A COMPUTATIONAL ANALYSIS OF THE THYROID IMAGING AND REPORTING DATA SYSTEM

By

Olivia Luisi, Bachelor of Science, Computer Science

A thesis submitted to the Graduate Committee of

Ramapo College of New Jersey in partial fulfillment

of the requirements for the degree of

Master of Science in Computer Science

Spring, 2025

Committee Members:

Lawrence D'Antonio, Advisor

Sourav Dutta, Reader

Scott Frees, Reader

COPYRIGHT

Dedication

The work of this Thesis is dedicated to my family, my loved ones and my dearest friends. I complained often, I leaned on you and I cherish what I could do because of you. To my mother who has given me the kindness and strength to push forward, and to my father whose unwavering support has guided my confidence and perseverance.

Acknowledgments

I'd like to acknowledge the Thyroid Cancer Survivors Subreddit where I spent many nights trying to read information on cases like mine, results like mine, and experiences like mine. The community of cancer survivors there gave me a level of curiosity away from my worries that allowed me to do my favorite thing: read about a topic until I'm no longer afraid of it. At the end of this thesis work I will be starting preparations to undergo radiation therapy, satisfied but eager to continue learning about this topic of research I found myself a part of.

May we all be healthier tomorrow.

Table of Contents

List of Tables

List of Figures

Glossary

FNA - A Fine-Needle Aspiration is a biopsy procedure used to collect cells, tissues and fluid from a malignancy suspicious lump or nodule.

Calcification - Calcification is the occurrence of calcium deposits in an area of tissue.

Echogenicity - Echogenicity refers to the ability of an area to reflect the sound waves of an ultrasound. It is used to determine a solid or fluid filled tissue.

Margin - Margins of a thyroid nodule are classified as protruding into adjacent tissue, irregularly angled, invasive to nearby tissue or unable to be determined.

Shape - The shape of nodules are classified as Wider-than-Tall or Taller-than-Wide, Taller-than-Wide nodules are heavily associated with malignancy.

Composition - Composition refers to the physical internal structure of a nodule. Spongiform or predominantly cystic composition is a clear sign of a benign nodule whereas a mixed or fully solid nodule is indicative of needing further assessment.

Echogenic Foci - Referring to calcification or calcium deposits on the nodule or surrounding area.

Comet Tail Artifact - A comet tail artifact refers to a reverberation artifact seen on ultrasound imaging. It is considered a sign of a benign nodule.

TI-RADS -The Thyroids Imaging Reporting and Data System is a standardized system coined by the American College of Radiology.

TR Level System - The TI-RADS score system refers to its levels as TR#, where TR1 refers to benign and TR5 refers to the highest possible risk.

Abstract

The detection of thyroid cancer is uniquely based upon a standardized system of numerical analysis. After a nodule is detected on a patient's thyroid, the ultrasound images are analyzed to determine the level of need for a biopsy. The majority of first world countries follow the basis of the Thyroid Imaging Reporting and Data System created by the American College of Radiology [1]. Each country however has their own rating system to determine the level of danger or suspicion surrounding the nodule, this has led to some country's systems being more or less sensitive in electing for a biopsy of the nodule. While the United States has a numerical value system, the European Union and South Korea have an algorithmic flow chart to determine the nodules rating, and the newer Chinese system focuses on dominant features of likely malignancy [1][3]. Each has their own strengths and weaknesses and in an attempt to better explain their differences, comparing their rates of positive identification will allow for a greater understanding. Patients of Thyroid Cancer are rarely given such insight into the mechanisms which declare the safety of their own health, this project seeks to allow patients to see the data behind what they are being told in their reports and compare their own cases against the systems handling of cases like their own.

Introduction

1.1 Background

Thyroid cancer detection rates have rapidly increased in the past decade of improvement in detection technology. To a casual observer it may seem as though people are becoming more likely to develop Thyroid Cancer, however the increasing rate is actually due to the system of detection as well as the technology of ultrasounds improving [5]. Such an increase has even drawn criticism that the FNA biopsy and cancer may be over diagnosed. This has left tens of thousands of people grasping at information regarding their own ultrasound results, struggling to understand the Thyroid Imaging Reporting and Data System or how it applies its level of "suspicion". This system, better known as TI-RADS, accepts the perceived features from a thyroid ultrasound and calculates the possible levels of suspicion of cancer before outputting whether the patient should receive a biopsy [1]. The system that the United States uses is based out of the American College of Radiology, though many countries and the European Union have their own personalized version of the TI-RAD system; each with differing levels of sensitivity and specificity. This information is not easy to come by as a patient going through ultrasound testing, but would give a more robust understanding to them. Thus this thesis seeks to explore the statistical value in a handful of TI-RADS.

1.2 Problem Statement

Given our current technological decade of using search engines for medical questions, it believes it would benefit the vast majority of patients and practitioners to understand the key

differences in TI-RAD systems around the world, as well as the analysis of the systems key points combined. Medical data is not easy to find, reputable data even less so, therefore a clear description with example data will give the best chance of understanding for patients like myself. This project does not seek to be as informative, definitive, or absolute as a medically lead project would be. But it does seek to set a groundwork for exploring multiple systems' sensitivities to give a more comprehensive view than a suspicion percentage. By showing the system they are being rated into, the differences it may have with other country's systems, as well as how cases like their own were determined by these cases, we provide patients with information allowing them to understand more of their own medical situation. This project is not ethically intact as medical adjacent programming rarely is. As such, the medical ethicacy will be an important element spoken on throughout the project's planning and execution.

1.3 Significance of the Study

The online sphere is filled with an influx of information with varying degrees of correctness and applicability. It is the foundation of design for the intention to have as many people as possible interact with a given product. This leads to some of the most vital pieces of information for Thyroid Cancer patients to be spread out across the vast ocean of internet articles, websites or thinkpieces. As many countries have their own version of the TI-RAD system it is important to show why these distinctions exist and how they can differ in answer between different cases. This research is designed for patients like myself: unable to find specific information with answers to the many uncertainties that come with a system that gives you a percentage score rather than a false or positive. There is no certainty in cancer until the tumor has

been removed and the pathology studied. Ergo the TI-RAD system cannot provide any more information than its suspicion. This paper believes that the system would benefit from including a depth of understanding how the individual systems of America, EU and China differ from one another.

The TI-RADS Calculator is an online website which allows users to input their features reported on their ultrasound report and view the TI-RADS level which is also already included in their thyroid ultrasound report [1]. The information here is obviously reductive. The patient must have already received their ultrasound report in order to input this information; including the TR level, features, and level of suspicion. The TR levels themselves are a simple risk level calculation where starting at TR1 the nodule is considered benign. The risk percentage increases up to TR5, which is considered the greatest suspicion level of malignancy [1]. There is no bridge from the system to similar cases or easily available datasets. The datasets themselves are another unfortunate but important aspect of the study. The vast majority of posted data is specifically for deep learning on the ultrasound images themselves. They have no information pertaining to the features, the outcome or the rating system. Datasets posted outside of the United States never fully contain the specifications required for the ACR TI-RADS numerical system. The data needed for comparison has to be of the TI-RADS standards from each region, specifically addressing each region's own way of describing feature points. That data simply did not exist in my months of researching. Even when repositories reported that the data was held for request, the data would end up being inaccessible or pulled entirely leaving a blank entry. Ultimately this project had to use a 1,000 entry dataset and a 332 entry dataset, requiring a level of abstraction over some of the terms which did not apply smoothly to the US system. Part of this project's

importance ought to highlight how the data acquisition relied on countries with released healthcare datasets for the general public.

1.4 Objectives of the Study

This thesis aims to address the following objectives:

- To obtain datasets of a sizable amount of patients with data pertaining to the TI-RAD system features. This data must contain each feature and if the case was determined to be malignant or benign.
- To write and produce code exploring linear regression on a benign and malignant dataset, exploring the sensitivities in the different features of thyroid cancer
- To write and produce code exploring linear regression on a TI-RADS level dataset for the ACR & C TI-RADS score system.
- To determine sensitivities or irregularities of the point-score systems viewed algorithmically side by side.
- To conceptualize an algorithm which takes in the positives and negatives of like systems.

Literature Review

2.1 Introduction to Diagnostic Statistics

The many Thyroid Imaging Reporting and Data Systems are simply algorithmic decision systems at their core. Each is held against disease screening statistical measures of Sensitivity and Specificity, a statistic used for how an algorithm correctly diagnoses a disease [2]. These values are what ultimately determine the usefulness and drawbacks of each algorithm, as it identifies the under or over diagnosis which in this case leads to a Fine Needle Aspiration biopsy, the final step of malignancy testing a patient can receive before surgery. Sensitivity refers to the number of correctly diagnosed positives within the system, categorized by the equation [12]:

$$Sensitivity = TruePositives(A) \,/\, ((TruePositives(A) + FalseNegatives(B))$$

It determines that of the full number of people with the disease, this many were correctly found. Specificity similarly refers to the number of correctly diagnosed negatives. In our case it will refer to the percentage of people incorrectly referred to an FNA biopsy procedure. The equation for Specificity is [12]:

$$Specificity = TrueNegatives(A) \,/\, ((TrueNegatives(A) + FalsePositives(B))$$

These are the guidelines with which I reviewed each TI-RADS system and began my analysis of each.

2.2 An Overview of the TI-RADS

The American College of Radiation TI-RADS is a score system which adds points of suspicion for feature categories: Composition, Echogenicity, Shape, Margin, and Echogenic Foci. Composition refers to the makeup of the nodule tissue, if it is cystic or solid. Echogenicity refers to the sound reflection of the tissue. Shape refers to the nodule either being Wider-than-Tall or Taller-than-Wide, the later being a sign of malignancy. Margin refers to the shape and possibly invasive tissue of the nodule. And Echogenic Foci refers to calcification or calcium deposits on the nodule. These points are added and classified as TR1-5, on a 0-7+ point scale. Each level in the American system displays a different malignancy rate percentage from 0.3% to 35% and a preferred action for FNA biopsy. An important goal of this tool, besides identifying malignancy, is to reduce the number of unnecessary FNA biopsies; the American system in this regard has a fantastic history of overall performance compared to EU or South Korean systems but this has a very interesting reason and cost.

The EU TI-RADS operates very differently from the ACR TI-RADS. Rather than assign point values to the features, the EU TI-RADS determines the risk level simply by the appearance of low, intermediate, or high risk features [3]. There is no added numerical score, if microcalcifications appear on the nodule, the risk is immediately determined to be TR5. The percentage risk also varies widely from the ACR system: EU TI-RADS levels 1 and 2 are a 0% risk, but level 5 has a variance of 26-87% [3]. This system has a higher sensitivity and lower specificity, which increases its number of unnecessary biopsies [3]. Often the size of the nodule is used to assess the need for biopsy as well. The K TI-RADS is similar in design to the EU TI-RADS but suffers even higher rates of unnecessary FNA biopsies [6]. The system is highly sensitive, perhaps in part to how its standards for nodule size are much lower than the EU. It is

however worth noting that while the comparison of these systems aims to create a more accurate version, the EU and K TI-RADS referring a higher number of patients to biopsies is not necessarily negative when working with medical data. For the majority of computer science work, these algorithms would be called too finicky to provide the correct assessment. But as the reference is for a non-invasive procedure which would produce a more concrete assessment of the nodule it can be argued that the higher sensitivity has more to gain than lose in some environments.

The final and most important version of the TI-RADS system to be discussed is one based in China. The C TI-RADS differs from the other discussed systems as it has the highest level of specificity across studies, meaning it has the lowest number of false positives [10]. The system similarly to the ACR works on a point based system though is much less complex. There are five features which are worth +1 point each and one feature which is worth -1. The suspicion risk for the levels spans from 0% to over 90%, heavily dramatizing the increase in each level [10]. This is perhaps why the system can achieve such high specificity, as by horizontally limiting the point system there is less room for a nodule score to reach higher and higher levels.

An interesting discussion of these models is the healthcare that each country provides. In countries with higher access to healthcare like the EU or South Korea, their models reflect a great ease to recommend a biopsy. Whereas America has a private healthcare system and focuses more on accuracy in order to limit the amount of necessary biopsies. China is an interesting cross section between these two groups. It is a country with government healthcare so it should logically look to diagnose risk levels at a similar rate to the EU or South Korea. FNA biopsies

however are not widely available across China, leading them to take on a more specific approach similar to the ACR system [10]. This is a large reason for understanding why I chose to research and work specifically with the C and ACR TI-RADS. As discussed prior, the overdiagnosis of FNA biopsy to determine malignancy is not necessarily negative enough to invoke change in a system, but for the ACR and C TI-RADS the specificity is extremely important. It also is important to note that the datasets this project utilizes are both from Chinese institutions of research which only strengthens the decision to specifically look at the C and ACR TI-RADS.

2.3 Data Assessment

After concluding that the ACR and C TI-RADS systems were best suited to one another it became necessary to study the best ways to implement improvements between the two. Each had clear higher and lower values which the other mirrored. Where ACR was best in sensitivity, C had a much higher specificity score. I decided to compare and contrast first with my own ultrasound and ACR TI-RADS test results to see how the system differed in my own case as it was a freely available dataset with a known outcome. The ACR TI-RADS listed my case as TR5 suspicion, the highest possible level with a large variance of risk percentage. The C TI-RADS on the other hand had me listed at C-TR 4, still requiring an FNA biopsy but bringing my risk percentage down to 50-95%. This is consistent with the initial theory that the C TI-RADS horizontally shortens the point metric in order to cut down on unnecessary biopsy referrals, cutting off the fat of extra points. This was used as a base idea of how to go about combining the two systems, the ACR TI-RADS has a fantastic accuracy of positive detections but the C TI-RADS can serve as a weight to curb false positives.

It was primarily important to research a few different implementation methods by reading through publications similarly attempting to apply logistic regression to the ACR TI-RADS. For the purposes of this project, the regression model would be used on both the ACR and C TI-RADS categorical system in order to see the key differences between them. For this, two different datasets were needed to correctly assess both points of key features desirable for a report. One dataset needed to have the information required for ACR TI-RADS & C TI-RADS so that the regression models would be easily compared. The other dataset needed to have cases which resulted in both benign and malignant outcomes in order to be useful if viewed by a patient researching outcomes of similar cases to their own. Oftentimes patients like myself would think to search their own case description online to see potential outcomes, this has a compositional fallacy as people who do not end up having cancer do not go onto cancer forms to post that they do not have cancer. Therefore producing the viewpoint of cases which do end up benign is of particular importance. As stated towards the beginning of this project the use of medical data for research or informing a public carries the risk of unethical use, the skewing of any presented data can lead to a negative reaction or outcome if the viewer is uninformed of the data's pitfalls.

Methodology

The research required for this project is about as much work as the project itself. It was an extremely lengthy process of finding data which suited the needs of the work. The main entry point was figuring out what data was available, what it contained and if it would cover the majority of areas which I wanted to use. This methodology will touch first on the dead ends of the project before going into discussion of the actual development points. Unfortunately the medical landscape is not filled with datasets that anyone can use to research their own interests and instead the project ended up relying on Chinese research institutions which posted their datasets alongside their research.

3.1 Introduction to Data Acquisition

Going into this project there was an assumed level of difficulty in locating the data necessary for the analysis desired. However it became quickly apparent that this would be more than just difficult to obtain but rather impossible in the specific aspects which were desired. A quick look throughout many databases will return a false sense of security in the many Thyroid Imagining datasets posted to the web for image-based deep learning [8]. Once these datasets are obtained and explored it becomes apparent that they do not meet the standard for research which the project required. Without the clear features weighted in the TI-RADS system, the area of analysis to be done shrunk quickly. Most were datasets which recorded only parts of the full picture, and while emails continued to be sent to the point of contact or database owners of many

different research papers to find something more detailed, almost all were left unreplied. This research ended up with just two datasets that had *most* of what was required.

The two datasets this project worked with were both publicly available research datasets from Chinese based researcher groups. Because of this both datasets were not necessarily created with ACR TI-RADS analysis in mind, though one of the datasets did contain detailing of the C TI-RADS score alongside of the ACR TI-RADS score. The drawback of this dataset however was that the entire set contained malignant cases of thyroid cancer and that it did not contain clear documentation for three of the columns. It was published with the intent of research into the autoimmune and marker genes that affect many Thyroid Cancer cases, not TI-RADS analysis. The primary dataset I would use for linear regression contained both benign and malignant cases with additional patient data, but the columns were again not directly translatable to the ACR point system. The primary and secondary dataset is displayed below:

id	age	gender	FT3	FT4	TSH	TPO	TGAb	site	echo_pattern	multifocality	size	shape	margin	calcification	echo_strength	blood_flow	composition	mal	multilateral
1	46	1	4.34	12.41	1.677	0.43	0.98	0	0	0	4.6	0	0	0	4	0	1	1	1
2	61	1	5.40	16.26	2.905	0.45	1.91	0	0	0	4.2	0	1	1	4	1	2	1	1
3	44	1	3.93	13.39	1.823	9.15	26.25	0	0	0	0.7	0	1	0	4	0	2	0	1
5	29	0	3.70	13.98	1.293	0.15	0.81	0	0	1	1.0	1	1	1	4	0	2	1	1
6	37	1	3.60	14.56	0.938	0.13	21.22	0	0	0	0.7	0	1	1	4	0	2	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
743	41	0	4.84	16.23	0.531	0.00	1.14	2	0	0	3.6	0	0	0	4	0	1	0	1
745	28	1	4.68	18.17	1.350	0.23	2.25	2	0	0	0.8	0	1	1	4	1	2	0	0
748	48	0	6.26	17.41	1.270	0.00	1.68	2	0	0	0.5	0	0	0	4	0	2	0	1
749	31	1	4.85	17.34	0.171	80.90	52.00	2	0	0	0.5	0	0	0	4	0	2	0	1
756	46	1	4.00	14.00	0.421	5.00	4.00	2	0	0	0.6	0	1	1	4	0	2	0	1

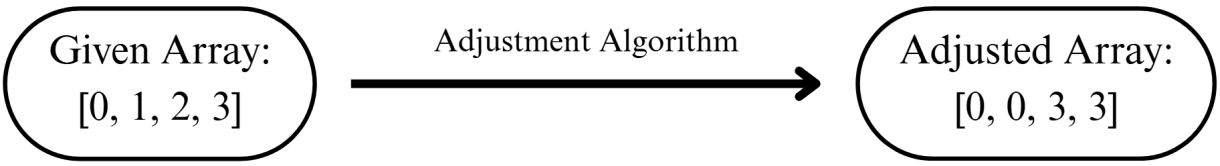
(Table 1: Malignant Dataset)

ID	sex	age	BRAF	Autoimmune	pBilateral	pmultifoci	pDiameter	T	pVI	USmultifoci	USDiameter	TIRADS	composition	AT	Echo	Calcification	Margin	BloodPart	cNodules	cTIRDAS	UScDiameter	ccomposition	cAT	cEcho	cCalcification	cMargin	cBloodPart
3	0	49	1	0	1	1	7	1	0	0	5.5	5	4	1	1	0	0	0	2	3	11.4	4	1	2	1	1	1
10	0	44	1	0	1	1	3	1	0	1	4.6	6	4	0	1	1	1	0	1	1	86	4	1	3	1	1	1
11	0	22	1	0	1	1	12	2	0	0	18	5	4	0	1	2	0	2	2	3	5	3	1	1	1	1	1
16	0	61	1	1	1	1	6	1	0	1	7.7	7	4	1	2	0	1	1	1	3	50.2	4	1	1	1	1	3
17	0	35	1	0	1	1	9	1	1	1	20.5	7	4	0	2	1	0	1	2	1	40.2	4	1	1	1	1	3
18	1	35	1	0	1	1	20	2	1	1	11.2	7	4	0	1	1	1	1	1	5	14.4	3	1	1	3	1	3
19	0	29	0	0	1	1	1.5	1	0	2	9.5	7	4	0	1	1	0	2	1	2	47.3	4	1	1	1	1	4

(Table 2: ACR & C TI-RADS Dataset)

The second of the two datasets required quite a bit of cleaning and adjustment. It contained 332 rows in total across two CSV files. The files were combined rather than used as separate training-testing files as the data would need heavy manipulation to be functional to the needs of the model. This data had no documentation and while there was a research paper which the dataset was created for, it did not provide additional information on what exactly columns like “AT” or “pVI” represented. It did however have both the C TI-RADS and ACR TI-RADS score to compare to one another and the definitive outcome that each case ended in malignancy. Ergo it was pieced together which column represented which as well as what needed to be adjusted. It was quickly noticed that the “TIRADS” column was not the TR level but instead was the total score collected from the features. The numbers in the “TIRADS” column varied between five, six and seven representing the two highest levels of suspicion TR4 and TR5. It needed to be assumed that if any case reached more than seven points, the minimum for the TR5 level, it would simply be scored at a 7. Next it was imperative that values consistent with the system's features were being used. This came about by manually adding together the points of the features to eventually find out which column represented which features. For example, there is no clarity on if the ‘T’ column represents the Taller-than-Wide risk feature. The column is

simply [1,2], ergo through as many columns as could be checked the ‘1’ value was applied as three points and the ‘2’ value as zero. This seemed to equal out to the “TIRADS” column score but in such a large data set it was decided to move forward, dropping rows if absolutely necessary. This approach continued until the data layout was satisfactory, then it was needed to match point value to features. For example, the ‘Margins’, ‘composition’, and ‘T’ columns all represented the array values of the features rather than the point values. The numbers were swapped, again paying attention to if they matched our point score, and proceeded to finish cleaning the dataframe. Below is the simplification of the point system process, converting the feature array into the designated point value:

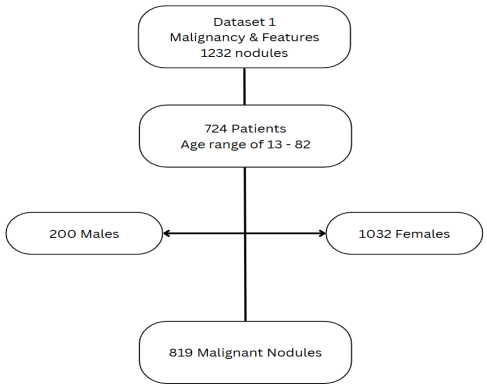


(Figure 1: ACR Point Adjustment)

After this it was decided to copy the data frame into an ACR and C data frame in order to give better organization for future modeling. Additionally a row was added which reported the TR level so that it would not overwrite the total point score. It provided a much easier look at whether or not the point calculations were correct as well as separated out which data was being used for modeling versus which could be used for graphing the statistical makeup of the dataset.

3.2 Linear Regression

With two sets of varying data the comparing and contrasting began with the R^2 scoring on the malignancy dataset. This score represents the overarching value of the relationship between the features and outcome. The first model created was used to predict malignancy based on the appearance of certain features within a group. It started first with the original features included within the TI-RADS feature listing before more or less advanced features in attempts to push the prediction functionality to be closer and closer to the malignancy outcome. The basic information of the malignancy dataset is below:

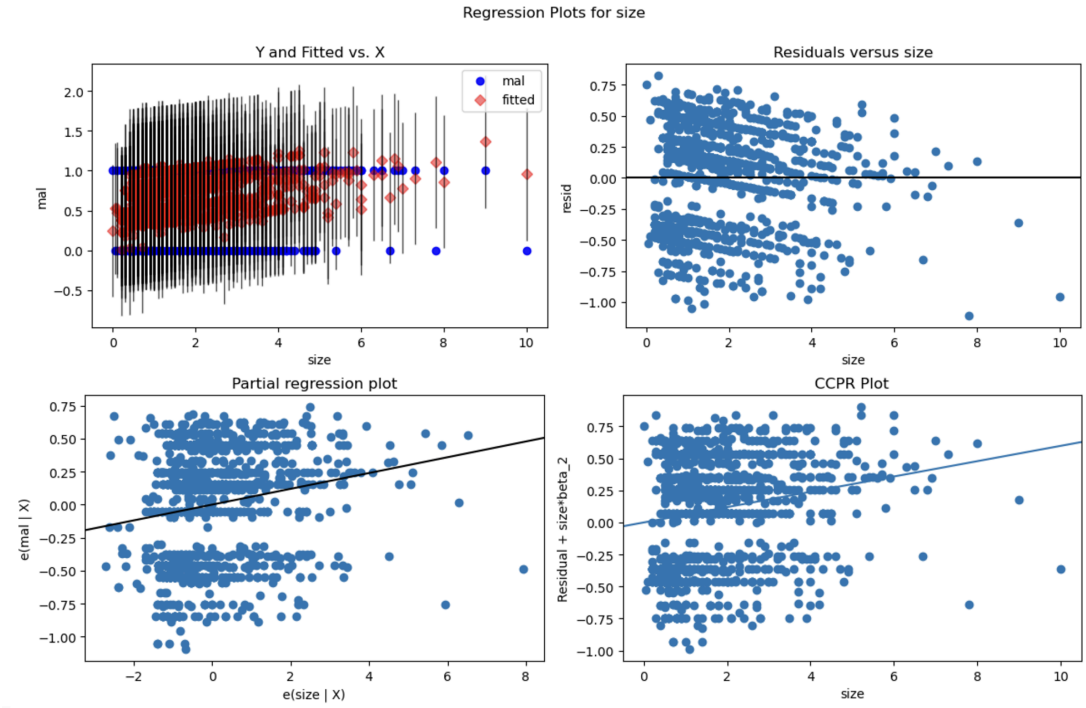


(Figure 2: Malignancy Dataset Breakdown)

The model struggled slightly with only features which are included in the five TI-RADS score pillars (Composition, Margins, Echogenicity, Shape, Echogenic Foci) as the data was not one to one. While it maintained a 22% variance based on the features of the TI-RADS information it was unable to predict consistently, assumedly because multiple columns referred to a single TI-RADS feature and the model needed those features to determine a score before using that score as a prediction value. This was fairly interesting as while the dataset variables were greater than five columns and presumably messed with the variation, they all represented information from the score system. This route ended up a backwards way of attempting to

recreate the TI-RADS five pillar system and provides a contextualization of how even in a model outside of the TI-RADS organizational features, it begins to mimic the same contextual structure of it through the variables value on the malignancy outcome.

After finishing the data for the first model it was determined to not continue with SciKit Learn's Linear Regression functionality and instead moved to learn how to work with Statsmodels Python module for Linear Regression. The module was entirely new to the workflow of the project but was integral for the amount of information being made from creating and running these models. Through this functionality it became clear to see the coefficients for each predictor variable within the model. For example, for each feature of marked composition which holds a 0, 1 or 2 point value the model boosted the expected binary outcome of malignancy by 0.1. Echogenicity on the other hand had a coefficient value of 0.04, this feature is worth up to three points within the TI-RADS system but has the lowest coefficient of the feature group. Calcification was easily the strongest predictor variable with the highest coefficient scoring. The $P > |T|$ value determines if the regression model relationship is statically reveavent and exists. This told me which features within the model were statistically significant. Shape, Margin, Calcification and Composition held the most significance within the group of variables. As the features of the nodules are rated on an arbitrary 0-3 scale representing features across a binary outcome, I found it most interesting to plot the nodule size in the regression model and see how it affected the model as a data piece:



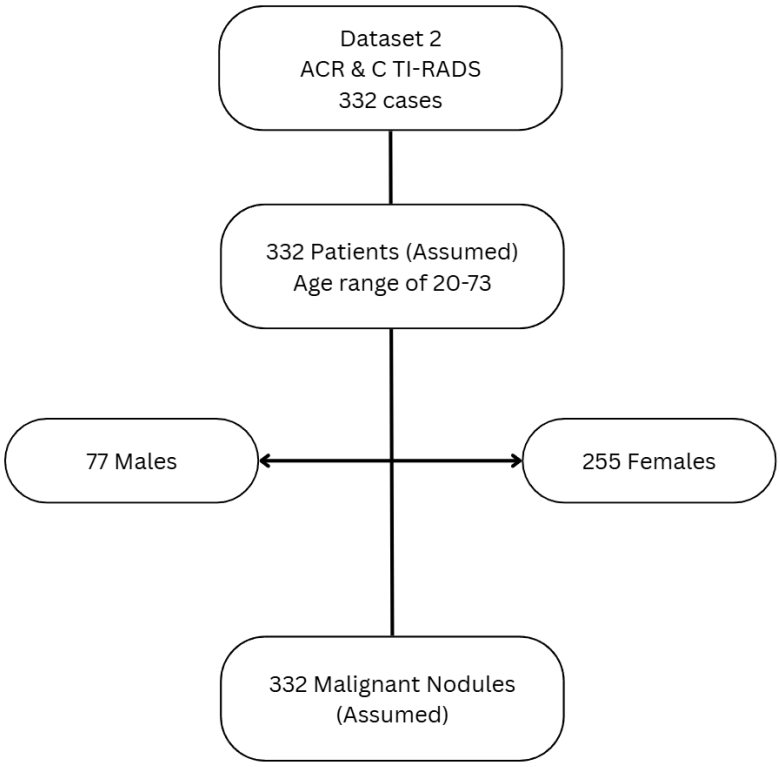
(Figure 3:Malignancy Dataset: Regression Plots for Size feature)

The large growth of thyroid nodules is heavily attributed to malignancy as thyroid cells themselves are naturally slow growing [5]. The size of the nodule is a very important factor which can indicate an immediate need for biopsy, it is very interesting how the model prioritized size in determining a malignancy response, whether a false or true positive.

The project then moved to the alternative dataset to explore how the performance of the model would change with predicting a score level rather than an actual malignancy. Logically speaking it would be assumed that this would not directly communicate the information which the first model had. Instead it would simply recreate the point system which each feature

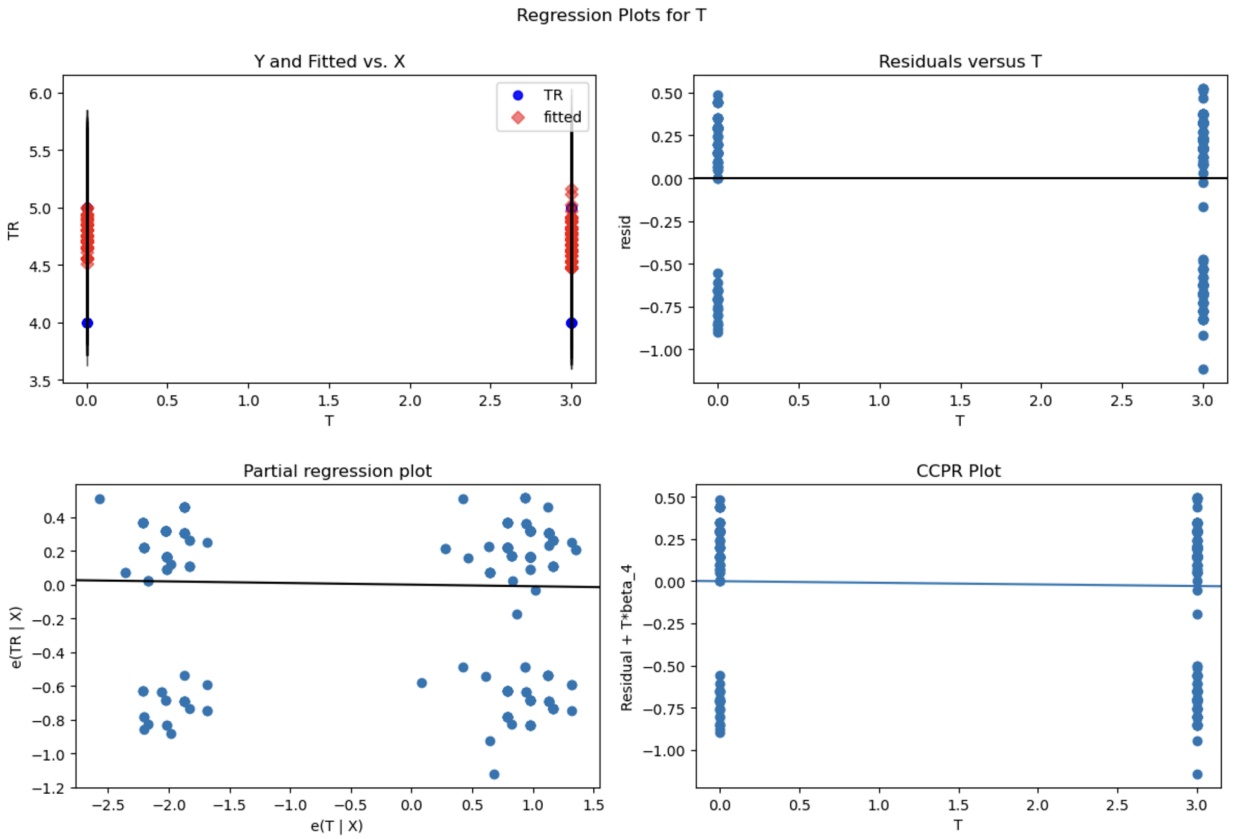
contributed to. Nevertheless the model was set up in hopes that it would provide a methodology for future publicly available datasets.

As spoken on in the previous section, the C and ACR TI-RADS dataset did not have clear documentation. Therefore the findings discussed are inherently flawed as the data is only as good as the interpretation. The data within the C and ACR TI-RADS dataset is entirely made up of malignant cases with both ACR and C features to describe the nodules. Additionally it holds the ages and genders of the patients. The dataset distribution is shown below:



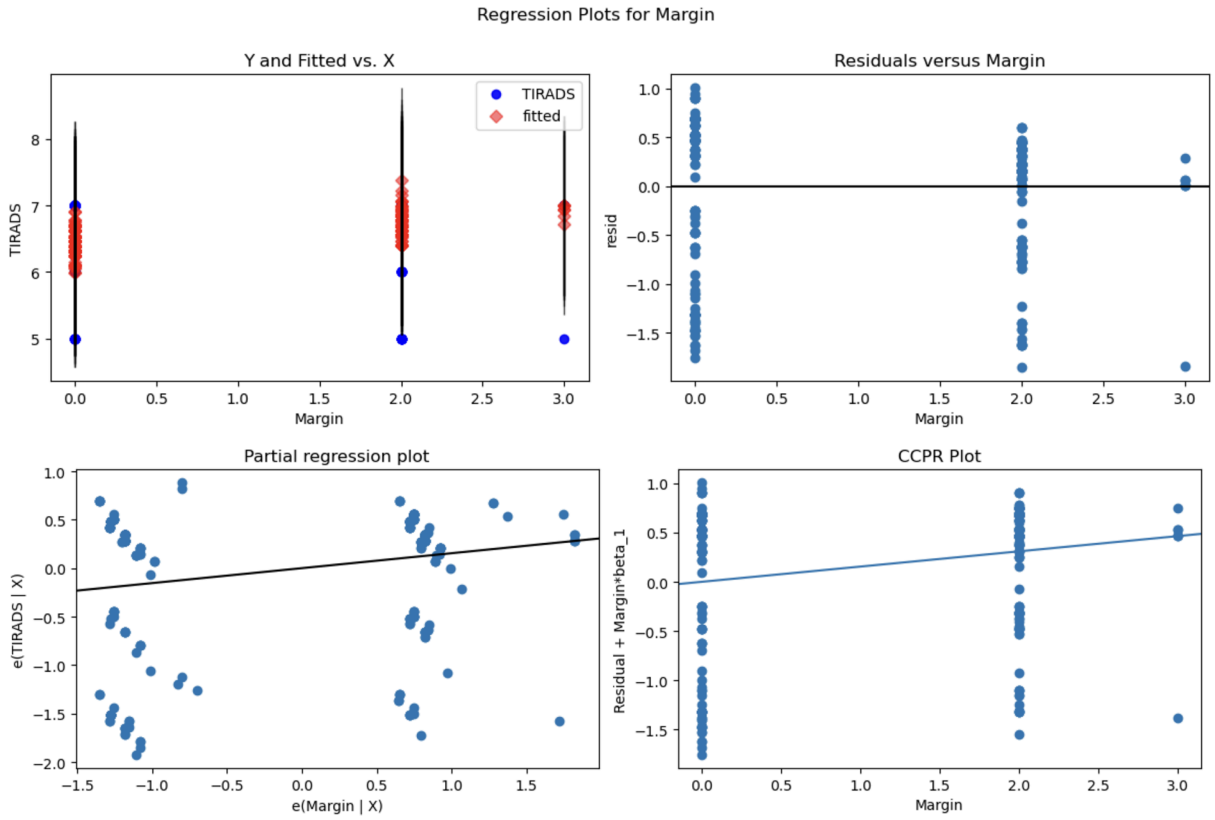
(Figure 4: ACR & C TI-RADS Data Distribution)

There are two assumptions being made on the dataset's information. The first is that there are 332 patients in total and that it is not a case by case breakdown. The second is that given the lack of documentation on the dataset it must be assumed that there are 332 malignant nodules in total for evaluation. There is a column named "cNodules" which contains values 0, 1, or 2 but two nodules being the maximum number out of 332 cases is too unlikely to draw a concrete conclusion on what the data represents. Ergo, it was decided to drop this table from any future analysis regardless of how useful it would have been to explore multiple nodules affecting the risk rating. For the first model produced the dataset from the manipulation done in Figure 1 was used. It received the TI-RADS point accumulation as the *y* value and the five TI-RADS features as *x* values. It was predicted that the model would easily realize that the data was a simple point system adding up the features and that features with similar point totals would function the same while outlier features like "Taller-than-Wide" would be hard to compute. It was true that the "Taller-than-Wide" attribute came back with a -0.0195 coefficient, a strong indication that it did affect the final point score negatively. I reran the model, this time using the TR score instead of the full point scoring, it once again returned with a negative coefficient but was a slightly larger number. While this would be interesting in a complete dataset, this data only contained TR4 and TR5 level nodules so a three point feature being slightly more useful between only two possible outcomes had no significance. While a rare feature, it has the point value to immediately place any nodule into the "Moderately Suspicious" risk group which is why it was expected to be more effective in calculating the higher risk of nodule or at least not have a negative association. The regression plots are displayed below:

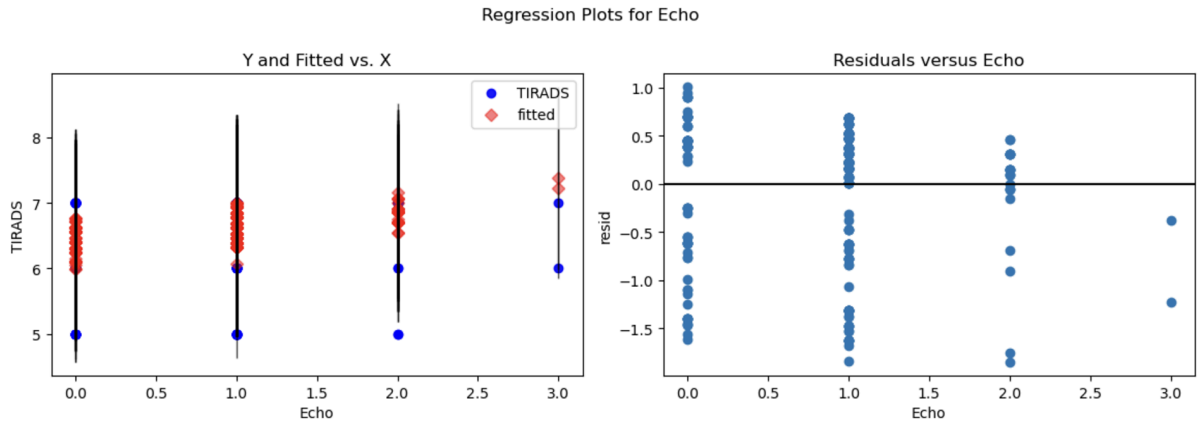


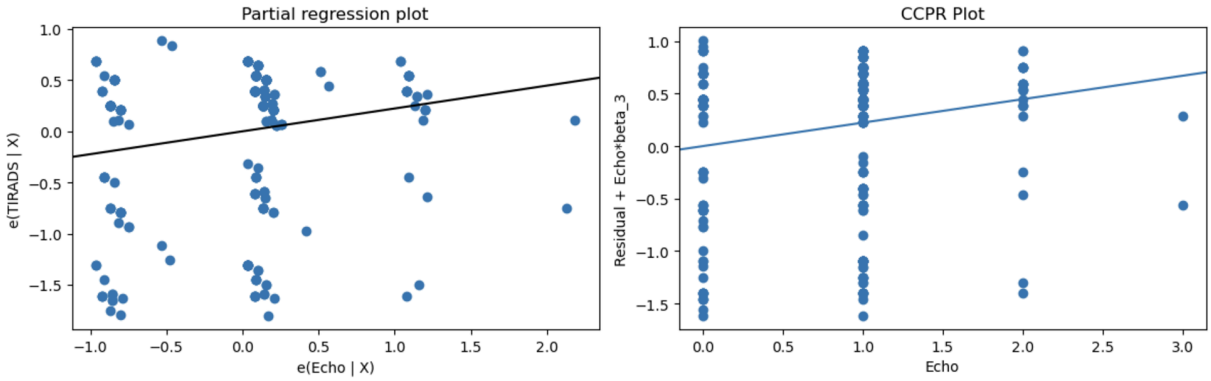
(Figure 5: ACR & C TI-RADS: T Regression Plot)

I then ran a regression plot on the Margins and Echogenicity features as they both had the lowest $P > |t|$ values of approximately ~0.00. Both displayed interesting yet disappointing value at the relationship of the features to the TI-RADS point outcome:



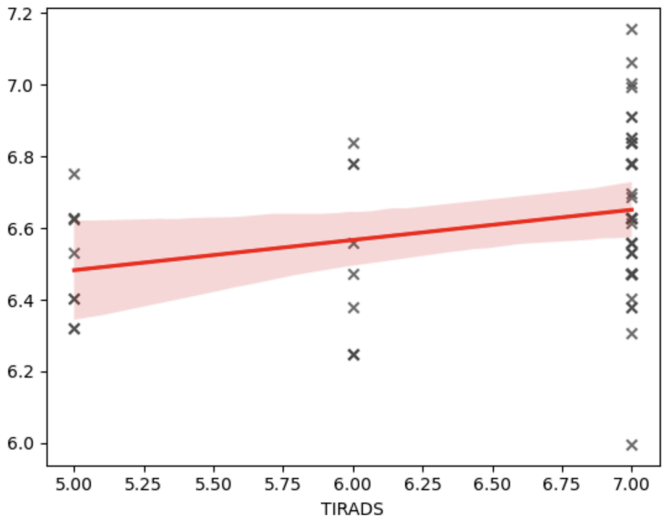
(Figure 6: ACR& C TI-RADS: Margin Regression Plot)





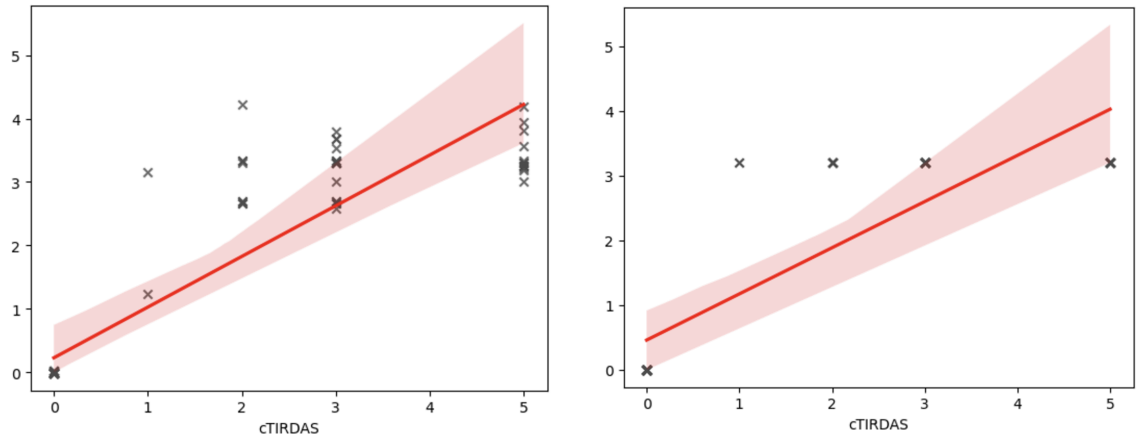
(Figure 7: ACR & C TI-RADS: Echogenicity Regression Plot)

The information displayed in the regression plots made it abundantly clear that the data I had displayed only a small window of information. Similarly to the first dataset there was a very small amount of possible outcomes which hindered the models ability to determine a correlation. Each graph feature had an array of points which it could deal out, but if the dataset only contains the most suspicious nodules it is only fair to assume that the nodules will display the higher value point features. I calculated an Actual vs. Predicted Points graphic on the data so far:



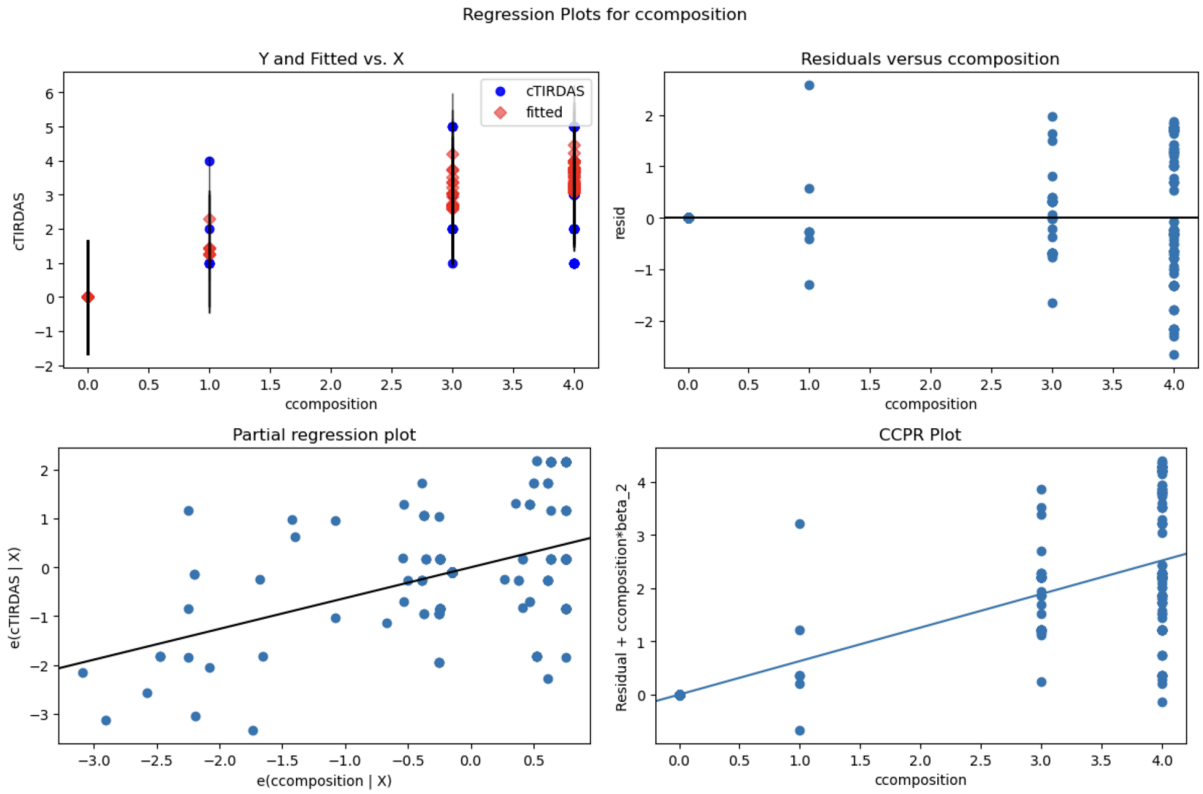
(Figure 8: ACR & C TI-RADS: Actual vs. Predicted ACR Plot)

Next was to start graphing the data specifically associated with the C TI-RADS. Again the data was not documented and it was decided to use the column for Taller-than-Wide which had been used for the ACR data analysis. For this it was a similar workflow as with the ACR dataframe, and slowly adjusted the value of every feature to reflect the set up of the C TI-RADS. This included adjusting every data variable point to [0, 1] based on the features existence and setting up the "cAT" feature as the Comet Tail Artifact feature at [-1, 0]. For clarity's sake an Actual vs. Predicted graph was kept from the original data values to compare to my data edits. This had been done with the ACR data but the graphics did not change nearly as much as the point system still had the same amount of possible values per feature. The data for the C TI-RADS changed a lot as unfortunately it was unknown what each value in the features column represented for a system that only counts the appearance of the feature, not the description. Below are the two prediction graphics, the left is before the point value was edited and the right is after:



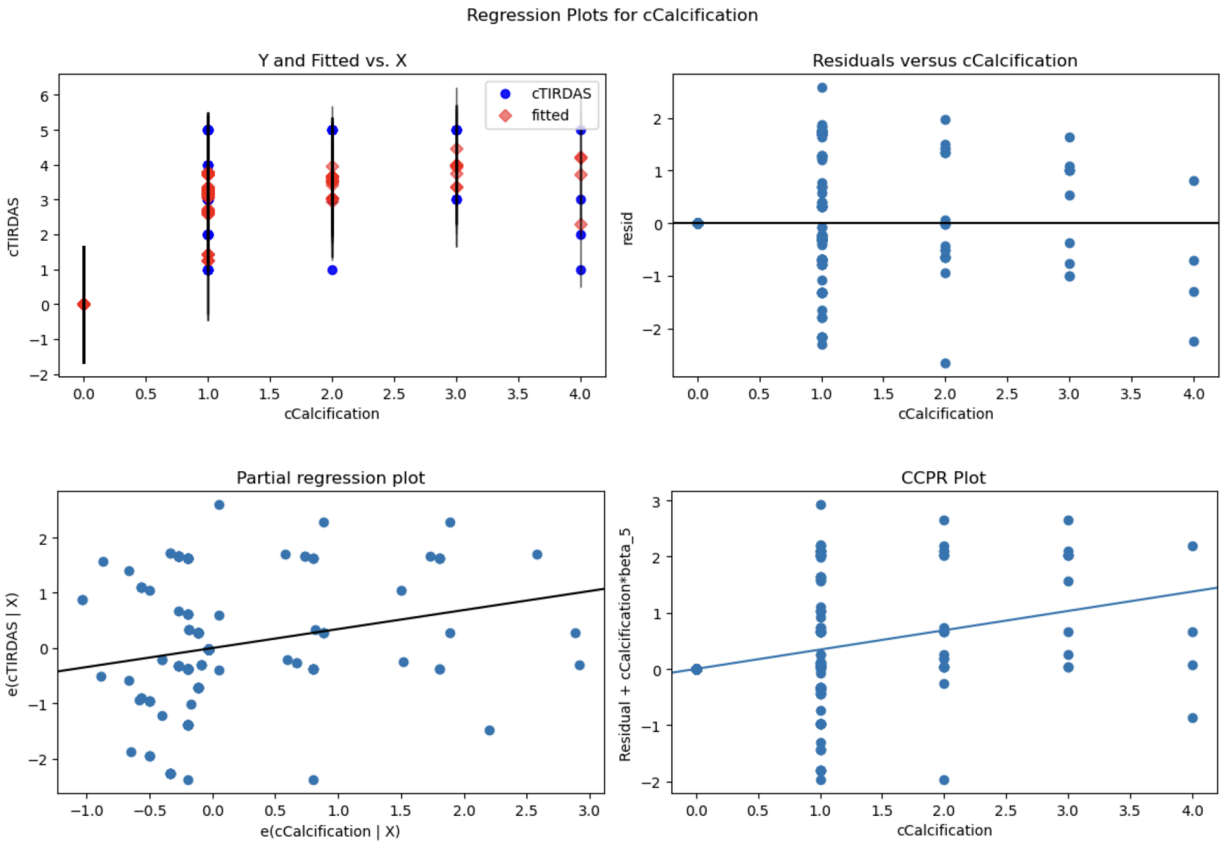
(Figure 9: ACR & C TI-RADS: C TI-RADS Point-value Prediction Difference)

The immediate variance in the model was a bit confusing at first before looking at the regression statistics. The R^2 value was 0.740 which was much higher than the ACR model. This was momentarily explained by the simplicity in the data points all being [-1, 0, 1], all of the data had the same coefficient values and strength over the model. When resetting the variables back to their original values the R^2 value actually increased slightly and showed $P > |t|$ values which had similar strength in the ACR model. The values with the highest influence were graphed below:

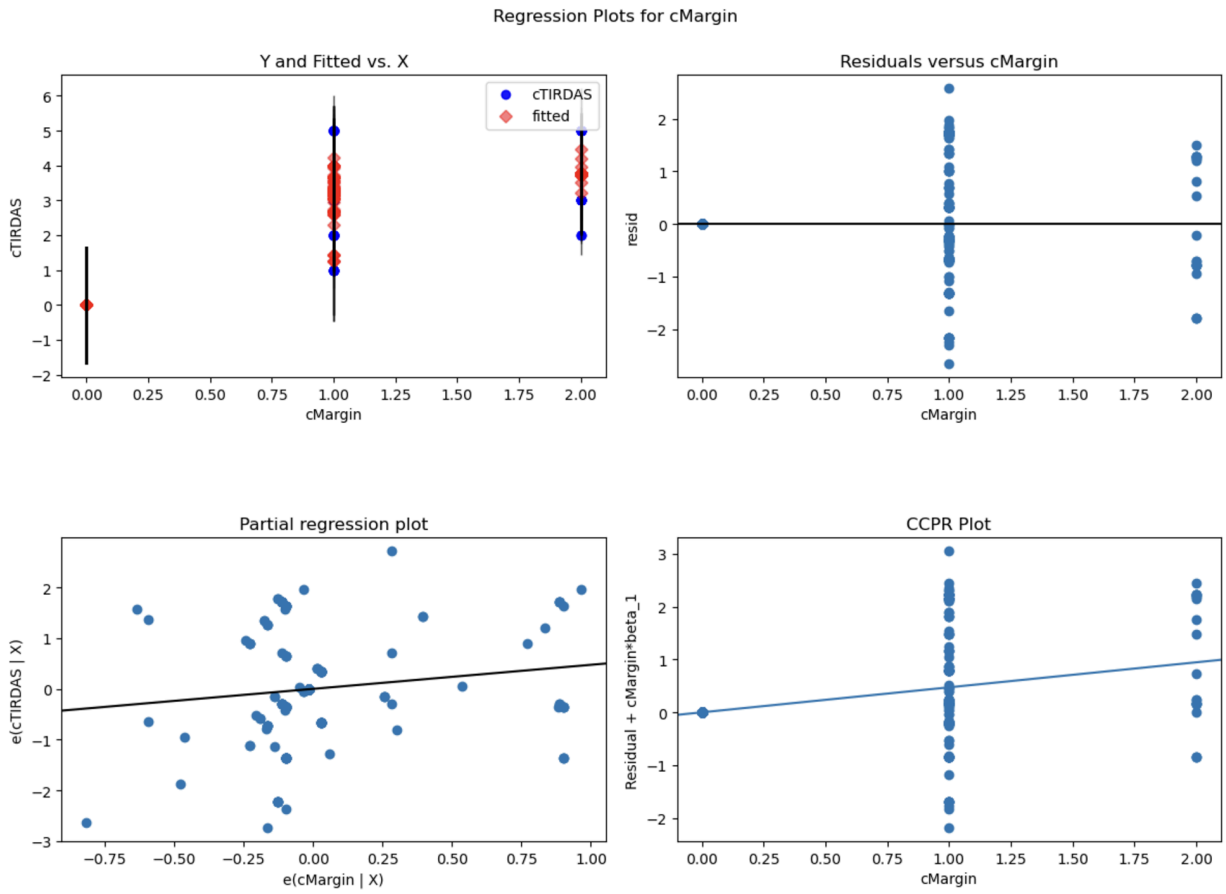


\

(Figure 9: ACR & C TI-RADS: C Composition Regression Plots)



(Figure 9: ACR & C TI-RADS: C Calcification Regression Plots)



(Figure 9: ACR & C TI-RADS: C Margin Regression Plots)

3.3 Combining Models

As shown in the models, regression plots, and prediction variance the three data groups were not strong models but did often show similarities in which features affected the general outcome of the data the most. The Margins feature showed significant strength in every data group's model and the Echogenicity was important in both models within the ACR & C dataset. The regression plots for these values in the malignancy data set probably would have provided a lot of interesting information if the predicted outcome had more variance than the

binary benign or malignant. If there was a wider array of data which included multiple TR levels along with a malignancy variable, it can be assumed that the data would continue to show the strong effects of the Margins and Echogenicity. Perhaps enough so to heavily weigh that feature and return others to the C TI-RADS weighting system.

Analysis and Discussion

The vast majority of research projects came across when working on this project seeked to rely on computational intelligence to determine malignancy at a higher outcome than current professionals and systems. When starting this project I knew I did not want to assume the same ideas but could not put into words exactly why. The TI-RAD system is not meant to determine cancer, it is meant to determine the risk and provide steps for the patient. When looking at these datasets it did not seem important to reinvent the wheel with fancier tools but rather reinforce it through examining the factors of malignancy which have a higher impact than other features within the system.

4.1 Challenges and Solutions

As stated many times throughout this project, the data used renders the information gained as incomplete to the full picture of how TI-RADS features interact with the assumed risk level. That being said, it seemed best to treat the data as if it was the full picture and make assumptions afterwards. There are five pillars in the TI-RADS score system, six when counting the C TI-RADS negative Comet Tail feature, ergo all of them of course have a constant level of importance in determining risk. Of the three linear regression models created to more effectively view the importance of each feature, three of the five stood out amongst the others. Composition is an interesting feature within the TI-RADS family as it holds a feature which is commonly associated with benign nodules: purely cystic or spongiform composition [16]. The data treated composition differently between the benign vs malignant dataset and the high risk assessment

dataset. This is assumedly because the high risk nodules were all malignant, ergo it would not make sense for any of them to have a benign cystic feature. As the datasets were unable to be combined due to the difference in feature documentation, I started to theorize about an if-then model of the TI-RAD system and how it would boost larger indicators of cancer while leaving the middle of the pack nodules to be supervised depending on size. With this in mind we can examine the importance of two other features: Calcification and margins. Both of these features had a large impact on the predicted outcome of cases across datasets and it can be hypothesized from this research that there can exist a weighting of malignant features to help produce risk factor results.

4.2 Comparative Analysis

The majority of computer science work within the medical field revolves around how computational force can replace the work of medical professionals. As spoken on during the literature review portion, models which review the ultrasound of nodules are very popular right now. Research determining which TI-RADS is superior is also very popular. This research argues that this direction is not as prolific as perfecting the TI-RADS algorithm itself. The TI-RADS does not need to diagnose cancer, it just needs to do its job well depending on the healthcare environment of the region it represents. It's why working with EU or K TI-RADS data is unnecessary when the research is for a country like the United States which needs specificity because the patients the system serves do not have easy access to biopsies. EU and K TI-RADS do not attempt to curb their false positives and because of the healthcare environment of those regions do not have the same drive to do so as the United States or China does [3][15]. This is an

important and complex take away when comparing the design of this research to the majority of papers which were reviewed in preparation of the hypothesis.

4.3 Future Research Directions

Using the structure of the analysis which I started within this project I aim to one day revisit it with whatever new data may be made available. I would like to revisit the hypothesis and idealism of a system aimed in the specificity needs of the region it resides over rather than attempting to create the perfect prediction of malignancy. That the prediction rating can have the outcome of a risk percentage rather than fruitlessly grasping towards an ultrasound's ability to view malignancy. Ultimately the treatment direction of thyroid cancer is not always a straight line to surgery. It involves a patient's informed decision to monitor the nodule, remove the half of the thyroid with the nodule, or completely remove the thyroid. Ergo, I find it much more important to be able to communicate the risk factor of features instead of training a model to report a cancer prediction which will not be proven until after the surgery. The point system does a great job of combining features in a presentable way but as shown in the data gathered, there are certain feature points not being fully represented such as how composition, margin and calcification lead the majority of strength across models yet do not pull ahead in point value on either the ACR or C TI-RADS models. Again, this research is easily disproved due to the lack of concrete data behind it. But I really do think that the relationships would continue to be expressed given a wider array of features within the data.

Conclusions

This paper concludes with a clear outline of TI-RADS improvements which would show themselves if given the data to do so. The deadends found throughout this project only emboldened the ending opinion that the current state of medical information access inhibits the uses of computer science in medical fields. It is not the belief of this paper that programming can replace the specialization and understanding of medical professionals but that in cases of abstract simplification computer science has a strong reason to be used as a helpful scoring tool. This project was challenging and required a lot of research into the topics discussed, the datasets, and if those datasets were even helpful to the specific point of direction.

5.1 Summary of Findings

While the datasets accessed left much to be desired, the project itself still serves as a basis for what could be done if said data were one day freely available. There are clear relationships between the datasets, regardless of their vast differences, in the features which were continuously considered important to the end point of risk level or determined malignancy. This to a certain degree does prove that the features of TI-RADS can be individually weighed in order to draw more specific conclusions of risk and therefore adjusted to fit a better specificity and sensitivity. It is the hope of this research that breakthroughs in data significance could be much more common if there was a freedom of such data. The ACR and C TI-RADS face the risk assessment problem in two different ways which means there is a third way between them which may ultimately improve the system if only there existed the data to prove it.

Ultimately there are many research projects describing the differences between the ACR, C, K, and EU systems. All attempted to prove which was the best but this project seeks to provide a differing route to combine the strengths of similar systems within these groups. TI-RADS does not need to diagnose cancer, it needs to recommend true positives to biopsy and lower the number of unnecessary false positives.

5.2 Contributions to the Field

I would like to believe that this work is a proof of concept for the groundwork of discussion that could occur on the adjustments of the TI-RADS. Rather than replace the ACR system which is arguably the best in the world, I want to make clear that there are more options such as reviewing how the features weighting can increase or decrease sensitivity vs specificity. Each of the TI-RADS algorithms has some sort of worth which works for the specific region it is used in. China and the United States are in a unique position together where both systems match the others weakness. While my data is incomplete I hope to a certain extent that the ideas I am attempting to convey are worth something to a group of computer scientists who may also not have any medical knowledge but do know how to adjust values in order to narrow the output of a system.

5.3 Final Thoughts

In conclusion, this project involved substantial effort to produce data analysis on a medical system which is not heavily documented. The acquisition of data was a substantial roadblock at each turn of the project, and there were many datasets which were repeatedly scrapped because of the lack of concrete information in them. It was increasingly necessary to be

strict with what data was acceptable vs what data would ultimately be fruitless. Overall, this project was created with patients of thyroid cancer in mind. It can absolutely be pushed further and it is the hope of this paper to continue its research as more data becomes publicly available for testing. This is not a topic I would have chosen a year ago, but I am really glad I did now for all the catharsis it provided.

References

[1] ACR thyroid imaging reporting and data system (ACR ti-RADS) | radiology reference article|radiopaedia.org.(n.d.-a).https://radiopaedia.org/articles/acr-thyroid-imaging-reporting-and-data-system-acr-ti-rads?lang=us

[2] Disease screening - statistics teaching tools - New York State Department of Health. (n.d.-b). https://www.health.ny.gov/diseases/chronic/discreen.htm

[3] European Thyroid Association Tirads | Radiology reference article | radiopaedia.org. (n.d.-c). https://radiopaedia.org/articles/european-thyroid-association-tirads?lang=us

[4] Executable Books Community. (2020). Jupyter Book (v0.10). Zenodo. https://doi.org/10.5281/zenodo.4539666

[5] *Key statistics for thyroid cancer*. American Cancer Society. (n.d.). https://www.cancer.org/cancer/types/thyroid-cancer/about/key-statistics.html#

[6] Korean society of Thyroid Radiology Thyroid Imaging, reporting and Data System (K-TIRADS) | radiology reference article | radiopaedia.org. (n.d.-d). https://radiopaedia.org/articles/korean-society-of-thyroid-radiology-thyroid-imaging-reporting-and-data-system-k-tirads

[7] Lin, Peiliang (2022), “CIMF”, Mendeley Data, V1, doi: 10.17632/cnr4nbf8bb.1

[8] Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., & Romero, E. (2015, January 28). *An open access thyroid ultrasound image database*. SPIE Digital Library. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9287/92870W/An-open-access-thyroid-ultrasound-image-database/10.1117/12.2073532.short?SSO=1

[9] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

[10] Qi, Q., Zhou, A., Guo, S., Huang, X., Chen, S., Li, Y., & Xu, P. (2021a, October 27). *Explore the diagnostic efficiency of Chinese thyroid imaging reporting and Data Systems by comparing with the other four systems (ACR TI-rads, Kwak-TIRADS, KSTHR-TIRADS, and EU-TIRADS): A single-center study*. Frontiers in endocrinology. https://pmc.ncbi.nlm.nih.gov/articles/PMC8578891/#s2

[11] Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference.

[12] Shreffler, J. (2023, March 6). *Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios*. StatPearls [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK557491/

[12] The Matplotlib Development Team. (2025). Matplotlib: Visualization with Python (v3.10.3). Zenodo. https://doi.org/10.5281/zenodo.15375714

[14] The pandas development team. (2024). pandas-dev/pandas: Pandas (v2.2.3). Zenodo. https://doi.org/10.5281/zenodo.13819579

Ti-Rads Calculator. TIRADS Calculator. (n.d.). https://tiradscalculator.com/

[15] Van Den Heede, K., Tolley, N. S., Di Marco, A. N., & Palazzo, F. F. (2021). Differentiated Thyroid Cancer: A Health Economic Review. Cancers, 13(9), 2253. https://doi.org/10.3390/cancers13092253

[16] Venkatesh, N., & Ho, J. T. (2021). Investigating thyroid nodules. Australian prescriber, 44(6), 200–204. https://doi.org/10.18773/austprescr.2021.055

[17] Xi, N. M., Wang, L., & Yang, C. (2022). Improving the diagnosis of thyroid cancer by machine learning and clinical data. Scientific reports, 12(1), 11143. https://doi.org/10.1038/s41598-022-15342-z

Appendices

Code & data for replication can be found in this Github repository here.

- Linear Regression Data.ipynb - Main workspace
- Thyroid_clean.csv - Benign vs Malignant dataset
- Testing.csv - Malignant only ACR and C TI-RADS Dataset
- Training.csv - Malignant only ACR and C TI-RADS Dataset, second file

The original datasets can be found:

- Testing.csv & Training.csv here.
- Thyroid_clean.csv here.