random Subspace looks like an alternative tool for classification of PBRM1 mutation status and stage status, nevertheless ANN yielded better results to predict stages. It was observed that ANN improved the results which were acquired from CL for predicting stage status and it seems that ANN could be an option to use for stage classification in the future.

Moreover, we also consider the accuracy may have affected because of the reg-

istration faults in ROI, and lack of the number of observations. Also, it could occur due to the data acquisition process in TCIA or diversity of contrast agents. More data are needed for further research. In addition, the studies indicate that the focus of radiogenomics may merge with 3D imaging and deep learning in the future. It is necessary to work with more data for the development of radiogenomics field and its progress using machine learning. Our work shows that implementing a single slice with the biggest tumor with quantitative 2D CT features have the potential to predict VHL and PBRM1 mutation status and stages of ccRCC. This could be used as an alternative solution to the storage problem in the future. Working with a single slide for each patient could help reduce costs and improve the performance of computers. Therefore, it can encourage investors to invest in this field and to further research.

5. CONCLUSION

We conclude that ML-based quantitative 2D CT analysis with the biggest tumor for each patient's axial CT image may be helpful for predicting VHL and PBRM1 mutation status and stage status in patients with ccRCC. Both Fine Gaussian SVM and k-NN with Random Subspace models in CL and ANN were useful in classification. In the future, our method could be investigated for other ccRCC patients' CT images. A certain pathway could be generated to use daily. Optimal selection of therapy depends on genetic data (precision medicine) could be available with these types of studies. Since RCC is a type of disease which generally does not respond well to radiotherapy, by combining studies like ours, drugs related to specific gene mutation could be developed to stop or change the gene mutations. Therefore, the prognosis of the disease like ccRCC could be changed or disease progression can be slowed down. During the process of this study, there was no similar study in the literature which used 2D and single slice with the biggest tumor for each patient. Considering the operating and CPU capacities of computers, there may be more advantages to working with a single-slice axial CT over full CT scanning. More models, features and optimization schemes may be researched in the future.

APPENDIX A. HYPERPARAMETERS FOR CLASSIFICATION MODELS

A.1 Hyperparameter Options For Classification Learner

Table A.1

Table For Hyperparameters Of Each Mutation Status And Stage.

Final Hyperparameters For CL	Case
Preset: Fine Gaussian SVM, Kernel Function: Gaussian, Kernel Scale: 1.3, Box Constraint Level:3.	VHL status
Fit template for classification KNN: Number of Neighbour: 3, Distance: Euclidean, Distance Weight: Equal, Ensemble Method: Subspace, Learner Type: Nearest Neighbours, Number of Learners: 33, Subspace Dimension:6.	PBRM1 Status
Fit template for classification KNN: Number of <u>Neighbour</u> : 3, Distance: Euclidean, Distance Weight: Equal, Ensemble Method: Subspace, Learner Type: Nearest <u>Neighbours</u> , Number of Learners: 36, Subspace Dimension:5.	Stage Status

A.2 Hyperparameter Options For ANN

Table A.2

Table For Hyperparameters for The ANN Model.

Number Of Hidden Neurons	Performance	Training
Default:10, Chosen:15	Cross-Entropy	Scaled Conjugate Gradient

APPENDIX B. RADIGENOMICS FEATURES

B.1 ImageJ

Image J was employed for both segmentation and radiogenomics feature extraction in this project. The first-order statistical distribution of the voxel intensities within the tumor was described by the intensity features. The patterns or the second and high-order spatial distributions of the intensities were described by the texture features [41]. Besides the first order features, texture feature plugin which computed several of the texture parameters described by Haralick [42]. The name of the extracted features was listed in figure B.1 and B.2.

B.2 MIPAV

In MIPAV software, features were extracted from images by using Statistic Generator section. ROI, which were drawn in ImageJ, was converted to mask and then in MIPAV the mask converted to VOI. The name of the extracted features was listed in figure B.3 and B.4.

Descriptions of some features were explained below:

Eccentricity: the geometric shape of the VOI as an ellipse, with 0 indicates a circle and 1 indicates a straight line.

Std. dev. of voxel intensity: calculating the standard deviation of the intensity of the voxels in the VOI.

For more extensive information about other features, MIPAV web site (<https://mipav.cit.nih.gov/>) could be reviewed.

Extracted Feature

Names
Area
Mean
Std Dev
Mode
Min
Max
X
Y
XM
YM
Perim.
BX
BY
Width
Height
Major
Minor
Angle
Circ.
Feret
IntDen
Median
Skew
Kurt
%Area
RawIntDen
Slice
FeretX
FeretY
FeretAngle
MinFeret
AR
Round
Solidity
MinThr
MaxThr

Figure B.1 List Of Extracted Features From ImageJ.

GLCM_Texture_Features

Angular Second Moment

Contrast

Correlation

Inverse Difference Moment

Entropy

Sum of all GLCM elements

Figure B.2 List Of Extracted GLCM Texture Features From ImageJ.

Extracted Features From MIPAV- Part1

of Voxels

Area (Millimeters²)

Perimeter (mm)

Circularity

Solidity

Min Intensity

Max Intensity

Avg Voxel Intensity

Std Dev of Intensity

Sum Intensities

Geometric centerX

Geometric centerY

Center of MassX

Center of MassY

Principal Axis with Minimum Moment of Inertia (degrees)

Eccentricity

Major axis length (mm)

Minor axis length (mm)

Coefficient of skewness

 Coefficient of kurtosis

Figure B.3 List Of Extracted Features From MIPAV.

Extracted Features From MIPAV- Part2
Largest slice distance
Largest slice distance
Median Intensity
Mode Intensity
Mode Count
Mean Curvature
Std Dev of Curvature
Mean Negative Curvature
Number of Indentations Curvature
Number of Indentations Hull
Asymmetry Index
Fractal Dimension Box Count
Fractal Dimension Euclidean Distance
Invariant Moment 1
Invariant Moment 2
Invariant Moment 3
Invariant Moment 4
Invariant Moment 5
Invariant Moment 6

Figure B.4 List Of Extracted Features From MIPAV.

B.3 LifeX

The software was written in Java. Results are exported in Excel files format. Figure B.5 shows the setting of feature extraction section in LifeX software which includes the radiomic features were extracted for this study.

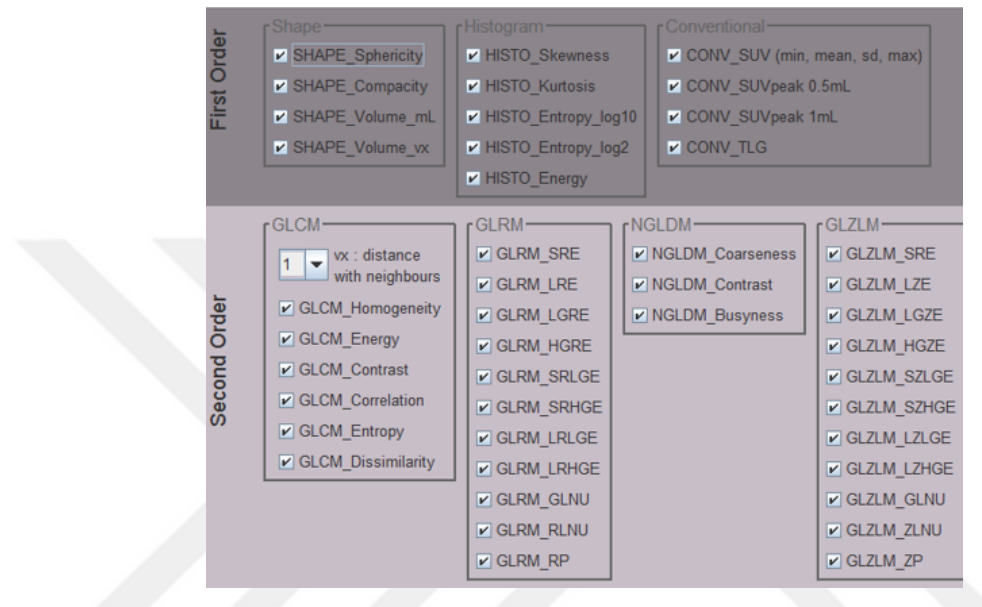


Figure B.5 Table For Advance Options For Radiomic Features In LifeX.

There are 43 quantitative features which include First Order Features - Histogram, First Order Features- Shape, and Second Order Features - GLZLM, GLRLM, NGLDM, and GLCM. A comprehensive description of texture features for radiomics can be found in [43].

Also, information about the features used as input for classification are listed below:

GLRLM_SRHGE: It is the distribution of the short homogeneous runs with high grey-levels.

GLZLM_SZE: It is the distribution of the short homogeneous zones in an image.

GLZLM_SZHGE: It is the distribution of the short homogeneous zones in an

image.

GLRLM_LRLGE: It is the distribution of the long homogeneous runs with low grey-levels.

GLRLM_LRE: It is the distribution of the long homogeneous runs in an image.

GLZLM_ZP: The homogeneity of the homogeneous zones was measured by it.



REFERENCES

1. Shinagare, A. B., K. M. Krajewski, M. Braschi-Amirfarzan, and N. H. Ramaiya, "Advanced Renal Cell Carcinoma: Role of the Radiologist in the Era of Precision Medicine," *Radiology*, Vol. 284, no. 2, pp. 333–351, 2017.
2. Choueiri, T. K., S. Halabi, B. L. Sanford, O. Hahn, M. D. Michaelson, M. K. Walsh, D. R. Feldman, T. Olencki, J. Picus, E. J. Small, *et al.*, "Cabozantinib versus sunitinib as initial targeted therapy for patients with metastatic renal cell carcinoma of poor or intermediate risk: the alliance a031203 cabosun trial," *Journal of Clinical Oncology*, Vol. 35, no. 6, p. 591, 2017.
3. Amin, M. B., F. L. Greene, S. B. Edge, C. C. Compton, J. E. Gershenwald, R. K. Brookland, L. Meyer, D. M. Gress, D. R. Byrd, and D. P. Winchester, "The eighth edition ajcc cancer staging manual: Continuing to build a bridge from a population-based to a more personalized approach to cancer staging," *CA: A Cancer Journal For Clinicians*, Vol. 67, no. 2, pp. 93–99, 2017.
4. Godman, H., "Renal Cell Carcinoma Prognosis: Life Expectancy and Survival Rates." <https://www.healthline.com/health/renal-cell-carcinoma-prognosis>, 2018.
5. Copeland, C., "Precision Medicine: Harnessing the Extraordinary Growth in Medical Data for Personalized Diagnosis and Treatment," *Healthcare Journal of New Orleans*, p. 20, 2016.
6. Karlo, C. A., P. L. Di Paolo, J. Chaim, A. A. Hakimi, I. Ostrovnaya, P. Russo, H. Hricak, R. Motzer, J. J. Hsieh, and O. Akin, "Radiogenomics of Clear Cell Renal Cell Carcinoma: Associations between CT Imaging Features and Mutations," *Radiology*, Vol. 270, no. 2, pp. 464–471, 2013.
7. Shinagare, A. B., R. Vikram, C. Jaffe, O. Akin, J. Kirby, E. Huang, J. Freymann, N. I. Sainani, C. A. Sadow, T. K. Bathala, D. L. Rubin, A. Oto, M. T. Heller, V. R. Surabhi, V. Katabathina, and S. G. Silverman, "Radiogenomics of clear cell renal cell carcinoma: preliminary findings of The Cancer Genome Atlas Renal Cell Carcinoma (TCGA-RCC) Imaging Research Group," *Abdominal Imaging*, Vol. 40, no. 6, pp. 1684–1692, 2015.
8. Kuo, M. D., and N. Jamshidi, "Behind the Numbers: Decoding Molecular Phenotypes with Radiogenomics—Guiding Principles and Technical Considerations," *Radiology*, Vol. 270, no. 2, pp. 320–325, 2014.
9. Cowey, C. L., and W. K. Rathmell, "VHL gene mutations in renal cell carcinoma: role as a biomarker of disease outcome and drug efficacy.," *Current Oncology Reports*, Vol. 11, no. 2, pp. 94–101, 2009.
10. Rodríguez-Antona, C., and J. García-Donas, "Constitutional genetic variants as predictors of antiangiogenic therapy outcome in renal cell carcinoma," *Pharmacogenomics*, Vol. 13, no. 14, pp. 1621–1633, 2012.
11. Patard, J. J., N. Rioux-Leclercq, D. Masson, S. Zerrouki, F. Jouan, N. Collet, C. Dubourg, B. Lobel, M. Denis, and P. Fergelot, "Absence of VHL gene alteration and high VEGF expression are associated with tumour aggressiveness and poor survival of renal-cell carcinoma," *British Journal of Cancer*, Vol. 101, no. 8, pp. 1417–1424, 2009.

12. Pantuck, A. J., Trinh, Q., Karakiewicz, P. I., Fergelot, P., Rioux-Leclercq, N., Figlin, R., ... & Patard, J., "Use of carbonic anhydrase IX (CAIX) expression and Von Hippel Lindau (VHL) gene mutation status to predict survival in renal cell carcinoma," *Journal of Clinical Oncology*, Vol. 5042-5042, p. 25, 2007.
13. Le, V. H., and J. J. Hsieh, "Genomics and genetics of clear cell renal cell carcinoma: a mini-review," *Journal of Translational Genetics and Genomics*, 2018.
14. Nargund, A. M., H. U. Osmanbeyoglu, E. H. Cheng, and J. J. Hsieh, "SWI/SNF tumor suppressor gene PBRM1/BAF180 in human clear cell kidney cancer," *Molecular and Cellular Oncology*, Vol. 4, no. 4, 2017.
15. Wang, Z., S. Peng, L. Guo, H. Xie, A. Wang, Z. Shang, and Y. Niu, "Prognostic and clinicopathological value of PBRM1 expression in renal cell carcinoma," *Clinica Chimica Acta*, Vol. 486, no. April, pp. 9–17, 2018.
16. Kim, J. Y., S. H. Lee, K. C. Moon, C. Kwak, H. H. Kim, B. Keam, T. M. Kim, and D. S. Heo, "The impact of PBRM1 expression as a prognostic and predictive marker in metastatic renal cell carcinoma," *Journal of Urology*, Vol. 194, no. 4, pp. 1112–1119, 2015.
17. Piva, F., M. Santoni, M. R. Matrana, S. Satti, M. Giulietti, G. Occhipinti, F. Massari, L. Cheng, A. Lopez-Beltran, M. Scarpelli, G. Principato, S. Cascinu, and R. Montironi, "BAP1, PBRM1 and SETD2 in clear-cell renal cell carcinoma: Molecular diagnostics and possible targets for personalized therapies," *Expert Review of Molecular Diagnostics*, Vol. 15, no. 9, pp. 1201–1210, 2015.
18. Brugarolas, J., "Pbrm1 and bap1 as novel targets for renal cell carcinoma," *Cancer journal (Sudbury, Mass.)*, Vol. 19, no. 4, p. 324, 2013.
19. Pawłowski, R., S. M. Mühl, T. Sulser, W. Krek, H. Moch, and P. Schraml, "Loss of PBRM1 expression is associated with renal cell carcinoma progression," *International Journal of Cancer*, Vol. 132, no. 2, 2013.
20. Mane, S. M., R. A. Kambli, F. S. Kazi, and N. M. Singh, "Hand motion recognition from single channel surface EMG using wavelet & artificial neural network," *Procedia Computer Science*, Vol. 49, no. 1, pp. 58–65, 2015.
21. Kocak, B., E. S. Durmaz, E. Ates, and M. B. Uluhan, "Radiogenomics in clear cell renal cell carcinoma: machine learning-based high-dimensional quantitative ct texture analysis in predicting pbrm1 mutation status," *American Journal of Roentgenology*, Vol. 212, no. 3, pp. W55–W63, 2019.
22. Chen, X., X. Mou, Z. Zhou, R. Hannan, J. Wang, K. Thomas, I. Pedrosa, P. Kapur, and J. Brugarolas, "Reliable gene mutation prediction in clear cell renal cell carcinoma through multi-classifier multi-objective radiogenomics model," *Physics in Medicine and Biology*, Vol. 63, no. 21, 2018.
23. Noor, N. Q. M., N. N. A. Sjarif, N. H. F. M. Azmi, S. M. Daud, and K. Kamardin, "Hardware Trojan Identification Using Machine Learning-based Classification," *Journal of Telecommunication, Electronic and Computer Engineering*, Vol. 9, no. 3-4 Special Issue, pp. 23–27, 2017.
24. Hussain, M. A., G. Hamarneh, and R. Garbi, "Noninvasive determination of gene mutations in clear cell renal cell carcinoma using multiple instance decisions aggregated cnn," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 657–665, Springer, 2018.

25. Zhu, Z., E. Albadawy, A. Saha, J. Zhang, M. R. Harowicz, and M. A. Mazurowski, “Deep learning for identifying radiogenomic associations in breast cancer,” *Computers in Biology and Medicine*, Vol. 109, pp. 85–90, 2019.
26. Li, S., E. J. Harner, and D. A. Adjeroh, “Random KNN feature selection - a fast and stable alternative to Random Forests,” *BMC Bioinformatics*, Vol. 12, no. 1, p. 450, 2011.
27. Akin, O., Elnajjar, P., Heller, M., Jarosz, R., Erickson, B. J., Kirk, S., . . . Filippini, J., “Radiology Data from The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma [TCGA-KIRC] collection. The Cancer Imaging,” 2016.
28. Clark, K., B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository.,” *Journal of Digital Imaging*, Vol. 26, no. 6, pp. 1045–57, 2013.
29. Schneider, C. A., W. S. Rasband, and K. W. Eliceiri, “NIH Image to ImageJ: 25 years of image analysis.,” *Nature Methods*, Vol. 9, no. 7, pp. 671–5, 2012.
30. Cabrera, J. E., “GLCM Texture Analyzer on ImageJ.” <https://imagej.nih.gov/ij/plugins/texture.html>, 2003.
31. McAuliffe, M. J., F. M. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. L. Trus, “Medical image processing, analysis and visualization in clinical research,” in *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001*, pp. 381–386, IEEE, 2001.
32. Nioche, C., F. Orlhac, S. Boughdad, S. Reuze, J. Goya-Outi, C. Robert, C. Pellot-Barakat, M. Soussan, F. erique Frouin, and I. Buvat, “Lifex: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity,” *Cancer Research*, Vol. 78, no. 16, pp. 4786–4789, 2018.
33. He, H., Y. Bai, E. A. Garcia, and S. Li, “Adasyn: Adaptive synthetic sampling approach for imbalanced learning,” in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328, IEEE, 2008.
34. Hinton, G. E., and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, Vol. 313, no. 5786, pp. 504–507, 2006.
35. Ladha, L., and T. Deepa, “Feature selection methods and algorithms,” *International Journal On Computer Science And Engineering*, Vol. 3, no. 5, pp. 1787–1797, 2011.
36. Yang, W., K. Wang, and W. Zuo, “Neighborhood component feature selection for high-dimensional data,” *Journal of Computers*, Vol. 7, no. 1, pp. 162–168, 2012.
37. “Matlab classification learner.” <https://www.mathworks.com/help/stats/classification-learner-app.html>, 2019.
38. Beale, M. H., M. T. Hagan, and H. B. Demuth, “Neural network toolbox user’s guide,” *The MathWorks Inc*, 1992.
39. Turhan, M., D. Şengür, S. Karabatak, Y. Guo, and F. Smarandache, “Neutrosophic weighted support vector machines for the determination of school administrators who attended an action learning course based on their conflict-handling styles,” *Symmetry*, Vol. 10, no. 5, p. 176, 2018.

40. MATLAB, *version 9.6 (R2019a)*, Natick, Massachusetts: The MathWorks Inc., 2019.
41. Ferreira, T., and W. Rasband, “Imagej user guide,” *ImageJ/Fiji*, Vol. 1, pp. 155–161, 2012.
42. Haralick, R. M., K. Shanmugam, *et al.*, “Textural features for image classification,” *IEEE Transactions On Systems, Man, And Cybernetics*, no. 6, pp. 610–621, 1973.
43. Orlhac, F., C. Nioche, and I. Buvat, *Technical appendix-local image features extraction-LIFE_x*. IMIV, CEA, Inserm, CNRS, Univ. Paris-Sud, Universite Paris Saclay, Orsay, France, version 4.nn ed., 2016.

