Melenoma

New Mexico

Supercomputing Challenge

Final Report

April 6, 2025

Team Immortals

New Mexico School For The Arts

Team Members

- Elisea Jackson
- Megan Odom

Teachers

- Sarah Rowe
- Acacia McCombs

Mentor

- Felina Rivera Calzadillas

Table of Contents

Executive Summary

We are working to find a solution to help raise awareness of the damage of UV rays and prevent skin cancer before it is diagnosed. Some of the most critical factors in determining an individual's risk include solar incidence angle (“UV Index | NASA”, n.d.) and Fitzpatrick Skin type (Barrington 2022). We have gathered several data sets of risk factors for analyzing our models. We are currently working on predictive models to help calculate an individual's risk, to encourage protective measures against skin cancer for those who are at risk. We will be comparing results from different types of models, and refining them to improve the effectiveness of the model. Some different types of models we will be comparing the results from are Random Forest, Support Vector Regression, and Neural Netwok models. We hope using multiple models will give us greater insight into the risk factors of skin cancer but all will help us determine the most effective predictive model.

We hypothesized that we are going to see that people who are one on the Fitzpatrick are more susceptible to damage by the UV rays than someone who is three or five. We also expected to see a higher risk of skin cancer for people near the equator because there are more UV rays. While we didn’t prove our initial hypothesis to be untrue we were able to demonstrate the importance of demographic information contributing to risk factors in an imaging based data set through our research, data, and models.

Introduction

Skin cancer is a leading cause of death in the U.S. Roughly 9,000 people in the US die from Melanoma (one of the deadliest forms of skin cancer) each year (Office of the Surgeon General 2014). According to the National Cancer Institute, around 2.1% of people in the U.S. will be diagnosed with Melanoma in their lifetime. Our team knows that more can be done to prevent skin cancer in the U.S..

Skin cancer is only treatable once diagnosed. However, radiation treatment can cause many negative side effects including fatigue and pain. Chemotherapy, another form of cancer

treatment, can cause red and white blood cell count to drop, resulting in an increased risk for anemia and infection. It can also lead to nausea and vomiting and damage to nerves, causing numbness and pain in your hands and feet (“Effects of Skin Cancer Treatment” 2015).

About 90% of skin cancer is linked to sun exposure (“Skin Cancer” n.d.). Another well-established risk factor for melanoma include a high number of nevi and the presence of atypical nevi ("Moles and Melanoma Risk." n.d.). Tanning beds that use artificial light can also cause skin damage and are associated with a 20% increased risk of melanoma ("Sunbeds and Skin Cancer Risk." n.d.). Another significant risk factor is the amount of times you have been sunburned. However, there are a variety of other factors that also contribute to an individual’s increased risk such as genetics, diet, and age (American Cancer Society). Through sun protection and lifestyle improvements, individuals can reduce their chances of melanoma (Health and Human Services).

Physics of UV Radiation

UV-a, UV-b and UV-c

Ultraviolet light lies between visible light and X-rays on the electromagnetic spectrum, with wavelengths spanning from 10 to 400 nanometers (nm). This range is segmented into three types of UV rays: UV-a, UV-b, and UV-c. The majority of Ultraviolet light (wavelengths of 200 nm and lower) is filtered out by our earth's Atmosphere, completely blocking UV-c, which has the shortest wavelength and highest energy, ranging from 100-280 nm. On the other hand, UV-a wavelengths range from 320-400 nm and UV-b range from 280-320 nm. (“Ultraviolet (UV) Radiation” 2017) Though both UV-a and UV-b rays reach past our ozone, UV-b’s shorter wavelengths are easily obstructed by clouds and windows, resulting in only 5% of our UV radiation being attributed to UV-b rays. Though they make up such a small portion of UV radiation, UV-b rays cause a majority of the effects one may normally associate with UV radiation including sun burns and blisters (Sullivan 2019). Because of this, the sun protectant factor (SPF) for sunscreen is based on the lotion's effectiveness on UV-b rays. This is because, unlike UV-a rays which only damage DNA indirectly, UV-b rays directly affect DNA (Chien, n.d.). UV exposure can cause skin damage causing aging and dark spots and, over time, it can

lead to skin cancer. It is considered to contribute the most out of all other risk factors in determining whether or not an individual will develop skin cancer.

Cyclobutane-pyrimidine dimers and 6-4 photoproducts

With DNA as the primary chromophore of cells absorbing sunlight energy, the short wavelength of UV-b rays proves to be detrimental as the aromatic (nitrogenous) ring structure of DNA readily absorbs the radiation ("Focus on UV-Induced DNA Damage and Repair—Disease Relevance and Protective Strategies", n.d.). This causes a reaction in a cell's DNA between thymine molecules, resulting in mutagenic and cytotoxic DNA lesions such as cyclobutane-pyrimidine dimers (CPDs) and 6-4 photoproducts (6-4PPs). These are referred to as thymine dimers, which are DNA lesions formed when adjacent thymine bases become covalently linked and cause a kink in the DNA, making it hard for the cell to replicate and transcribe, often resulting in incorrect repair or a "missed" dimer. Despite the fact 6-4PPs are much more detrimental to DNA structure itself, CPDs have been observed to inflect over 75% of cellular damage, leaving 25% of cellular damage attributed to 6-4PPs (You et al. 2001, and Kciuk et al. 2020). UV light may also cause indirect damage from the generation of free radicals from photodynamic reactions, including one of the most reactive oxygen species, hydroxyl radical (Kciuk et al. 2020 and Rastogi et al. 2010). If the damage disrupts cellular processes, the cell will carry out one of two functions depending on the severity of the damage. If the cell is deemed viable, gene products such as TP53 will send repair machinery, usually including photolyase (a DNA repair enzyme which utilizes blue light to repair UV induced damage) (Garinis et al. 2005). If the cell does not resume transcription after 72 hours, the cell will undergo programmed apoptosis (Andrade-Lima et al. 2015). Though cells are usually able to successfully repair, UV damage is cumulative and recurrent damage becomes harder and harder to repair. The probability that a cell will recover, die, or become cancerous/precancerous is dependent on the radiation dose, patient history, and cell type.

Cancerous cells

If the DNA in a cell is not correctly repaired, or if a cell evades apoptosis, the mutated genome becomes unstable and causes further damage. The transcription and replication of cells

containing cancer-causing CPDs results in a prolific group of mutated cells (Zheng 2020). These mutations may accumulate around critical genes that regulate the cell cycle, DNA repair, and apoptosis such as proto-oncogenes, DNA repair genes, and tumor suppressor genes. Proto-oncogenes genes aid in cell growth and division. When one of these genes mutates, it becomes activated, now being called an oncogene, and will begin to duplicate out of control. Tumor suppressor genes, such as TP53, CDKN2A, and BAP1, regulate cell division. When the DNA of a tumor suppressor is compromised, it can lead to unchecked cellular growth. Though mutations in tumor suppressor genes are commonly inherited, most are acquired during a patient's lifetime. Finally, DNA repair genes, such as BRCA1 and BRCA2, encode proteins that detect and fix DNA damage. When DNA repair genes become mutated, they can no longer perform the job of fixing genetic mistakes, preventing unwanted mutation, and prompting apoptosis in non-viable cells ("Oncogenes, Tumor Suppressor Genes, and DNA Repair Genes" 2022). When this accumulation happens, the affected cells may become a clonal population, growing out of control. Then, depending on the location and formation, the clonal population may develop into different types of skin cancer including basal cell carcinoma, squamous cell carcinoma, or the most dangerous, melanoma. The population will then undergo angiogenesis (the formation of new blood cells) which is crucial for the growth's access to nutrients and water, and for further growth and metastasization. Over time, the cancerous cells may invade nearby tissues, further promoting unchecked growth.

Risk Factors and Machine Learning Models

Random Decision Tree Forest Machine Learning

Random Forests Models (RFMs) are more resistant to overfitting, a phenomenon when machine learning models are so well trained on a sample data set that it fails to make correct predictions with new data. Similarly, they perform better on unseen data when compared to individual Decision Trees, as they are less likely to be overly specialized to the training data. Essentially, RFMs build a more reliable model using smaller and easier-to-understand models. These Random Forest Models consist of a collection of Decision Trees that are randomized in two ways. One of which is called "bootstrapping" where each Decision Tree uses a different

random sample of the data while being trained. Additionally, each Decision Tree is trained on select variables or features (a few dimensions of each point). When a new data point is introduced, it is classified using every Decision Tree in the Forest. Then the final classification is determined by selecting the majority vote from each of the Decision Trees. (Zivkovic 2022) If you are using a regression model the final result will be an average result of all the decision trees. (“What Is Gradient Boosting?” 2023)

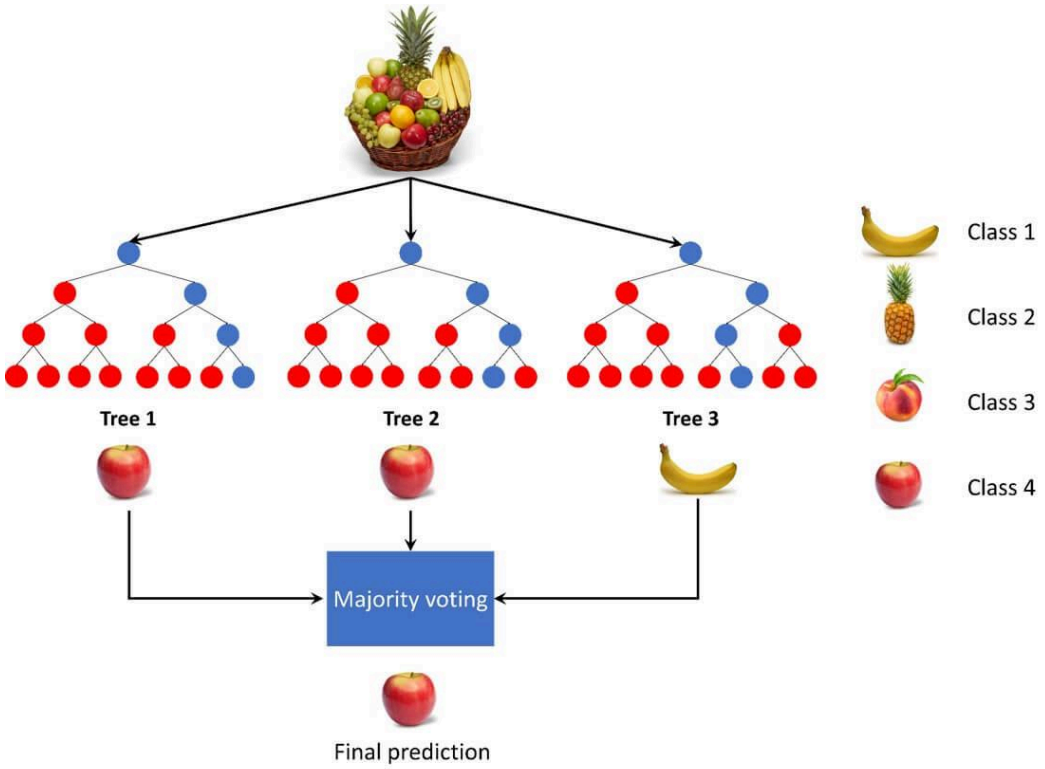


Fig 1. A diagram demonstrating the voting process of Decision Trees in the Random Decision Tree Forest Machine Learning Model. (Zivkovic 2022)

Support Vector Machine Learning

A Support Vector Machine (SVM) determines the best hyperplane that divides the data into classes. If we can't easily divide the data, we will have to use kernelling. This means that it will set the data into higher dimension planes until it finds the best one that allows it to separate

it using a hyperplane, which is one less dimension than the transformed data. This means that if we couldn't separate data that was on a line with a point, we would transform it into a two dimensional graph and try separating the data with a line. If we couldn't neatly separate the data in a two dimensional space with a line we could transform it with an equation into a three dimensional space and separate it with a plane. The transformation of this data can combine different equations to come up with more complex shapes in the graphs supporting more noisy data.

The model works to make sure all points are as far away from the hyperplane as possible; the further the values are the more confidence there is in a solution. The support vectors are the data points closest to the hyperplane. If the support vectors were removed it would change the position of the hyperplane, therefore they are considered critical.

SVM is a good model if you're looking for something that tends to be accurate. It works well with small clean data sets, and can be efficient because it uses a smaller training set of data. However SVM models can take a long time to train with large data sets, and are less accurate with noisy data sets (Quantexa, n.d.).

Hyperparameter Tuning and Parameter Searching

A Hyperparameter is a parameter that governs the learning process of a machine learning model. They are set before learning the model digests data. Hyperparameters include things like the kernel size (the relationship between data points and their separation), C (ratio of resulting errors and acceptable margin of error), and gamma (the amount of influence a support vector has on the hyperplane) in a SVM model. Also known as Hyper Optimization, this process helps improve the accuracy, performance, and efficiency. (“What Is Hyperparameter Tuning?” 2024)

Evaluating Accuracy

Regression models and the classification models have different ways for evaluating accuracy. Our regression model used an out of bag score (OOB), mean squared error(MSE), and r2. The OOB score for the model uses data points that weren’t used for certain random trees and runs it through them evaluating the accuracy as the OOB score. The lower OOB score indicates more accuracy. The r2 score is the proportion of how far the actual values are from the average

predicted value for the independent variables. For R2 we are looking for values closer to 1 as this would mean 100% correct. The MSE is the average of the difference between the predicted values and the actual values squared. We are looking for a lower MSE as a 0 would mean the model is perfect. (Rowe, 2018)

Meanwhile the SVM model and Random Forest Classifier model use Accuracy, Precision, recall, F1 Score, and ROC AUC scores (Shalev 2019). Accuracy calculates the number of accurately predicted values divided by the total number of predicted values. The F1 score combines the recall and Recall metric to help tak into count boteh false positives and negatives to help with uneven class distributions. Precision each value was correctly evaluated for your category, divided by the number of values it predicted to be that class. Recall takes the number of values that were accurately divided by the number of values that should have been in that category.

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision} = 2 * \frac{0.333 * 0.125}{0.333 + 0.125} \approx 0.182$$

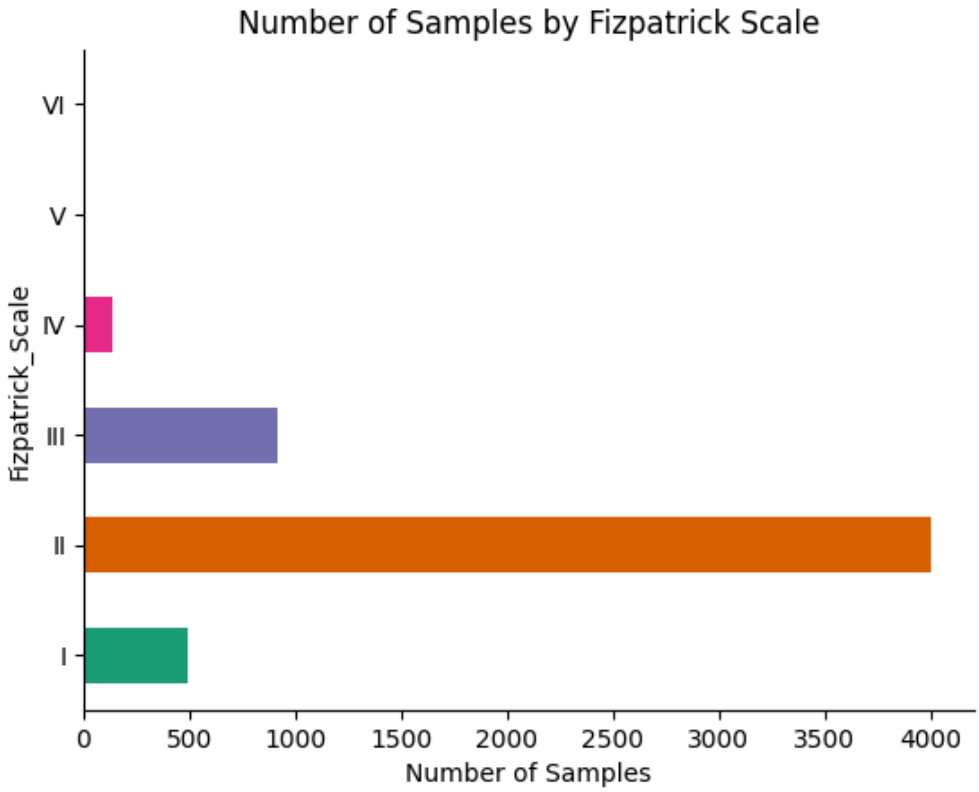
The ROC curve is determined by how well your model can determine different thresholds or ways to divide your data. It is created by plotting the precision rate for all of the different categories with different threshold settings. By plotting precision data called True Positive Rate against the False Positive rate at differing thresholds for all the categories you get the ROC curve. By taking the area under the threshold line you get your ROC AUC. For this score we want a value closer to 1. (Shalev, 2019)

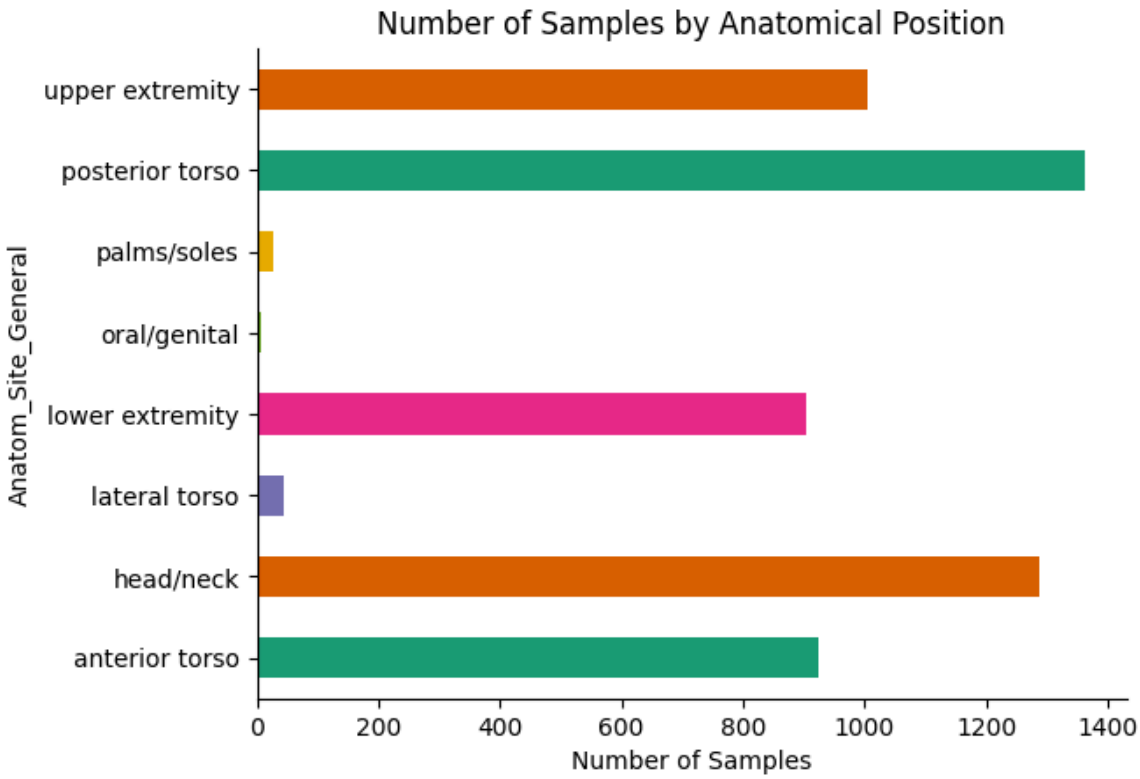
Our Data Set

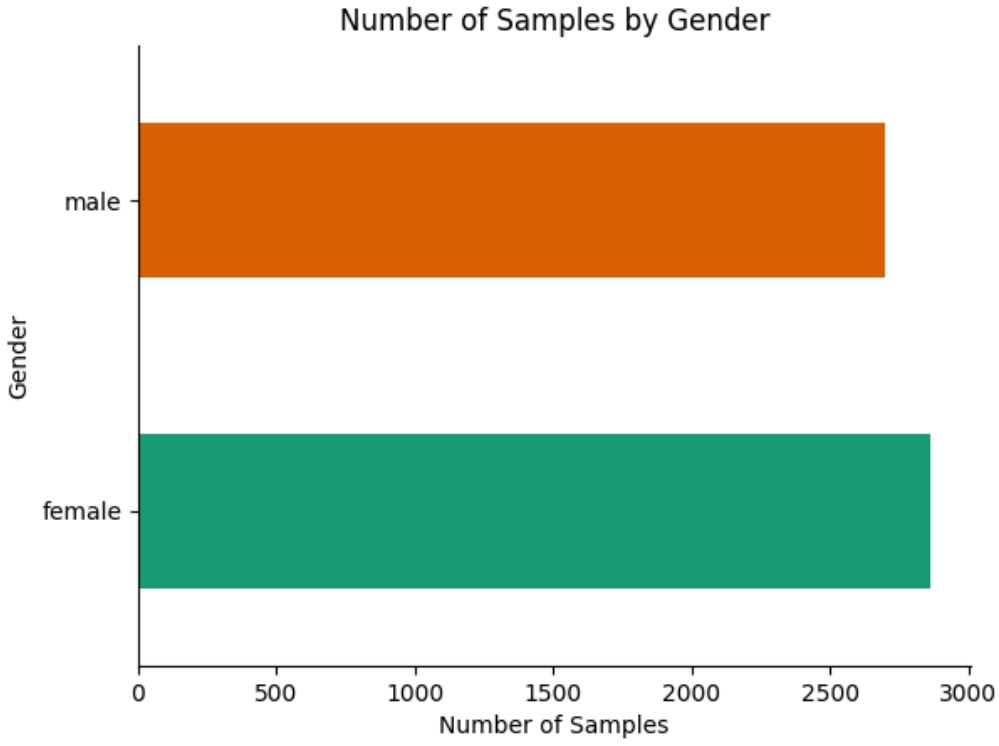
The data set we acquired was from the International Skin Imaging Collaboration. We filtered through about 500,000 publicly accessible data points, downloading all the data that included fitzpatrick skin type and approximate age, narrowing our data to around 5635 values. The data set included images of the lesions, diagnostic attributes, clinical attributes, technological attributes, and the different licensing for each image. We only analyzed the

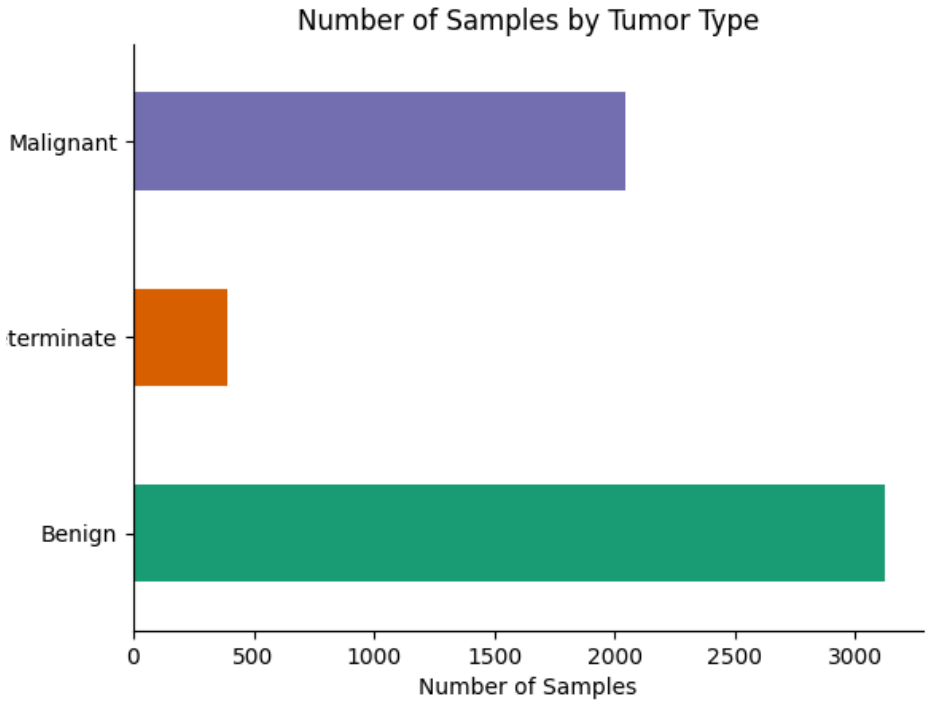
influence of four of the clinical attributes - age, sex, fitzpatrick skin type, and general anatomic site - in relation to one of the diagnostic attributes - lesion diagnosis.

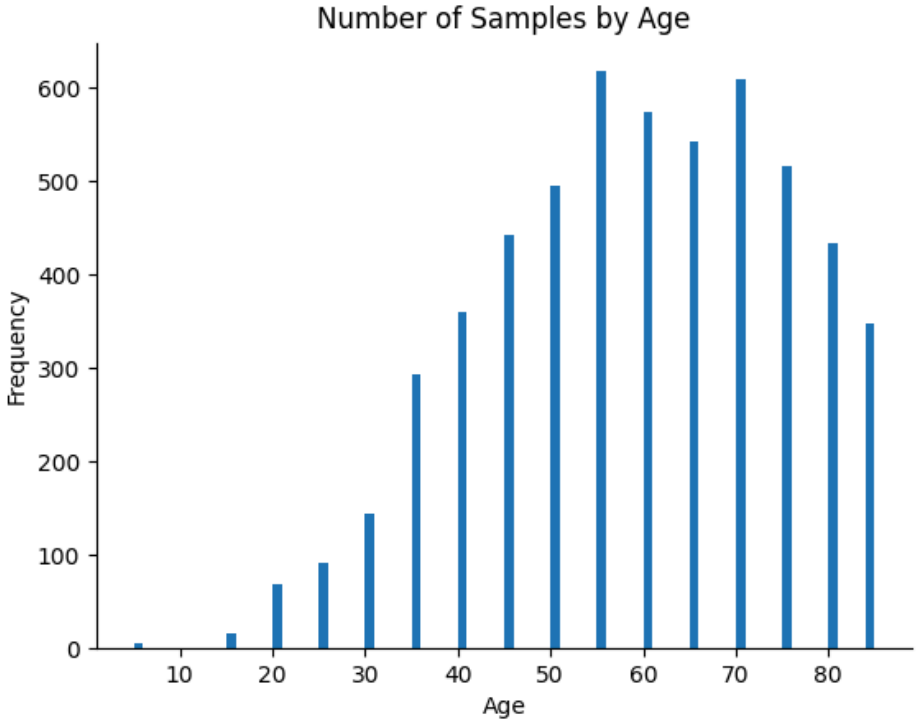
The data set contained more points for certain groups of people than others. Below are some graphs demonstrating the difference in the sample distribution for all our features.




Number of Samples by Anatomical Position
upper extremity
posterior torso
palms/soles
oral/genital
lower extremity
lateral torso
head/neck
anterior torso
Anatom_Site_General
0
200
400
600
800
1000
1200
1400
Number of Samples


Number of Samples by Gender
male
female
Gender
0
500
1000
1500
2000
2500
3000
Number of Samples


Number of Samples by Tumor Type
Malignant
terminate
Benign
0
500
1000
1500
2000
2500
3000
Number of Samples


Number of Samples by Age
Frequency
600
500
400
300
200
100
0
10
20
30
40
50
60
70
80
Age

Code

In our code we select the data we are using from the larger data set and encode the different features as numerical results.

```
# defining data - posibly melanin lever and canser rates
X = melanindata.iloc[0:5634, 12].values
y = melanindata.iloc[0:5634, 19].values
a = melanindata.iloc[0:5634, 4].values
g = melanindata.iloc[0:5634, 30].values
b = melanindata.iloc[0:5634, 5].values

# Sample data (replace this with your actual data)
melanindata = pd.DataFrame({
   'Tumor_Type': X,  # Example values  # used to be [X]
   'Fizpatrick_Scale': y,  # Example values
   'Age': a,
   'Gender': g,
   'Anatom_Site_General': b,
   # Add other features if needed
})

melanindata = melanindata.dropna()

# Step 1: Label encoding for Tumor_Type (x-values)
label_encoder = LabelEncoder()
melanindata['Tumor_Type_Encoded'] =
label_encoder.fit_transform(melanindata['Tumor_Type'])


# Step 2: Map Roman numerals to integers for Fizpatrick_Scale (y-values)
fizpatrick_map = {
   'I': 1, 'II': 2, 'III': 3, 'IV': 4, 'V': 5, 'VI': 6, 'VII': 7
}

Anatom_Site_General_map = {
```

```
    'oral/genital': 0, 'head/neck': 1, 'upper extremity': 2,
'palms/soles': 3, 'lower extremity': 4, 'anterior torso': 5, 'lateral
torso': 6, 'posterior torso': 7,
}
melanindata['Fizpatrick_Scale_Encoded'] =
melanindata['Fizpatrick_Scale'].map(fizpatrick_map)
melanindata['Gender_Encoded'] =
label_encoder.fit_transform(melanindata['Gender'])
melanindata['Anatom_Site_General_Encoded'] =
melanindata['Anatom_Site_General'].map(Anatom_Site_General_map)
```

Before using the different models we split the data into testing and training data with 20% of the data being used for testing and 80% being used for training.

```
X_train, X_test, y_train, y_test = train_test_split(x_values, y_values,
test_size=0.2, random_state=42)
```

We ran a Random Forest Regression using basic hyperparameters, as well as optimized hyperparameters using grid search.

Basic regressor set up:

```
basic_regressor = RandomForestRegressor(n_estimators=500, random_state=0,
oob_score=True)
basic_regressor.fit(x_values, y_values)
```

Regressor grid search set up:

```
param_grid = {
   'n_estimators': [100, 300, 500],
   'max_depth': [None, 10, 20, 30],
   'min_samples_split': [2, 5, 10],
   'min_samples_leaf': [1, 2, 4],
   'max_features': ['auto', 'sqrt']
}

grid_search = GridSearchCV(
   estimator=RandomForestRegressor(random_state=0, oob_score=True),
```

```
    param_grid=param_grid,
    cv=5,
    scoring='neg_mean_squared_error',
    n_jobs=-1,  # Use all available cores
    verbose=1
)

# Fit the grid search to the data
grid_search.fit(X_train, y_train)
```

Regressor random parameter search set up:

```
param_distributions = {
    'n_estimators': randint(10, 200),
    'max_depth': [None] + list(randint(5, 25).rvs(2)),
    'min_samples_split': randint(2, 15),
    'min_samples_leaf': randint(1, 10),
    'max_features': ['auto', 'sqrt', 'log2', None],
    'bootstrap': [True, False]
}

random_search = RandomizedSearchCV(
    estimator=RandomForestRegressor(random_state=0, oob_score=True),
    param_distributions=param_distributions,
    n_iter=50,  # Number of parameter settings sampled
    cv=5,
    scoring='neg_mean_squared_error',
    n_jobs=-1,  # Use all available cores
    verbose=1,
    random_state=42
)

random_search.fit(X_train, y_train)
```

Instead of rerunning the parameter search models, their values were imputed as a variable in our code to allow for a faster overall running time. We don't have to worry about the results changing in the random search section unless the data itself changes as it has a set random state.

We then evaluated the accuracy using the results from the OOB, MSE, and R2 scores.

Code for the grid search optimized regressors accuracy scores:

```
ob_score2 = regressor2.oob_score_
print(f'Out-of-Bag Score: {ob_score2}')

predictions2 = regressor2.predict(X_test)

mse2 = mean_squared_error(y_test, predictions2)
print(f'Mean Squared Error: {mse2}')

r22 = r2_score(y_test, predictions2)
print(f'R-squared: {r22}')
```

For the Random Forest Classifier and the SVM models we used the same data, encoding, and testing/training split. We set up and ran a basic Random Forest Classifier, a grid search optimized Random Forest Classifier, a basic SVM, and a optimized hyperparameter SVM using a grid search.

Random Forest Classifier set up:

```
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
```

Hyperparameter optimized Random Forest Classifier:

```
param_grid = {
    'n_estimators': [200, 500 ],
    'max_depth': [None, 20, 40],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2],
    'max_features': ['sqrt', 'log2', None],
    'bootstrap': [True, False],
    'criterion': ['gini', 'entropy']
}
```

```
# Create the grid search with cross-validation
grid_search = GridSearchCV(
   estimator=RandomForestClassifier(random_state=0, oob_score=True),
   param_grid=param_grid,
   cv=5,
   scoring='accuracy',
   n_jobs=-1,  # Use all available cores
   verbose=1,
)
try:
 # Fit the grid search to the data
 grid_search.fit(X_train, y_train)

 # Get the best parameters and best estimator
 best_params = grid_search.best_params_
 best_classifiers = grid_search.best_estimator_
```

SVM set up:

```
svm_pipeline = Pipeline([
   ('scaler', StandardScaler()),  # SVMs require scaled features
   ('svm', SVC(kernel='rbf', C=1.0, probability=True, random_state=42))
])
```

Hyperparameter optimized SVM:

```
# Define the parameter grid
param_grid = {
   'svm__C': [0.1, 1, 10, 100],
   'svm__gamma': ['scale', 'auto', 0.1, 0.01],
   'svm__kernel': ['rbf', 'linear', 'poly', 'sigmoid']
}

# Set up GridSearchCV
grid_search = GridSearchCV(
   svm_pipeline,
   param_grid,
   cv=5,
   scoring='accuracy',
   verbose=1,
```

```
    n_jobs=-1
)
```

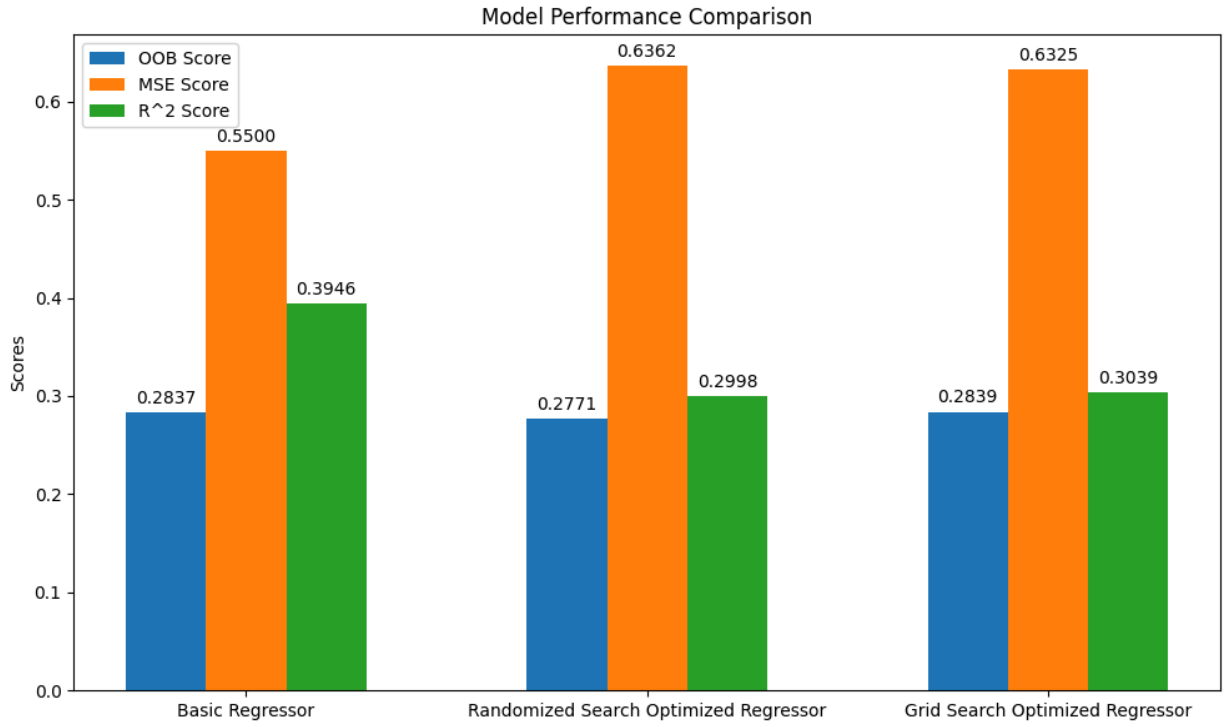
After running the models we were able to calculate the probability and compare different accuracy ratings.

```
# Calculate evaluation metrics
models = ['Basic RF', 'Optimized RF', 'SVM']
predictions = [rf_predictions, rf_grid_predictions, svm_predictions]
probabilities = [rf_basic_prob, rf_opt_prob, svm_prob]

# Accuracy metrics
metrics = {
    'Accuracy': [],
    'Precision': [],
    'Recall': [],
    'F1 Score': [],
    'ROC AUC': []
}

for i, model in enumerate(models):
    metrics['Accuracy'].append(accuracy_score(y_test, predictions[i]))
    metrics['Precision'].append(precision_score(y_test, predictions[i],
average='weighted', zero_division=0))
    metrics['Recall'].append(recall_score(y_test, predictions[i],
average='weighted', zero_division=0))
    metrics['F1 Score'].append(f1_score(y_test, predictions[i],
average='weighted', zero_division=0))
    metrics['ROC AUC'].append(roc_auc_score(y_test, probabilities[i],
multi_class='ovr'))
```

Results

Using the result from the Random forest regressor didn't lead to the best results. We tried 2 methods of optimizing the Random Forest regressor. The first was the Randomized Search and

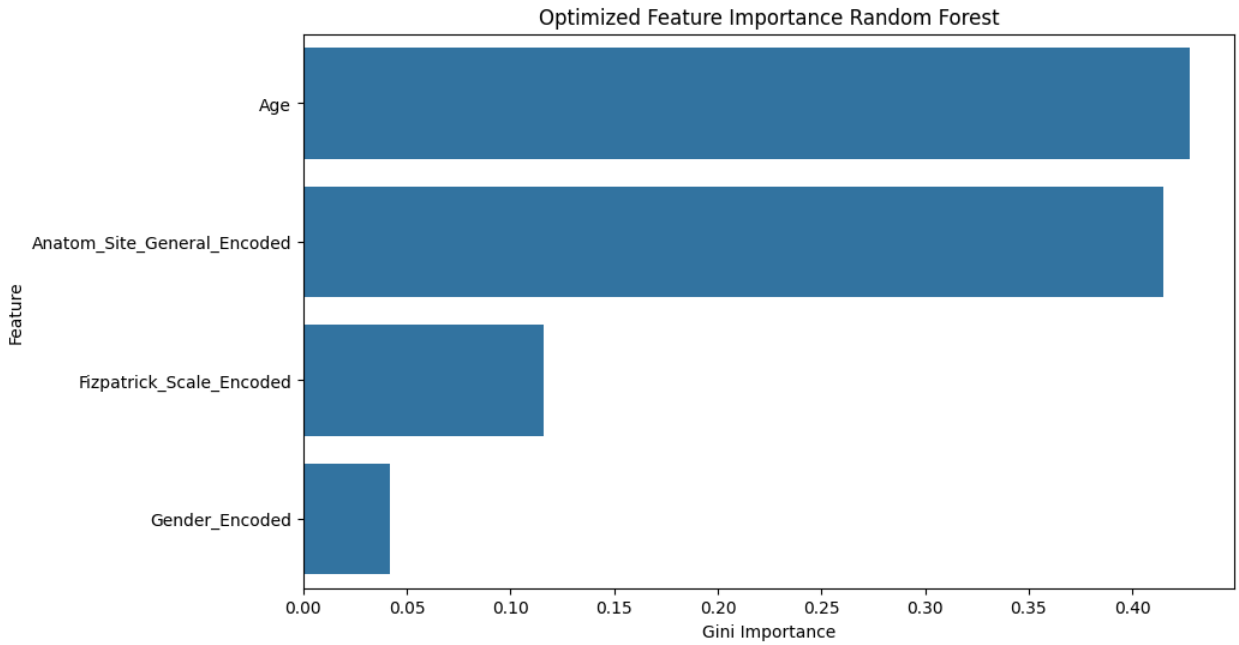
the Grid Search for the best parameters. The model that didn't use a parameter search did better than the models that did.

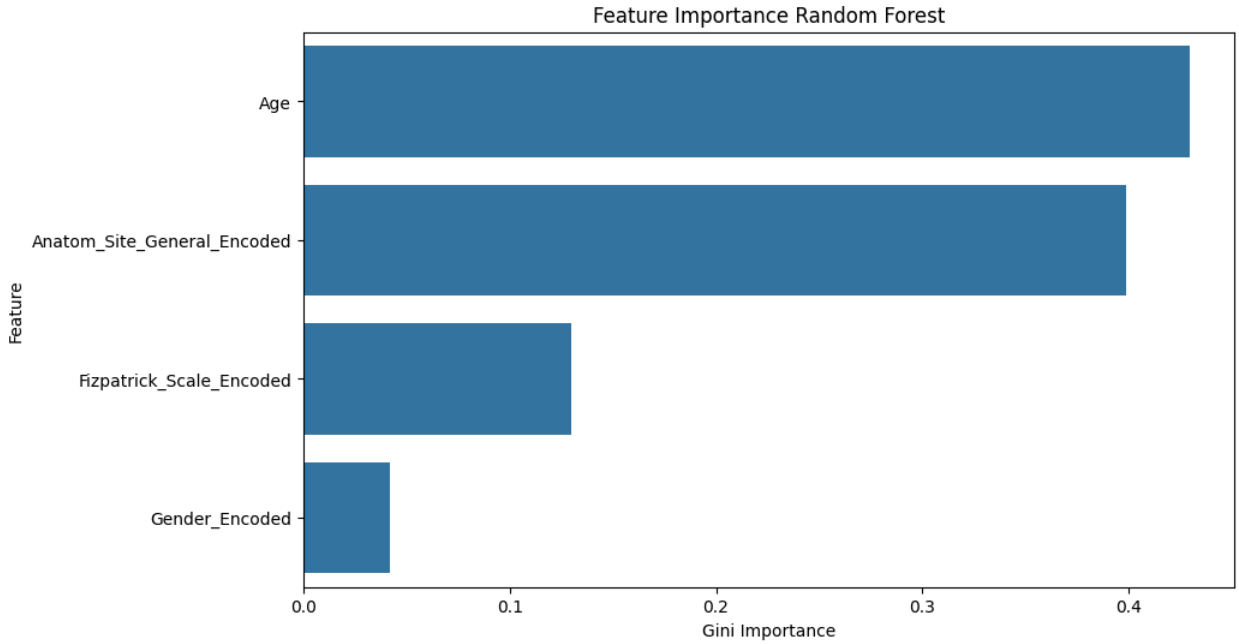


In the context of the problem it makes more sense to use the Random Forest Classifier as opposed to the Random Forest as our y values are either malignant (2), indeterminate (1), or benign (0). The random forest regressor will return float values indicating a prediction of something between our possible diagnosis. However in the real world it would be one of the 3 possible values which the classifier does identify.

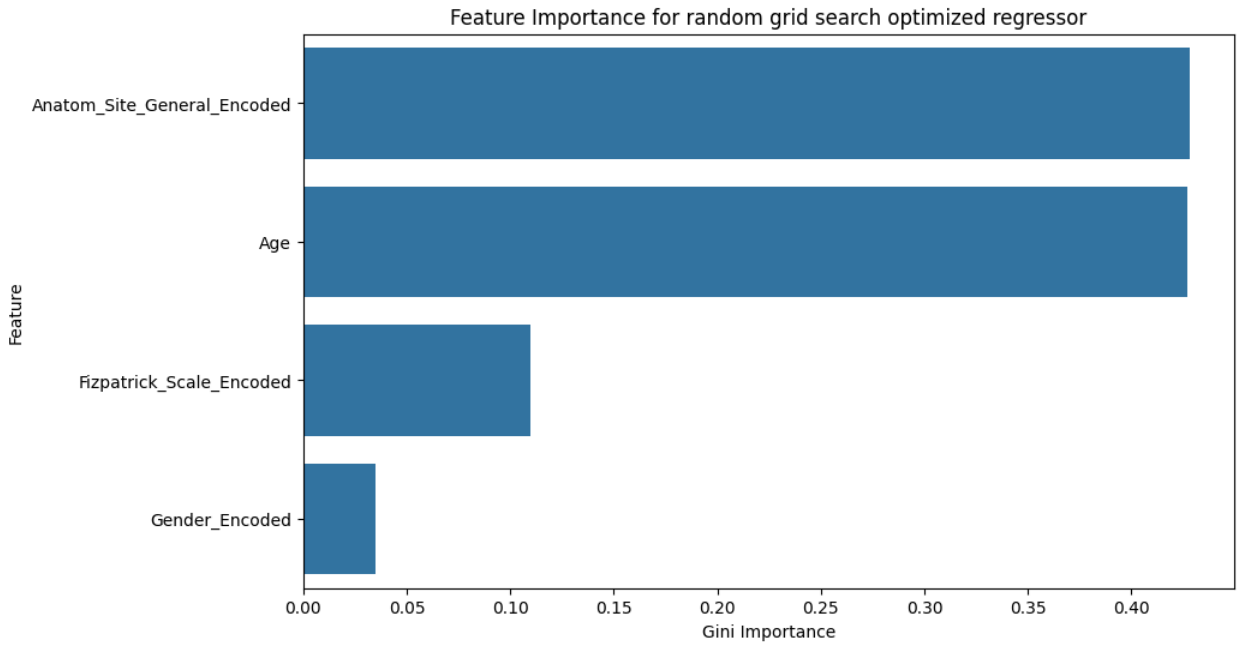
The Random Forest Regression and Classifier models all produced similar feature importance results with some variability.

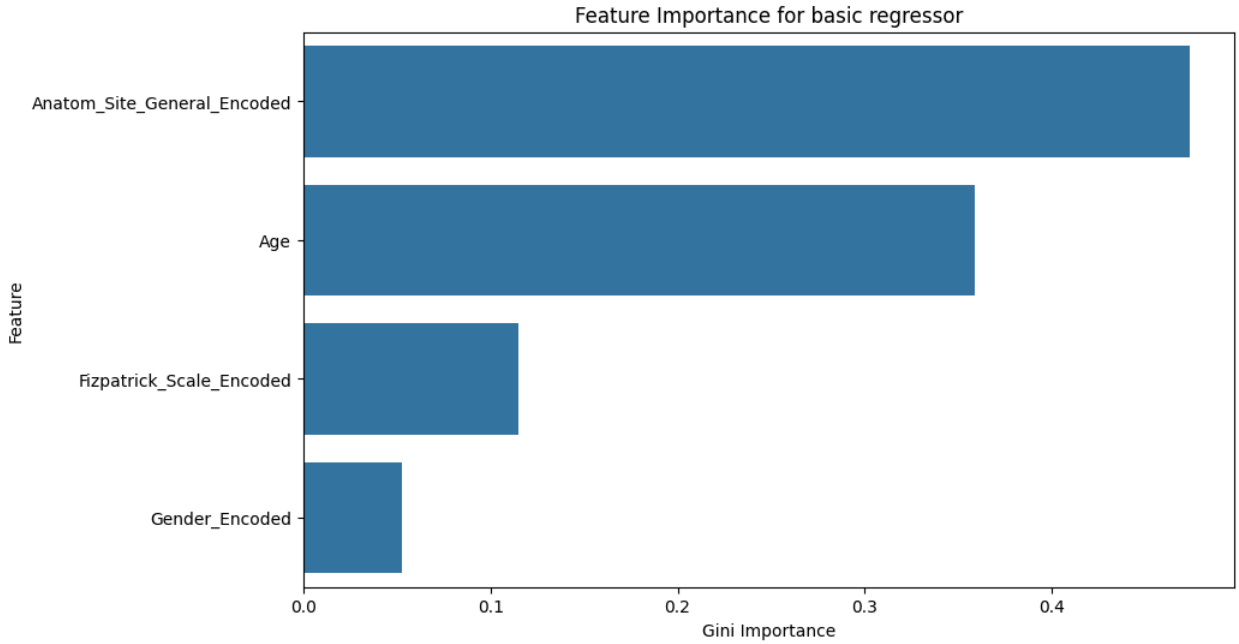
Classifiers Feature Importance:

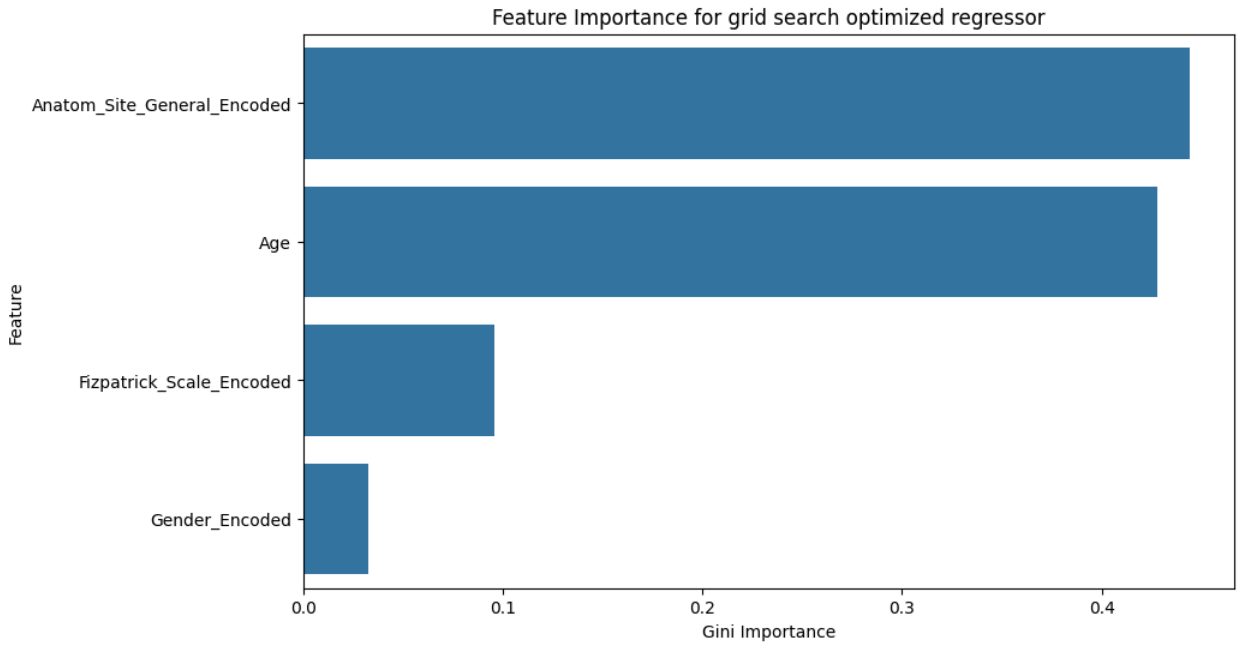




Regression Feature Importance:


Feature Importance for random grid search optimized regressor
Anatom_Site_General_Encoded
Age
Fizpatrick_Scale_Encoded
Gender_Encoded
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Gini Importance


Feature Importance for basic regressor
Anatom_Site_General_Encoded
Age
Fizpatrick_Scale_Encoded
Gender_Encoded
Feature
0.0
0.1
0.2
0.3
0.4
Gini Importance



The data for the classifiers accuracy, precision, recall, F1 Sc was compared.

	Accuracy	Precision	Recall	F1 Score	ROC AUC
Basic RF	0.6972	0.6741	0.6972	0.6773	0.7908
Optimized RF	0.6918	0.6721	0.6918	0.6725	0.7912
Tuned SVM	0.6972	0.6475	0.6972	0.6633	0.7479

Conclusion

The accuracy we were able to get our model to may suggest that even though this data was for an image recognition system it would greatly benefit from entering demographic information considering we were able to get such accurate results. Indicated by the accuracy evaluation using the oob, mse and R^2 values none of the regression models were viable. On the other hand using the classifier models the accuracy rate for all of them was around 70% and a ROC AUC of 79%. Considering images recognition training programs using VGG19 that used this data set in 2020 have been able to achieve 80% accuracy, our program suggests that the demographic information may significantly contribute to further improving image recognition accuracy (Cassidy et al. 2022).

Limitations and Restrictions

One of the biggest limitations for our project was finding suitable data sets. Even after connecting with local field experts, we couldn't find any publicly accessible data set that contained information pertaining to lifetime UV exposure. This made it difficult to properly analyze the risk factor contribution of UV exposure especially compared to other features we analyzed.

Another issue was data privacy concerns for information. Meaning that data we did have access to in general was made to be very generic. For example age in our data set was rounded only showing general age, skip counting by five.

The data set we used contains more data on certain groups of people than others. For example when analyzing the number of individuals in the different fitzpatrick categories we have an overwhelming number of individuals who fall into the fitzpatrick skin type 2 group compared to those who have higher fitzpatrick skin types like categories 5 and 6 where the number of individuals isn't visible, but still present. We also see patterns in the data showing more prominent numbers of sampling for certain groups than others in age and anatomical position. This can lead to less accuracy for values that don't have as many samples.

SPF Experiment

To study the deterioration of sunscreen, we tested various sunscreens on pieces of paper to help us understand the protective power of SPF. We will then record the data using UV imaging. Also to help us supplement for missing data, we will be using this data to help us model the degradation of sunscreen throughout the course of the day.

Materials

For this experiment, we used the following materials:

- UV pass filter and compatible camera
- Printer Paper
- Sunscreen Lotion SPF 50
    - Lot: E4B1904

    - Expiration: 3/2027
    - Brand: Supergoop Play! Spf 50 everyday lotion
    - Active ingredients: Avobenzone 3%, Homosalate 10%, Octisalate 5%, Octocrylene 7.5%
- Sunscreen Spray SPF 50
    - Lot: 0341723A
    - Expiration: 6/2026
    - Brand: Up&up sport sunscreen spray SPF50
    - Active ingredients: Avobenzone 3%, Homosalate 10%, Octisalate 5%, Octocrylene 4%
- UV light
- ½ teaspoon measuring spoon

Procedure

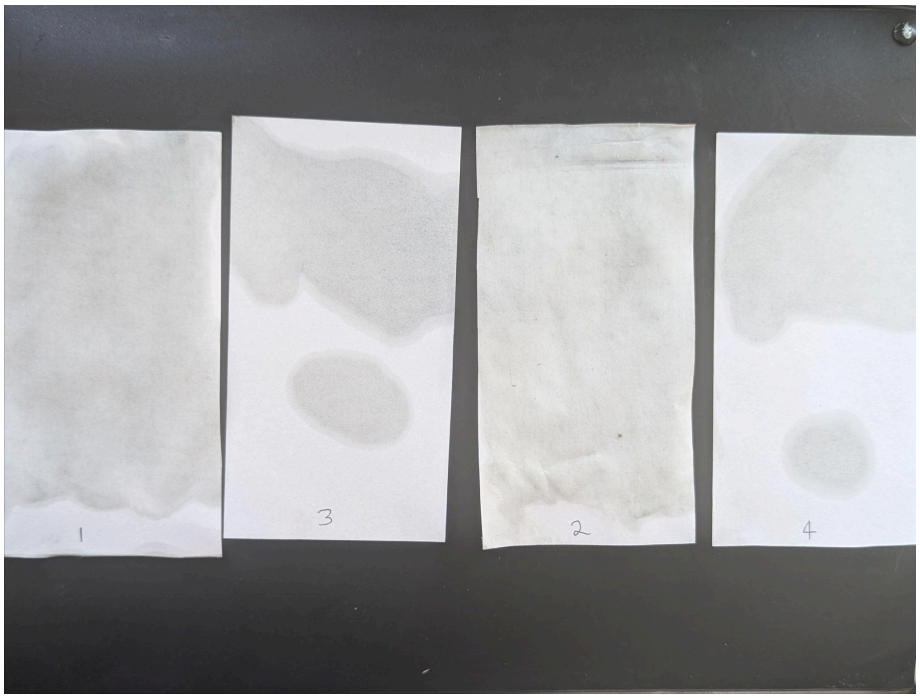
First, we took into account and recorded any significant weather factors or unforeseen circumstances that could affect the integrity of the experiment. Then, we set a black piece of paper as our background for the experiment and set up a dry flat surface with space for two 2x5 pieces of paper both inside and outside. We then set up a station with a UV pass filter and compatible camera facing nearly straight downwards, and a UV light that can be shone in the same place each time on the pieces of paper while images are being taken. Then, in a darkroom indoors (with enough light to see) we cut four 2x10 pieces of paper out of the printer paper and labeled each respectively 1-4. We then applied the sunscreen to each paper, putting ½ teaspoon of sunscreen lotion to papers 1 and 2, and putting ½ teaspoon of sunscreen spray to papers 3 and 4.

Once the papers were dry, we took UV images of each piece of paper in the dark room with the camera under the UV light. After we recorded the strength of the local UV index, we quickly took each paper to their designated locations, leaving papers 1 and 3 inside and placing the papers 2 and 4 outside in direct sunlight. After that, we recorded the strength of the local UV index. Every 20 minutes, we would quickly take images of each piece of paper until sunset, returning them to their locations immediately after images are taken. We kept a record of the

images with their names and their time. For the most accuracy, this process should be repeated multiple times with different time variables, sunscreen, and sunscreen amount.

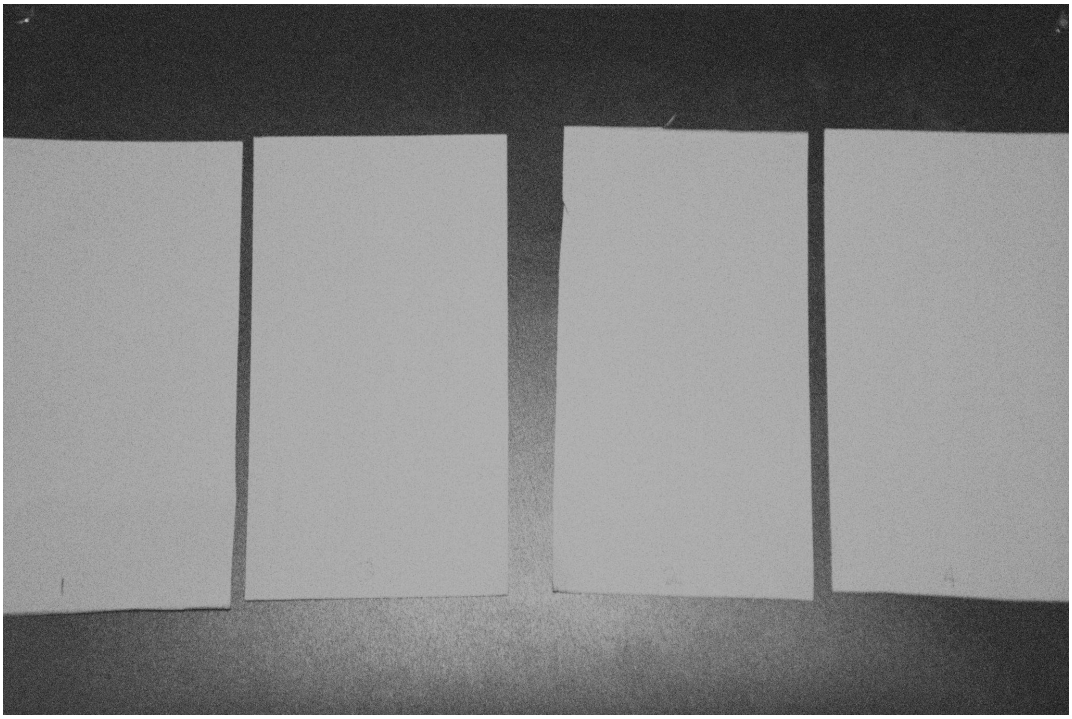
Results

The results of our experiment displayed the degradation of SPF from both time and sun exposure. As seen in the photos below, paper 1, which has SPF lotion applied and was kept inside, remains presenting darker on our UV camera, indicating the integrity of the sunscreen.

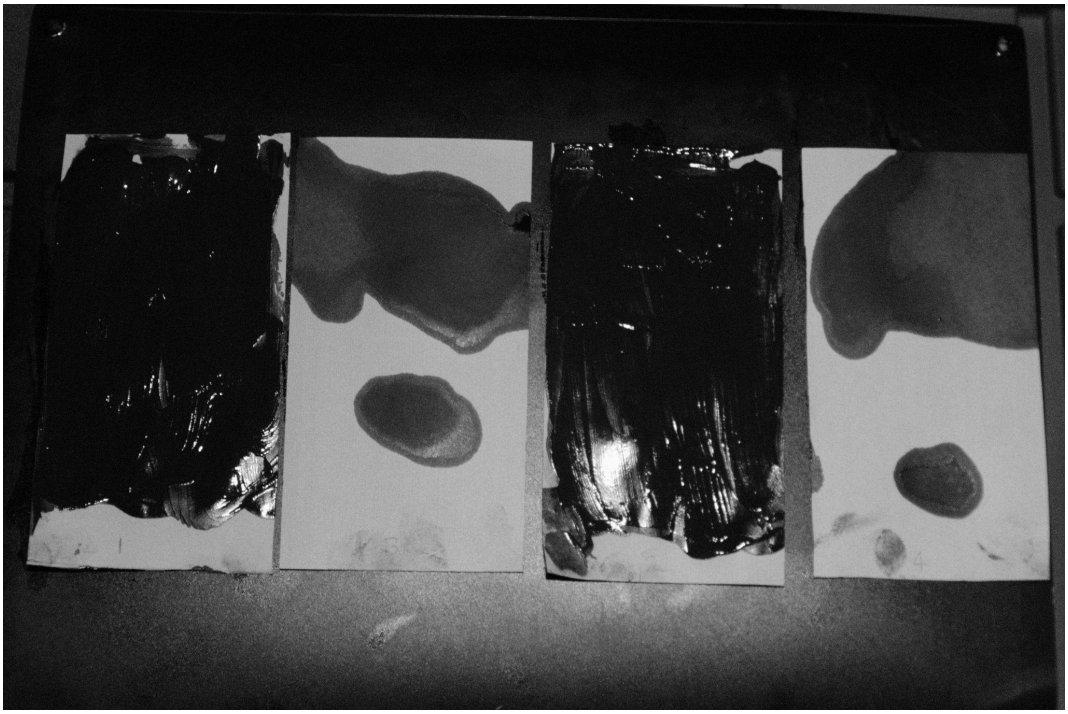


Papers 1-4 with dry sunscreen, not taken by our UV camera.

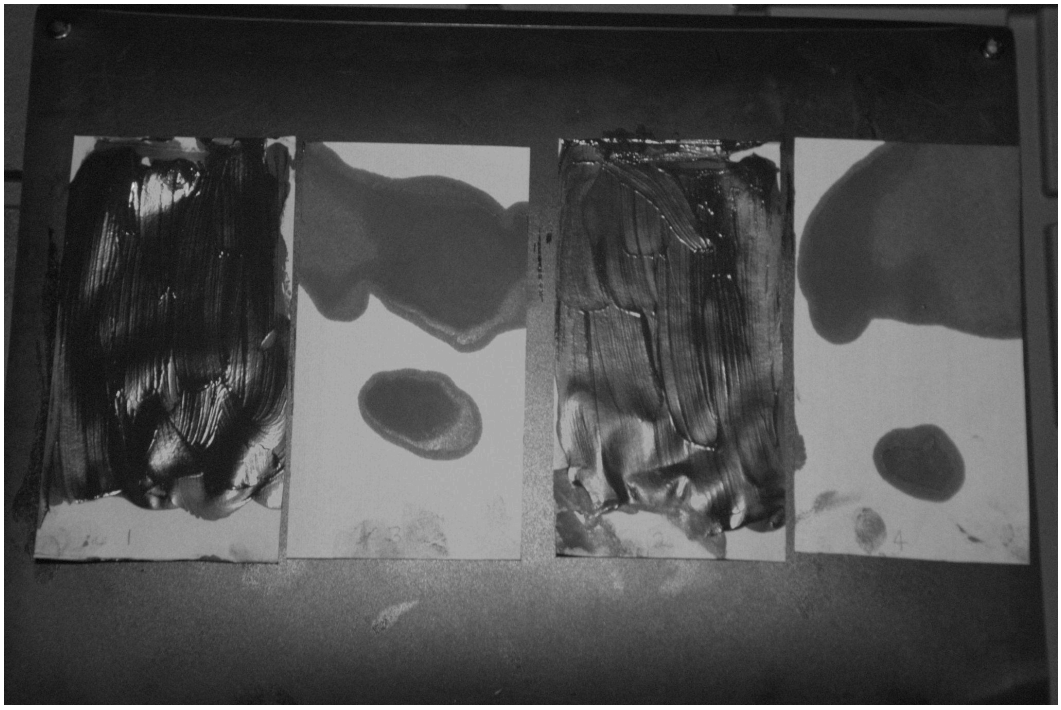
UV camera positioning



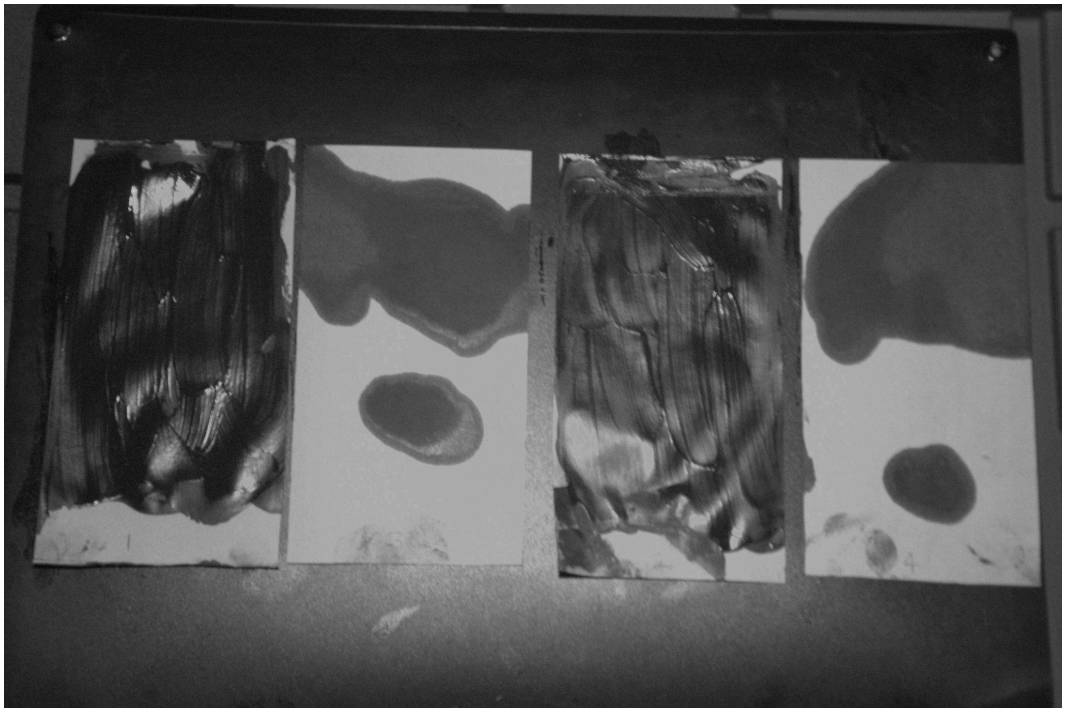
Papers 1-4 before application of sunscreen, taken by UV camera.



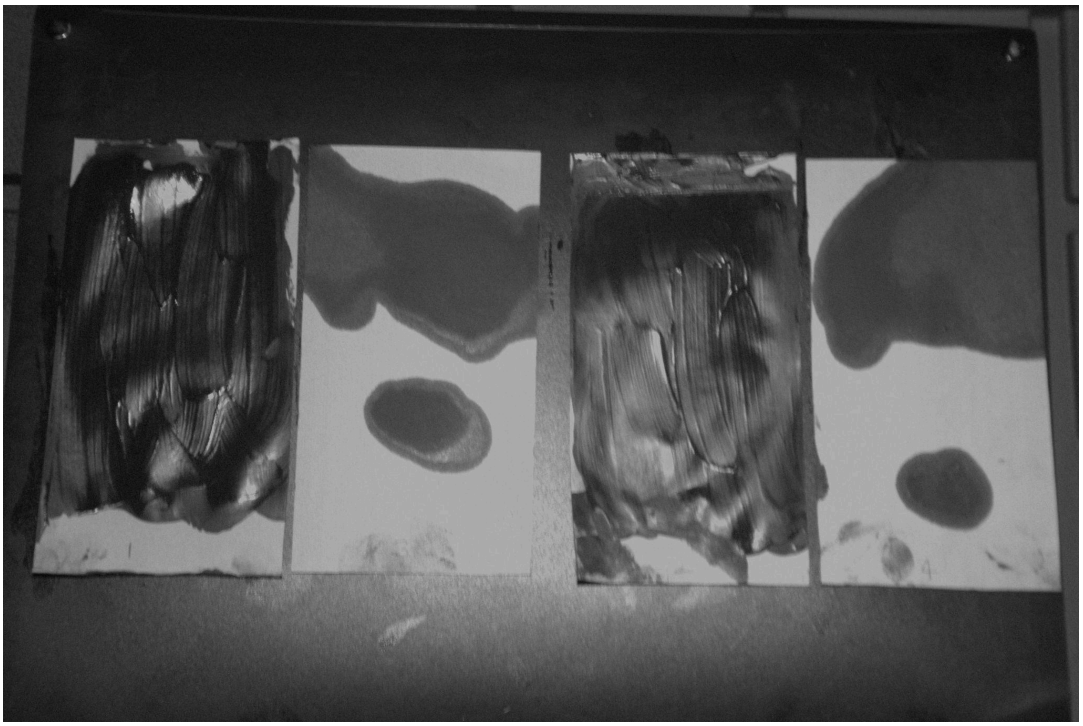
Papers 1-4, before sun exposure, taken by UV camera.



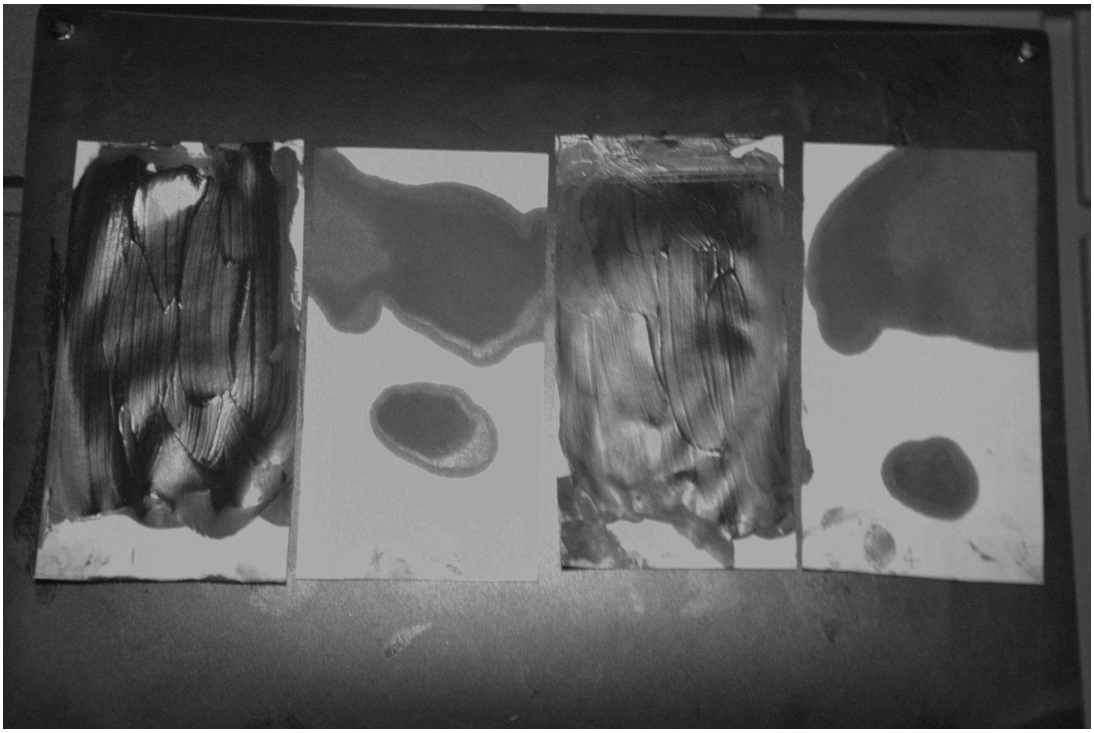
Papers 1-4, after first 20 minute increment, taken by UV camera.



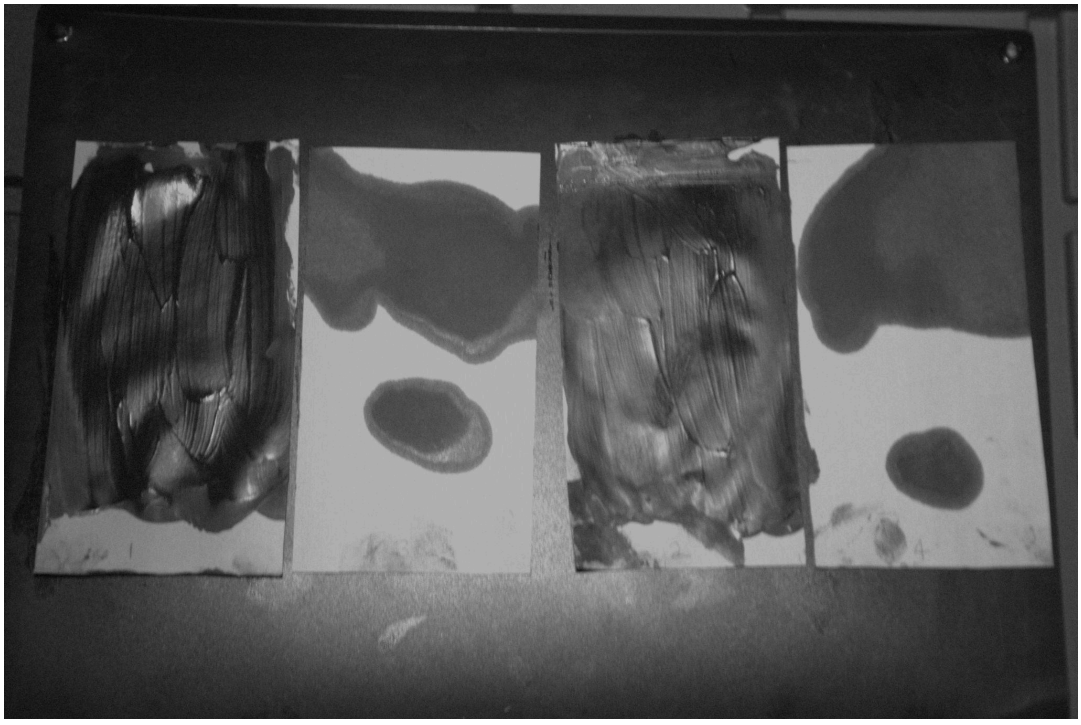
Papers 1-4, after second 20 minute increment (40 minutes total), taken by UV camera.



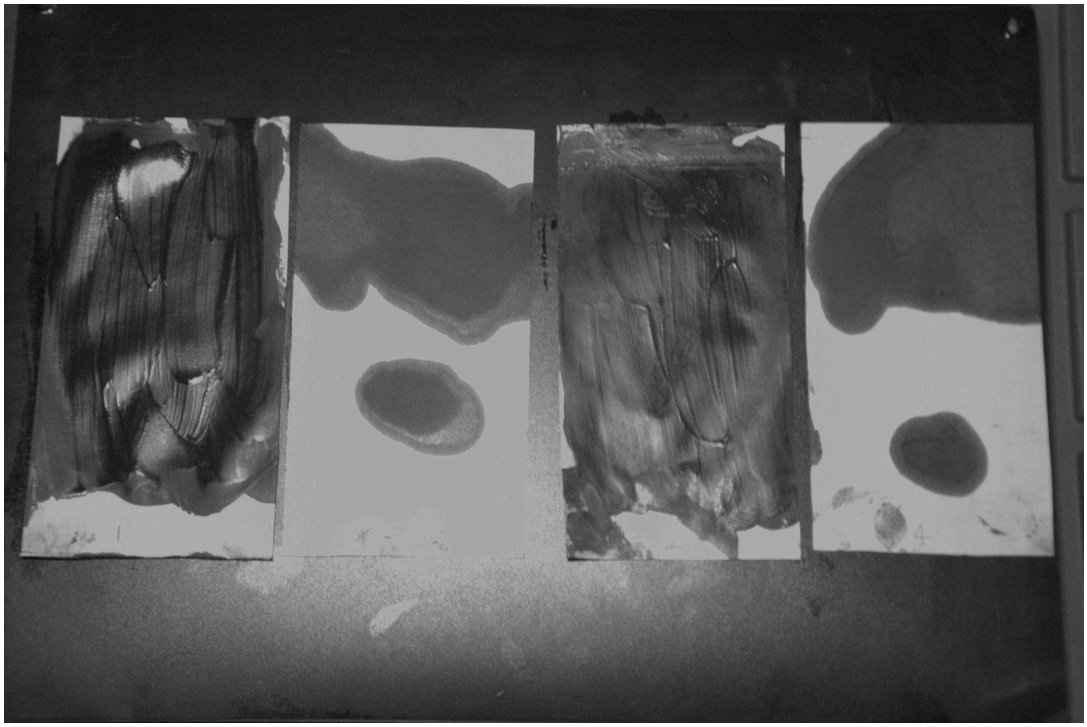
Papers 1-4, after third 20 minute increment (60 minutes total), taken by UV camera.



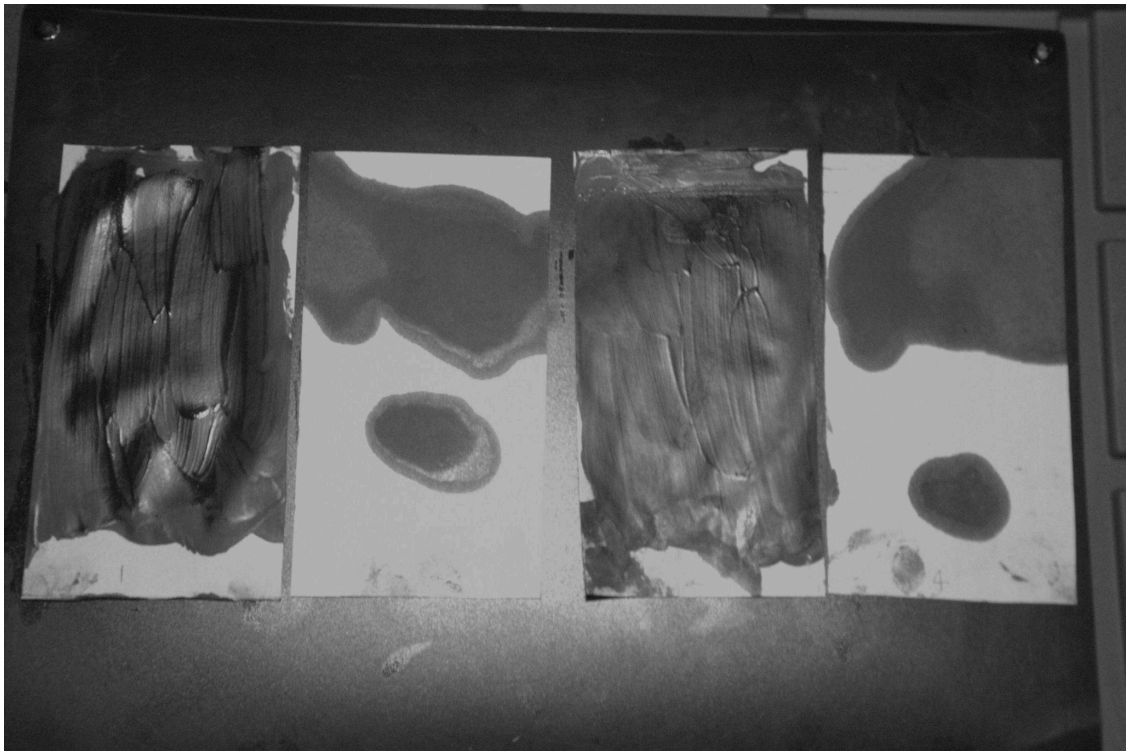
Papers 1-4, after fourth 20 minute increment (80 minutes total), taken by UV camera.



Papers 1-4, after fifth 20 minute increment (100 minutes total), taken by UV camera.



Papers 1-4, after sixth 20 minute increment (120 minutes total), taken by UV camera.



Papers 1-4, after seventh and final 20 minute increment (140 minutes total), taken by UV camera.

Over the two hour and twenty minute time period, we see the integrity of the sunscreen was compromised by two factors. First, the spray sunscreen performed much worse than the sunscreen lotion. When comparing papers 1 and 2, which were kept inside, the sunscreen spray appears lighter on camera, indicating less protection from UV radiation. This difference is especially visible after the first 20 minute increment. Then, when comparing papers 1 and 2 which stayed indoors to papers 3 and 4 which were placed outdoors, the papers that were left inside appear slightly darker to the UV camera when compared to the others. This difference is especially visible when comparing papers 1 and 3 which both had sunscreen lotion applied. At the end of the experiment, papers 2, 3, and 4 seem to be a similar shade under the UV camera, showing an approximate equality in sun protection. The integrity of paper 1 remains the strongest through the time, as it is noticeably darker on the UV camera at the end of the experiment.

If we were to do this experiment again, we would replicate it multiple times in a more controlled environment. The weather was not optimal when we performed this experiment as it was notably cloudy, impacting the UV level. The next time we do this experiment, we would repeat the same process but we would also note the brand, ingredients, lot number, and expiration date of the sunscreen used.

Prevention Methods

Skin cancer, being the most common type of cancer in the US and worldwide, kills more than two people per hour. Even when non lethal, one in every five Americans will deal with some form of skin cancer in their lifetime. Most cases of skin cancer stem from overexposure to UV rays from the sun, tanning beds, and or sunlamps. (“Skin Cancer Facts & Statistics”, n.d.) Due to the cumulative nature of UV skin damage, what may seem like a few harmless sunburns will add up leading to premature aging, skin texture changing, precancerous cells, and sometimes cancerous cells. (“Skin Cancer Basics” 2024)

Tanning beds

The first step to reducing your personal risk of skin cancer is to stop using tanning beds. Using a tanning bed can increase an individual's risk of squamous cell carcinoma by 58% and basal cell carcinoma by 24%. ("Indoor tanning", n.d.) When use begins before the age of 35, an individual's risk of developing melanoma increases by up to 59% to 75%. (Fisher and Hall, n.d.) Additionally, those who use tanning beds ten or more times in their lives increase their risk factor for melanoma by 34%. Though young women who use tanning beds are more likely to develop skin cancer than virtually any other group, nearly 70% of tanning salon clientele are caucasian women, generally ranging between the ages of 15 to 29. ("Indoor Tanning", n.d.) Alongside increasing an individual's risk of skin cancer, tanning beds increase risk of eye damage and eye cancer, actinic or solar keratoses, and immune system suppression. ("The Risks of Tanning" 2023)

Sun Safety

Aside from avoiding tanning beds, the CDC recommends practicing sun safety, especially when the UV index is 3 or higher. Practicing sun safety can look like covering exposed skin with wide brim hats, UV protectant clothing, and UV protectant sunglasses. Any remaining exposed skin should be covered with 1 ounce of broad spectrum (UV-a and UV-b) sunscreen of SPF 15 or higher 30 minutes prior to exposure. ("Reducing Risk for Skin Cancer" 2024) Additionally, it is advised to stay out of direct sun during peak UV index hours, generally around 10am to 4pm. ("Skin Cancer Prevention", n.d.)

The type of clothing worn is a factor in how protected one is against UV radiation. Tightly woven fabrics such as denim, satin, and polyester blend clothing will block more UV rays than loosely woven fabrics such as chiffon, lace, and linen. When choosing a hat, one should also look for tightly woven fabrics such as canvas and avoid any hat that features holes in its design such as straw hats and open back trucker caps. As a general rule, the darker the clothing, the more UV rays it will soak up, preventing them from reaching skin. ("Sun Safety Facts" 2024)

The type of sunscreen worn is another factor in UV protection. One should look for a broad spectrum SPF that protects against both UV-a and UV-b rays. Sunscreen may wear off

after water activities like swimming, or activities that cause increased sweat such as running. One should reapply sunscreen every two hours, or after one of the aforementioned activities. (“Sun Safety Facts” 2024)

Future Work

Improving the Machine Learning Model

Our team plans to further our exploration into skin cancer risks, expand our machine learning data set to improve accuracy, utilize additional models, and explore auxiliary projects. We plan to further research documented factors that most significantly impact skin cancer rates, along with their relative importance. Our initial focus would be increasing the accuracy of our machine learning model. We plan to do so by including more predictor variables and training it on additional types of regression models. With increasing the veracity and consistency of our model, we will be able to reliably predict risk factors. Additionally we would like to try using Neural Networks and including images in the future.

Global UV Risk Simulation

During our project, we faced difficulty in obtaining an expansive and reliable data set which included all of our factors of interest. Moving forward, we plan to reach out to hospitals, universities, and professors specializing in skin cancer. Obtaining multiple data sets would improve the accuracy of our machine learning model and aid in finding the greatest risk factors. We then plan to model these factors using an object-oriented Python code and create a proposition to help protect people against skin cancer. With this code, our team intends to use publicly available geological and meteorological data to simulate the global climate and create pre-set environments and time scales to allow for precise situational results. Using this data, we could then investigate the penetration of UV rays and see how different UV rays penetrate the atmosphere in different locations around the world. Additionally, we plan to find a way to incorporate our data on personal risk factors such as age, gender, and fitzpatrick scale score from this project to predict and simulate global effects of UV Rays on people with different risk factors. It is in this model where we would use result from our live experiment where we would

incorporate results from our live experiment using the results from our live experiment in sunscreen degradation evaluating the sunscreen degradation from UV images. We additionally predict that this process may raise new factors of interest which we would further investigate.

Our team plans to make this data easily accessible to the public by converting our code into a resourceful app which people could use to research and explore the application of prevention methods. Creating this simulation would inform the public of individualized risks of skin cancer and aid in personal protection. In order to accomplish such a goal, we would have to extract current global UV radiation data, geological information such as elevation and sun coverage, then code consideration for how heavily each prominent factor plays a role in skin cancer risk.

We could also implement the object oriented code into our app. Using it to model skin damage and evaluate skin damage severity from images. we could allow individuals to upload pictures of themselves. Using UV imaging equipment to analyze their skin damage to help evaluate risk and possibly identify cancerous lesions.

Acknowledgements

Elisea Jackson and Megan Odom would like to express our deepest gratitude for the individuals who have supported and guided us through this process. First and foremost, we would like to acknowledge Felina Rivera Calzadilla who was integral to the process. Felina's expertise in coding, generous provision of UV imaging equipment, and guidance in writing this report proved pivotal in our process.

We would also like to note our teachers Acacia McCombs and Sarah Rowe. We would like to thank them for their unwavering support, especially shown by Mrs. Rowe who stayed active in the project even after departing from their job at our school. Ms. McCombs has been crucial to our project as her encouragement and mentorship helped us navigate the process. We would like to give a big thank you to Marianne Birweck, a field specialist from the University of New Mexico, who lent her time to help in our project.

We also acknowledge the assistance of Claude, an AI assistant developed by Anthropic, which provided guidance on code implementation and data visualization techniques for the model comparison analysis. We would also like to acknowledge the use of Chat GPT, which we

used for formatting guidance on our object oriented code model. Finally, we would like to note our use of Gemini in Google Collab to debug our programming and create graphs of our data and results.

References

Andrade-Lima LC, Veloso A, Paulsen MT, Menck CF, Ljungman M. DNA repair and recovery of RNA synthesis following exposure to ultraviolet light are delayed in long genes. Nucleic Acids Res. 2015 Mar 11;43(5):2744-56. doi: 10.1093/nar/gkv148. Epub 2015 Feb 26. PMID: 25722371; PMCID: PMC4357734.

Barrington, Kate. "Fitzpatrick Scale: What It Is, How It Works, & How to Explain It to Your Clients." Professional Skincare Guide, 28 Sept. 2022, professionalskincareguide.com/fitzpatrick-scale/.

Cassidy, Bill, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, Moi Hoon Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations" Medical Image Analysis, Volume 75, 2022, 102305, ISSN 1361-8415, https://doi.org/10.1016/j.media.2021.102305.

Department of Health & Human Services. "Melanoma." Better Health Channel, www.betterhealth.vic.gov.au/health/conditionsandtreatments/melanoma.

"Effects of Skin Cancer Treatment." *Siteman Cancer Center*, 2015, siteman.wustl.edu/treatment/cancer-types/skin/effects-of-treatment. Accessed 20 Dec. 2024.

Fisher, David E., and Bartlett Hall. n.d. "Indoor UV tanning and skin cancer: health risks and opportunities." PubMed Central. Accessed April 6, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC2913608/.

Garinis et al. "Transcriptome analysis reveals cyclobutane pyrimidine dimers as a major source of UV‐induced DNA breaks" The EMBO Journal (2005) 24: 3952 - 3962 https://doi.org/10.1038/sj.emboj.7600849

"Indoor Tanning." n.d. AIM at Melanoma Foundation. Accessed April 6, 2025. https://www.aimatmelanoma.org/melanoma-101/understanding-melanoma/melanoma-risk-factors/indoor-tanning/.

The International Skin Imaging Collaboration. (2025, February 7). ISIC archive data set [Data set]. Retrieved from https://www.isic-archive.com/

Kaiser, Isabelle, et al. "Risk Prediction Models for Melanoma: A Systematic Review on the Heterogeneity in Model Development and Validation." International Journal of Environmental Research and Public Health, vol. 17, no. 21, 1 Jan. 2020, p. 7919, www.mdpi.com/1660-4601/17/21/7919, https://doi.org/10.3390/ijerph17217919. Accessed 24 May 2023.

Kciuk M, Marciniak B, Mojzych M, Kontek R. Focus on UV-Induced DNA Damage and Repair-Disease Relevance and Protective Strategies. Int J Mol Sci. 2020 Oct 1;21(19):7264. doi: 10.3390/ijms21197264. PMID: 33019598; PMCID: PMC7582305.

"Moles and Melanoma Risk." MedlinePlus, U.S. National Library of Medicine, https://medlineplus.gov/genetics/understanding/traits/moles/#:~:text=In%20susceptible%20individuals%20(those%20with,increasing%20their%20risk%20of%20becoming.

Office of the Surgeon General, Assistant Secretary for Health (ASH). "Skin Cancer: Quick Facts from the Surgeon General." *HHS.gov*, 7 Aug. 2014, www.hhs.gov/surgeongeneral/reports-and-publications/skin-cancer/fact-sheet/index.html

"Oncogenes, Tumor Suppressor Genes, and DNA Repair Genes." 2022. American Cancer Society. https://www.cancer.org/cancer/understanding-cancer/genes-and-cancer/oncogenes-tumor-suppressor-genes.html.

Quantexa. n.d. “Support Vector Machines for dummies; A Simple Explanation.” aylien.com. https://aylien.com/blog/support-vector-machines-for-dummies-a-simple-explanation

Rastogi RP, Richa, Kumar A, Tyagi MB, Sinha RP. “Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair.” J Nucleic Acids. 2010 Dec 16;2010:592980. doi: 10.4061/2010/592980. PMID: 21209706; PMCID: PMC3010660.

“Reducing Risk for Skin Cancer | Skin Cancer.” 2024. CDC. https://www.cdc.gov/skin-cancer/prevention/index.html.

"Risk Factors for Basal and Squamous Cell Skin Cancer." American Cancer Society, www.cancer.org/cancer/types/basal-and-squamous-cell-skin-cancer/causes-risks-prevention/risk-factors.html.

Rowe, W. (2018, July 5). Mean Square Error & R2 Score Clearly Explained. BMC Software. Retrieved April 6, 2025, from https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/

“The Risks of Tanning.” 2023. FDA. https://www.fda.gov/radiation-emitting-products/tanning/risks-tanning.

Shalev, Ofir. 2019. “Recall, Precision, F1, ROC, AUC, and everything | by Ofir Shalev (@ofirdi) | The Startup.” Medium. https://medium.com/swlh/recall-precision-f1-roc-auc-and-everything-542aedf322b9.

"Skin Cancer." Illinois Department of Public Health, https://dph.illinois.gov/topics-services/diseases-and-conditions/cancer/type/skin.html.

“Skin Cancer Basics | Skin Cancer.” 2024. CDC. https://www.cdc.gov/skin-cancer/about/index.html.

"Skin Cancer Facts & Statistics." n.d. The Skin Cancer Foundation. Accessed April 6, 2025. https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/.

"Skin Cancer Prevention." n.d. The Skin Cancer Foundation. Accessed April 6, 2025. https://www.skincancer.org/skin-cancer-prevention/.

Sullivan, Debra. 2019. "UVA vs. UVB Rays: What's the Difference?" Healthline. https://www.healthline.com/health/skin/uva-vs-uvb#about-uvb-rays.

"Skin Cancer Prevention." n.d. The Skin Cancer Foundation. Accessed April 6, 2025. https://www.skincancer.org/skin-cancer-prevention/.

"Sunbeds and Skin Cancer Risk." Irish Cancer Society, www.cancer.ie/cancer-information-and-support/cancer-types/skin-cancer/sunbeds-and-skin-cancer-risk#:~:text=Sunbeds%20and%20cancer,-The%20World%20Health&text=Even%20one%20sunbed%20session%20can,your%20risk%20by%20almost%2060%25. Accessed 1 Mar. 2025.

"UV Index | NASA." UV Index | NASA, 3 Apr. 2024, neo.gsfc.nasa.gov/view.php?datasetId=AURA_UVI_CLIM_M.

"Ultraviolet Radiation and Melanoma." *Elsevier*, Seminars in Cutaneous Medicine and Surgery, 2011, cdn.mdedge.com/files/s3fs-public/issues/articles/Vol30_i4_Kanavy.pdf. Accessed 20 Dec. 2024.

Ultraviolet (UV) Radiation | Center for Science Education. (2017). UCAR Center for Science Education. Retrieved April 5, 2025, from https://scied.ucar.edu/learning-zone/atmosphere/ultraviolet-uv-radiation

"What Is Hyperparameter Tuning?" 2024. IBM. https://www.ibm.com/think/topics/hyperparameter-tuning.

Zheng, Catherine. 2020. “How Sun Damage Really Causes Cancer: The effect of UV radiation on cancer-causing DNA mutations – Yale Scientific Magazine.” Yale Scientific Magazine. https://www.yalescientific.org/2020/03/how-sun-damage-really-causes-cancer-the-effect-of-uv-radiation-on-cancer-causing-dna-mutations/.

Zivkovic, Strahinja. 2022. “#012 Machine Learning – Introduction to Random Forest – Master Data Science.” Master Data Science. https://datahacker.rs/012-machine-learning-introduction-to-random-forest/.

The International Skin Imaging Collaboration. (2025, February 7). ISIC archive data set [Data set]. Retrieved from https://www.isic-archive.com/