Unilever
Explore the National Health and Nutrition Examination Survey

Siyu Shen (ss6669), Xinran Chen (xc2660), Ying Hong (yh3538),
Shiyuan Xu (sx2311), Tianyu Han (th2987)

1 Project Overview

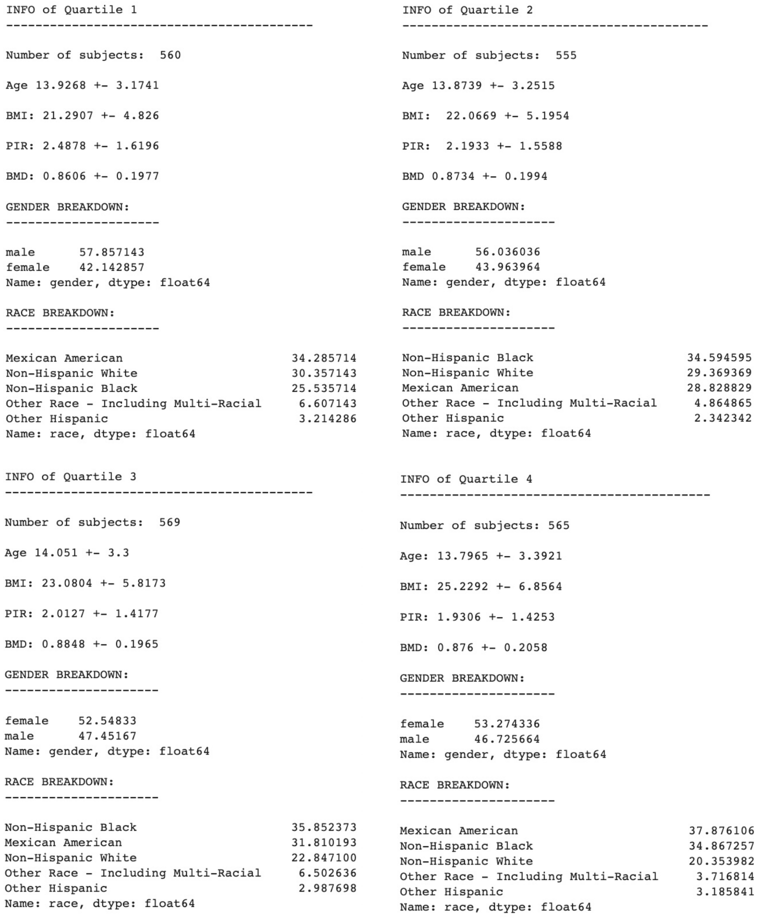
Skin diseases, such as psoriasis, not only impact the physical health of millions worldwide, but they also bear significant psychological and social implications for those affected. By researching skin diseases, we seek to understand their underlying causes, progression, and impact on quality of life. Delving into this realm offers opportunities to develop better diagnostic tools, more effective treatments, and comprehensive care strategies. Furthermore, the insights gained can shed light on the complex interplay between genetics, environmental triggers, and immune system responses, which can be pivotal for medical advancements beyond dermatology.

With the goal of understanding skin diseases, we explored datasets from the National Health and Nutrition Examination Survey (NHANES), a comprehensive initiative that provides detailed health and nutritional data from a diverse cross-section of the U.S. population. The NHANES datasets offer a wealth of information, including but not limited to, demographic details, medical histories, and specific dermatological assessments. Recognizing the profound impact of psoriasis on a significant portion of the population, we've chosen to begin our research by investigating this condition with the help of classification models. By starting with psoriasis, we aim to uncover patterns, trends, and potential risk factors, paving the way for a deeper understanding of skin diseases as a whole.

2 Re-implementation of past NHANES study

To gain more experience working with NHANES data before delving into our psoriasis study, we took the advice of our mentor and decided to reimplement a past study focusing on the relationship between bone mineral density and Vitamin E. Since Vitamin D level is an important feature that is commonly associated with psoriasis, the re-implementation of a past study on it will help us gain insights into our study as well.

2.1 Exploratory Data Analysis



```
INFO of Quartile 1
------------------------------------------

Number of subjects:  560

Age 13.9268 +- 3.1741

BMI: 21.2907 +- 4.826

PIR: 2.4878 +- 1.6196

BMD: 0.8606 +- 0.1977

GENDER BREAKDOWN:
---------------------

male      57.857143
female    42.142857
Name: gender, dtype: float64

RACE BREAKDOWN:
---------------------

Mexican American                       34.285714
Non-Hispanic White                     30.357143
Non-Hispanic Black                     25.535714
Other Race - Including Multi-Racial     6.607143
Other Hispanic                          3.214286
Name: race, dtype: float64
```

```
INFO of Quartile 2
------------------------------------------

Number of subjects:  555

Age 13.8739 +- 3.2515

BMI:  22.0669 +- 5.1954

PIR:  2.1933 +- 1.5588

BMD 0.8734 +- 0.1994

GENDER BREAKDOWN:
---------------------

male      56.036036
female    43.963964
Name: gender, dtype: float64

RACE BREAKDOWN:
---------------------

Non-Hispanic Black                     34.594595
Non-Hispanic White                     29.369369
Mexican American                       28.828829
Other Race - Including Multi-Racial     4.864865
Other Hispanic                          2.342342
Name: race, dtype: float64
```

```
INFO of Quartile 3
------------------------------------------

Number of subjects:  569

Age 14.051 +- 3.3

BMI: 23.0804 +- 5.8173

PIR: 2.0127 +- 1.4177

BMD: 0.8848 +- 0.1965

GENDER BREAKDOWN:
---------------------

female    52.54833
male      47.45167
Name: gender, dtype: float64

RACE BREAKDOWN:
---------------------

Non-Hispanic Black                     35.852373
Mexican American                       31.810193
Non-Hispanic White                     22.847100
Other Race - Including Multi-Racial     6.502636
Other Hispanic                          2.987698
Name: race, dtype: float64
```

```
INFO of Quartile 4
------------------------------------------

Number of subjects: 565

Age: 13.7965 +- 3.3921

BMI: 25.2292 +- 6.8564

PIR: 1.9306 +- 1.4253

BMD: 0.876 +- 0.2058

GENDER BREAKDOWN:
---------------------

female    53.274336
male      46.725664
Name: gender, dtype: float64

RACE BREAKDOWN:
---------------------

Mexican American                       37.876106
Non-Hispanic Black                     34.867257
Non-Hispanic White                     20.353982
Other Race - Including Multi-Racial     3.716814
Other Hispanic                          3.185841
Name: race, dtype: float64
```

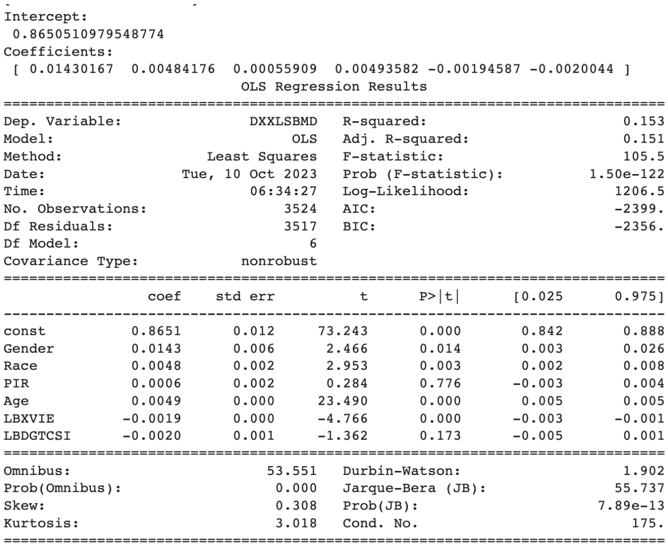
Table 1: The Population Distributions of Respondents

A series of preprocessing steps were conducted after we imported the NHANES data in XPT format. We joined the Vitamin E table and the BMD (bone mineral density) table with the demographic table using the unique survey participant ID. Then, we performed baseline characteristics analysis based on the quartile of alpha-tocopherol (a type of Vitamin E) to gain more understanding regarding the distribution of the features across the four quartiles. The four quartiles are created as follows:

Categories 1: 7.5–15.4 umol/L
Categories 2: 15.4–17.6 umol/L
Categories 3: 17.6–20.4 umol/L
Categories 4: >20.4 umol/L

We can see that there are more female participants in all four categories. Meanwhile, in the racial breakdown, Mexican Americans have the highest percentage in Category 1 and Category 4 while non-Hispanic blacks are the most prevalent race constituting Category 2 and Category 3. From the characteristic analysis, an increasing trend of BMD from Category 1 to Category 4 can also be witnessed, indicating that each individual's alpha-tocopherol level might have a positive correlation with his/her body mineral density.

2.2 Regression Analysis



```
Intercept:
 0.8650510979548774
Coefficients:
 [ 0.01430167  0.00484176  0.00055909  0.00493582 -0.00194587 -0.0020044 ]
                            OLS Regression Results
==============================================================================
Dep. Variable:               DXXLSBMD   R-squared:                       0.153
Model:                            OLS   Adj. R-squared:                  0.151
Method:                 Least Squares   F-statistic:                     105.5
Date:                Tue, 10 Oct 2023   Prob (F-statistic):          1.50e-122
Time:                        06:34:27   Log-Likelihood:                 1206.5
No. Observations:                3524   AIC:                            -2399.
Df Residuals:                    3517   BIC:                            -2356.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.8651      0.012     73.243      0.000       0.842       0.888
Gender         0.0143      0.006      2.466      0.014       0.003       0.026
Race           0.0048      0.002      2.953      0.003       0.002       0.008
PIR            0.0006      0.002      0.284      0.776      -0.003       0.004
Age            0.0049      0.000     23.490      0.000       0.005       0.005
LBXVIE        -0.0019      0.000     -4.766      0.000      -0.003      -0.001
LBDGTCSI      -0.0020      0.001     -1.362      0.173      -0.005       0.001
==============================================================================
Omnibus:                       53.551   Durbin-Watson:                   1.902
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               55.737
Skew:                           0.308   Prob(JB):                     7.89e-13
Kurtosis:                       3.018   Cond. No.                         175.
==============================================================================
```

Table 2: OLS Regression Output with Low Explanatory Power

After exploratory analysis, a regression model is used to predict the influence of alpha-tocopherol and gamma-tocopherol (LBXVIE & LBDGTCSI) on bone mineral density. We have also included demographic variables (gender, race, PIR, age) to adjust the linear regression. The resulting table indicates a negative correlation between alpha-tocopherol and BMD and also a negative correlation between gamma-tocopherol and BMD. This result aligns with the domain knowledge that bone mineral density should be inversely correlated with the Vitamin E level of the patient. In this reimplementation, we are able to find the inverse correlation present in the past study; however, due to a smaller sample size, we are unable to further adjust our regression model. Nevertheless, this recreation still serves as a solid foundation from where we can begin our study on psoriasis.

3 Psoriasis Data Overview, EDA, and Data Cleaning

Following the reimplementation of the existing study, we want to focus on finding new light on possible correlations with skin disease. Psoriasis is a chronic autoimmune skin disorder and it presents a complex interplay of genetic and environmental factors. While the precise cause of this disease is unclear, several potential influencing factors such as vitamin intake and dietary intake have been identified in past research. Thus, our next stage of exploring the NHANES comes with the goal of discovering any insights associated with psoriasis.

After exploring the NHANES dataset, we identified data containing information on psoriasis from the 2003-2004 Examination and Dermatology questionnaire. The examination dataset contains three readers' results on respondents' psoriasis presence and the questionnaire asks about whether the respondent has ever been told by a doctor that they had psoriasis. The criteria we set for the response variable "Psoriasis" is if any one of three readers gives positive results or the respondent used to have psoriasis according to the questionnaire, the label for psoriasis on the subject would be positive. Then, in order to gather as many skin disease-related features as possible, we researched psoriasis and its common co-occurrence. Based on these inferences, we explored the NHANES dataset and prepared features spanning 10+ datasets. The main categories of these variables include demographics (e.g. age, gender, race, BMI), smoking, drinking, dietary, Vitamin intake, other disease conditions, etc.

With the initial data cleaning, we gathered 2198 entries with around 20 features. Since several columns contain missing values, the first feature engineering step is imputation. For numerical values, mean imputation is used for simplicity. For categorical values, bootstrap imputation is employed to fill in missing values based on the observed distribution of non-missing values. On the other hand, collinearity may hinder the model interpretability, so we also conducted the correlation analysis on the features and removed highly correlated features with a correlation greater than 0.8. Here, body measures such as waist and thigh circumference and age measures with high correlation are dropped.

Below is the correlation heatmap of the data. At the initial analysis, there is no straightforward correlation between psoriasis and any other factors we have already identified, and more advanced modeling would be used to shed light on the study.

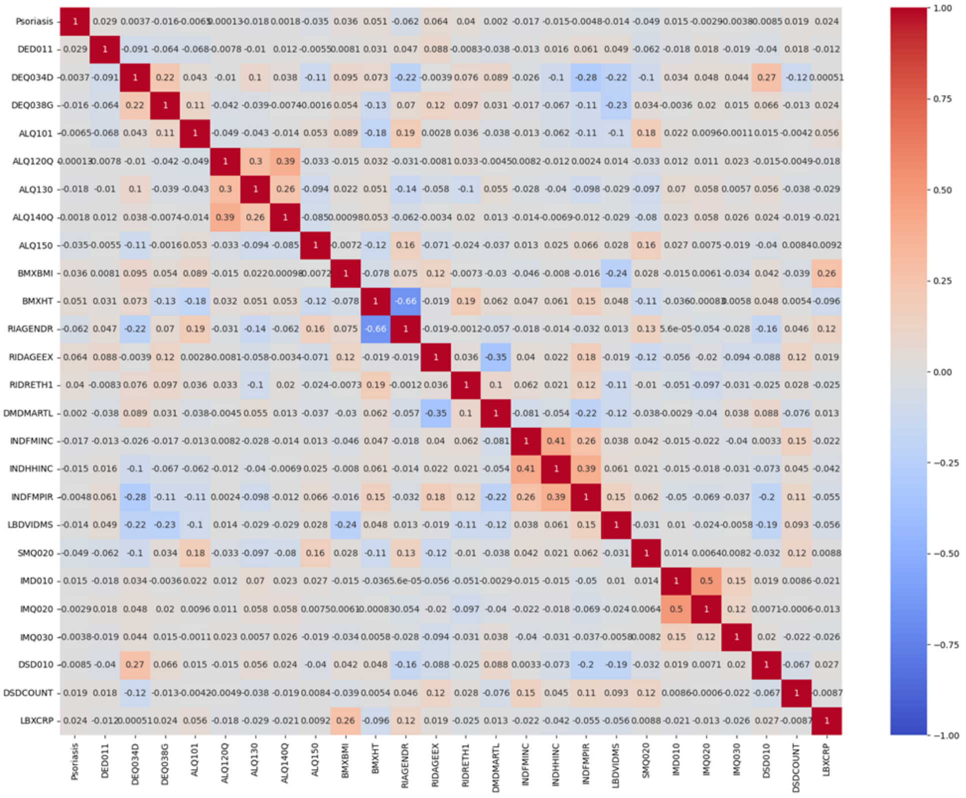


Table 3: Heatmap Representing Correlation Coefficients

In the second part of the study, we examined the larger NHANES dataset and included more features that may be a potential cause. Also, the previous stage study only contains data available from 2003 to 2004. So, we aggregate more years of data by carefully identifying matching columns. The aggregated dataset now contains psoriasis data and feature data (demographic, physical exam, dietary, vitamin supplements, etc.) spanning years 2003 - 2005 and 2011 - 2014 and the current number of rows is 24815. With the increasing amount of data, we can now make more confident inferences.

After examining the distribution of the new dataset, we discovered that the number of psoriasis observations is much larger than the number of healthy skin with a value count of:

0 (healthy)	*24130*
1 (psoriasis)	*685*

In order to mitigate the impact of the imbalanced dataset on the model performance, we experimented with oversampling techniques and data augmentation which will be introduced later.

Lastly, in order to enhance the interpretability of the model, we conducted another round of feature selection. After some time tuning the model, we realized the current data contains some features that exhibit high importance while hard to interpret. For example, WIMEC2YR: Full Sample 2 Year MEC Exam Weight and SDMVSIRA: Masked Variance Unit Pseudo-Stratum variable. These features cannot be collected in the future experiments and so are meaningless to include in our current model. After examining the close interpretation of each feature, we ended up with 39 features that were deemed most relevant. With these changes, the resultant feature importance also becomes more explanatory.

4 Classification Model Selection

4.1 Classification models

This section aims to discuss the critical progress of selecting suitable machine/deep learning models to complete the task, which is predicting whether a given individual has psoriasis. The successful deployment of such models has the potential to revolutionize the conventional way of diagnosing psoriasis at early stages, allowing for timely treatments and proactive patient care.

The goal of our project was to develop a classification model that can predict the likelihood of psoriasis occurrence in individuals based on the provided data. However, as previously introduced in the data overview section, the amount of data at hand is limited and extremely imbalanced, causing our major concern. Hence, the models of choice must accurately address these issues to meet the desired outcome.

Due to the imbalances of the dataset, accuracy alone does not serve as a sufficient metric. To mitigate the influence of class imbalance on our model selection process, we will also examine the precision and recall of each model candidate to provide a more comprehensive understanding of model performance.

Furthermore, while the image classification model achieves decent outcomes, we will mainly focus on the classic machine learning models rather than deep learning models with more sophisticated architectures due to the limitation of image data quality and the lack of computational power.

4.2 Model Candidates

For this project, we proposed five classification models:

4.2.1 Logistics Regression

Logistic regression models the probability of a binary outcome, making it suitable for binary classification tasks like psoriasis prediction. Its simplicity is essential for the model's interpretability, and its low computational complexity expedites the training process. However, the logistic regression model has obvious weaknesses such as its assumption of the linear relationship between the logit and features.

4.2.2 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It's known for its versatility and robustness. The model overcomes non-linearity as it can capture complex non-linear relationships in the data while providing a comprehensive and interpretable feature importance measurement. Nonetheless, random forest requires much more computational power than logistic regression, making the training process much more complex.

4.2.3 XGBoost

Similar to the random forest, XGBoost is an ensemble learning method. However, it introduces a gradient-boosting algorithm into the training process, which further improves the model performance. Due to its advanced structure, XGBoost requires even more time in the training process than random forest and has even more hyperparameters to tune.

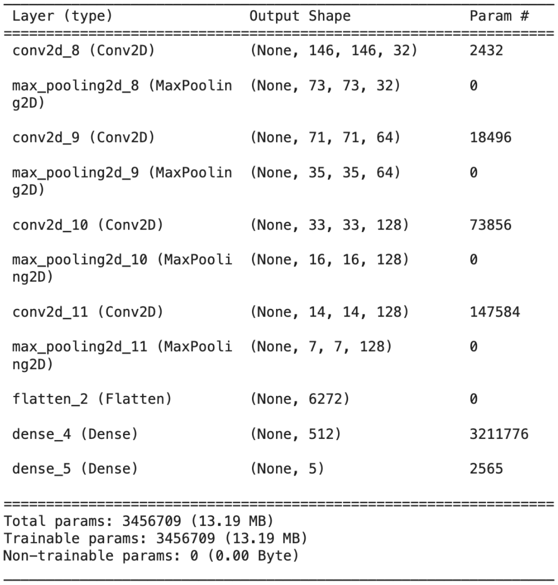
4.2.4 Support Vector Machine (SVM)

SVM is a strong choice for psoriasis prediction due to its effectiveness with high-dimensional and non-linear data. However, computational intensity and class imbalance require attention, and interpretability may need post hoc analysis.

4.2.5 Image Classification CNN

Convolutional Neural Network (CNN) is commonly used for image classification tasks. It relies on a "sliding window" mechanism to use filters and "slide" through the input images for parameter updates. At the end of the training process, the model will freeze the parameters in order to be able to take an unseen image and classify the model as either "healthy" or "psoriasis".

The image classifier model consists of 4 CNN layers followed by 2 dense layers with a total of over 3.4 million parameters. The model also required a large amount of data, thus image augmentation was introduced and quadrupled the original dataset, reaching a total of 800 images. The model summary is shown below:



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_8 (Conv2D)           (None, 146, 146, 32)      2432

 max_pooling2d_8 (MaxPoolin  (None, 73, 73, 32)        0
 g2D)

 conv2d_9 (Conv2D)           (None, 71, 71, 64)        18496

 max_pooling2d_9 (MaxPoolin  (None, 35, 35, 64)        0
 g2D)

 conv2d_10 (Conv2D)          (None, 33, 33, 128)       73856

 max_pooling2d_10 (MaxPooli  (None, 16, 16, 128)       0
 ng2D)

 conv2d_11 (Conv2D)          (None, 14, 14, 128)       147584

 max_pooling2d_11 (MaxPooli  (None, 7, 7, 128)         0
 ng2D)

 flatten_2 (Flatten)         (None, 6272)              0

 dense_4 (Dense)             (None, 512)               3211776

 dense_5 (Dense)             (None, 5)                 2565

=================================================================
Total params: 3456709 (13.19 MB)
Trainable params: 3456709 (13.19 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Table 4: Model Summary and Detailed Parameter Report

4.3 Hyperparameter Tuning

Logistics Regression: N/A

Random Forest: grid search on

i) Number of estimators
ii) Max depth of each decision tree
iii) Minimum samples required to split
iv) Minimum samples required in each leaf node.

XGboost: grid search on

i) Number of estimators
ii) Max depth of each decision tree
iii) Minimum sum of instance weight needed in a child node
iv) Learning rate

SVM: grid search on

i) Kernel (linear, sigmoid, polynomial)

CNN: fine-tuning on

i) Input image size
ii) batch size
iii) number of epoch
iv) number of layers

4.4 Data Splitting

The imbalanced nature of the dataset requires more than simply splitting the dataset into training and test subsets. For this project, we first decided to oversample the under-represented class, but when developing the model we discovered that using data augmentation for the under-represented class dramatically improved the model performance. Hence, we first augment the under-represented class (positive psoriasis labels) and use 20% of the data as a testing set. For the sake of being able to reproduce our result, random_state is set to 42.

5 Model Performance

5.1 Hyperparameters

We employed a 5-fold cross-validation strategy, which divided the dataset into five subsets, to find the optimal hyperparameters. We evaluated the best hyperparameters for each model.

Machine Learning Model	Best Hyperparameters Used
Logistic Regression	{‘C’: 0.1, ‘penalty’: ‘L2’, ‘solver’: ‘liblinear’}
Random Forest	{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 100}
XGboost	{'learning_rate': 0.1, 'max_depth': 6, 'min_child_weight': 1, 'n_estimators': 300}
Support Vector Machine	{'kernel': 'linear'}

Table 5: Best Hyperparameters

5.2 Evaluation Metrics

Since the dataset was imbalanced, we decided to use the following metrics to evaluate model performance. Precision measures the proportion of true positive predictions among all predicted positive instances. It assesses the model's ability to avoid false positives, which is crucial in a medical context. Recall measures the proportion of true positive predictions among all actual positive instances. In a medical context, it is important to obtain a high recall to avoid false negatives, minimizing the cases of psoriasis that are missed. F1-score is the harmonic mean of precision and recall, serving as an important metric when both false positives and false negatives are taken into consideration. The above three metrics are commonly used in imbalanced data classification model evaluation. Receiver Operating Characteristic Curve (ROC) can visualize the trade-off between true positive rate (TPR) and false positive rate (FPR) at different classification thresholds. Confusion Matrix provides insights into the actual counts of true positives, true negatives, false positives, and false negatives, which can be valuable for understanding the distribution of prediction errors such as Type I and Type II errors.

5.3 Model Results and Evaluations

Model	Precision	Recall	F1-score	Accuracy	AUC
Logistic Regression	0.92	0.63	0.73	0.63	0.62
Random Forest	0.98	0.98	0.98	0.98	1.00
XGboost	0.91	0.95	0.93	0.95	0.52
SVM	0.66	0.64	0.63	0.64	0.70

Table 6: Model Evaluation Metrics

Logistic Regression	Predicted Negative	Predicted Positive
Actual Negative	568	331
Actual Positive	20	23

Table 7: Logistic Regression Confusion Matrix

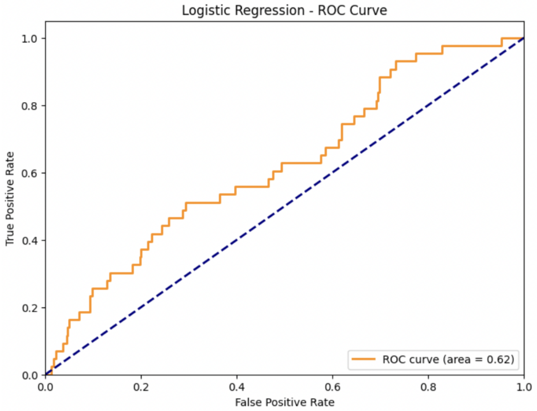


Figure 1: ROC Curve of Logistic Regression

Random forest	Predicted Negative	Predicted Positive
Actual Negative	603	0
Actual Positive	22	574

Table 8: Random Forest Confusion Matrix

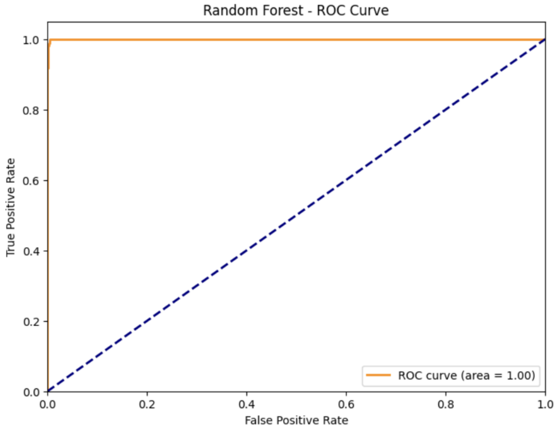


Figure 2: ROC Curve of Random Forest

XGboost	Predicted Negative	Predicted Positive
Actual Negative	891	8
Actual Positive	43	0

Table 9: XGBoost Confusion Matrix

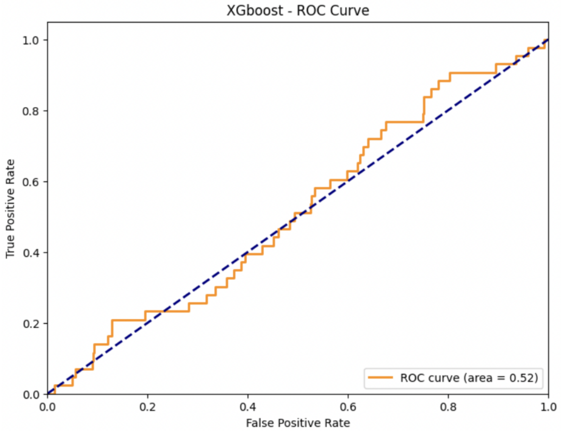


Figure 3: ROC Curve of XGBoost

SVM	Predicted Negative	Predicted Positive
Actual Negative	458	174
Actual Positive	477	689

Table 10: SVM Confusion Matrix

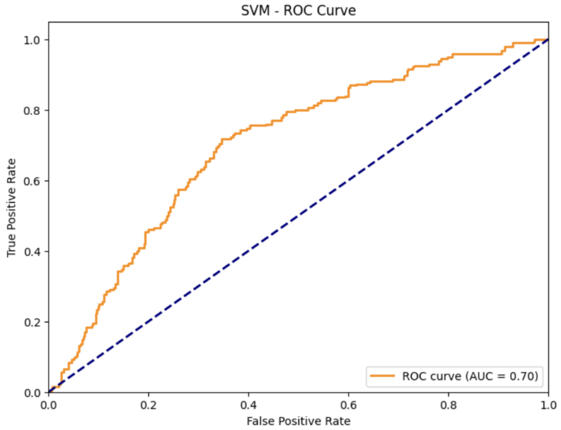


Figure 4: ROC Curve of SVM

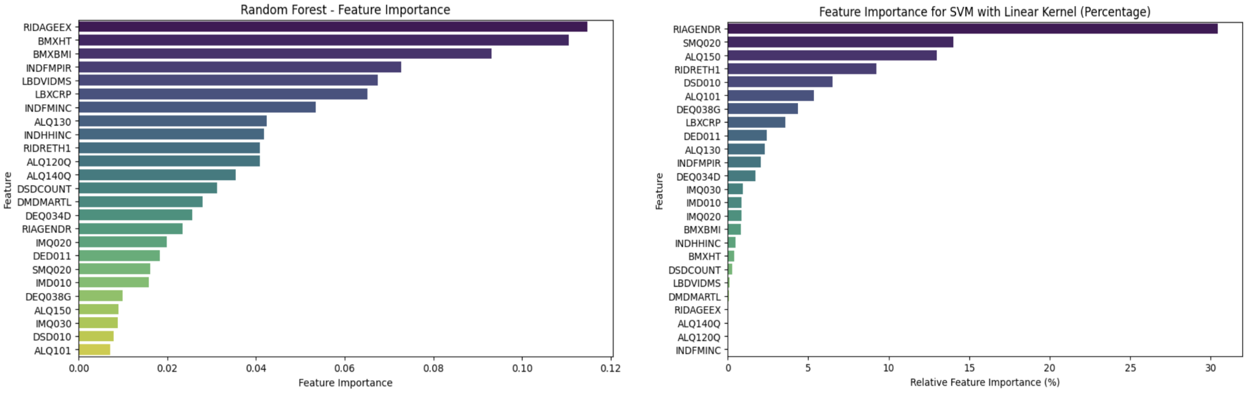


Figure 5: Feature Importance from Random Forest and SVM

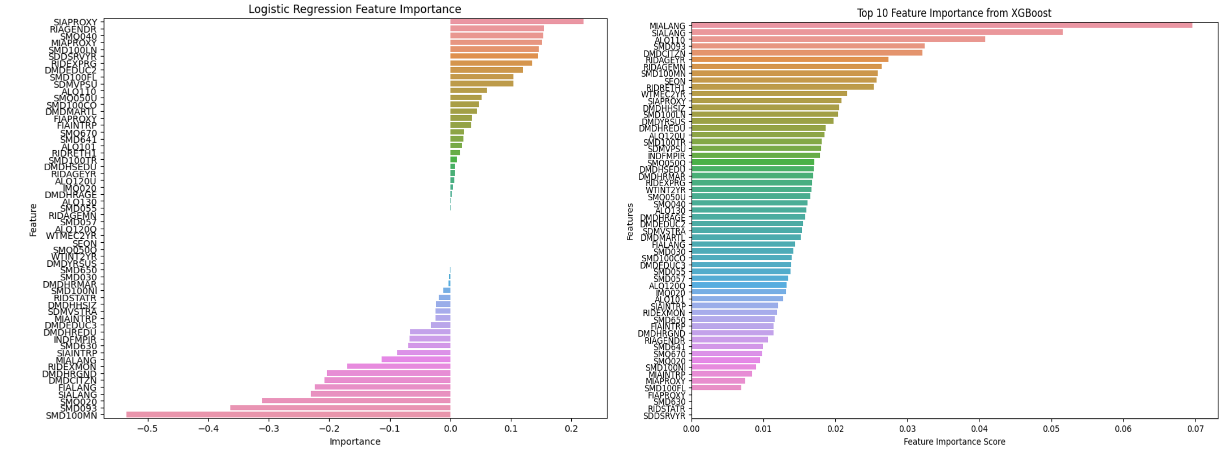


Figure 6: Feature Importance from Logistic Regression and XGBoost

Logistic Regression achieves moderate precision, recall, and F1-score, indicating a relatively balanced but not strong performance. AUC of 0.62 suggests that the Logistic Regression model performs slightly better than random guessing, but the ability to distinguish between classes is still weak. Random Forest demonstrates excellent precision, recall, and F1-score, indicating outstanding overall performance. The ROC curve, along with an AUC of 1.00, indicates that the Random Forest model achieves nearly flawless classification performance. The Random Forest model can correctly differentiate between the two classes, making it a highly accurate and reliable classifier for this dataset specifically. XGBoost achieves very high precision, recall, and F1-score, but a relatively low AUC of 0.52, suggesting strong overall performance with some room for improvement in differentiating between classes. SVM demonstrates moderate precision, recall, and F1-score, with a reasonable AUC of 0.70, indicating moderate but not exceptional performance.

We also examined feature importance from the models to check if the important features align with domain knowledge. The top 5 most important features of the Random Forest model are as follows. RIDAGEEX, Exam age in months; BMXHT, Standing height (cm); BMXBMI, Body mass index (kg/m^2); INDFMPIR, Family PIR; LBDVIDMS, Vitamin D (nmol/L). The top 5 most important features of the SVM model are as follows. RIAGENDR, Gender; SMQ020, Smoked at least 100 cigarettes in life; ALQ150, Ever have 5 or more drinks every day; RIDRETH1, Race/Ethnicity; DSD010, Any dietary supplements taken. The top 3 most important features from the Logistic Regression model are as follows. RIAGENDR, Gender; SMD100LN, Cigarette product length; RIDEXAGM, Age in months at the exam- 0 to 19 years. The top 3 most important features of the XGBoost model are as follows. RIDAGEYR, Age at Screening; RIDRETH1, Race/Ethnicity; WIMEC2YR, Full sample 2-year MEC exam weight.

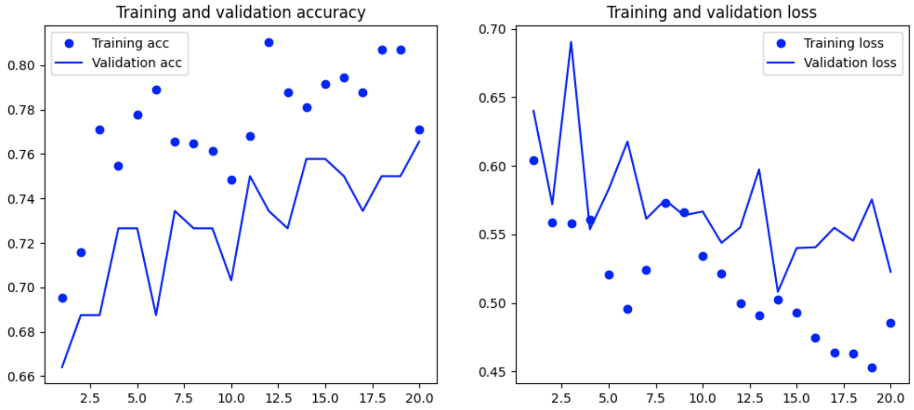


Figure 7: Accuracy and Loss from Training and Validation Set in Image Classifier

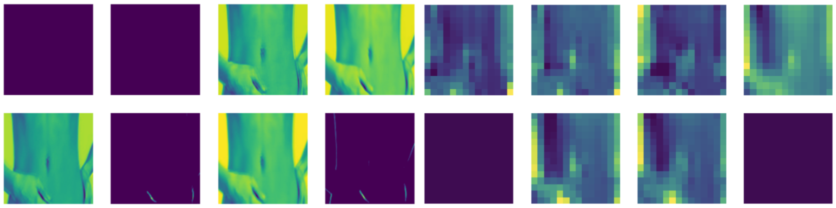


Figure 8: Feature Map of Image Classifier

In addition, the image classifier model was trained and tested on a dataset consisting of 800 images, which were augmented from an initial pool of 200 images to enrich the diversity of the dataset and improve model generalization. The model-tuning process involved adjusting hyperparameters, fine-tuning layers, and optimizing the learning rate. Through a systematic approach to hyperparameter tuning, the model's performance was significantly improved. The primary metric used to evaluate the model's performance was accuracy, which measures the proportion of correctly classified images out of the total. After rigorous tuning and training, the model achieved an accuracy of approximately 75%. This indicates that the model correctly classified roughly 75% of the skin images as either healthy or having psoriasis.

6 Conclusion

In conclusion, our comprehensive study on skin diseases, primarily focusing on psoriasis, represents a significant stride in the intersection of dermatology and machine learning. Utilizing a diverse array of methods, including Convolutional Neural Networks, Logistic Regression, XGBoost, Random Forest, and Support Vector Machines, we analyzed the NHANES dataset to uncover critical insights into skin conditions. This multifaceted approach not only enhanced the accuracy of our findings but also paved the way for innovative diagnostic and treatment strategies. Our research illustrates the immense potential of data science in revolutionizing healthcare, offering a new lens through which to view and tackle dermatological issues. The depth and breadth of our study signal a promising future for medical research, where data-driven techniques can lead to more personalized, effective, and comprehensive care for patients with skin diseases.

Building on our current research, the next stage could involve expanding our analytical scope to other health conditions present in the NHANES dataset, such as sunburn and micronutrient deficiencies. This would entail employing similar machine learning and statistical techniques to uncover patterns and correlations. Additionally, a more nuanced statistical modeling of the dataset could offer insights into variations across different demographics like race, gender, and geographic regions. Another promising avenue is comparing laboratory data with self-reported data to assess their reliability and coherence, potentially revealing critical aspects of patient-reported outcomes versus clinical measurements. This approach would not only deepen our understanding of various health conditions but also refine our methodologies in data analysis and health informatics.

Contributions

Shiyuan Xu: Model selection, model implementation, literature review, and progress report
Tianyu Han: Model selection, model implementation, literature review, and progress report
Siyu Shen: Literature review, data preparation, cleaning, exploratory data analysis, and progress report
Xinran Chen: Literature review, data preparation, cleaning, and exploratory data analysis, and progress report
Ying Hong: Literature review, model metrics, model evaluation, and progress report

References

Lim RK, Woo S, El Raheb S, Qureshi A, Cho E. Association of serum vitamin D levels and psoriasis severity: An Analysis of the US National Health and Nutrition Examination Survey. Presented at: Nutrition 2023; July 22-25, 2023; Boston, MA. Accessed Thursday, July 27, 2023.

Large study shows link between Vitamin D and psoriasis severity. News release. Newswise. July 17, 2023. Accessed July 27, 2023. https://www.newswise.com/faseb/large-study-shows-link-between-vitamin-d-and-psoriasis-severity/

Zhang R, Huang Q, Su G, Wei M, Cui Y, Zhou H, Song W, Di D, Liu J, Wang Q. Association between multiple vitamins and bone mineral density: a cross-sectional and population-based study in the NHANES from 2005 to 2006. BMC Musculoskelet Disord. 2023 Feb 10;24(1):113. doi: 10.1186/s12891-023-06202-6. PMID: 36765290

Cui A, Xiao P, Fan Z, Zeng Y, Wang H, Zhuang Y. Associations between vitamin E status and bone mineral density in children and adolescents aged 8-19 years: Evidence based on NHANES 2005-2006, 2017-2018. PLoS One. 2023 Mar 16;18(3):e0283127. doi: 10.1371/journal.pone.0283127. eCollection 2023. PMID: 36928218