# B. TECH. PROJECT REPORT

On

# Explainable AI For Thyroid Nodule Classification

BY

**Shiva Nunemunthala**

**190001041**

**DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING**

**INDIAN INSTITUTE OF TECHNOLOGY INDORE**

**November 2022**

# Explainable AI For Thyroid Nodule Classification

**A PROJECT REPORT**

*Submitted in partial fulfilment of the*
*requirements for the award of the degrees*

***of***
**BACHELOR OF TECHNOLOGY**
**in**

**COMPUTER SCIENCE AND ENGINEERING**

*Submitted by:*
**Shiva Nunemunthala**
**190001041**

*Guided by:*
**Prof. Kapil Ahuja**

**INDIAN INSTITUTE OF TECHNOLOGY INDORE**
**November - 2022**

## CANDIDATE’S DECLARATION

We hereby declare that the project entitled **“Explainable AI for Thyroid Nodule Classification”**, submitted in partial fulfilment for the award of the degree of Bachelor of Technology in 'Computer Science and Engineering' completed under the supervision of **Prof. Kapil Ahuja** IIT Indore is an authentic work.

Further, I declare that I have not submitted this work for the award of any other degree elsewhere.

Shiva

Shiva Nunemunthala

29/11/21

**Signature and name of the student with the date**

---

## CERTIFICATE by BTP Guide(s)

It is certified that the above statement made by the students is correct to the best of my/our knowledge.

KA

Prof. Kapil Ahuja
Professor in CSE
30/11/2022

**Signature of BTP Guide(s) with dates and their designation**

# Preface

This report on “Explainable AI for Thyroid Nodule Classification" is prepared under the guidance of Prof. Kapil Ahuja.

*(Through this report, we have tried to give a detailed design of explainable AI for classifying thyroid nodules and attempted to cover every aspect of the system.*

*I have tried to the best of my abilities and knowledge to explain the content lucidly. I have also added graphs, tables and figures to make it more illustrative.)*

**Shiva Nunemunthala**  
B.Tech. IV Year  
The discipline of Computer Science and Engineering  
IIT Indore

# Acknowledgements

It is my privilege to express my gratitude to several persons who helped me directly or indirectly to conduct this research project work. I express my heart full indebtedness to my BTP guide **Prof. Kapil Ahuja** for his sincere guidance and inspiration in completing this Project.

I am extremely thankful to **Mr. Saurabh Saini** for his coordination and cooperation and for his kind guidance and encouragement.

I also thank my friends who have more or less contributed to the making of this project. This study has indeed helped me to explore more knowledgeable avenues related to this topic and I am sure that it will help me in future.

**Shiva Nunemunthala**
B.Tech. IV Year
The discipline of Computer Science and Engineering
IIT Indore

# Abstract

Researchers have created a computer-aided diagnosis (CAD) system to aid professionals in medical industry diagnosing thyroid nodules and lessen the errors associated with older techniques. Physician experience is the basis of traditional diagnostic methods. Therefore, such systems' performance plays an important role in improving the quality of diagnostic tasks. But trusting black box models for no reason has many negative consequences. Using a complex model for accuracy comes at the expense of model interpretability. Here, we have worked on the explainability of black-box models using LIME and SHAP models. Our experiments were performed using a common dataset, namely the Thyroid Digital Image Database (TDID), and used HOT, PB-DCT texture exploiting descriptors.

# CONTENTS

**CANDIDATE'S DECLARATION………………………………………………………….3**

**CERTIFICATE BY BTP GUIDE……………………………………………………...3**

**PREFACE………………………………………………………………………………….4**

**ACKNOWLEDGEMENTS…………………………………………………………….5**

**ABSTRACT……………………………………………………………………………..6**

**CONTENTS……………………………………………………………………………..7**

**LIST OF FIGURES………………………………………………………………………9**

**LIST OF TABLES……………………………………………………………………….10**

**INTRODUCTION……………………………………………………………………..11**

EXPLAINABLE AI (XAI)…………………………………………………………13

INTERPRETABILITY……………………………………………………………...14

XAI VS INTERPRETABILITY …………………………………………………15

**LITERATURE REVIEW……………………………………………………………..16**

**TIRADS DATASET ………………………………………………………………..20**

IMAGE BINARIZATION……………………………………………..….21

IMAGE NORMALIZATION…………………………………………….....22

IMAGE ENHANCEMENT………………………………………………...23

**XAI METHODS…………………………………………………………………….25**

LIME………………………………………………………………………………...25

SHAP…………………………………………………………………………...35

**EXPERIMENTAL RESULTS……………………………………………………...39**

DATASET AND EXPERIMENTAL RESULTS……………………………………..39

RESULTS OF LIME……………………………………………………………….40

RESULTS OF SHAP………………………………………………………………41

**CONCLUSION AND FUTURE WORK…………………………………………….42**

**REFERENCES…………………………………………………………………...43**

# List of Figures

**Figure 1**- Illustration of explainable AI ………………………………………………..15

**Figure 2**- Taxonomy of explainable AI……………………...………………………17

**Figure 3**- (a) An input ultrasound thyroid image ; (b) The binarized image ; (c) The region of interest………………………………………………………….……………………22

**Figure 4**- Example of Thyroid image preprocessing; (a) An input image; (b) A cropped image; (c) The enhanced image…………………………………………………...24

**Figure 5**- The dashed line is the learnt explanation that is locally faithful………..………..26

**Figure 6**- Illustration of Felzenszwalb's superpixel algorithm ;(a) Enhanced and preprocessed thyroid nodule image ;(b) Felzenszwalb segmentation of the enhanced image.30

**Figure 7**- Quickshift segmentation…………………………...………………………...31

**Figure 8**- SLIC segmentation………..……………………………………………...32

**Figure 9**- Perturbed image when masking has been done with p=0.5……………………....33

**Figure 10a**- Segmented image by doctor ; 10b- Segmented image by LIME……………….40

**Figure 11a**- Segmented image by doctor ; 11b- Segmented image by SHAP……...……….41

# List of Tables

**Table 1**- Summary of outcome-based explanations of the black box model……………….19

**Table 2**- TIRADS score and its associated suspicion………………………………………..21

**Table 3**- Distribution of benign and malignant images according to TIRADS scores………39

# INTRODUCTION

Explainable Artificial Intelligence (XAI) is a set of procedures and techniques that enables users of machine learning algorithms to comprehend and trust the outcomes and results of those algorithms. The need to interpret machine learning algorithms is useful in many safety-critical applications. One application we have explored is the classification of thyroid nodules.

Thyroid nodules (or lumps), which occur in the human thyroid gland, are a disease in which cells proliferate abnormally and are likely to spread to other parts of the body.  [1]. However, the presence of thyroid nodules may or may not be a sign of thyroid cancer.

When a thyroid nodule is discovered, an ultrasound of the thyroid area is performed to determine whether the nodule is, in fact, cancerous or non-cancerous, which are known, respectively, as benign and malignant nodules in medical terminology. Fortunately, the majority of thyroid nodules found are benign. However, the existence of the nodule (whether benign or malignant) results in a number of health issues for the patient, including breathing and swallowing difficulties [2]. Additionally, thyroxine, a second hormone produced by malignant thyroid nodules, can have serious negative effects on a patient's health and even be fatal. Consequently, diagnosing these nodules at an early stage can decrease the likelihood that the patient will pass away.

The thyroid can be examined using a variety of imaging methods, including a CT scan, an ultrasound, an X-ray, and others[1]. Ultrasound imaging, which uses high-frequency sound waves to produce an image of the internal organs, is the best technique for early thyroid cancer identification.[1][2].

Although utilising artificial intelligence in computer-aided diagnostics offers a promising technique to improve the diagnosing process [3], trusting a black box model is difficult for most medical experts and patients. Because of this, AI methods have yet to achieve significant deployment in the medical industry. Here we provide explanations behind why our algorithm is predicting a particular output.

Our black box model consists of the histogram of Oriented Texture (HOT) and Pass Band - Discrete Cosine Transform (PB-DCT) as texture exploiting descriptors [4].

## XAI (EXPLAINABLE ARTIFICIAL INTELLIGENCE)

Explainable Artificial Intelligence (XAI) is a set of procedures and techniques that enables users of machine learning algorithms to comprehend and trust the outcomes and results of those algorithms. The need to interpret machine learning algorithms is useful in many safety-critical applications, not only in safety-critical applications but also in various sectors like finance, recommendation systems, computer vision etc.

XAI is used to describe an artificial intelligence model and stress its impact and potential biases. Using XAI, we can characterize model accuracy, fairness, transparency and outcomes in artificial intelligence-powered decision-making.

It is one of the rapidly emerging fields in artificial intelligence as XAI is becoming more and more crucial for an organization in building confidence and trust when you want to deploy our artificial intelligence model. As the state-of-the-art models are becoming very less interpretable on the cost of accuracy, We humans are challenged to explain why our algorithm has made such a decision.

Apart from the advantages mentioned above, explainability can help us developers to ensure that our model is working as expected and corresponding regulation standards.

## INTERPRETABILITY

Interpretability is the extent to which we can consistently predict what will happen if there is a small change in input or what will happen if there is a change in weights, aka parameters. It is also the degree to which we understand why our black box model makes such predictions. Interpretability is needed when the model metrics, such as accuracy, F1 score etc., aren't sufficient.

## XAI vs INTERPRETABILITY

Though these two terms appear to be similar, there is a subtle difference between these two. For simplicity, let's take an example of linear regression; here, to understand the output, one can go and check weights assigned to different features used in linear regression to understand why we got this output - This is interpretability.

Understanding why certain features got more weight than others deals with explainability. In literature, both these terms are used interchangeably, but we need to note the subtle differences between these two.

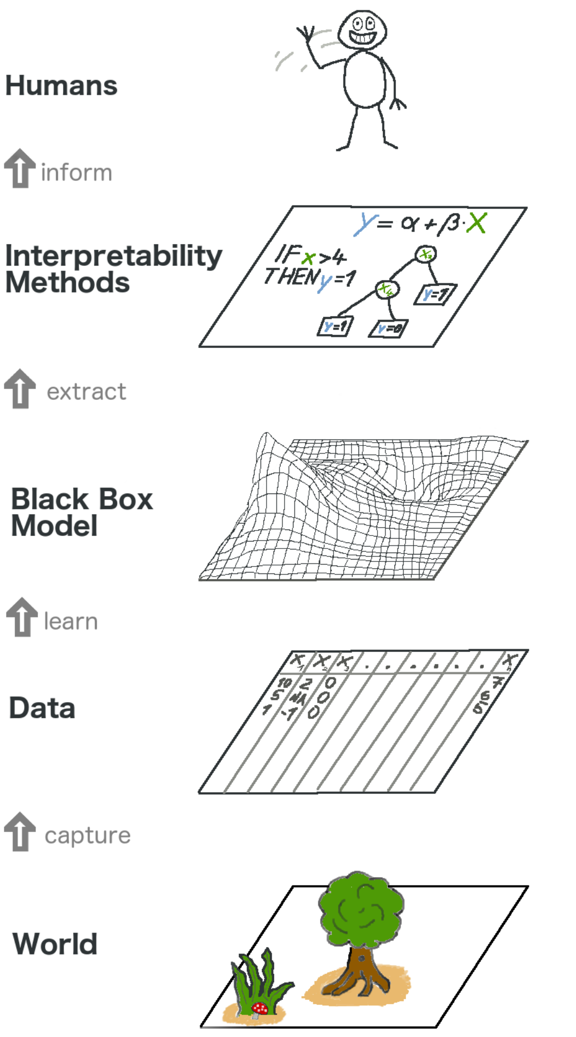


*Figure 1 - Illustration of explainable AI* [5]

In the above figure, the black box artificial intelligence model is interpreted as a decision tree model, which is intrinsically interpretable [5].

# LITERATURE REVIEW

Systems for making decisions have become more prevalent and opaque during the last ten years. These "black box" systems forecast critical data using cutting-edge machine learning methods. Examples include insurance risk, creditworthiness, and health status. We carried out a comprehensive literature survey on the existing techniques of explainable AI models.[3]

## TAXONOMY OF INTERPRETABLE METHODS

There are various XAI classifications, including Opening Black Box Model, Model-Specific Interpretability versus Model-Agnostic Interpretability, Local Interpretability vs Global Interpretability, and Intrinsic vs Post Hoc Interpretability.[5]

In The first classification, i.e. "Opening black box model ", we further classify into "Transparent box design" and "Black box explanation".

In Transparent box design, we use only intrinsically interpretable models as our black box models, such as decision trees, linear regression, logistic regression, etc. But this comes at the expense of accuracy as these intrinsically interpretable models don't achieve performance comparable to state-of-the-art models.

There are three types of "black box model explanations": model explanations, outcome-based explanations, and model inspection-based explanations. We explain why we need to trust our model in the Model explanation. This was accomplished by approximating it to the closest interpretable model, and it can therefore be confirmed.[6]

Model inspection examines how the prediction changes when input data or model parameters change.

The outcome-based explanation is yet another type of explanation. We emphasize on this topic and attempt to explain why the algorithm is making such a decision. The illustration below depicts a metaphorical categorization of XAI methods.

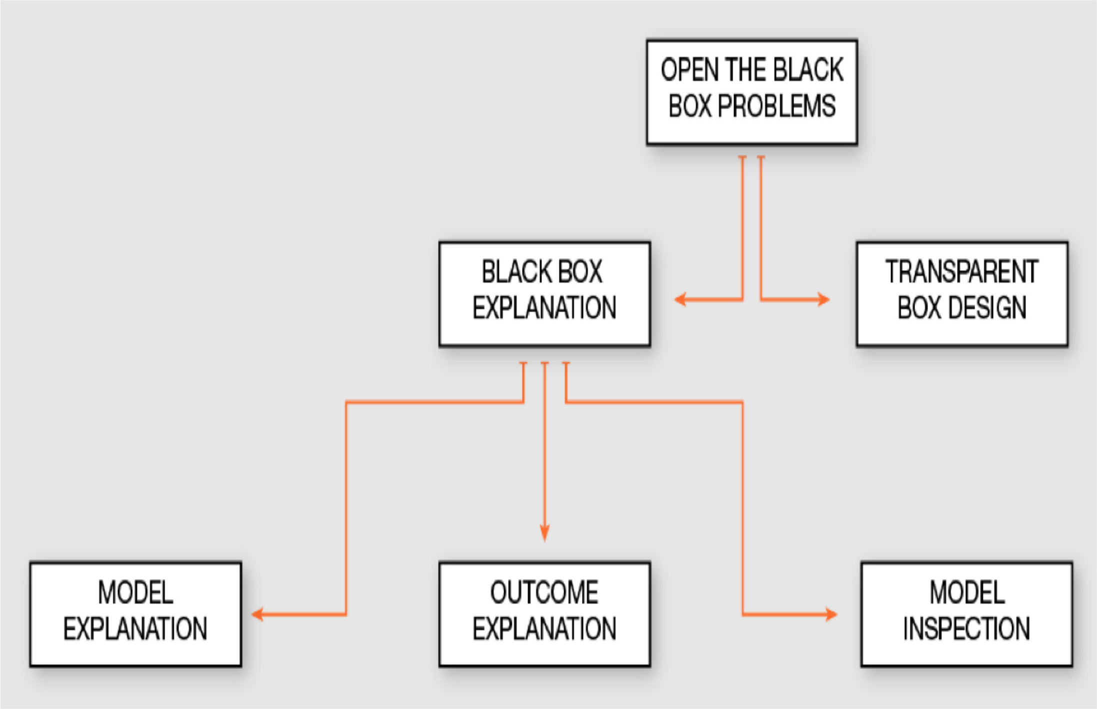


*Figure 2 - Taxonomy of XAI [6]*

The second kind of classification is Model-specific vs Model-Agnostic taxonomy, as the name suggests; model-agnostic in the sense that it is independent of the model or separates the explanations from the model. A model-agnostic explanator has a significant advantage over model-specific ones. The kind of interpretability technique to use depends on the problem we deal with. Some of the desirable aspects that come with model-agnostic system explanations are model flexibility, explanation flexibility, and representation flexibility[5].

Another kind of classification of interpretability techniques is global vs local interpretability [7]. Most of the famous interpretable techniques come under local interpretation. Basically, Local interpretation methods explain individual predictions. Whereas in global, we try to explain the behaviour of the whole black box model. LIME and SHAP are local interpretable models. We also worked on model-specific interpretability by visualizing the HOT feature vector.

| **NAME** | **AUTHOR** | **YEAR** | **EXPLANATOR** | **BLACK BOX** | **DATA TYPE** |
|---|---|---|---|---|---|
| - | Xu et al. | 2015 | SM | DNN | IMG |
| CAM | Zhou et al. | 2016 | SM | DNN | IMG |
| GRAD-CAM | Selvaraju et al. | 2016 | SM | DNN | IMG |
| LIME | Ribiero et al. | 2016 | FI | AGN | ANY |
| MES | Tuber et al. | 2016 | DR | AGN | ANY |
| ANCHORS | Ribiero et al. | 2018 | DR | AGN | ANY |
| LORE | Guidotti et al. | 2018 | DR | AGN | TAB |
| VBP | Bojarski et al. | 2016 | SM | DNN | IMG |
| SHAP | Lundberg et al. | 2017 | SM | AGN | ANY |

*Table 1 - Summary of outcome-based explanation of a black box model* [6]*

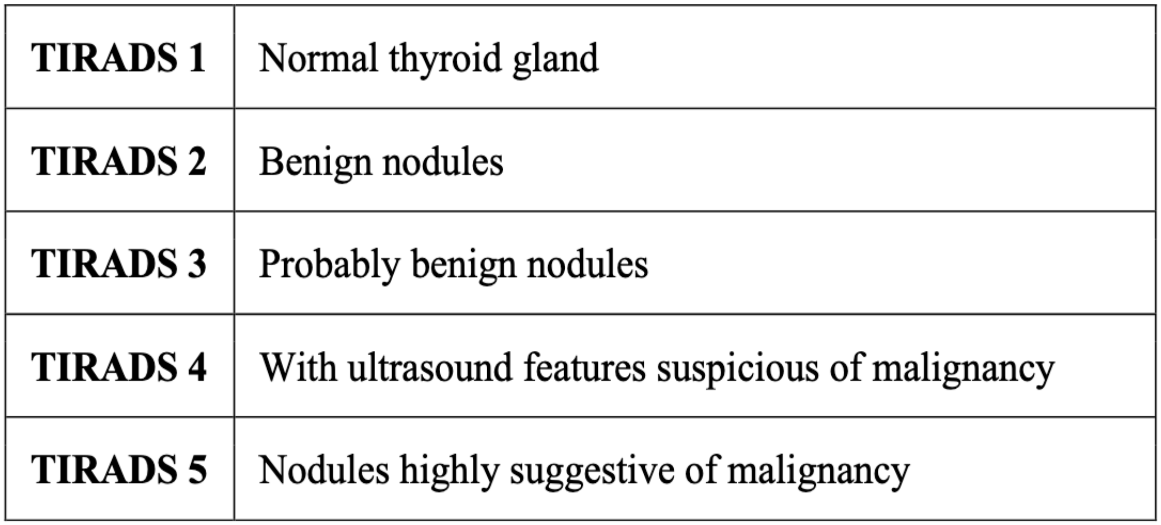
* SM=Saliency mask , DNN= Deep Neural Network , AGN = Agnostic , IMG = Image , TAB = Tabular , DR = Decision Rules

In this work, we work on LIME and SHAP explainable AI models for our black box algorithm. Further, we explore model-dependent explanation by visualizing the HOT feature vector as an image by separately calculating magnitude and slope. Further, we touch upon how Case-Based reasoning techniques can be applied to our model [7][8].

# TIRADS DATASET

This work worked on how Local Interpretable Model Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) can be used to explain our thyroid classification algorithm. We also explore model-dependent models by exploring the HOT feature vector by visualizing its respective gradients and magnitude. [10]

This work used Thyroid Imaging Reporting and Data Systems. TIRADS is a five-point scale for determining cancer risk in thyroid nodules. Nodules are classified as unsuspected, probably benign, with one or more suspicious features, and potentially malignant. These categories are represented by his TIRADS scores of 2, 3, 4, and 5, respectively. Based on this, ultrasound images with TIRADS scores of 2 or 3 are considered benign cases, and ultrasound images with TIRADS scores of 4 and 5 are considered malignant cases [9].



| | |
|---|---|
| **TIRADS 1** | Normal thyroid gland |
| **TIRADS 2** | Benign nodules |
| **TIRADS 3** | Probably benign nodules |
| **TIRADS 4** | With ultrasound features suspicious of malignancy |
| **TIRADS 5** | Nodules highly suggestive of malignancy |

*Table 2 - TIRADS scores and their associated suspicion*

## IMAGE BINARIZATION

The thyroid area and the background (an edge region with low light and some extra artefacts) are the two main components of the ultrasound thyroid images that were recorded (bright inner part capturing details of the thyroid area). It is clear that the background areas do not carry any information identifying thyroid nodules in an image as benign or cancerous. Indicators for the radiologist such as the patient's data or the capturing equipment used during the picture capture phase are also included as artefact information.

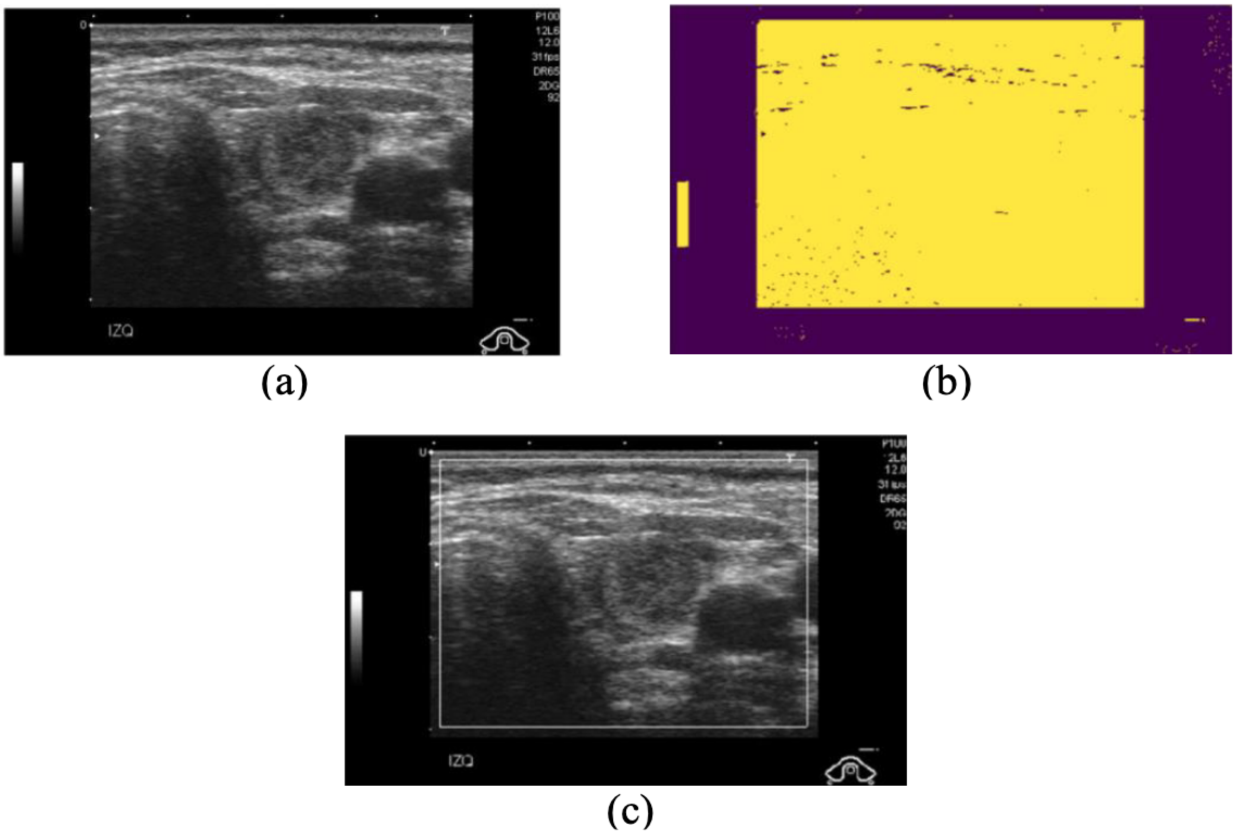


*Figure 3 - (a) An ultrasound thyroid image; (b) The binarized image; (c) The region of interest*

## IMAGE NORMALIZATION

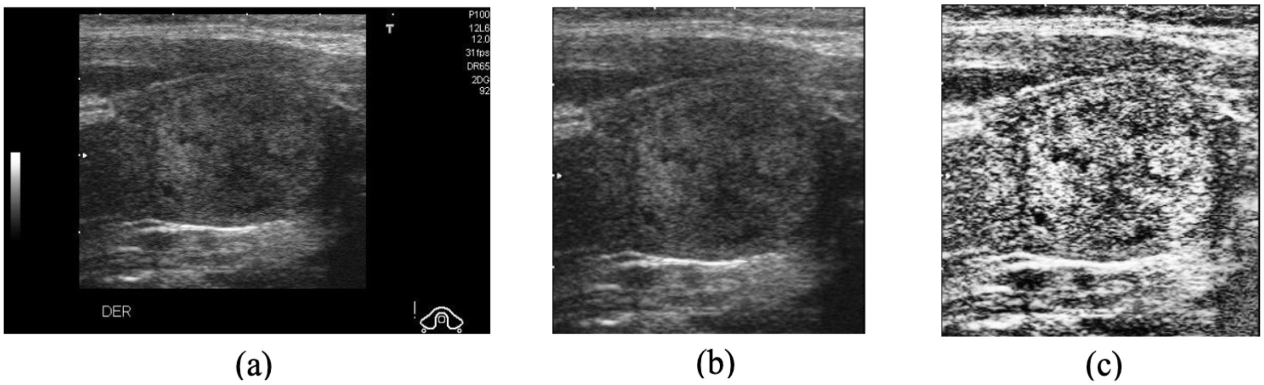
After this, we perform image normalization because while capturing the thyroid nodule images, the colour intensity and illumination conditions might be different, As a result, the range of the grey level varies for various thyroid pictures. Thus, we employ the normalisation algorithm that is most frequently employed, which normalises pixel intensity values between 0 and 1.

$$I'(x, y) = \frac{I(x,y) - min(I)}{max(I) - min(I)}$$

I' (x, y) is the normalised pixel intensity, I(x, y) is the actual pixel intensity, min(I) is the minimum intensity across all pixels, and max(I) is the maximum intensity across all pixels. Where (x, y) is the pixel position.

## IMAGE ENHANCEMENT

After image normalization we perform enhancement of tissues of thyroid nodules. Histogram equalisation is one of the most fundamental strategies utilised here, compressing the contrast in the low histogram regions while stretching the contrast in the high histogram region.As a result, if an image's region of interest is only a small percentage, it will not be improved during histogram equalisation. This gives rise to extremely complex improvement techniques such as i) Contrast Limited Adaptive Histogram Equalization (CLAHE), ii)Adaptive Histogram Equalization (AHE), iii)Two-Stage Adaptive Histogram Equalization (TSAHE), iv)Unsharp Masking (UM), v)Nonlinear Unsharp Masking (NLUM), and so on. [11]. For tissue augmentation in thyroid nodules, CLAHE is more appropriate[4]. We employ a combination of CLAHE and TSAHE since tissues are where most malignant cells are formed and the HOT and PB-DCT descriptors are highly linked to tissue texture. On thyroid nodules, we use two steps of CLAHE in a cascaded fashion. Histogram equalisation is first performed on 8x8 sized blocks, then on 4x4  sized blocks.



(a) (b) (c)

*Figure 4- Example of thyroid image processing (a) an input image; (b) a cropped image; (c)enhanced image*

# XAI METHODS

## LIME

Local Interpretable Model Agnostic Explanations (LIME) [7] is a local surrogate explanatory model that are generally used for explaining individual outputs of black box models. To approximate the predictions of the underlying black box model, substitute models are trained. Instead of building a global surrogate model to explain the whole black-box model, LIME focuses on training local surrogate models to explain particular predictions.

LIME works as follows: it first segments an image and performs n-number of random perturbations with each segment having a probability p of getting picked and that newly generated dataset, and it predicts the output using our black box model. On this newly generated dataset LIME trains an interpretable model whose weights are the distance between the perturbed image and the original image. Let's explore each of the steps in detail.

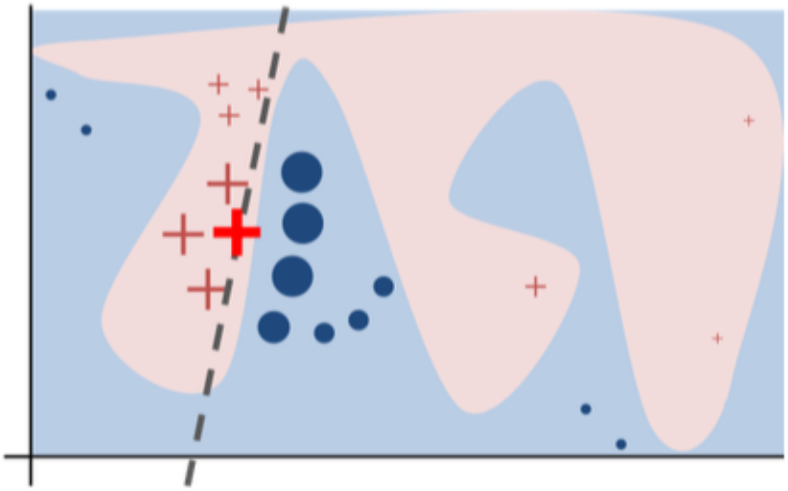


*Figure 5:- The dashed line is the learned explanation that is faithful locally.*

A toy example to demonstrate LIME's concepts. A blue-pink background represents the complex choice of the black-box model, which cannot be well approximated by a linear model. The example being explained is the striking red cross. LIME gathers predictions on the newly generated dataset, and then weighs them in relation to the instance being explained (represented here by size).

## LOCAL SURROGATE MODELS

Black-box machine learning models are explained using specific predictions by employing interpretable local surrogate models. Substitute models are trained to approximate the predictions of the underlying black box model. LIME focuses on training local surrogate models to explain specific predictions rather than developing a global surrogate model.

Mathematically we can define local surrogate models(here LIME) as the one shown below.

$$explanation(x) \ = \ arg \min_{\mathbf{g} \in \mathbf{G}} L(f, g, \ \pi x) \ + \ \Omega(g) \ \text{ - (1)}$$

The above equation can be interpreted as giving an instance x and model g (in our case, locally weighted linear regression) and an explanatory model for given instance x, which minimizes the loss function L, which measures how close our prediction is to the original black-box model f. And $\Omega(g)$ denotes the complexity of the model, which we keep low. And G is the family of all possible explanations (for, e.g.:- all possible locally weighted LRs). And the proximity measure $\pi x$ denotes how large the neighbourhood should be for explaining an instance x. [6]

Suppose G is the group of linear models, such that g(z′) = $wg \cdot z'$. The locally weighted square loss is represented as L , as defined in Eq. (2), where we let $\pi x(z) = exp(- D(x, z)^2 / \sigma^2)$

$$L(f, g, \pi x) = \sum_{z,z' \in Z} \pi x(z)\,(f(z) - g(z'))^2 \qquad - (2)$$

In the above equation, $L(f, g, \pi x)$ represents the aware locality loss without any assumption about $f$. As we need our explainer to be model agnostic.

The general recipe for training a local surrogate model is as follows.

i) Select an instance on which you want explanations.
ii) Generate a list of perturbations of a given instance and find black box predictions.
iii) Weight the generated instances according to proximity measure.
iv) Train an intrinsically interpretable model on the newly generated dataset with the above weights.
v) Explaining the predictions using our interpretable model.

Let's discuss each of the steps in detail. First, we will discuss how to generate perturbations of a given instance x (here, our x is an image, so we speak in the context of the image. )[5]

### Segmentation and Superpixel algorithms

These segmentation and superpixel algorithms segment an image into a couple of regions by using various similarity measures. Here we will explore three kinds of most used segmentation algorithms.

### Felzenszwalb's efficient graph-based segmentation

In the field of computer vision, this quick 2D picture segmentation approach, proposed in [14], is well-liked. The segment size of the method is affected by a single scale parameter. Depending on the local contrast, the number of segments and the actual size can differ significantly.

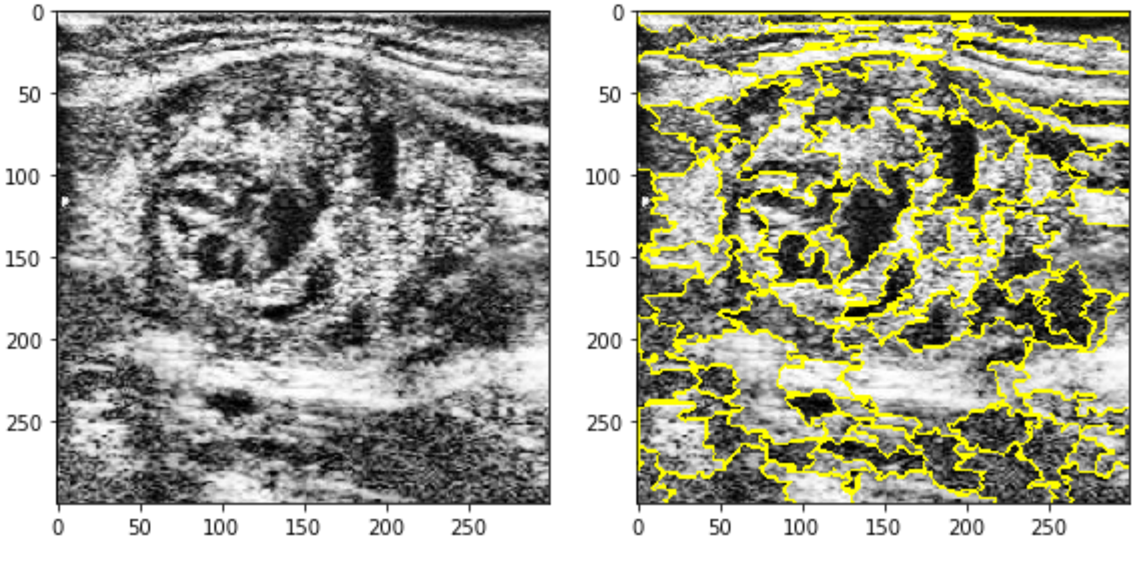


*Figure 6:- Illustration of Felzenszwalb's superpixel algorithm. (a) Enhanced and preprocessed thyroid nodule image. (b) Felzenszwalb segmentation of the enhanced image.*

**Quick-shift image segmentation**

A kernelized mean-shift approximation serves as the foundation for the recent 2D picture segmentation technique known as Quick-shift. It belongs to the family of local mode-seeking algorithms since it utilises the 5D space made up of colour information and picture position.[15]

Quickshift really computes hierarchical segmentation on many scales at once, which is one of its advantages.

The two primary parameters of Quickshift are sigma, which regulates the size of the local density approximation, and max distance, which chooses a level in the generated hierarchical segmentation. Additionally, there is a trade-off between the ratio-determined distance in colour space and the distance in picture space.

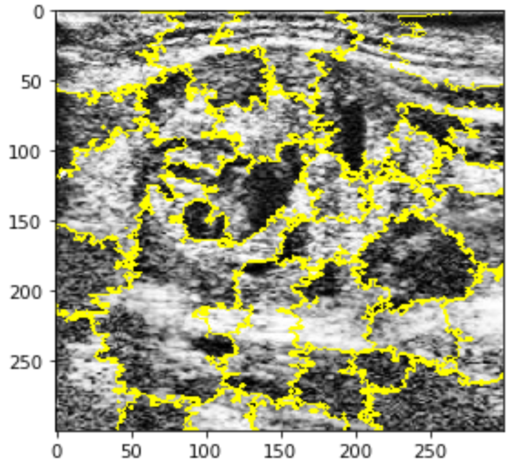


*Figure 7: - Quickshift segmentation*

### SLIC - K-Means-based image segmentation

This approach is comparable to a quick-shift in simply applying K-means to colour information and image location in the 5d space. The clustering method is particularly effective because it is less complicated. For this method to produce good results, it must operate in Lab colour space. The algorithm gained popularity fast and is currently frequently utilized. The compactness parameter, like Quickshift, trades off colour similarity and proximity, while the n segments parameter sets the number of centres for kmeans.[16]

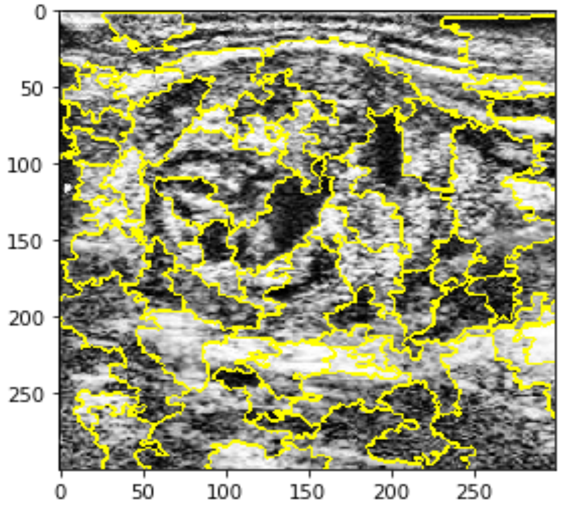


*Figure 8:- SLIC segmentation*

### Perturbation of segmented image

After segmenting the sample of interest, we generate an n-number of perturbations by randomly turning off(masking) each segment with a probability 'p'. Where p and n are hyperparameters, we need to tune. One good hyperparameter for the value of p is 0.5, and n is the length of the image. The below figure represents a perturbed image.

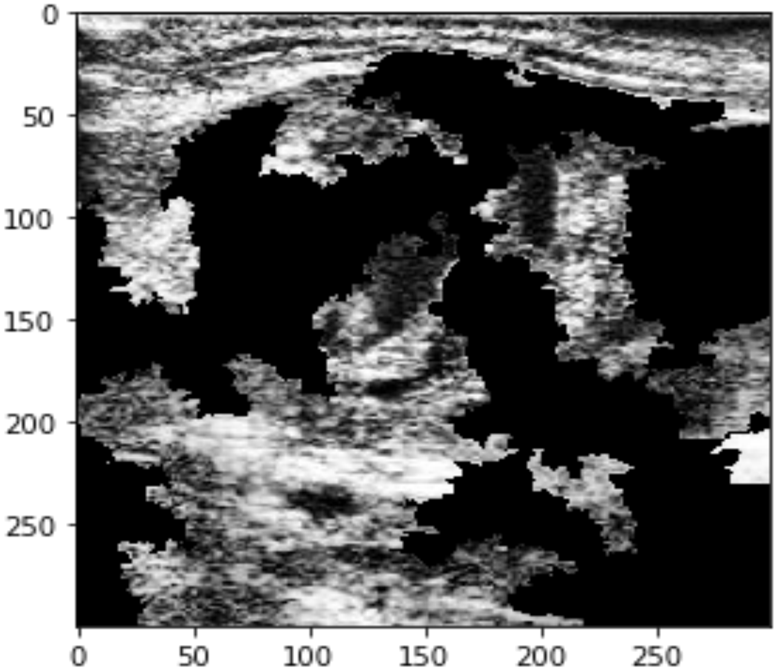


*Figure 9:- Perturbed image where masking has been done with p=0.5*

Next we find the pairwise distance between the two images in the next step. Here, we compare the altered image to the original image using cosine similarity.

Cosine similarity between 2 matrices X and Z is given as K(X,Z) = $\frac{<X,Z>}{(|X| * |Z|)}$ . On L2-normalized data, this function is equivalent to the linear kernel. Here X denotes our sample of interest and Y denotes the single perturbation, and we represent K(X, $Y_i$) as $\boldsymbol{w^i}$, where $w^i$ denotes the proximity measure of the perturbation to the sample of interest X.

Normalizing the weights is the most common technique in the field of ML. We normalize the weights using a kernel function which brings them between 0 and 1.

### Sparse Linear Explanations using LIME [7]

```
Require: Classifier $f$, Number of samples $N$
Require: Instance $x$ and its interpretable version $x'$
Require: Similarity kernel $\pi x$, Length of explanation $K$
$Z \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ do
        $z_i^{'} \leftarrow$ sample around($x^{'}$)
        $Z \leftarrow Z \cup \langle z'_i, f(zi), \pi x(zi) \rangle$
end for
$w \leftarrow$ K-Lasso($Z$, $K$) . with $z'$
i as features, $f(z)$ as target
return $w$
```

## SHAP

SHAP stands for Shapley Additive Explanations. It is based on the Shapley values which was derived from coalition game theory. SHAP is inspired by Local Surrogate models (which we have defined in LIME section). Before going to SHAP, let's discuss Shapley values and additive feature attribution methods.[8]

### Shapley Values

A prediction can be expressed by supposing that each feature value of the instance is a "player" in a game where the prediction is the prize. Using Shapley values, a technique from coalitional game theory, we may determine how to equitably distribute the "payout" across the features.

The main purpose of these Shapley values in the explanation of an instance is we want to know how each feature contributes to the prediction. For example, let's consider a linear model in (3)

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots.. + \beta_p x_p \qquad - (3)$$

Here x is the instance, and we need the contributions of each feature. And let $p$ denotes the number of features of $x$, and $\beta_j$ is the weight of feature j. Then we have the contribution of feature j as

$$\phi_j(f) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j) - (4)$$

Here $E(\beta_j X_j)$ in eq - (4) means the mean effect estimate of feature j. Intuitively, the contribution should be equal to the feature effect - average effect.

$$\sum_{j=1}^{p} \phi_j(f) = \sum_{j=1}^{p} (\beta_j x_j - \beta_j E(X_j)) = f(x) - E(f(x)) - (5)$$

This sum of all feature contributions should be equal to the prediction-base measure, which can be seen from eq- (5).

The value function val of players in S is used to define the Shapley value. We define the contribution of feature $\phi_j$ of val in S as

$$\phi_j(val) = \sum_{S \subseteq \{1,2...p\} \setminus \{j\}} \frac{|S|!\,(p-|s|-1)!}{p!} (val(S \cup \{j\}) - val(S)) \quad - (6)$$

Here $S$ denotes the subset of features that were being used in the black-box model and, $x$ is the list of features of an instance, p is the number of features. Further, let's define $val_x(S)$ as the prediction of feature values in set S

$$val_x(S) = \int f(x_1, x_2, x_3 ...... x_p)\, dP_{x \in S} - E_x(f(x)) \quad - (7)$$

For estimating a Shapley value, all possible coalitions should be evaluated without considering feature $j$.The number of coalitions grows exponentially as the number of features grows, and we employ approximation with Monte Carlo Sampling.[18]

$$\phi_j = \frac{1}{M} \sum_{m=1}^{M} (\hat{f}(x_{+j}^{m}) - (\hat{f}(x_{-j}^{m})) - (8)$$

Where $\hat{f}(x^{m}_{+j})$ represents the prediction of x but with feature values from a random data point z aside from the specific feature value j. similarly, we define $\hat{f}(x^{m}_{-j})$ as almost identical to $\hat{f}(x^{m}_{+j})$ but the value $x^{m}_{j}$ is also taken from sampled z. And thus, we estimate the Shapley value corresponding to each feature.

**SHAP AS ADDITIVE FEATURE ATTRIBUTION METHODS**

A basic model is easiest to understand when used alone because it accurately depicts itself. Because the original model is difficult to comprehend, we cannot utilize it as the best explanation for sophisticated models like ensemble methods or deep networks. An easier explanation model must be used in its place, which we define as any comprehensible approximation of the original model.

Let $f$ be the initial/original black-box model that needs to be explained, and let $g$ be the explanatory model, as we are focussing on local surrogate models; that is, we are trying to explain individual instances. Here in additive feature attribution methods, we use simplified inputs. Let's call them $x'$ and let these inputs map through a mapping function. Let's call it $h_x(x')$. Then we pick $h_x(x') \cong x$. And the local methods also makes sure that $g(z') \cong f(h_x(x'))$ whenever $z' \cong x'$ .

Additive feature attribution methods have an explanatory model that is a linear function of binary variables

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i^1 \quad - \ (9)$$

Where $z_i^1 \in \{0,1\}^M$ , M is the simplified input features, and $\phi_i$ is the associated Shapley value.

The prediction f (x) of the original black-box model can be approximated by methods with explanatory models that satisfy equation (9) and attribute an effect $\phi_i$ to each feature.

# EXPERIMENTAL RESULTS

## Dataset and Experimental Setup

As mentioned earlier, we use the TDID database for our experiments, which consists of 349 images. Each original image is of size 360 × 560, which becomes 300 × 300 after preprocessing. Out of these, 61 are benign, while 288 are malignant. Table 3 lists the number of images in TDID based on the TIRADS classification.

| TIRADS | No. of Images | Classification (Total Images) |
|---|---|---|
| 2 | 42 | Benign (61) |
| 3 | 19 | |
| 4 | 243 | Malignant (288) |
| 5 | 45 | |

*Table 3:- Distribution of malignant and benign images with their TIRADS scores*

Experiments are carried out on MATLAB-2020 and python 3.9 with an intel i5 processor on a MacBook air 2019.

## RESULTS OF LIME

The below figure shows the top 5 features which are leading to malignancy of a given thyroid nodule image. We have performed 300 perturbations with probability p=0.5 of picking a superpixel.

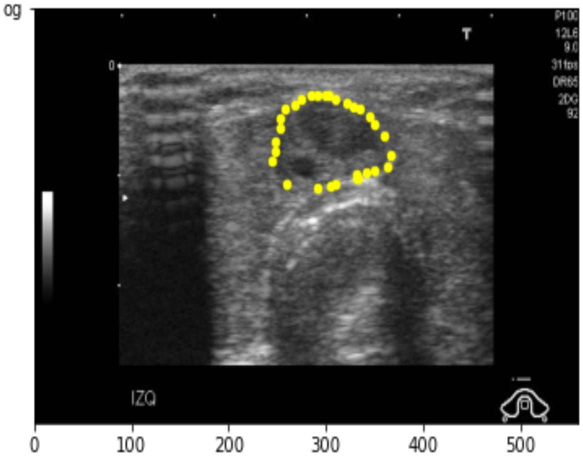


*Figure 10 a:- Segmented image by doctor*

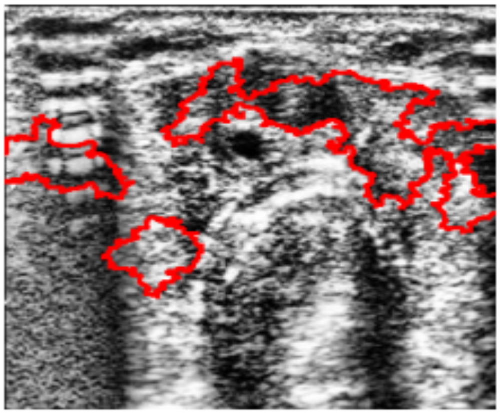


*Figure 10 b:- Segmented image by LIME*

## RESULTS OF SHAP

The below figure shows the importance of each segment to malignancy. The below number line represents the SHAP value of the associated mask/features. The higher the SHAP value, the more it contributes to the corresponding label.

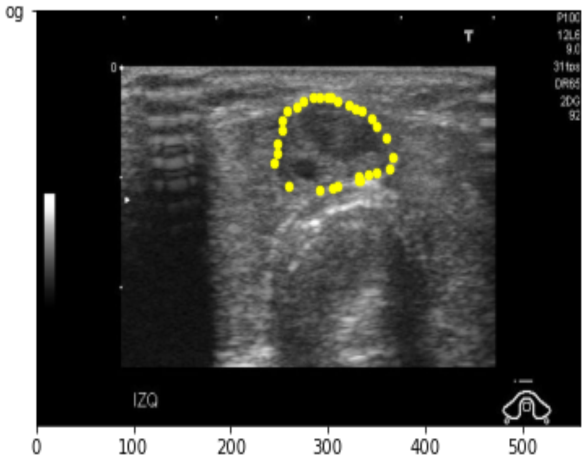


*Figure 11a:- Segmented image by doctor*

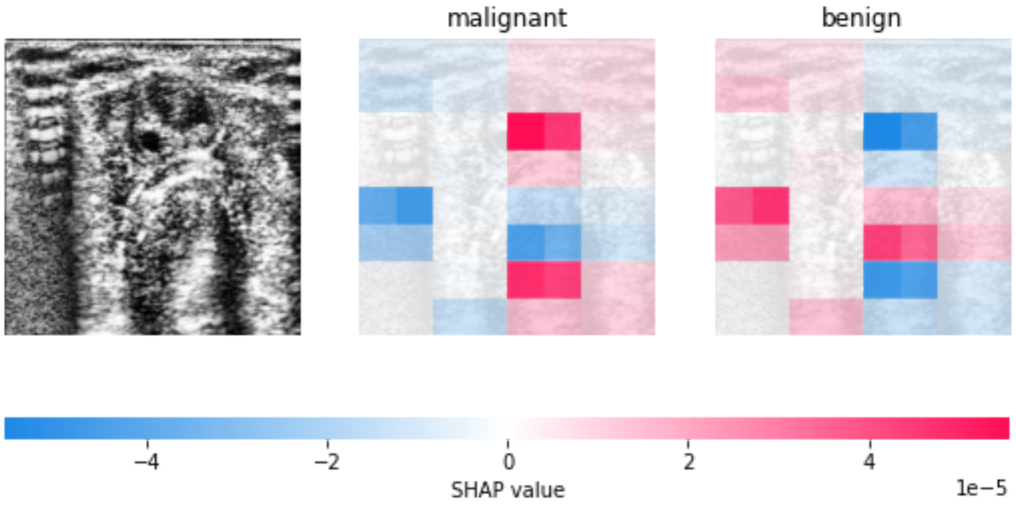


*Figure 11b:- Segmented image by SHAP. As we can see, the thick red portion in the second figure corresponds to higher malignancy, which was segmented by the doctor.*

## CONCLUSION AND FUTURE WORK

We present the working of LIME and SHAP and the literature review corresponding to interpretability/XAI as detailed as possible. We achieved a decent result using multiple superpixel techniques for LIME. We used Shapley Additive Explanations to determine which portion corresponds to malignancy and which does not.

In future, this work can be extended by working on case-based reasoning (CBR) models, which are inherently more interpretable than LIME and SHAP. This may also be extended to work on cutting-edge models in the field of interpretability, such as XRAI [17], SmoothGrad-based XAI techniques etc.[19]

Further, this work can be extended for studying model-specific interpretability and interpretability of our texture exploiting descriptors such as HOT and PB-DCT.

# References

[1] “Thyroid cancer: Diagnosis,” American Society of Clinical Oncology, [Online]. Available: https://www.cancer.net/cancer-types/thyroid-cancer/diagnosis. [Accessed October 2020].

[2] D. Nguyen, J. Kang and T. Pham, “Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence,” Sensors, p. 20, 2020.

[3] D. Nguyen, T. Pham and G. Batchuluun, “Artificial intelligence-based thyroid nodule classification using information from spatial and frequency domains,” Clinical Medicine, p. 11, 2019.

[4] A. Shastri, “Novel Statistical and Probabilistic Machine Learning Algorithms for Genotype Clustering and Cancer Classification," PhD. Thesis, IIT Indore, 2020.

[5] C. Molnar, “A Guide for Making Black Box Models Explainable “, [Online].

[6] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, "A Survey of Methods for Explaining Black Box Models".ACM Computing Surveys 51(5), 2018.

[7] M.T. Ribiero, S. Singh, C. Guistren, "Why Should I Trust You - Explaining the predictions of any classifier.”In *Proceedings of the 2016 Conference of the North American*

*Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics, 2017.

[8] S. Lundberg, "A Unified Approach to Interpreting Model Predictions." NIPS ,2017.

[9] E. Romero, "An open access thyroid ultrasound-image Database ."10th International Symposium on Medical Information Processing and Analysis ,2015.

[10] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection."2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 05), 2005.

[11] S. Anand , S. Gayathri, “Mammogram image enhancement by two-stage adaptive histogram equalization.” Optik-International Journal for Light and Electron Optics, pp. 3150-3152, 2015.

[12] A. Melis, David , S. Jaakkola. “On the robustness of interpretability methods.” arXiv preprint arXiv:1806.08049 (2018).

[13] Slack, Dylan, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. “Fooling lime and shap: Adversarial attacks on post hoc explanation methods.” In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180-186 ,2020.

[14]  P.F. Felzenszwalb ,D.P. Huttenlocher . “Efficient graph-based image segmentation, Felzenszwalb.”International Journal of Computer Vision, 2004.

[15] A. Vedaldi ,S. Soatto .”Quick shift and kernel methods for mode seeking.”,  European Conference on Computer Vision, 2008.

[16] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Suesstrunk, “SLIC Superpixels Compared to State-of-the-art Superpixel Methods.” TPAMI, 2012.

[17] A. Kapishnikov, T. Bolukbasi, M. Terry " XRAI: Better Attributions Through Regions ." 2019 IEEE/CVF International Conference on Computer Vision (ICCV) , 2019.

[18] Štrumbelj, Erik, I. Kononenko. “Explaining prediction models and individual predictions with feature contributions.” Knowledge and information systems 41.3 (2014): 647-665, 2014.

[19] D. Smilkov , Nikhil T , M. Wattenberg. “SmoothGrad: removing noise by adding noise.” ICML , 2017.