ASSESSING THE IMPACT OF PRINCIPAL COMPONENT ANALYSIS ON ACCURATELY PREDICTING MELANOMA DIAGNOSIS APPLIED ON DIFFERENT CLASSIFICATION MODELS

by

JUAN CRISTOBAL OLMEDO RIVERA

THESIS

Submitted in partial fulfillment of the requirements for the
degree of Master of Science in Industrial Engineering at
The University of Texas at Arlington
December, 2019

Arlington, Texas

Supervising Committee:

Aera LeBoulluec, Supervising Professor
Paul Componation
Jaime Cantu

ABSTRACT

Assessing the impact of Principal Component Analysis on accurately predicting Melanoma diagnosis on different classification models

Juan Cristobal Olmedo Rivera, Master of Science in Industrial Engineering

The University of Texas at Arlington, 2019

Supervising Professor: Aera LeBoulluec

With huge amounts of data at our disposal in the medical field, mathematical models are built to diagnose diseases. This study focuses on melanoma because it's the type of skin cancer that accounts for most deaths, up to 7,230 in 2019 according to the American Cancer Society. The study focuses on the effectiveness on diagnosing melanoma and how Principal Component Analysis (PCA) impacts the performance of four models being assessed, which are: K Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). Each model evaluates the melanoma dataset before and after performing the PCA transformation. Results show that PCA does not impact performance in this case. Even though PCA does not improve performance, the modeled results achieve better results when compared to dermatologist and other algorithms.

DEDICATION

I dedicate my thesis work to my family. A special thanks to my parents, Juan and Jaqueline whose support and encouragement has helped me not just on my thesis but throughout my live. And to my brothers Andres and Fernando who were always by my side.

I also want to give special thanks to my grandma, Maria de Lourdes who supported me prior to start my master’s degree.

Table of Contents

List of Tables

List of Figures

Introduction

Now more than ever data is at our disposal to assist in the diagnosis of diseases. This study will focus on the effectiveness of diagnosing melanoma and how Principal Components Analysis impacts the performance of several models being assessed. This paper will focus on four machine learning and deep learning models which are as follows: k-nearest neighbor, logistic regression, support vector machines, and artificial neural networks.

The reason why there is a strong interest in using modeling techniques in diagnosing melanoma it's because is the most aggressive form of skin cancer and incidences continue to rise worldwide [7]. The American Cancer Society estimates in the 96,480 cases of melanoma for 2019 in the United States 7,230 people are expected to die [8]. But if melanoma is diagnosed early (early stage melanoma) it remains very treatable with a high long-term survival rate [9]. Also, melanoma is unique when compared to other cancers because detection is performed through visual inspection [9]. Epiluminescence microscopy is a method that was developed to assist dermatologists in the diagnostic process, increasing expert performance [10]. The test uses a dermatoscope which may or may not use an oil immersion in the pigmented skin lesion, and an image is studied to obtain a distinction of benign and malignant melanocytic lesions [9-11]. These dermoscopy images can be used and are being used in the development of models and systems to aid dermatologists in melanoma diagnosis [2-5]. These new systems take the images and perform: lesion segmentation, feature segmentation, and finally, classification [4, 5]. For this study the focus will be in the last part, classification. The objective is to refine the models above to obtain or exceed the performance of current practices and methods. The interest on Principal Components Analysis is that the data transformation will result in a smaller data set of uncorrelated variables [12], which may improve model performance.

Materials and Methods

1. Classification Models

1.1 K-Nearest Neighbor

The first model is k-nearest neighbor (KNN), a widely used algorithm. It basically involves two main steps: an initial step for building a classification model from the data, and a deducible step for applying the previous model to a new test of examples [13]. Unlike other models, KNN is both a lazy learner and non-parametric. Lazy learner's just stores the data and holds up until new examples need to be evaluated. Also KNN is a non-parametric model so it makes no assumption to the underlying data distribution [14]. The KNN classifier represents each observation as a data point in a d-dimensional space, where d is the number of features [14]. When given a new data point (test set), the proximity to the rest of the observations in the training set are computed, using a proximity measured [14]. Finally that test data point is classified depending on the number of closest neighbors determined by k [14]. In medicine, usually the function of KNN is to be a benchmark of other models [15,16]. That is the reason why KNN was picked for this study.

1.2 Logistic Regression

Logistic regression (LR) like linear regression produces the logit(Y) equation that explains the relationship between the dependent variable and the independent variables [17]. Unlike regression LR evaluates the outcome variable when it is binary or dichotomous [18]. In LR the general method of estimation is called maximum likelihood (MLE), which yields values for the unknown parameters that maximize the probability of obtaining the observed set of data [18]. The LR model outputs a number between 0 and 1 which are the estimated probabilities. If the output is greater or equal to 0.5 for a binary classification it will result in a class of 1; otherwise the class will be 0. In the medical field logistic regression has been widely used because of the ease of interpreting the impact of the parameters of the model (odds ratio) [19-21]. Also, logistic regression performs well on small data sets which is the case in this study. Another advantage of logistic regression is variable selection which improves the model by selecting the most significant features.

1.3 Support Vector Machines

Support vector machines (SVM) create a set of decision boundaries that separate the classes (in a binary classification problem) by maximizing the margin between instances. One essential innovation link to SVM is the “kernel trick”, which consists of observing that many algorithms can be written in terms of dot product, and enabling them to be rewritten to improve performance [22, 23]. The kernel trick is powerful for two main reasons: first, it enables the algorithm to learn models that are nonlinear meaning the algorithm views, the decision function as being linear in a different space; second, the kernel function implementation allows it to be substantially more computationally efficient [23]. A simple SVM can only separate data linearly but thanks to the kernel functions the algorithm can separate nonlinear data. For this specific study three versions of the SVM are compared: normal SVM (linear), SVM utilizing the polynomial kernel, and SVM utilizing the Gaussian kernel. In medicine these types of algorithms are becoming more popular [24-26]. SVM tends to perform well on small and medium size data sets and is a more powerful algorithm as it can deal with data sets that are not linearly separable.

1.4 Artificial Neural Networks

Artificial neural networks (ANN) usually outperforms other machine learning models on large data sets and complex problems. ANN is composed of neurons called linear threshold units (LTU). A single layer of LTUs compose a perceptron, and each neuron is connected to all the inputs [23]. “These chained structures are the most commonly used structures of neural networks” [23]. In this case the first structure is called the first layer, the next one second layer, and so on; the overall length of the chain gives the depth of the model [23] (deep learning name comes from). These layers, when the model is trained, do not display the desired output for each layer. That is why they are called hidden layers [23]. ANN like SVM has the capacity to model nonlinear functions giving it the chance to model more complex problems as stated before. Researches have been using ANN in medicine since the nineties [27, 28], and it is the interest of the study to see if it can outperform the other models and give a better prediction.

2. Data

The data set was obtained for this study came from the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) and Bioinformatics and Bioengineering Technical Committee (BBTC). The data set contains only 298 observations (small data set) and 17 features explaining positive and negative diagnosis of melanoma (binary classification: 1 results in a positive diagnosis, 0 results in a negative diagnosis). The following Table 1.1 contains the features used for modeling and predicting diagnosis.

Table 1.1

Features extracted for model prediction
Geometric features:
Solidity
Filled area
Equivalent diameter
Perimeter
Area over perimeter
Eccentricity
Euler number
Entropy
Imaging features:
Maximum red (RGB)
Maximum green (RGB)
Maximum blue (RGB)
Mean red (RGB)
Mean green (RGB)
Mean blue (RGB)
Mean standard deviation red (RGB)
Mean standard deviation green (RGB)
Mean standard deviation blue (RGB)

The features can be divided into two groups: imaging features, which use dermoscopy images consisting, in the red, green and blue (RGB) colors [29] and geometric features, which physically describe the skin lesion. The next step is to take a closer look at the features themselves. Figure 1 shows the histogram of each feature, including *diagnosis.*

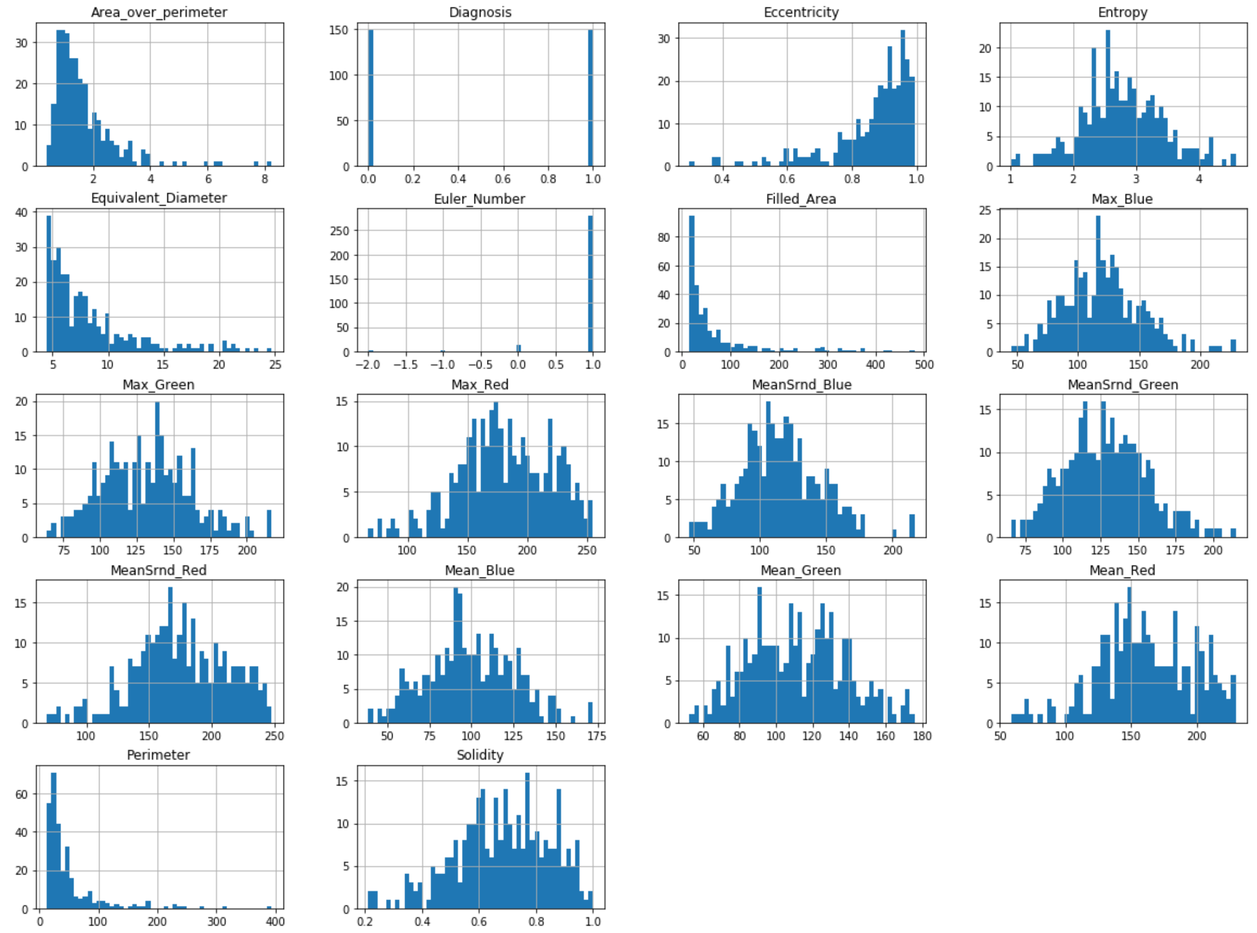


Figure 1: Histograms of data set features

The histograms represent the frequency distribution of each feature. Most of the features make (or follow) some kind of distribution except diagnosis and Euler number histograms. Both diagnosis and Euler number are dichotomous variables, meaning they represent a category or levels. When taking a closer view to diagnosis, it can be observed that the data set fortunately is completely balanced (reference Figure 2). Because the features belong to a certain group (imaging and geometric) there is a high chance of relation between them. Figure 3shows the correlation heat-map between features. It shows how the imaging features are highly correlated. Likewise the distance features (such as Perimeter, Equivalent Diameter, etc.) from the geometric group.

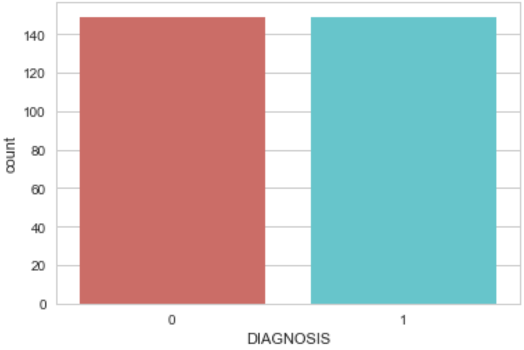


Figure 2: Diagnosis counts

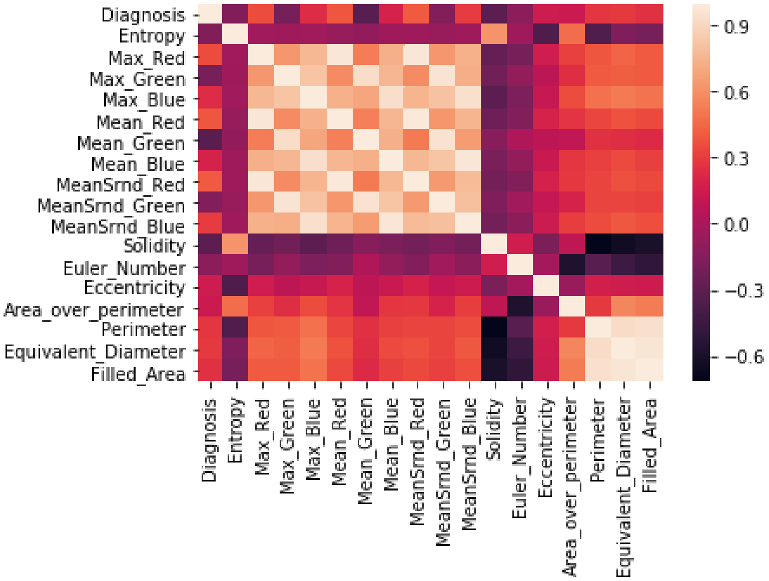


Figure 3: Correlation heat-map

Because some coefficients of determination values exceed 0.9; there is a strong correlation between features that may cause problems during modeling. Before analyzing, the data needs to be prepared. The data set is split randomly into a training set, which is composed of 80% of the observations and a test set which contains the remaining 20%. Then the training set is randomly split into a second training set (80%) and a validation test (20%). Then depending on the model being used, the data must go through a transformation for feature scaling. In this study two types of feature scaling are performed: min-max scaling and standardization. For min-max scaling the values are re-scaled so that all values are ranged from 0 to 1, while standardization subtracts the mean value and divides that result by its variance. Feature scaling is necessary because algorithms do not perform well when the numerical attributes are in different scales.

3. Modeling and Analysis

3.1 Modeling procedure

All models were analyzed in Python and followed the same steps: first the data must be scaled accordingly (min-max scaling or standardization). Then the model will be trained using the second training set and using the validation set to test its performance and experiment with the model's parameters. To get a better understanding how the data is organized, Figure 4 graphically shows the data split. Once a preliminary model is picked, that model is used on the training set by performing 10 fold cross validation [30]. The results of the model are compared to get an idea how the model is behaving, meaning if the model tends to be overfitting or underfitting the data [31]. The next step is to find the best model, applying 10 fold cross validation on the training set and using accuracy as the performance metric. Several models are evaluated over the specified parameters and the best model is chosen. That final model is used to evaluate the test set.

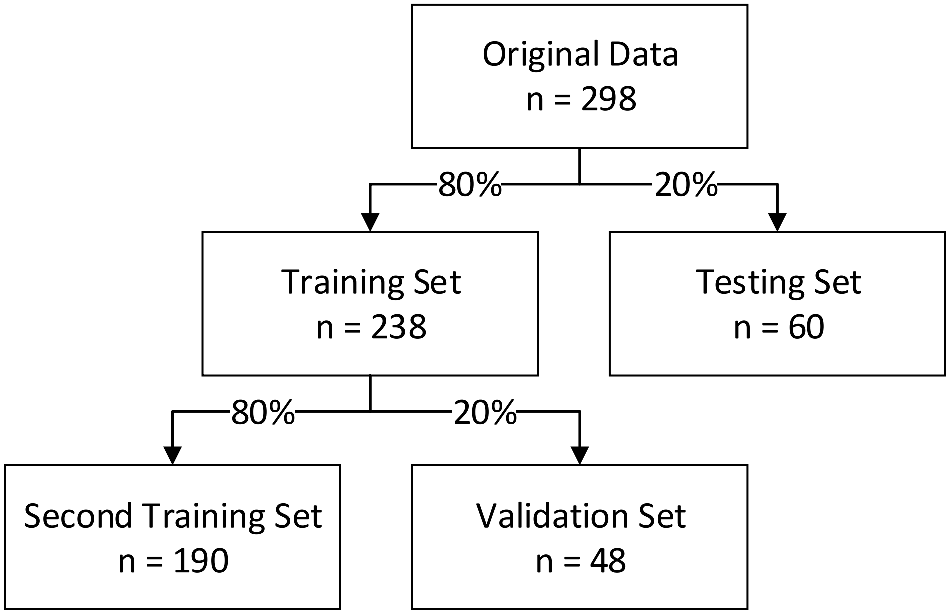


Figure 4: Melanoma Data Split

3.2 Initial Modeling

3.2.1 Initial KNN Modeling

To start modeling KNN a measure of distance needs to be picked. There is no optimal distance metric that can be used for all types of data sets [14]; for this model Euclidean distance is used. Also, the data must be standardized. Data standardization is necessary to avoid high variation of a specific feature to dominate the proximity measure. The parameter to determine is k, the number of neighbors. The initial k parameters for experimentation are the following: [3,6,9,12,15,18,21,24,27,30] in which k=9 gives the best performance.

3.2.2 Initial Logistic Regression Modeling

This LR model does not require the specification of any parameters but the model needs to be viable. By looking at the log likelihood ratio (LLR) p-value it can be determine if the model is feasible. In this case the LLR p-value is less than 0.5 making the LR model feasible. The preliminary model warns of possibly complete quasi-separation. When a model suffers from complete quasi-separation the model has high standard errors and generally but not always high coefficients [17]. The model does experience high standard errors corroborating complete quasi-separation, which may indicate problems in the data or analysis [17] being used. A problem with the data in LR can be multicollinearity, which results in large standard errors [17] (reference Figure 3). The next step is to delete insignificant features that do not contribute to the model. In this study backward elimination (BE) is used [32]. In regression when using BE or another type of variable elimination it is standard to use an alpha level of 0.05 to determine significance, but according to Lee & Koval 1997 [18] it can be too stringent when performing feature elimination using forward stepwise logistic regression. The same principle is followed when performing BE so the significance level used is 0.2.

3.2.3 Initial SVM Modeling

As explained before SVM is evaluated using three different kernels. The first one is linear, the second is polynomial and the third one is Gaussian. Like KNN, SVM is also sensitive to scaling so the data is standardized. In the linear SVM model, to make it more flexible, soft margin classification is used; therefore the parameter subject to change is C [33]. A small value of C tends to have a bigger margin causing more violations, while a large C will result in a smaller margin with fewer violations. The C parameters for experimentation are the following: [0.1,1,10,100] in which C=10 gives the best performance. The next SVM model uses the polynomial kernel trick, where first the degree of the polynomial needs to be specified [22, 33]. To avoid overfitting the model uses a 3rd degree polynomial kernel, and also like linear SVM, C needs to be specified. Two Cs are used [1,10] and the best model has a degree =3 and C=10. The final SVM model used the Gaussian kernel trick or "radial basis function" which has two parameters: gamma and C [22, 33]. The gamma parameter makes a ball-shape narrower around the class. As gamma grows bigger as a result each instance range of influence is smaller making the decision boundary irregular twisting around individual instances; on the other hand if gamma is small that bell-shaped curve is wider and the decision boundary is smoother giving greater influence to each individual instance. The values for C used are [1,10] and for gamma are [0.1,5,100] and the best model is C=10 and gamma=0.1.

3.2.4 Initial ANN Modeling

For the ANN model the parameters to determine are the number of neurons and the number of hidden layers. This artificial neural network or multi-layer perceptron employs the backpropagation training algorithm [34]. Alongside the backpropagation algorithm in our model the ReLU function is used as an activation function. The ReLU function was chosen because it speeds up convergence compared to other functions such as hyperbolic tangent function or sigmoid function [35]. For the initial model the ANN will have only one hidden layer while a different number of neurons are experimented with to obtain the initial model. The number of neurons tried are: [10,100,200,300,400]. The model with the best performance has number of neurons=200 with one hidden layer. Different strategies will be applied to obtain the final ANN model but for now the initial model is hidden layers=1 and number of neurons=200.

3.3 Initial Results

Table 1.2

Results: Initial model testing Validation Set

	KNN	LR	SVM	SVM Poly.	SVM Gauss.	ANN
Avg. accuracy	0.9375	0.8958	0.9792	1.0000	1.0000	0.9583
Sensitivity	0.9600	0.9600	1.0000	1.0000	1.0000	0.9600
Specificity	0.9130	0.8261	0.9565	1.0000	1.0000	0.9565

Table 1.3

Results: Initial model 10 fold cross validation Training Set

	KNN	LR	SVM	SVM Poly.	SVM Gauss.	ANN
Avg. accuracy	0.9286	0.9327	0.9412	0.9621	0.9416	0.9201
Sensitivity	0.9397	0.9310	0.9310	0.9569	0.9483	0.8966
Specificity	0.9180	0.9344	0.9426	0.9672	0.9344	0.9426

Looking at Table 1.2 and Table 1.3 the preliminary model results can be observed. Table 1.2 show results of training the models using the Second Training Set and testing them on the Validation Set. Table 1.3 evaluates the same preliminary algorithms on the Training Set employing 10 fold cross validation. The performance of the models dropped when using k fold cross validation because the models tend to be overfitting the data, except for logistic regression.

3.4 Final Model

From the preliminary results the parameters picked are revisited to reduce the range of the search to refine the model. The final model will evaluate the new parameters through a "grid-search"; meaning that all new parameters will be modeled using the Training Set (10 fold cross validation) and will be evaluated by the model performance. Whichever combination of parameters yields the highest accuracy will result in the final model to evaluate the Test Set.

3.4.1 KNN Final Model

For the final KNN model the new parameters are similar, just the range was reduced into [3, 7, 9, 13, 15, 17] after performing the grid-search the best model is with k=7

3.4.2 Logistic Regression Final Model

For the final LR model the best performance was obtained by performing backward elimination and the following features were dropped from the model [Solidity, Mean standard deviation green, Maximum green, Maximum blue, Euler number, Perimeter, Mean standard deviation red].

3.4.3 SVM Final Model

For linear SVM the new parameters for C are [0.001, 0.1, 1, 10,100]. For Polynomial SVM the new parameters for C are [0.1,1,10,50,100] and are tested in different degrees = [3,4,5,6,7]. For Gaussian SVM the new C values are [0.1,1,10,50,100] and are tested in the following gamma values [0.001, 0.01, 1, 5, 10]. The overall best performing model its linear SVM C=10.

3.4.4 ANN Final Model

ANN has great flexibility to deal with different kinds of problems and datasets but because of that flexibility there is one disadvantage which is adjusting all the different parameters to maximize performance. The preliminary model started with only one hidden layer and number of neurons=200. To find the best ANN model two strategies were followed. The first strategy tries a model with one hidden layer and increases it one by one, trying the same number of neurons per hidden layer until a “best” model is found [33]. The second strategy developed by Vincent Vanhoucke is to use a greater number of hidden layers with greater number neurons that the model would need and use early stopping [33, 36] which will stop the model when is not improving. Several models are tried and the first strategy outputs the best model resulting with: hidden layers=4 and number of neurons=200.

3.5 PCA Transformation and Analysis

Principal Component Analysis (PCA) is a technique that transforms the variables in a data set which could result in dimensionality reduction with a set of uncorrelated variables [12]. This data analytic technique attains a linear transformation of correlated variables and returns new variables named principal components (pcs) which are the new uncorrelated variables. This pcs will have mean zero and variance li, ith characteristic root (eigenvalues) [12].

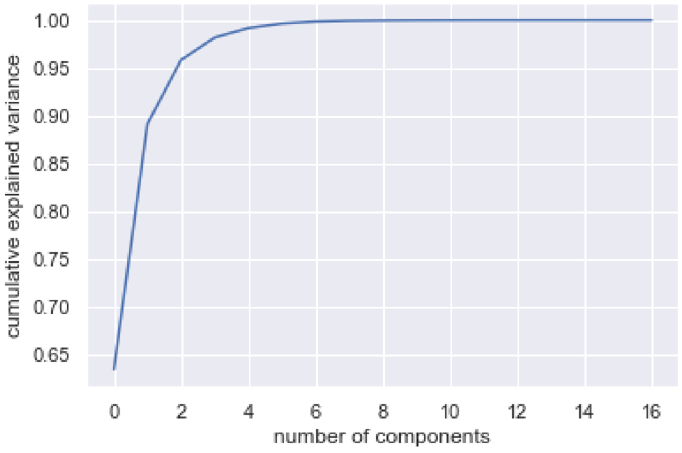


Figure 5: PCA Cumulative Explained Variance

Before performing the PCA transformation the data set must be scaled because the variables are recorded in different units [12]. After going through the transformation, the data is uncorrelated a test it's picked to determine which pcs are significant. In this study the SCREE test is used, this test plots all eigenvalues of the covariance matrix vs. pcs number (eigenvalue number). The "scree being defined as the rubble at the bottom of the cliff, the retain pcs are the cliff and the deleted ones are the rubble" [12]. What it means is that there are few pcs that are significant (which are the cliff) and break away from the rest of the pcs (rubble). It is suggested to use the most significant pcs and the first one of the latter group [12]. Using the previous steps the melanoma dataset it's scaled and transformed by performing PCA, Figure 5 shows the cumulative explained variance by the principal components. On Figure 5 it can be observed that roughly the first 4 to 5 components account for most of the explained variation.

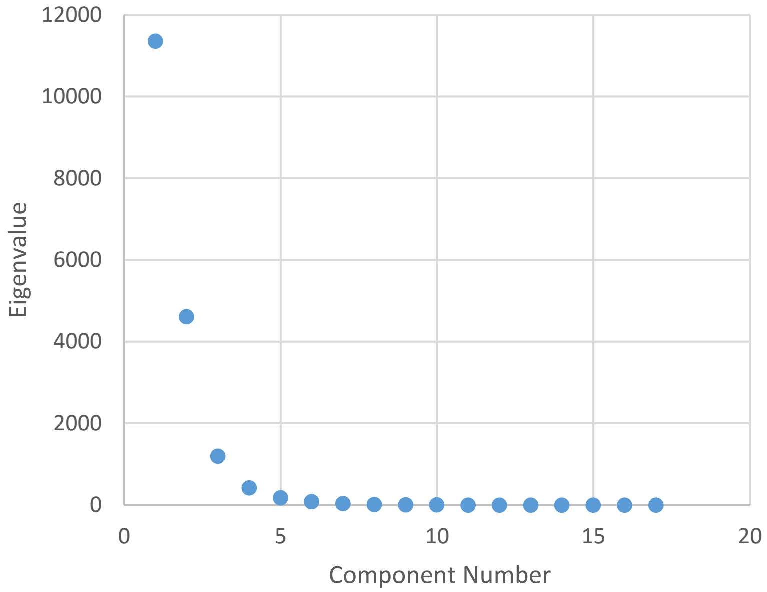


Figure 6: SCREE Test

Figure 6 is the SCREE test plot which shows the first three pcs as the "cliff" while the rest are the "rubble". Following the SCREE Test recommendations the first four pcs are used to model the transformed data set. The modeling procedure used in section 3.1 is applied the same way to the new PCA transformed data set. Because the procedure was explained in detail previously only the final PCA models are shown. The final model for KNN is having a k=3. For LR the model followed the same trend as the best model results from backward elimination. For SVM after comparing all three models (linear, polynomial, and Gaussian) the best model is the linear SVM with a C = 0.1, it's not a surprise as the data went through a linear transformation. For the final model, ANN the previous two strategies were followed; the same activation function (ReLU) was used and the best model has: hidden layer = 1 and number of neurons=30

Results

1. Final Results

The final models are trained on the Training Set and finally evaluate the Test set, Table 1.4 shows the results of all four models before going through the PCA transformation.

Table 1.4

Results: Final models evaluating Test Set

	KNN	LR ("BE")	SVM	ANN
Accuracy	0.7843	0.9086	0.8476	0.8443
Sensitivity	0.8148	0.9090	0.8485	0.8485
Specificity	0.7576	0.8889	0.8148	0.8148
AUC Score	0.8956	0.9845	0.9046	0.9371

The best performing model overall is LR after performing backward elimination (reference Figure 7). It has the highest accuracy in classifying positive and negative diagnosis of melanoma. The other important metric is sensitivity, it tells us the true positive rate of correctly diagnosing melanoma. In other words, correctly identifying melanoma cases which are actually suffering of melanoma. The focus is to not just finding the model that gives the highest accuracy but maximizes sensitivity. Whereas specificity tells us the true negative rate, identifying a negative melanoma diagnosis for a case that does not suffer of melanoma. If these models were to be fully implemented in practice, it is preferable to maximize sensitivity at the cost of specificity. The final metric is the AUC score, which is the area under the ROC (receiver operating characteristic) curve which compares the false positive rate vs. the true positive rate. SVM and

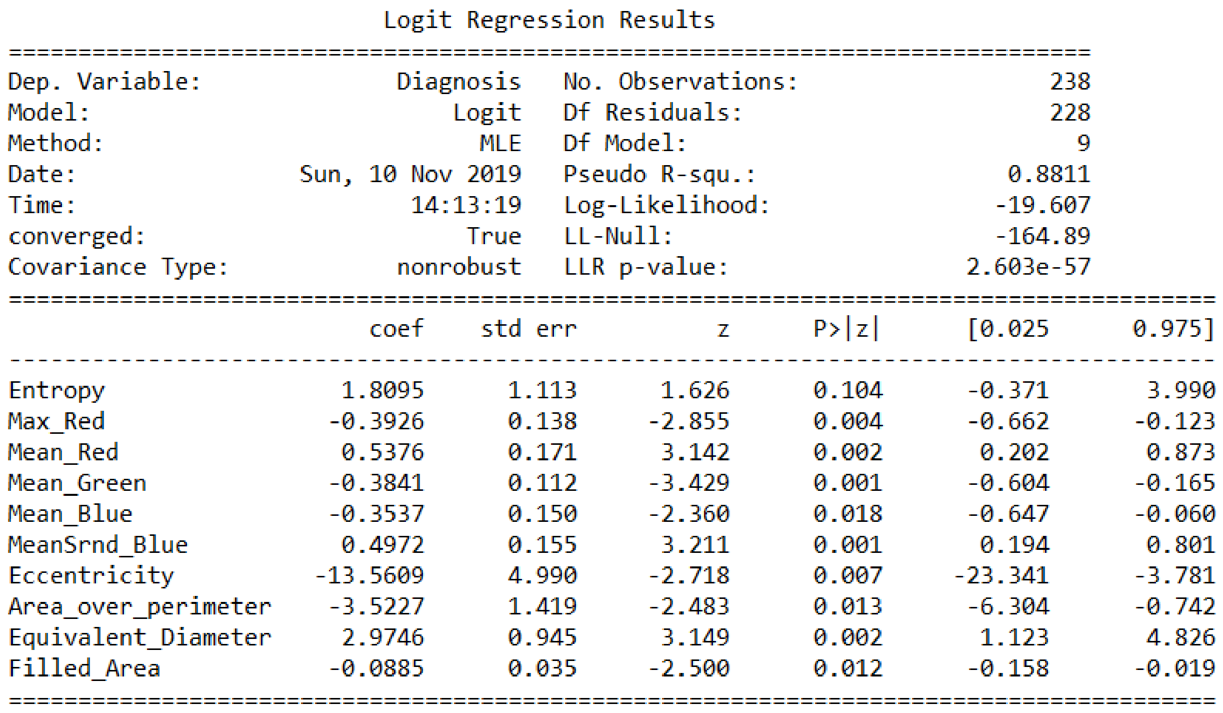
ANN have very similar results, while KNN is the worst performing model. Table 1.5 contains the results of the final models after performing the PCA transformation.

Table 1.5

Results: Final PCA models evaluating Test Set

	KNN	LR (“BE”)	SVM	ANN
Accuracy	0.8105	0.8838	0.8505	0.8329
Sensitivity	0.8182	0.8485	0.8485	0.8182
Specificity	0.8148	0.9259	0.8519	0.8519
AUC Score	0.8956	0.9405	0.9506	0.9125

Looking at the results, PCA did not improved LR. Also, it did marginally affect SVM and ANN. The only model that slightly improved thanks to PCA was KNN. The only marginal trend resulted from the PCA transformation is an increase on specificity across the models, resulting in the opposite direction of what is desired (increase sensitivity at the cost of specificity).



```
                           Logit Regression Results
==============================================================================
Dep. Variable:              Diagnosis   No. Observations:                  238
Model:                          Logit   Df Residuals:                      228
Method:                           MLE   Df Model:                            9
Date:                Sun, 10 Nov 2019   Pseudo R-squ.:                  0.8811
Time:                        14:13:19   Log-Likelihood:                -19.607
converged:                       True   LL-Null:                       -164.89
Covariance Type:            nonrobust   LLR p-value:                 2.603e-57
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Entropy                 1.8095      1.113      1.626      0.104      -0.371       3.990
Max_Red                -0.3926      0.138     -2.855      0.004      -0.662      -0.123
Mean_Red                0.5376      0.171      3.142      0.002       0.202       0.873
Mean_Green             -0.3841      0.112     -3.429      0.001      -0.604      -0.165
Mean_Blue              -0.3537      0.150     -2.360      0.018      -0.647      -0.060
MeanSrnd_Blue           0.4972      0.155      3.211      0.001       0.194       0.801
Eccentricity          -13.5609      4.990     -2.718      0.007     -23.341      -3.781
Area_over_perimeter    -3.5227      1.419     -2.483      0.013      -6.304      -0.742
Equivalent_Diameter     2.9746      0.945      3.149      0.002       1.123       4.826
Filled_Area            -0.0885      0.035     -2.500      0.012      -0.158      -0.019
=======================================================================================
```

Figure 7: Logistic Regression Final Model

Discussion

Melanoma is the most aggressive form of skin cancer [7], it is highly desirable to improve the accuracy in which dermatologists predict a positive diagnosis. In preliminary studies it was reported that dermatologist's accuracy in making a positive diagnosis was 64% [10]. As time and diagnosis improved a study using computerized database of skin lesions showed that dermatologists partaking in the study reached a sensitivity of 80.8% and a specificity of 99.2% [10]. Due to improved performance of models and algorithms and the use of dermoscopy images for melanoma diagnosis; melanoma diagnosis is experiencing a new interest in using models to aid with diagnosis. More recently a study asked dermatologists to classify melanoma just by looking at the dermoscopy image, the result showed an average sensitivity of 82%, an average specificity of 59% and an average AUC Score of 0.71 [37]. Even though these studies evaluate different lesions, both of them reach similar sensitivity levels. The second study also took those images and used three machine learning methods for classification (greedy ensemble fusion, linear binary SVM, and SVM using histogram intersection kernel) and compared them to the top 5 performance algorithms from 25 participants. The best performance algorithm was the SVM model using the histogram intersection kernel with a sensitivity of 70% and a specificity of 88% [37]. The LR model of this study performed with a sensitivity of 90% and a specificity of 89%. Even though the LR model greatly surpassed the SVM model they cannot be compared directly as one study modeled the data after going through lesion segmentation, feature segmentation before classification (this study concentrated on the latter part: classification) while the SVM model skips those steps and directly predicts diagnosis from the images. In another study a convolutional neural network was used to predict melanoma diagnosis directly from dermoscopy images and resulted with a sensitivity of 95%, a specificity of 80% [2]. Further research is needed to determine which strategy yields the best result: directly predicting diagnosis from the images or perform lesion segmentation, feature segmentation and classification [4, 5].

Study 1: Grin et al. 1990
Study 2: Marchetti et al. 2016
Study 3: Haenssle et al. 2018

Table 1.6: Study discussion comparison

Study	Sensitivity	Specificity
Study 1: Dermatologists	80.80%	99.20%
Study 2: Dermatologists	82%	59%
Study 2: SVM histogram intersection	70%	88%
Study 3: Convolutional Neural Networks	95%	80%
This Study: Logistic Regression (BE)	90%	89%

Conclusion

The preliminary results showed that KNN, SVM and ANN tended to be overfitting the data and in the case of LR the high correlation may pose issues in the model. The hypothesis of this study is to use PCA to transform the data resulting in a new uncorrelated data set which also reduces the dimensionality; resulting in a less complex model and improving the model's performance. The final results shows that PCA marginally improves KNN which was the worst performer. It is improved because KNN suffers the "curse of high dimensionality" so the PCA model with 4 features performed better compared to the original model with 17. ANN and SVM have insignificant changes, basically performing at the same rate. The original dimensionality of the model is not high enough to cause issues to the ANN and SVM models therefore a reduction of dimensionality did not improved the models. Finally the best model, LR is not improved by PCA. In the preliminary LR model there were warnings of multi-collinearity but it was not high enough to degrade the final LR model. Overall PCA does not have a significant impact in the models, other techniques and transformations have to be researched to see if they impact and improve model performance in the melanoma problem.

References

[1] Dreiseitl, Stephan, Lucila Ohno-Machado, Harald Kittler, Staal Vinterbo, Holger Billhardt, and Michael Binder. “A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions.” *Journal of Biomedical Informatics* 34, no. 1 (February 2001): 28–36. https://doi.org/10.1006/jbin.2001.1004.

[2] Haenssle, H A, C Fink, R Schneiderbauer, F Toberer, T Buhl, A Blum, A Kalloo, et al. “Man against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists.” *Annals of Oncology* 29, no. 8 (August 1, 2018): 1836–42. https://doi.org/10.1093/annonc/mdy166.

[3] Sboner, Andrea, Claudio Eccher, Enrico Blanzieri, Paolo Bauer, Mario Cristofolini, Giuseppe Zumiani, and Stefano Forti. “A Multiple Classifier System for Early Melanoma Diagnosis.” *Artificial Intelligence in Medicine* 27 (2003): 29–44.

[4] Mishra, Nabin K, and M Emre Celebi. “An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning.” *Cornell University*, January 28, 2016, 15.

[5] Gautam, Diwakar, Mushtaq Ahmed, Yogesh Kumar Meena, and Ahtesham Ul Haq. “Machine Learning-Based Diagnosis of Melanoma Using Macro Images: Machine Learning-Based Diagnosis of Melanoma Using Macro Images.” *International Journal for Numerical Methods in Biomedical Engineering* 34, no. 5 (May 2018): e2953. https://doi.org/10.1002/cnm.2953.

[6] Gilmore, Stephen, Rainer Hofmann-Wellenhof, and H. Peter Soyer. “A Support Vector Machine for Decision Support in Melanoma Recognition: Support Vector Machine and Melanoma.” *Experimental Dermatology* 19, no. 9 (September 2010): 830–35. https://doi.org/10.1111/j.1600-0625.2010.01112.x.

[7] *Melanoma*. Cham: Springer, 2016.

[8] “American Cancer Society | Cancer Facts & Statistics.” American Cancer Society | Cancer Facts & Statistics. Accessed October 7, 2018. http://cancerstatisticscenter.cancer.org/.

[9] O’Neill, Conor H., and Charles R. Scoggins. “Melanoma.” *Journal of Surgical Oncology* 120, no. 5 (October 2019): 873–81. https://doi.org/10.1002/jso.25604.

[10] Grin, Caron M., MD, Alfred W. Kopf MD, Bruce Welkovich MD, Robert S. Bart MD, and Marcia J. Levenstein. “Accuracy in the Clinical Diagnosis of Malignant Melanoma.” *Archives of Dermatology* 126, no. 6 (June 1990): 763–66.

[11] Pehamberger, Hubert, Michael Binder, Andreas Steiner, and Klaus Wolff. “In Vivo Epiluminescence Microscopy: Improvement of Early Diagnosis of Melanoma.” *Journal of Investigative Dermatology* 100, no. 3 (March 1993): S356–62. https://doi.org/10.1038/jid.1993.63.

[12] Jackson, J. Edward. *A User's Guide to Principal Components*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley, 1991.

[13] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson, 2005.

[14] Prasath, V. B. Surya, Haneen Arafat Abu Alfeilat, Ahmad B. A. Hassanat, Omar Lasassmeh, Ahmad S. Tarawneh, Mahmoud Bashir Alhasanat, and Hamzeh S. Eyal Salman. "Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier -- A Review." *Big Data*, August 14, 2019, big.2018.0175. https://doi.org/10.1089/big.2018.0175.

[15] Chen, Chin-Hsing, Wen-Tzeng Huang, Tan-Hsu Tan, Cheng-Chun Chang, and Yuan-Jen Chang. "Using K-Nearest Neighbor Classification to Diagnose Abnormal Lung Sounds." *Sensors* 15, no. 6 (June 4, 2015): 13132–58. https://doi.org/10.3390/s150613132.

[16] Murugan, A., S.Anu H. Nair, and K. P. Sanal Kumar. "Detection of Skin Cancer Using SVM, Random Forest and KNN Classifiers." *Journal of Medical Systems* 43, no. 8 (August 2019): 269. https://doi.org/10.1007/s10916-019-1400-8.

[17] Menard, Scott. *Applied Logistic Regression Analysis*. Second edition. Quantitative Applications in the Social Sciences, 2001

[18] Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. 3rd ed. Wiley Series in Probability and Statistics. Chicester: Wiley, 2013.

[19] Pecori, Biagio, Secondo Lastoria, Corradina Caracò, Marco Celentani, Fabiana Tatangelo, Antonio Avallone, Daniela Rega, et al. "Sequential PET/CT with [18F]-FDG Predicts Pathological Tumor Response to Preoperative Short Course Radiotherapy with Delayed Surgery in Patients with Locally Advanced Rectal Cancer Using Logistic Regression Analysis." Edited by Daniele Santini. *PLOS ONE* 12, no. 1 (January 6, 2017): e0169462. https://doi.org/10.1371/journal.pone.0169462.

[20] Wu, Mark, Satheesh Krishna, Rebecca E. Thornhill, Trevor A. Flood, Matthew D.F. McInnes, and Nicola Schieda. "Transition Zone Prostate Cancer: Logistic Regression and Machine-Learning Models of Quantitative ADC, Shape and Texture Features Are Highly Accurate for Diagnosis: Machine-Learning Diagnosis of PZ PCa." *Journal of Magnetic Resonance Imaging* 50, no. 3 (September 2019): 940–50. https://doi.org/10.1002/jmri.26674.

[21] Li, Hongxia, Xiaoyan Liang, Xuebing Qin, Shaohua Cai, and Senyang Yu. "Association of Matrix Metalloproteinase Family Gene Polymorphisms with Lung Cancer Risk: Logistic Regression and Generalized Odds of Published Data." *Scientific Reports* 5, no. 1 (September 2015): 10056. https://doi.org/10.1038/srep10056.

[22] Steinwart, Ingo, and Andreas Christmann. *Support Vector Machines*. 1st ed. Information Science and Statistics. New York: Springer, 2008.

[23] Googfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Massachusetts Institute of Technology, 2016.

[24] Ehrentraut, Claudia, Markus Ekholm, Hideyuki Tanushi, Jörg Tiedemann, and Hercules Dalianis. "Detecting Hospital-Acquired Infections: A Document Classification Approach Using Support Vector Machines and Gradient Tree Boosting." *Health Informatics Journal* 24, no. 1 (March 2018): 24–42. https://doi.org/10.1177/1460458216656471.

[25] Wang, Hui, and Gang Huang. "Application of Support Vector Machine in Cancer Diagnosis." *Medical Oncology* 28, no. S1 (December 2011): 613–18. https://doi.org/10.1007/s12032-010-9663-4.

[26] Vidić, Igor, Liv Egnell, Neil P. Jerome, Jose R. Teruel, Torill E. Sjøbakk, Agnes Østlie, Hans E. Fjøsne, Tone F. Bathen, and Pål Erik Goa. "Support Vector Machine for Breast Cancer Classification Using Diffusion-Weighted MRI Histogram Features: Preliminary Study: Machine Learning in DWI of Breast Cancer." *Journal of Magnetic Resonance Imaging* 47, no. 5 (May 2018): 1205–16. https://doi.org/10.1002/jmri.25873.

[27] Er, Orhan, Nejat Yumusak, and Feyzullah Temurtas. "Chest Diseases Diagnosis Using Artificial Neural Networks." *Expert Systems with Applications* 37, no. 12 (December 2010): 7648–55. https://doi.org/10.1016/j.eswa.2010.04.078.

[28] Moghimi, Fatemeh Hoda, and Nilmini Wickramasinghe. "Artificial Neural Network Excellence to Facilitate Lean Thinking Adoption in Healthcare Contexts." In *Lean Thinking for Healthcare*, edited by Nilmini Wickramasinghe, Latif Al-Hakim, Chris Gonzalez, and Joseph Tan, 13–27. New York, NY: Springer New York, 2014. https://doi.org/10.1007/978-1-4614-8036-5_2.

[29] Stanley, R. Joe, William V. Stoecker, and Randy H. Moss. "A Relative Color Approach to Color Discrimination for Malignant Melanoma Detection in Dermoscopy Images." *Skin Research and Technology : Official Journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI)* 13, no. 1 (February 2007): 62–72. https://doi.org/10.1111/j.1600-0846.2007.00192.x.

[30] Rodriguez, J.D., A. Perez, and J.A. Lozano. "Sensitivity Analysis of K-Fold Cross Validation in Prediction Error Estimation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, no. 3 (March 2010): 569–75. https://doi.org/10.1109/TPAMI.2009.187.

[31] Aalst, W. M. P. van der, V. Rubin, H. M. W. Verbeek, B. F. van Dongen, E. Kindler, and C. W. Günther. "Process Mining: A Two-Step Approach to Balance between Underfitting and Overfitting." *Software & Systems Modeling* 9, no. 1 (January 2010): 87–111. https://doi.org/10.1007/s10270-008-0106-z.

[32] Bursac, Zoran, C Heath Gauss, David Keith Williams, and David W Hosmer. "Purposeful Selection of Variables in Logistic Regression." *Source Code for Biology and Medicine* 3, no. 1 (2008): 17. https://doi.org/10.1186/1751-0473-3-17.

[33] Geron, Aurelien. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly, 2017.

[34] Hinton, E. “Learning Internal Representations by Error Propagation,” n.d., 49.

[35] Shakiba, Fatemeh M. “CMOS-Based Implementation of Hyperbolic Tangent Activation Function for Artificial Neural Network.” Southern Illinois University in Carbondale, 2018.

[36] Shao, Yang, Gregory N. Taff, and Stephen J. Walsh. “Comparison of Early Stopping Criteria for Neural-Network-Based Subpixel Classification.” *IEEE Geoscience and Remote Sensing Letters* 8, no. 1 (January 2011): 113–17. https://doi.org/10.1109/LGRS.2010.2052782.

[37] Marchetti, Michael A., Noel C.F. Codella, Stephen W. Dusza, David A. Gutman, Brian Helba, Aadi Kalloo, Nabin Mishra, et al. “Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging Challenge: Comparison of the Accuracy of Computer Algorithms to Dermatologists for the Diagnosis of Melanoma from Dermoscopic Images.” *Journal of the American Academy of Dermatology* 78, no. 2 (February 2018): 270-277.e1. https://doi.org/10.1016/j.jaad.2017.08.016.