DATS-6501 Data Science Capstone

Anomaly detection in Wood Fossil

Prof. Edwin lo

Final Report

Pranay Bhakthula

Table of Contents

Introduction:

The secret to humanity's success is probably in how well we are able to use tools to mediate environmental change. But practically all of our present understanding of early cultural evolution comes from research on stone tools and fossilized bones discovered in the archaeological record. Due to their inherent fragility, tools constructed of plants are almost never found in the earliest records of human material culture. The fact that plant materials are used for tools much more frequently than stone in contemporary human civilizations and among non-human primate species raises the possibility that a significant portion of ancient technology is being left out of current archaeological data.

Here, we provide methods for analyzing damage patterns in living primates' percussive wooden tools by performing analysis on images of those tools. Our research demonstrates that the harm done is irreversible and may endure throughout the fossilization processes. This study provides the opportunity to examine organic artifacts, a significant but overlooked part of the evolution of technology within the primate order.

This project is done under Professor Chen Zeng [Department of Physics and Data Science Program, George Washington University] as a part of a research paper which is continuation of previous research paper which prof. Chen co-authored 'Chimpanzee wooden tool analysis advances the identification of percussive technology'.

The aim of this project is to find a best model that can distinguish between the damaged and undamaged parts of a wood fossil.

Dataset Description:

Fieldwork to collect percussive tools used by chimpanzees was carried out during December 2017 and March 2018 in the North Group of the Tai Chimpanzee Research Project in the Tai National Park (Cote d'Ivoire).
Dataset link : http://cdna.eva.mpg.de/Organic_Tool_Data/

The dataset contains 8 high resolution images of wood fossils which are dated more than 1 million years. Of the dataset 7 images are of undamaged wood fossils while the target image we are performing our majority part of the project is "FW31_Damaged_14525_1206_8_18022019.png" has the damaged part. We perform the majority of the analysis on the FW31 image.

Theory:
Convolution Networks:

Convolution Neural Network -- [S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.]

A convolution network is a multilayer feedforward network that has two or three-dimensional inputs. It has weight functions that are not generally viewed as matrix multiplication operations. In general we use CNN to analyze and classify images given in the dataset.

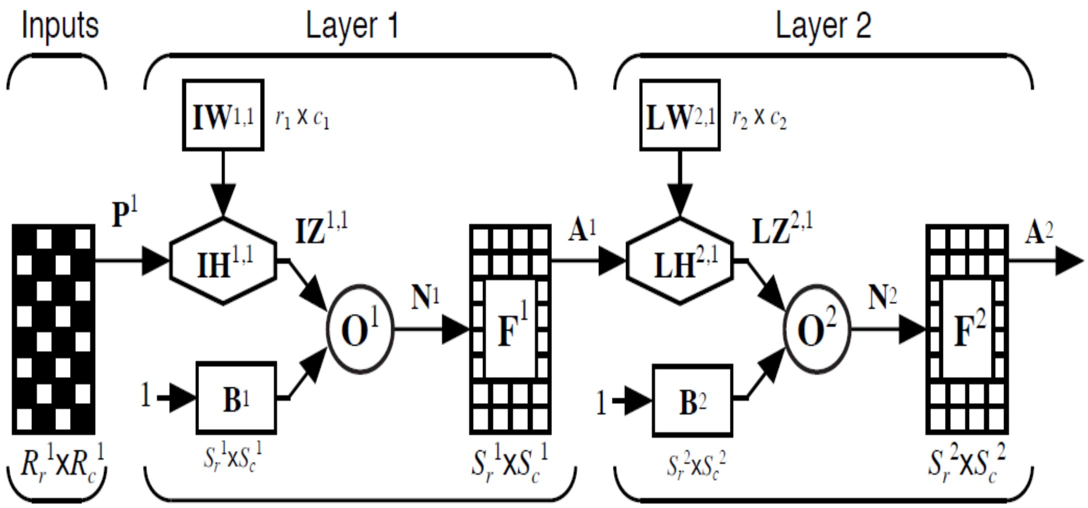


Figure1 : convolution network

Let the input image be represented by the Rr x Rc matrix V. The weight function for this layer performs a convolution kernel that is represented by the r x c matrix W.

$$z_{i,j} = \sum_{k=1}^{r} \sum_{l=1}^{c} w_{k,l} v_{i+k-1,j+l-1} \,.$$

Figure2 : weight function

In matrix form we write it as:

$$\mathbf{Z} = \mathbf{W} \circledast \mathbf{V}$$

Figure3 : weight function matrix

Resent50:

Restnet50 -- [Source: arXiv:1512.03385**,** Authors: Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun]

ResNet50 is a variant of ResNet model which has 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer. It has 3.8 x 10^9 Floating points operations. The Resnet50 model achieved 92.1% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes.

Resnet 50 Architecture:

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		[3×3, 64; 3×3, 64] ×2	[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_x	28×28	[3×3, 128; 3×3, 128] ×2	[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
conv4_x	14×14	[3×3, 256; 3×3, 256] ×2	[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512; 3×3, 512] ×2	[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure4: Resnet 50 Architecture

VGG19:

VGG19 -- [Source: arXiv:1409.1556**,** Authors: Karen Simonyan, Andrew Zisserman]

VGG19 is a simple and widely used convolution neural network. In this abbreviation 19 means that this model has 19 layers. It has 19.6 x 10^9 floating points operations. The VGG19 model achieved 90.0% top-5 test accuracy in ImageNet. It made improvements over AlexNet architecture by replacing large kernel-sized filters with multiple 3x kernel-sized filters one after another. VGG19 was trained for weeks using NVIDIA Titan Black GPUs. We can observe that since Resnet50 has lesser parameters than VGG19 it is faster and also has higher accuracy in ImageNet.

VGG Architecture:

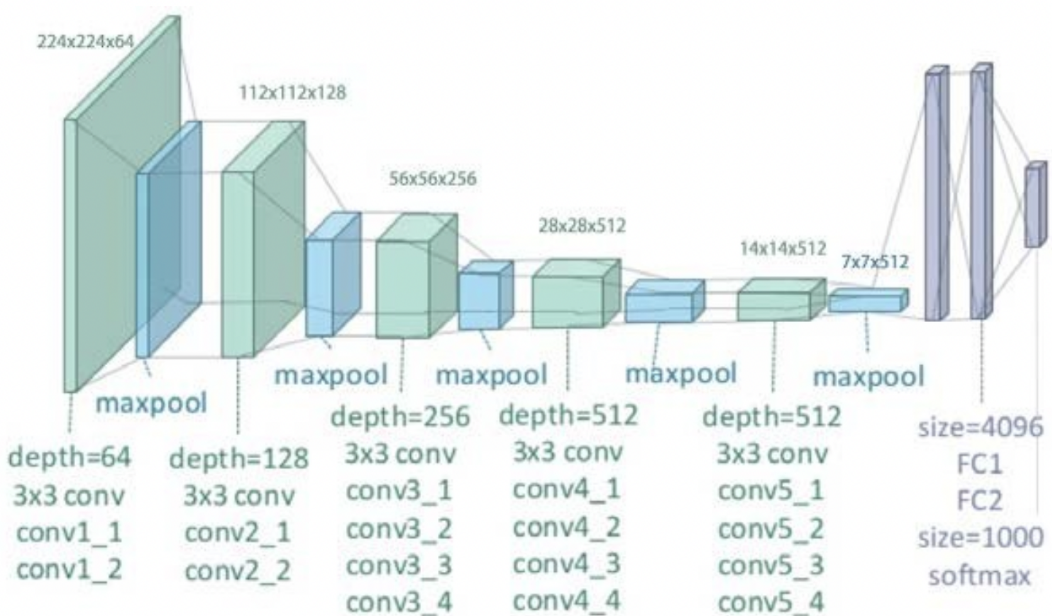


Figure5: VGG19 Architecture

Principal Component Analysis:

Principal Component Analysis – [Source: https://doi.org/10.1098/rsta.2015.0202, Authors: Ian T. Jolliffe, Jorge Cadima]

Principal component analysis (PCA) is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables. Its goal is to extract the important information from the table, to represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity of the observations and of the variables as points in maps.

In the project we use PCA for clustering purposes i.e; we cluster the outputs of pre-trained models.

Gram matrix:

[Source: https://arxiv.org/abs/1912.12510, Authors: Chandramouli Shama Sastry, Sageev Oore]

Gram matrix is the output of a matrix multiplied by its transpose. Gram matrix provides degree of correlation between vectors of a matrix so in this project we can use gram matrix to find correlation between parameters of different convolutional filters in a Convolutional Neural Network.

$$g_{ij} = \mathbf{v}_i^{\mathrm{T}} \mathbf{v}_j.$$

Figure6: gram matrix

Haralick Features:

[Source: https://ieeexplore.ieee.org/document/4309314, Authors: Robert M. Haralick, K. Shanmugam, Its'hak Dinstein]

Haralick features provide some of the easily computable textural features based on gray-tone spatial dependencies. There are 28 haralick features but in this project we use only 13 out of 28 features which are most commonly used.

Methods:

In this project we aim to find a method to train a machine to detect anomalies in the wood fossil; i.e, it can detect the damaged part of the fossil from the undamaged part.

For this problem we will analyze a few "features" of the wood fossil images to see if they can separate the damaged from the undamaged portion via unsupervised linear PCA or nonlinear AE/VAE (autoencoder or variational autoencoder).

The features that we use in this capstone project will be:

1. Haralick features
2. VGG's 4096 features normally used in transfer learning
3. Resnet50's Gram matrix used in neural style transfer

Finally, for each of the above features, we run PCA and AE/VAE to visualize data clustering, asking if the damaged crops can be somewhat separated from the undamaged crops

Some of the limitations of the project are the damage on the surface and internal structure will most likely be impacted by the physical properties of wood. These properties vary widely depending on tree species and water content at the time of use. Our study has focused on the materials that were selected by chimpanzees for nut cracking in the Tai Forest. As a result, we only investigated the damage pattern of the most prevalent wood species (Coula edulis).

The required resources for the project are usage of python libraries and cloud computing. Significant progress can be made in the following 2 months and come out with insightful findings from wood fossils.

The one real time usage of this project when fully implemented would be useful for archeologists in remote parts of the world to take pictures of fossils and anomalies/unusual/important parts can be detected instantly, instead of transporting the delicate fossil to labs for analysis.

Experimental Setup:

First, we crop the wood fossil image in the sizes of 1 x 1mm, 3 x 3mm, 5 x 5mm, 7 x 7mm and store them separately in each folder. We label the damaged part of the image as “damaged” and the undamaged part of image as “undamaged”.

We have 2 methods: one is by using haralick features and the other is using pre-trained models.

For the 1st method, we extract haralick features from all the images. We normalize the values and apply the PCA model on it. We do this process for all the various sizes of images.

For the 2nd method, since our aim is to use pre-trained models like Resnet50 and VGG19, we convert the image into size of 244x244x3 as it is the input sizes of mentioned pre-trained models.

We then apply pre-trained models on the re-sized images and store outputs after applying the gram matrix on 1^{st}, 22^{nd}, 23^{rd}, 24^{th}, 25^{th}, 26^{th}, 27^{th} and 49^{th} layers for Resnet50 model and 1st, 9th and 18th layers of VGG19 model.. We perform this step separately for all the various sizes of crops.

Finally, we apply the PCA model on outputs of all the above-mentioned layers for each of the various cropped images and plot a scatter plot of PC2 vs PC1. We color code damaged as ‘red’ and undamaged as ‘blue’ and check to see if the PCA model has separated them distinctively.

We also repeat the above process for other wood fossil images and check if the model still holds up by clustering these images in places other than our damaged image.

Results:

When we apply the PCA model on haralick features we get the following graphs.

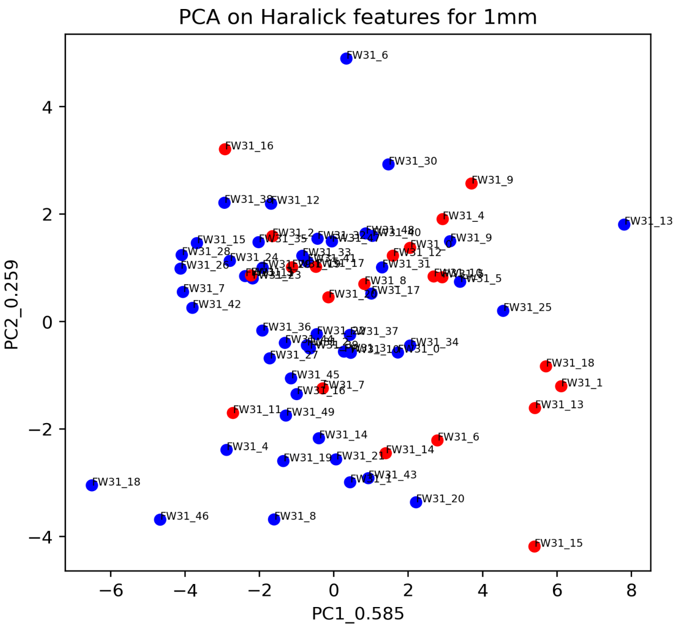


Figure7: 1mm

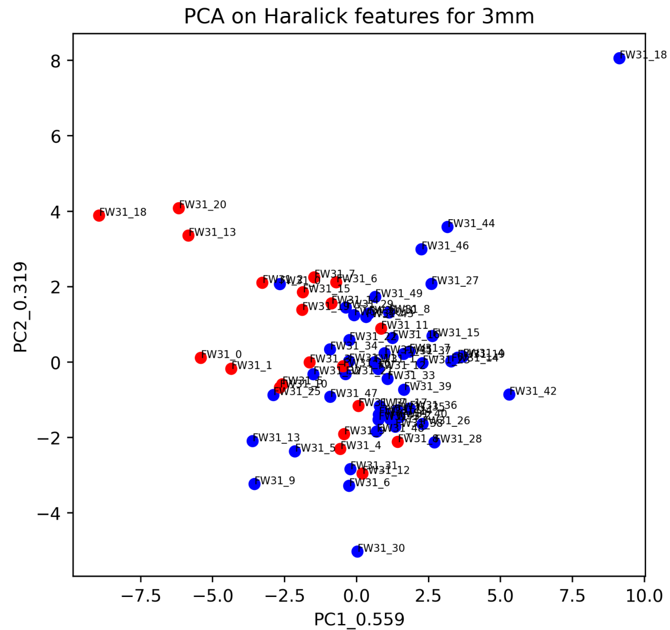


Figure8: 3mm

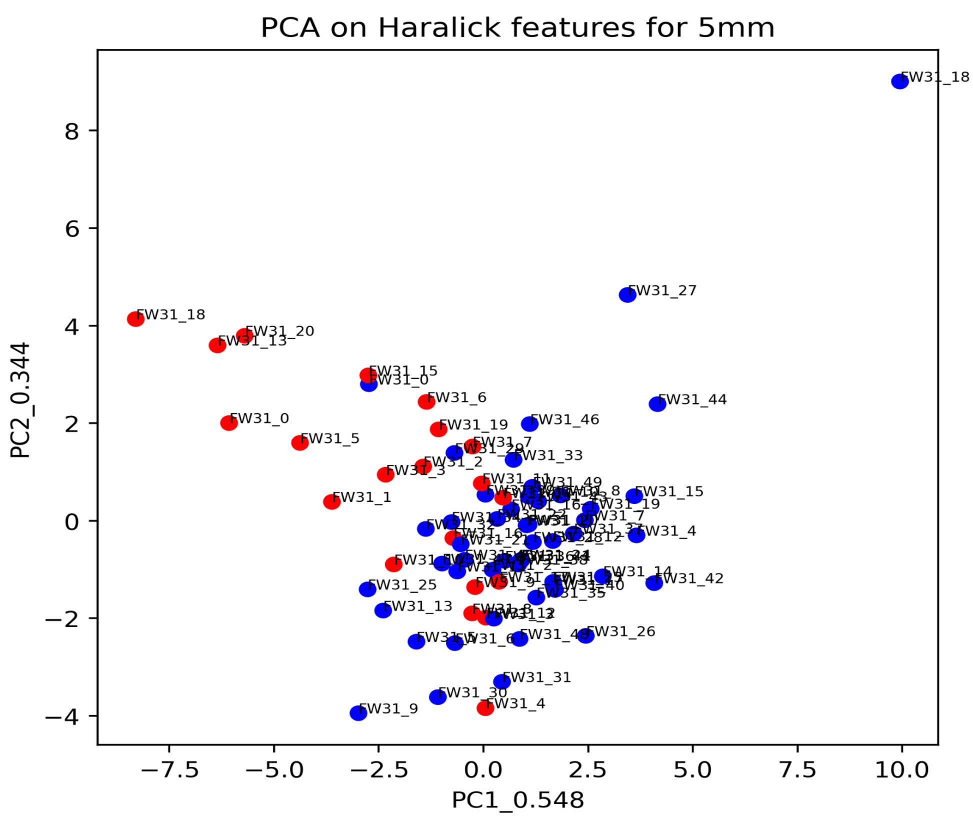


Figure9: 5mm

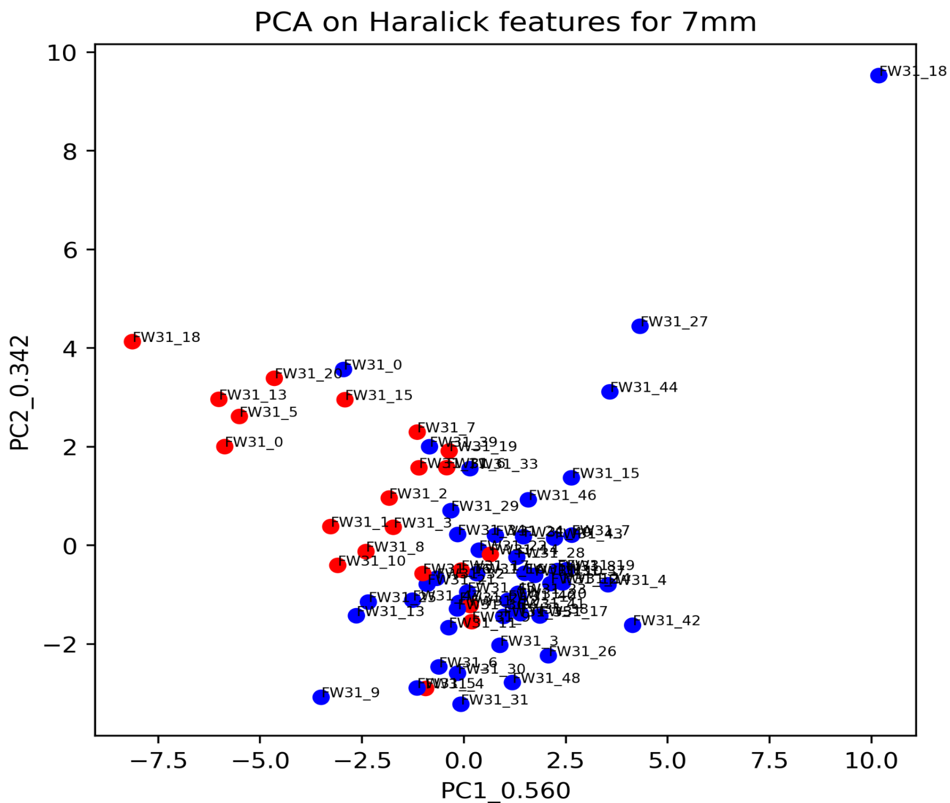


Figure10: 7mm

From the above graphs we can observe that the 1mm and 3mm PCA graphs do not do a good job of separating the damaged and undamaged images because the data points are all mixed up with each other. If we look at the 5mm and 7mm graphs we can see that some of the points representing damaged images are separated but many of the points are still mixed up.

When we apply the PCA model on output of various layers of Resnet50 we get the following graphs.

For 1mm cut:

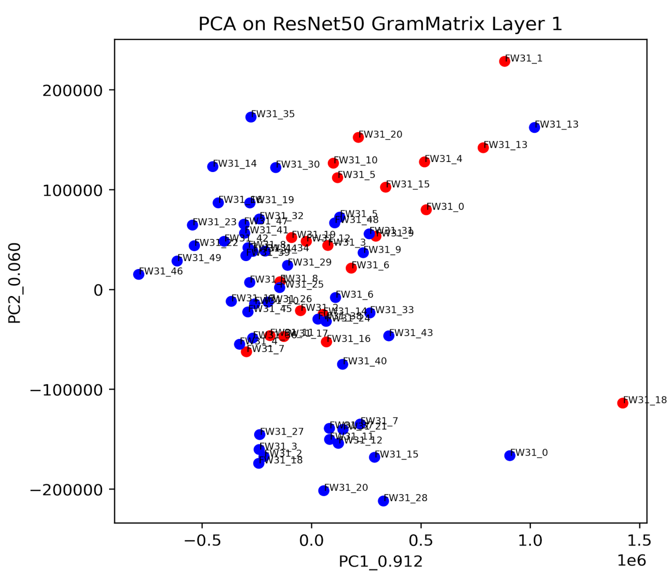


Figure11: 1st layer

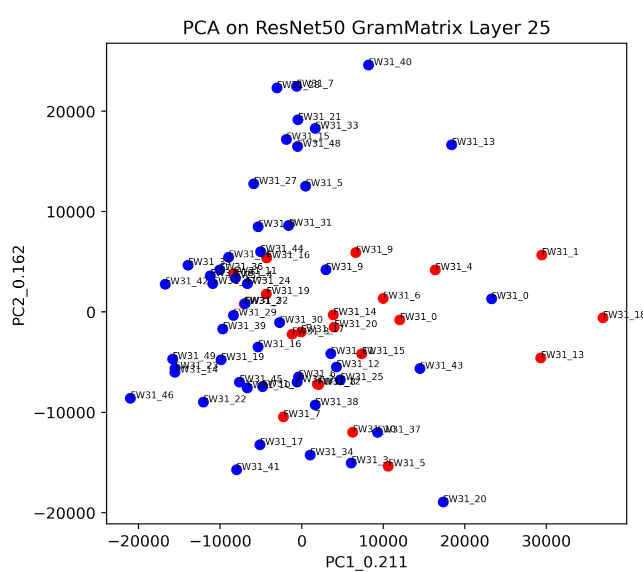


Figure12: 25th layer

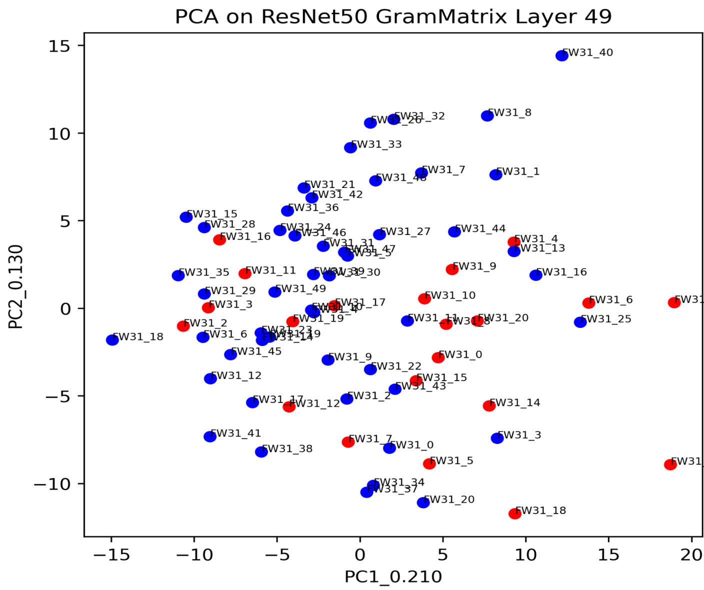


Figure13: 49th layer

For 3mm cut:

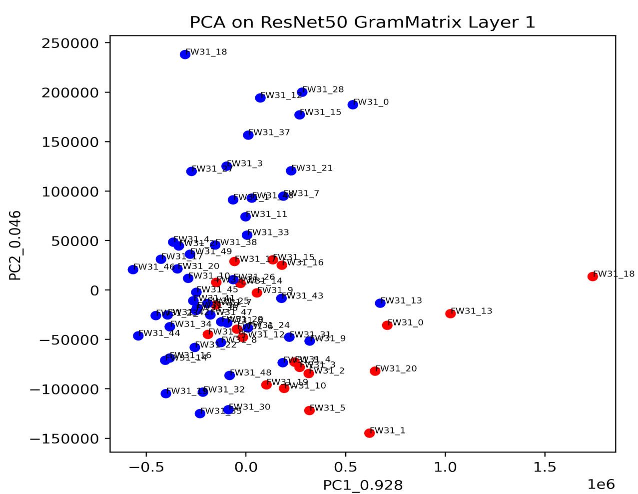


Figure14: 1st layer

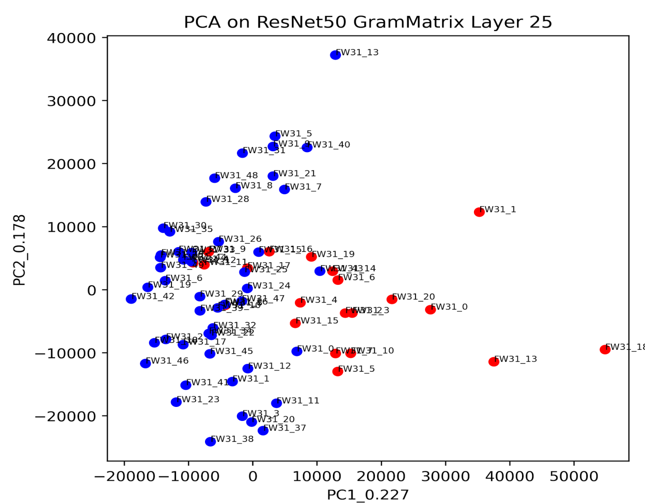


Figure15: 25th layer

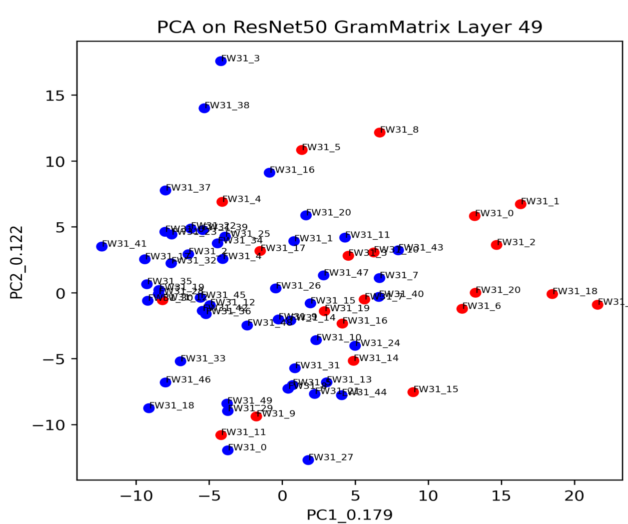


Figure16: 49th layer

For 5mm cut:

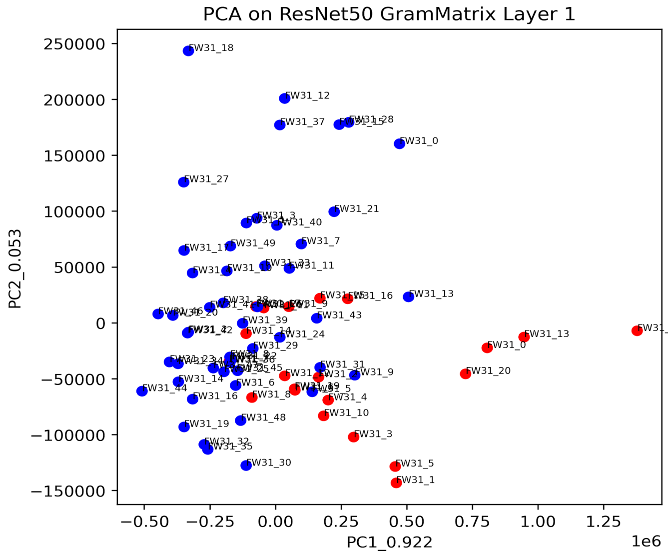


Figure17: 1st layer

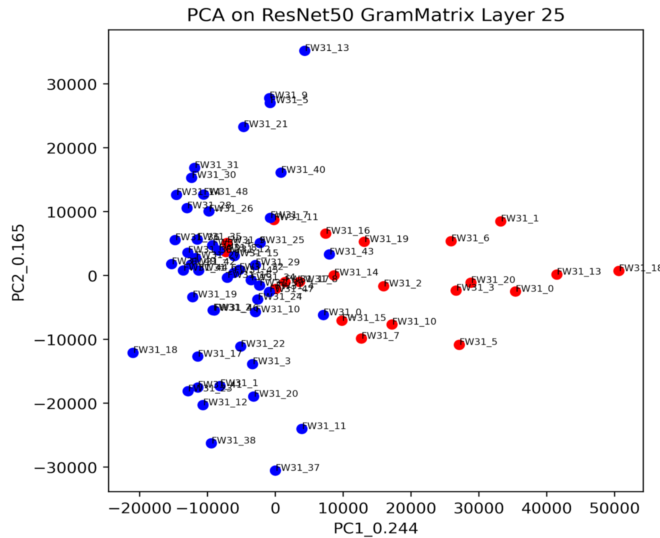


Figure18: 25th layer

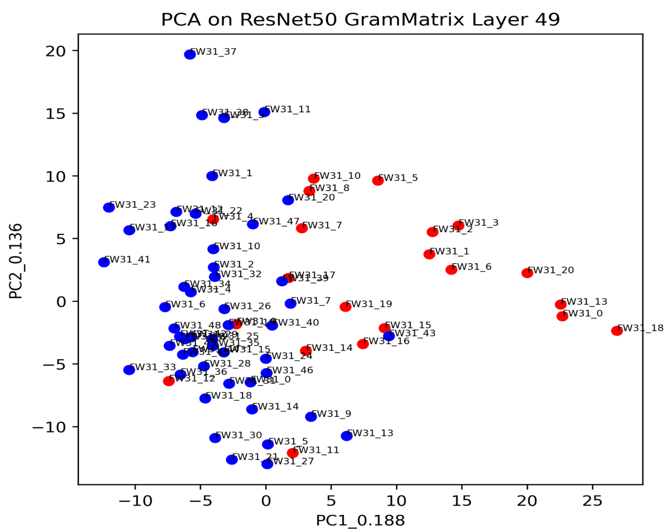


Figure19: 49th layer

For 7mm cut:

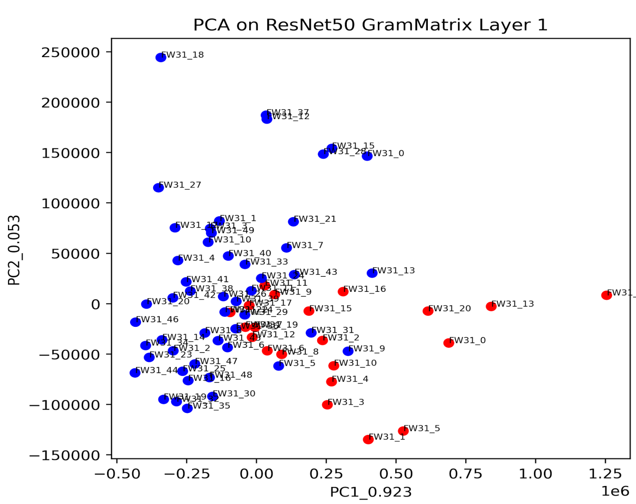


Figure20: 1st layer

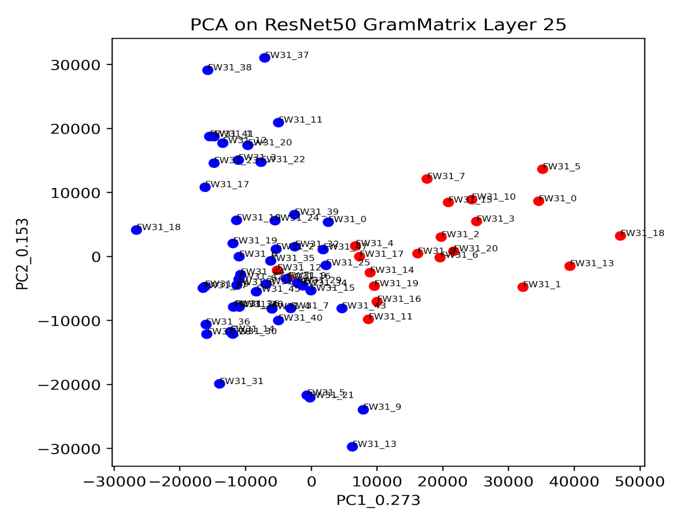


Figure21: 25th layer

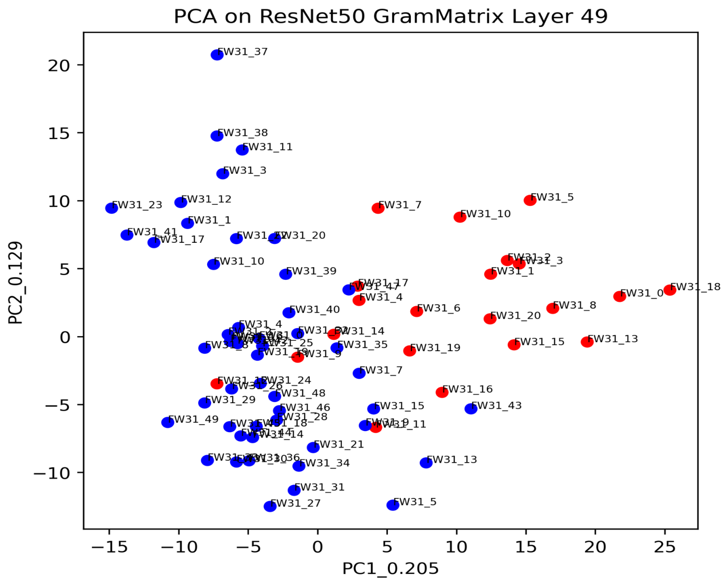


Figure22: 49th layer

Early in the process we observed that other layers like 22nd, 23rd, 24th, 26th, 27th did not give desired results and 25th layers were giving better results than them. So we proceeded to use only 25th layer outputs from there on. From above graphs we observe that the outputs of 1st and 50th layers are as expected of all the various cuts i.e; the damaged part of the image is mixed up with the undamaged part of the image since the initial and final layers represent a given image and not the patterns in it. When we look at outputs of 25th layer, we can observe that for cuts of 1mm and 3mm damaged part of the image is mixed up with the undamaged part of the image since it doesn't find proper patterns and conclude the models are looking at very zoomed in versions of image so it can't differentiate between damaged and undamaged part; which are essentially same i.e; image of wood. If we look at graphs of 5mm and 7mm cuts for the 25th layer, we can see that majority points of the damaged and undamaged images are separated. This goes to prove that the process we used in this project holds up pretty well for separating 2 parts of images if they are in different styles (damaged and undamaged).

We can also observe by comparing the PCA graph of haralick features to the PCA of resnet50 25th layer, we can see that the similar data points on both graphs are positioned in similar positions relative to their graph. For example, by looking at damaged images (red color) points like FW31_18, FW31_13, FW31_0, FW31_20, FW31_20, FW31_1 we can see that they are positioned almost similar in the both the graphs, which means that both haralick features and gram matrix are picking up similar patterns or features from the images, though gram matrix seems to be going into more details in picking up patterns as expected because it works with number of parameters while haralick features has only 13 features to analyze from. Hence many data points of damaged and undamaged images get mixed up in PCA plot of haralick features and resnet50 does better job at separating them, but main takeaway is that both the methods are picking up similar patterns but resnet50 model looks at the image in more detail which results in better classification.

When applied to the PCA model on output of various layers of VGG19 we get the following graphs.

For 1mm cut:

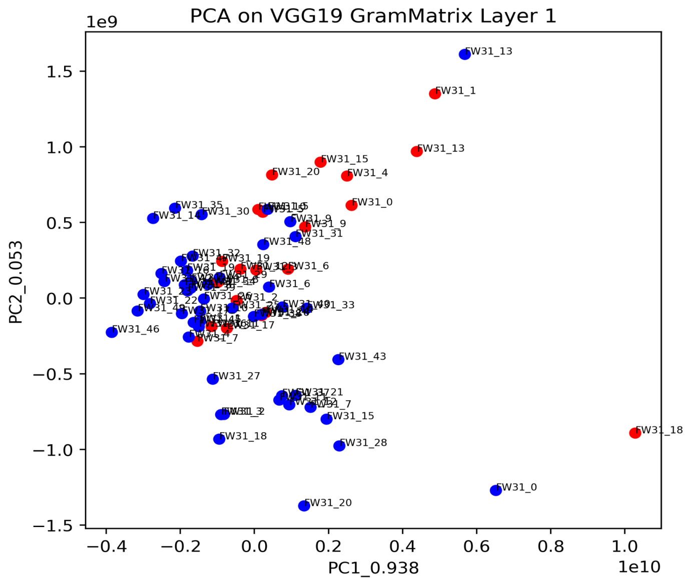


Figure23: 1st layer

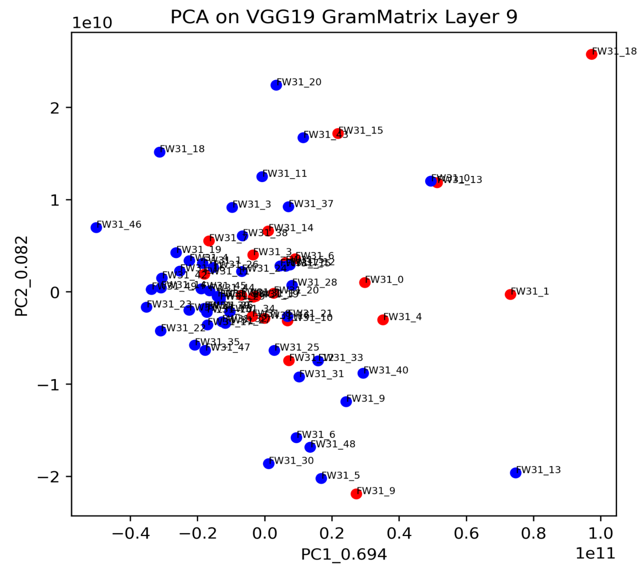


Figure24: 9th layer

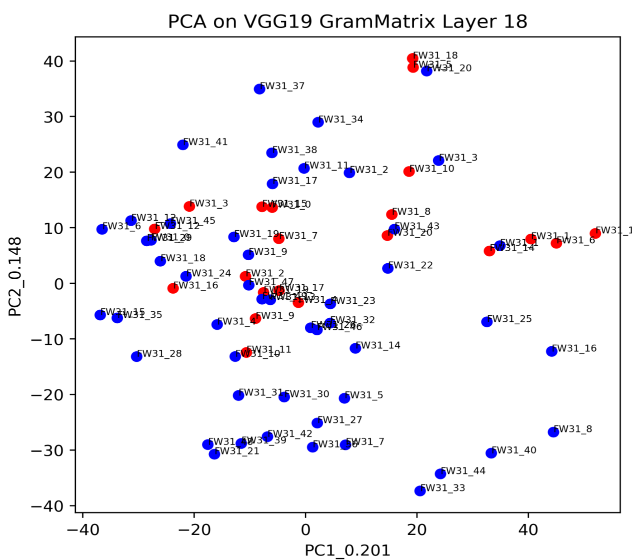


Figure25: 18th layer

For 3mm cut:

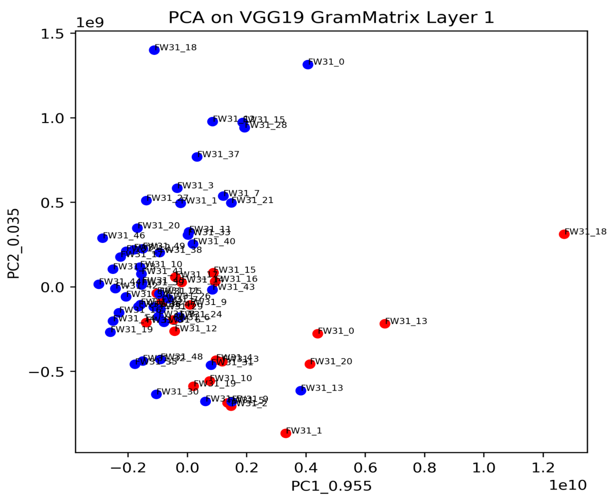


Figure26: 1st layer

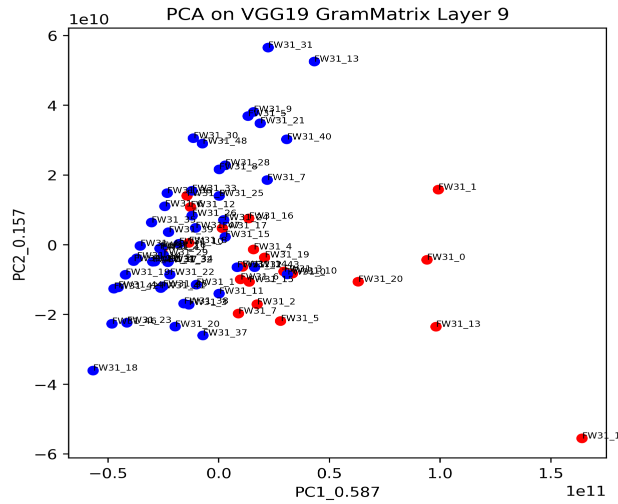


Figure27: 9th layer

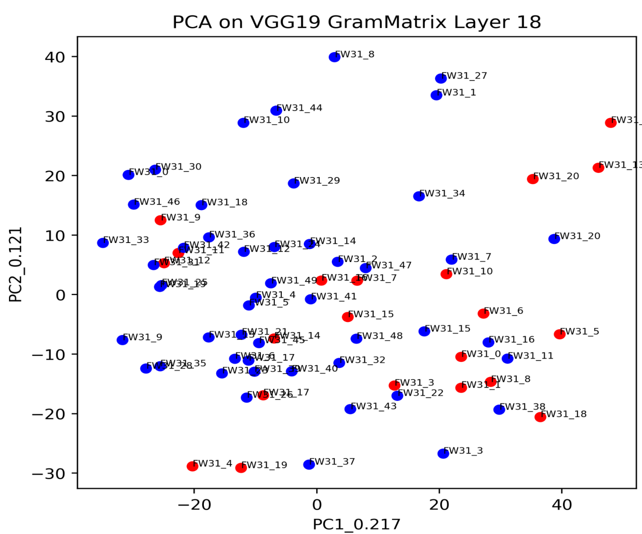
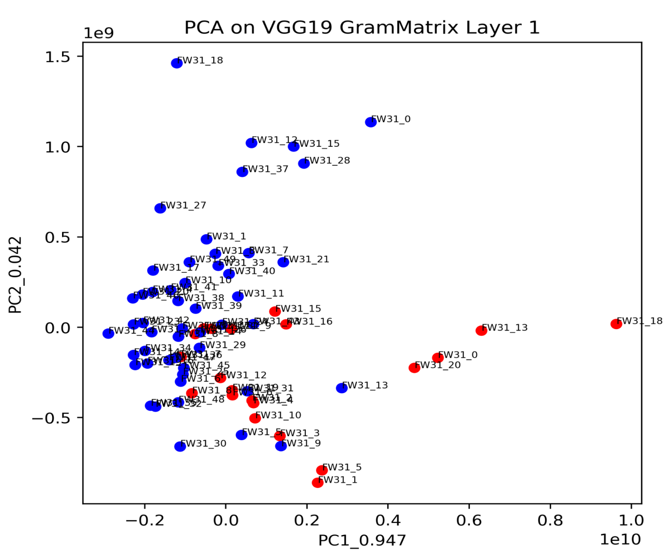


Figure28: 18th layer

For 5mm cut:



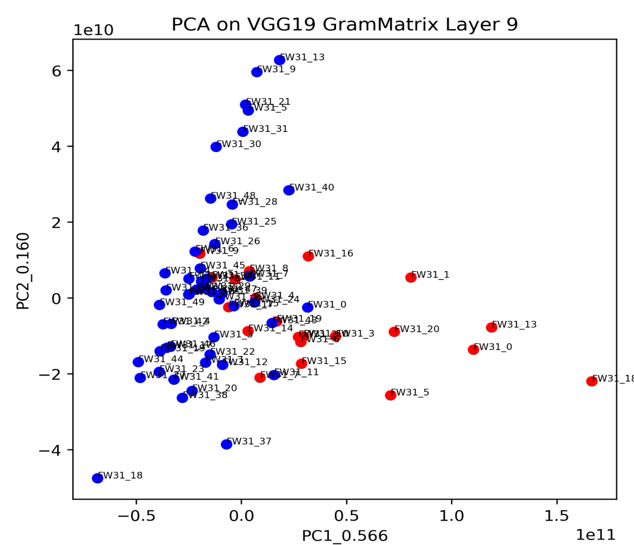


Figure29: 1st layer

Figure30: 9th layer

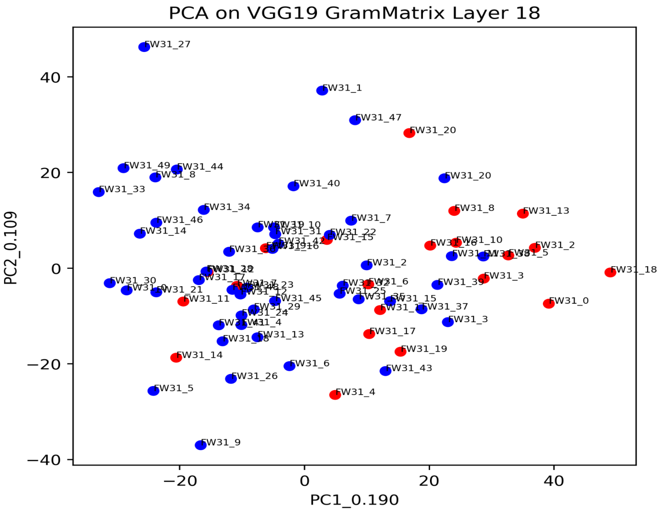
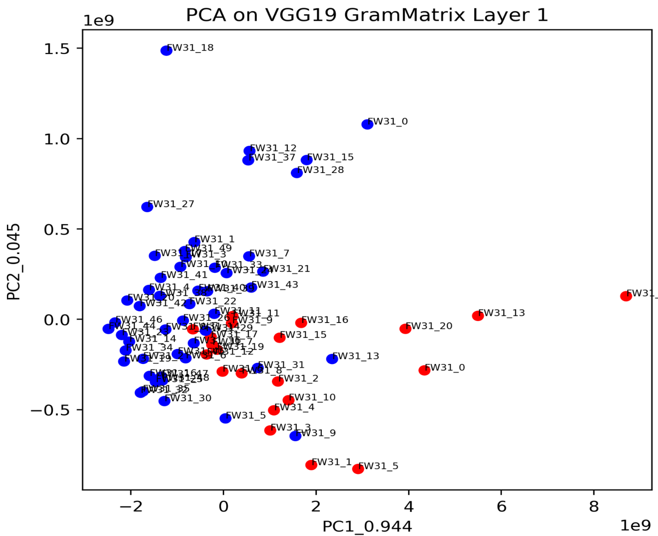


Figure31: 18th layer

For 7mm cut:



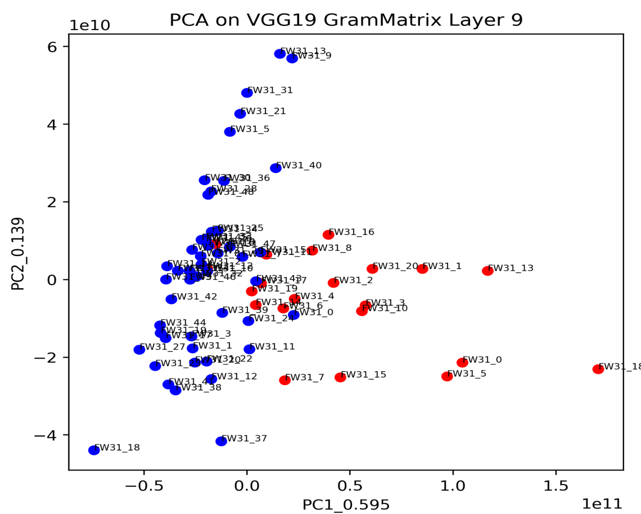


Figure32: 1st layer

Figure33: 9th layer

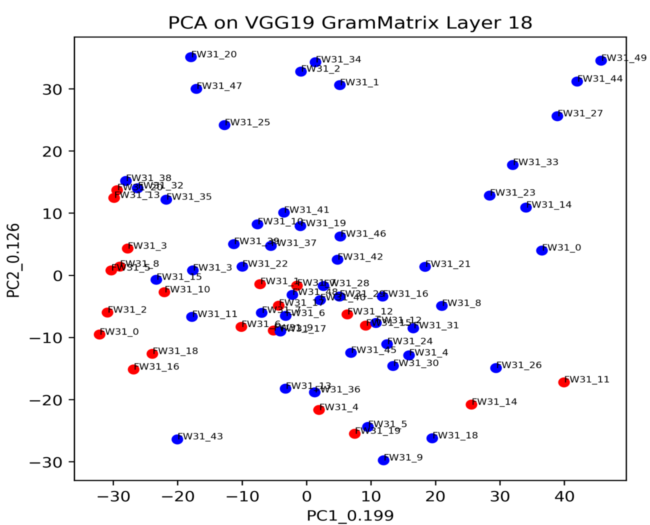


Figure34:18th layer

From above graphs we observe that the outputs of 1st and 18th layers are as expected of all the various cuts i.e; the damaged part of the image is mixed up with the undamaged part of the image since the initial and final layers represent the given image and not the patterns in it. When we look at outputs of 8th layer, we can observe that for cuts of 1mm and 3mm damaged part of the image is mixed up with the undamaged part of the image since it doesn't find proper patterns and conclude the models are looking at very zoomed in versions of image so it can't differentiate between damaged and undamaged part; which are essentially same i.e; image of wood. If we look at graphs of 5mm and 7mm cuts for the 8th layer, we can see that the majority points of the damaged and undamaged images are separated. This goes to prove that the process we used in this project holds up pretty good for separating 2 parts of images if they are in different styled (damaged and undamaged) by using gram matrix on output of middle layer of pre-trained model on image and using clustering process to separate different styled part of image.

But if we compare graphs of Resnet50 and VGG19, we can observe that the Resnet50 does a slightly better job at separating damaged and undamaged images. The reason might be due to Resnet50 having lesser parameters than VGG19 so Resnet50 does a better job at observing patterns rather than looking into details of the image in the middle layers.

We have performed t-test on PC1 and PC2 values to distinguish damaged and undamaged layers for all sizes of images. The results are tabulated below:

t-test (p-values) for 25th layer		
	PC1	PC2
1mm	0.00255	0.19394
3mm	0.0001	0.41366
5mm	4.12E-06	0.93813
7mm	3.35E-08	0.38144

Table1: t-test

We can observe that all the p-values for PC1 column are below significant value 0.05, which means that the means of damaged and undamaged are different along PC1 and p-values for PC2 column are above significant value 0.05, which means that the means of damaged and undamaged are same along PC2. This shows that the separation is done majorly along the PC1 and not along the PC2 for this dataset. We can also observe that as we increase the size of images, the p-values of the PC1 column are decreasing a lot, so we can conclude that as we increase the size of images the model is separating the damaged and undamaged images by greater difference.

In order to generalize the results we check the results by adding other undamaged images. We have previously concluded that 1mm and 3mm images do not yield good results. So, we have only used 5mm and 7mm images for this process. When we apply Resnet50 and VGG19 to all the images we get the following results:

Here each color indicates a different wood fossil image.

For Resnet50 with 5mm cut:

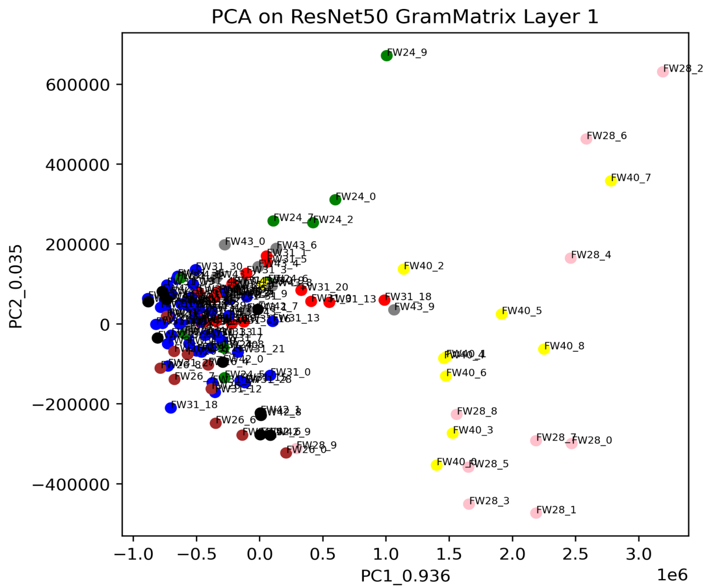


Figure35: 1st layer

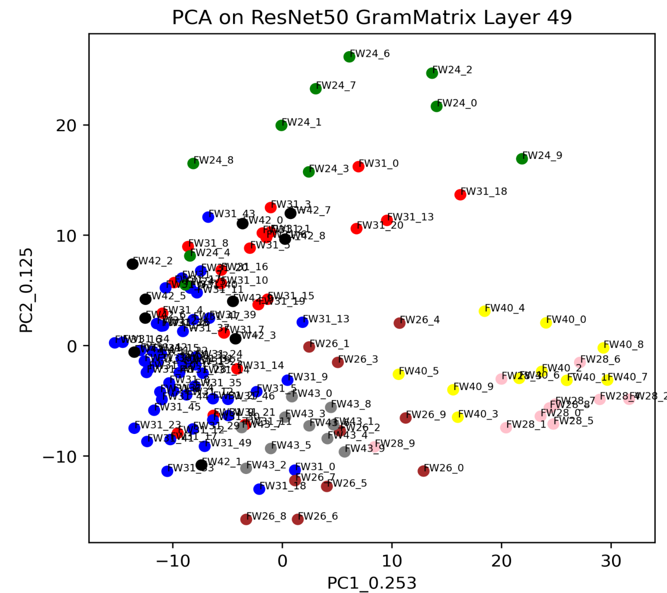


Figure36: 49th layer

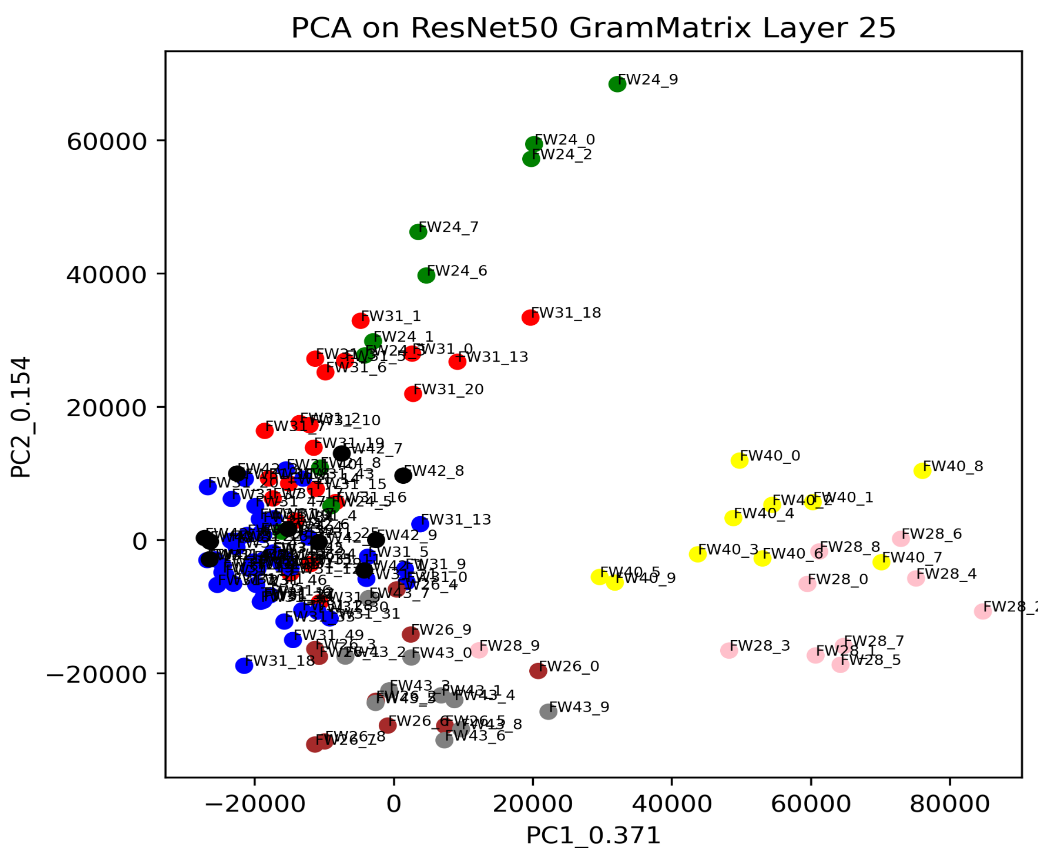
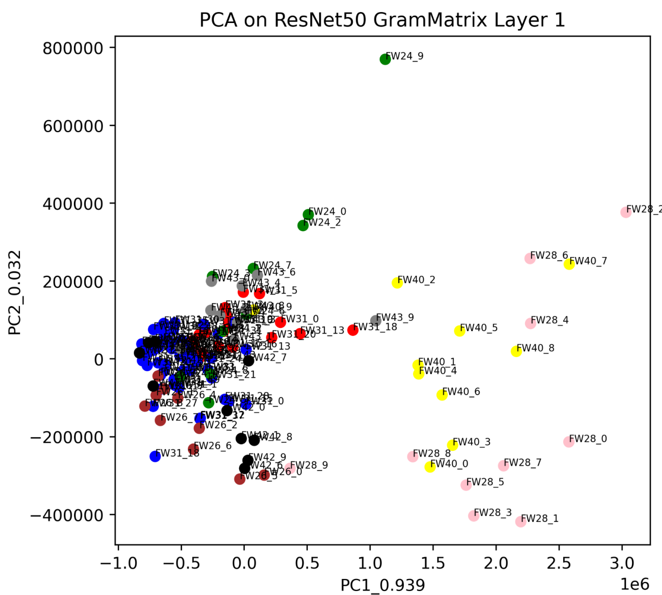


Figure37: 25th layer

For Resnet50 with 7mm cut:



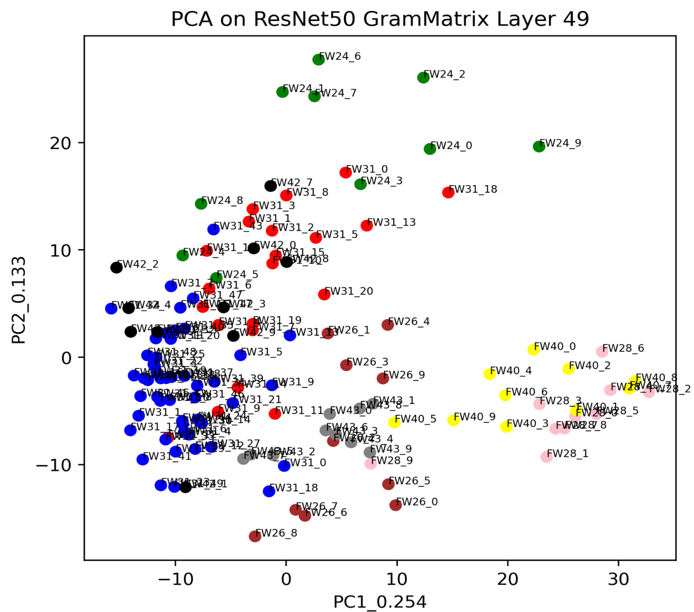


Figure38: 1st layer

Figure39: 49th layer

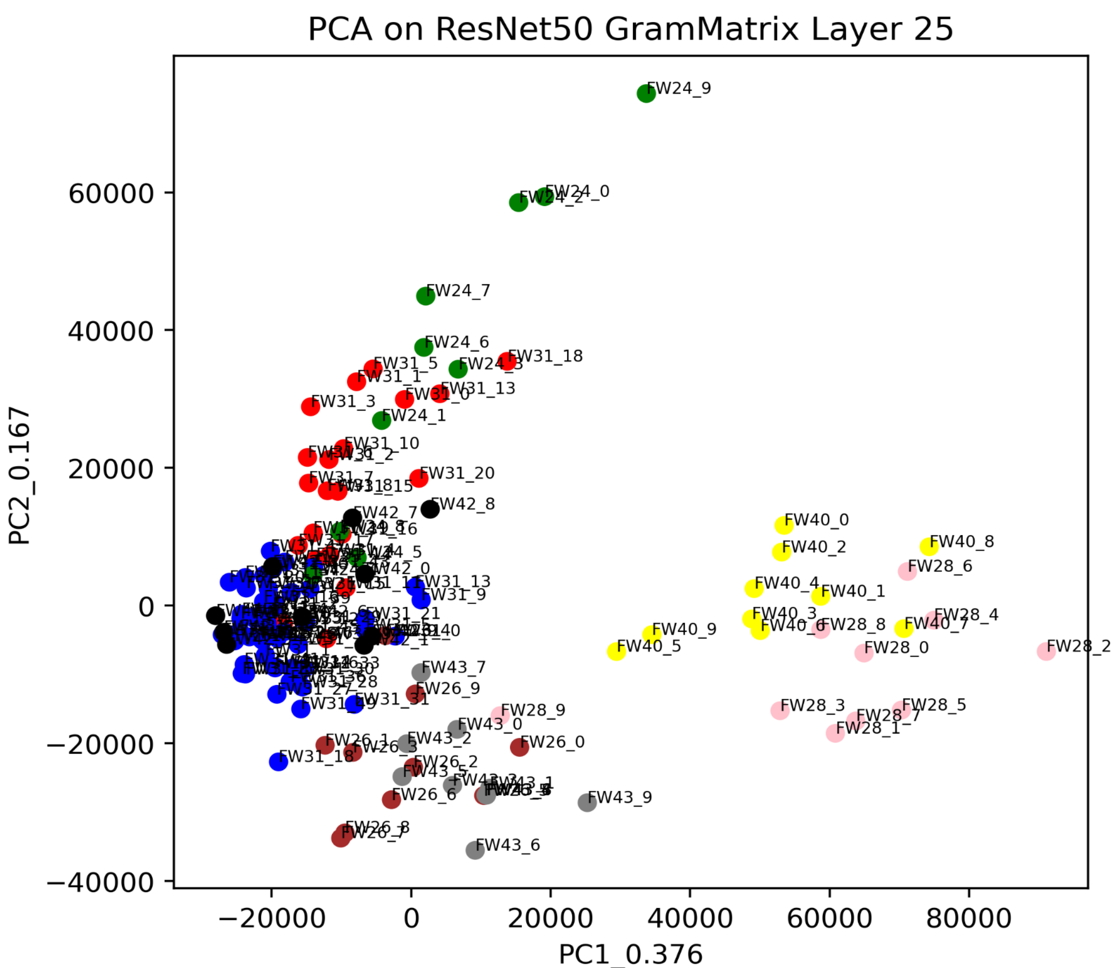


Figure40: 25th layer

For VGG19 with 5mm cut:

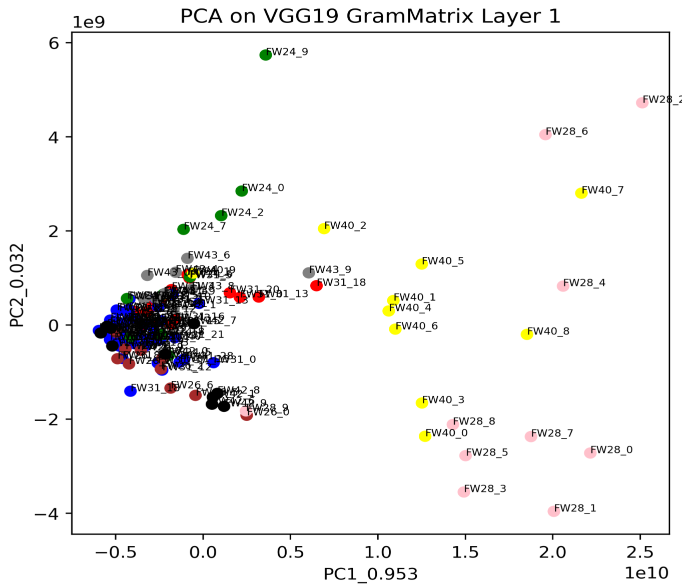


Figure41: 1st layer

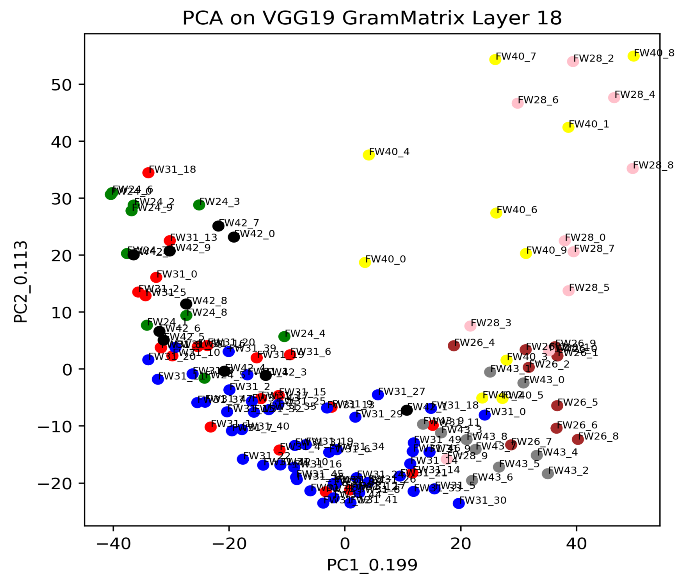


Figure42: 18th layer

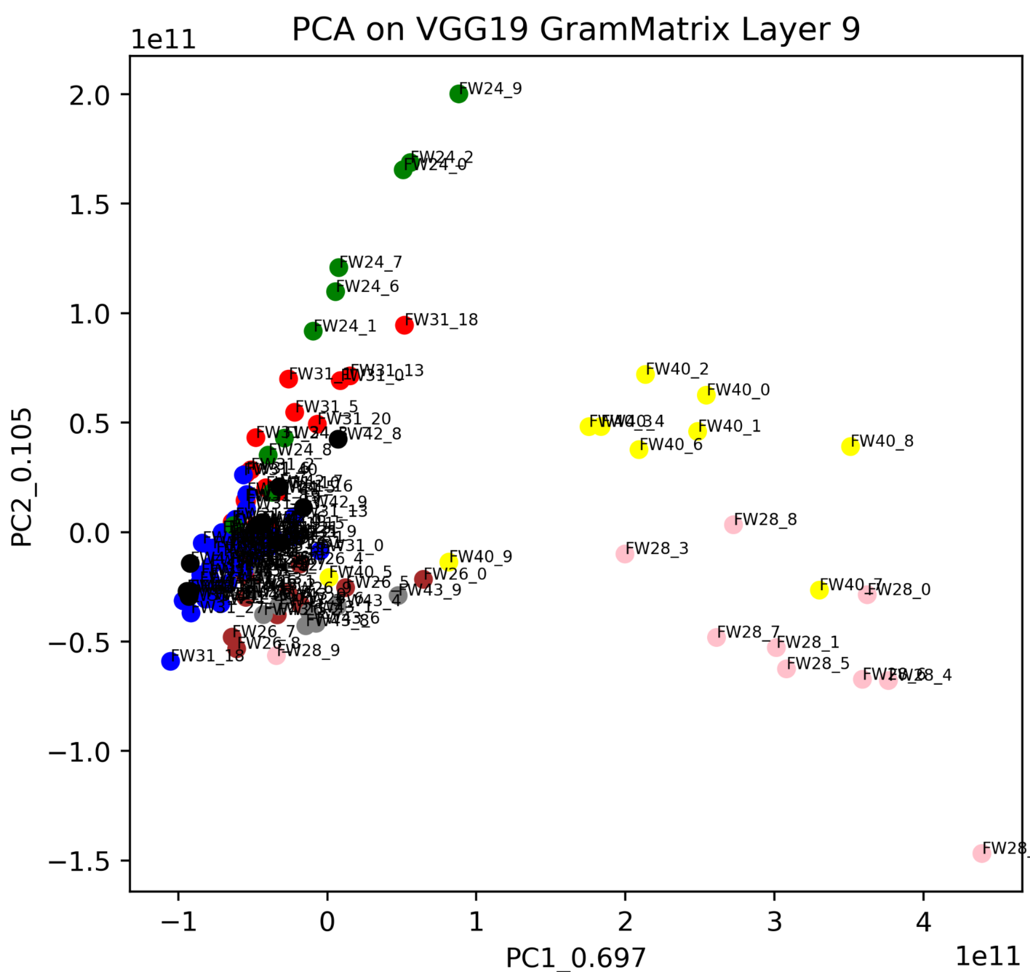


Figure43: 9th layer

For VGG19 with 7mm cut:

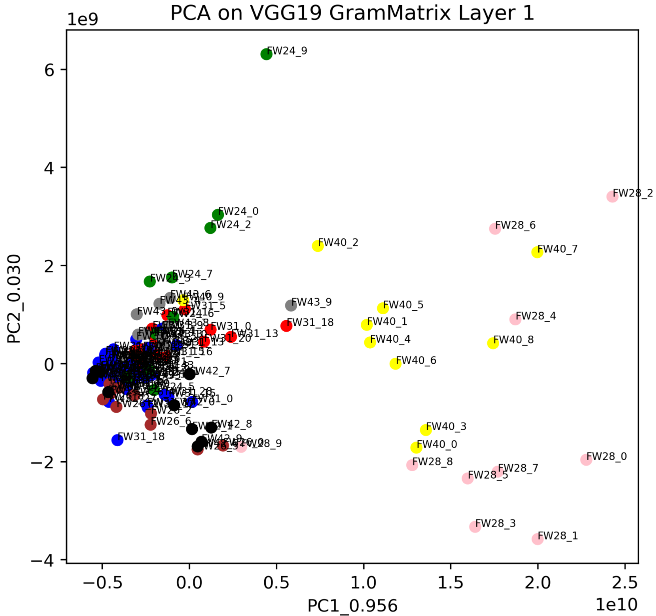


Figure44: 1st layer

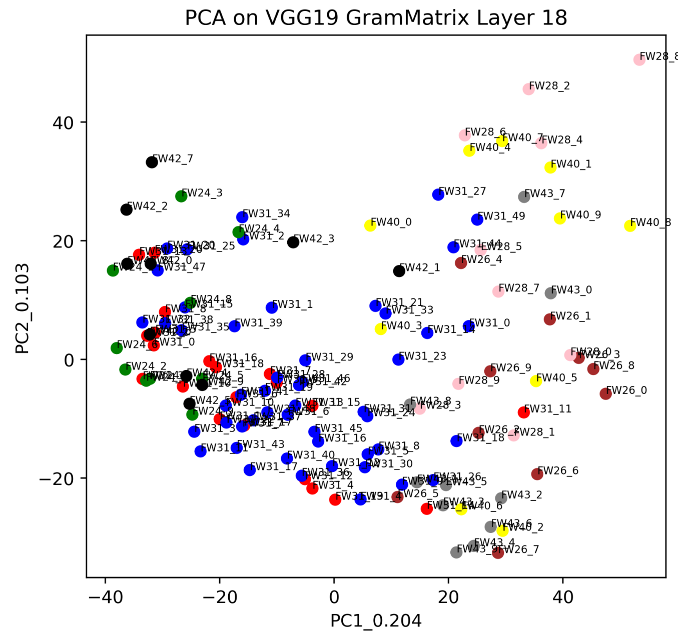


Figure45: 18th layer

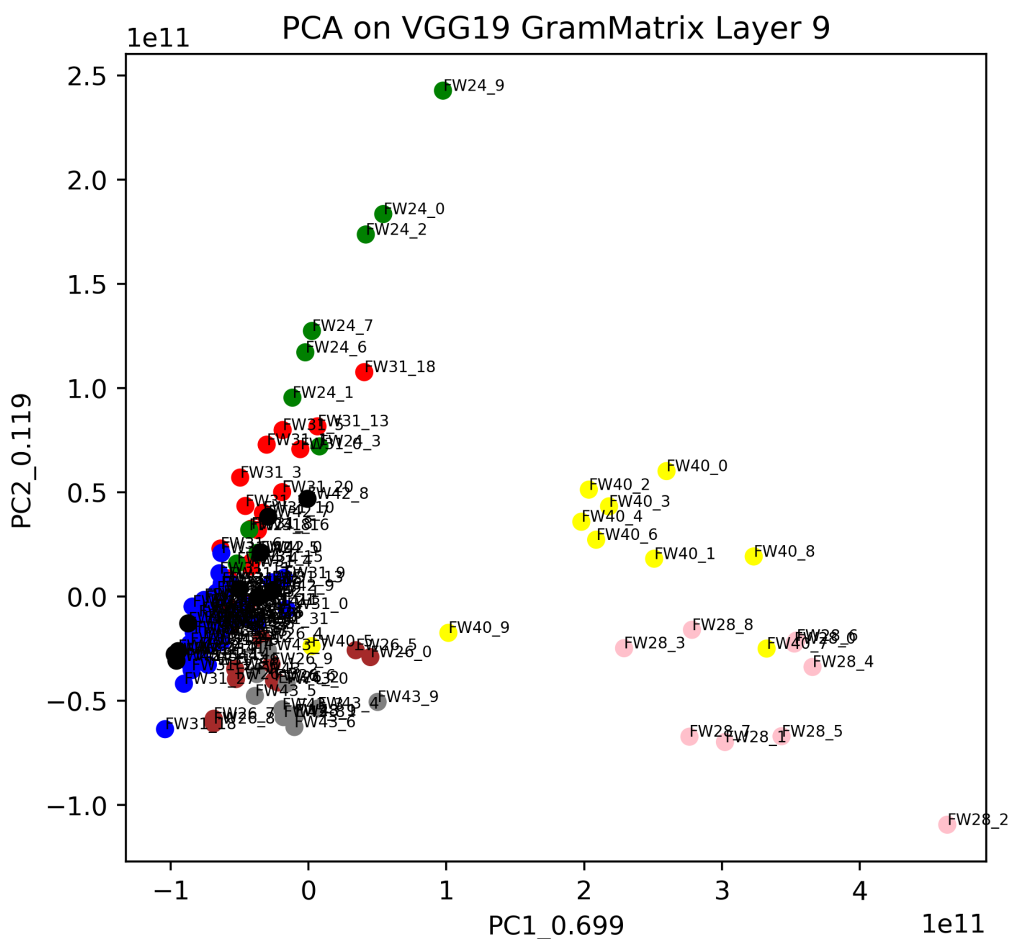


Figure46: 9th layer

We can observe from the outputs of the Resnet50 25th layer that the images representing the same wood fossil are clustered at different locations. This means that the model is able to distinguish the texture of different wood fossils. We can observe that the Resnet50 model outperforms the VGG19 model in clustering the similar images of wood fossils.

Conclusions:

1. Using PCA on pre-trained models is better than using PCA on haralick features to distinguish patterns in the image since pre-trained models have 1000's of features to work on and have more detailed information to analyze while haralick are only 13 features.
2. Middle layers such as 25th layer of Resnet50 and 9th layer of VGG19 are better than using 1st and the last layer for this analysis as middle layers recognise patterns or style of the image while 1st and last layer look at other information of the image.
3. We can observe that using larger images like 5mm, 7mm images give better results than those of 1mm and 3mm images because the larger images have more information in them for the model to recognise the patterns. From the t-test we concluded that by increasing the size of images the separation of means of damaged and undamaged values of PC1 further away.
4. Of the pre-trained models we used for the classification, we can observe from all the results that Resnet50 is better at distinguishing patterns than the VGG19 model.

References:

1. https://www.sciencedirect.com/science/article/pii/S2589004222015875
2. https://arxiv.org/abs/1409.1556
3. https://pubs.rsc.org/en/content/articlehtml/2014/ay/c3ay41907j
4. https://arxiv.org/abs/1512.03385
5. https://arxiv.org/abs/1508.06576
6. https://ieeexplore.ieee.org/abstract/document/8732370?casa_token=PR3bhlYgmmEAAAAA:p0EAcw0MALd_OJ1MbcVUkdV9fNdm6Am0VxPMEZH7CxBiY3p7RPBSvCz03VcG8J2cLvsV_UU
7. https://ieeexplore.ieee.org/document/4309314
8. https://arxiv.org/abs/1912.12510