Applied Machine Learning Model for Improved Diagnosis of Melanoma

WILL LAFORGE, BOB OWENS, and KYLE SHANNON, University of Tennessee, USA

Skin cancer is the most common form of cancer in the United States, with about 4.3 million adults receiving treatment each year. Reported global skin cancer occurrences are increasing each year, necessitating the need for more advances in modern medical detection methods. Machine learning can be implemented as a viable tool to aid in detecting forms of skin cancer. In this paper we will show how medical imaging, gathered from publicly held data sets can be utilized to build a machine learning pipeline that ultimately allows successful classification of skin cancer. This paper will also discuss the development of our machine learning pipeline, including data collection, pre-processing, feature extraction, and implementation of our machine learning model. Our group will show how identify features of melanoma such as uniformity, perimeter irregularity, color variance, and size can be used to train a machine learning model to effectively determine if a sample is showing skin cancer or not. The conclusion of this work will aid in demonstrating the viability of machine learning as an effective means for medical detection.

1 INTRODUCTION

Recent advancements have facilitated early detection for skin cancer allowing effective treatment and improving survivability rates, however the challenge of educating the populace on identify the early signs of skin cancer remains. Melanoma, though highly curable when detected early, can be a very dangerous form of skin cancer. Delayed detection of melanoma decreases the overall survival rate from 99% for cases diagnosed at a localized stage to 30% for distant-stage disease.[Can 2022]. Utilizing advancements in machine learning and the proliferation of smartphone access globally, this project's goal is to create a tool that accurately classifies potential melanoma with some degree of accuracy, increases melanoma literacy, and provides a means for affordable melanoma detection.

Authors' address: Will LaForge, wlaforge@vols.utk.edu; Bob Owens, rowens24@vols.utk.edu; Kyle Shannon, kshanno5@vols.utk.edu, University of Tennessee, Knoxville, TN, USA.
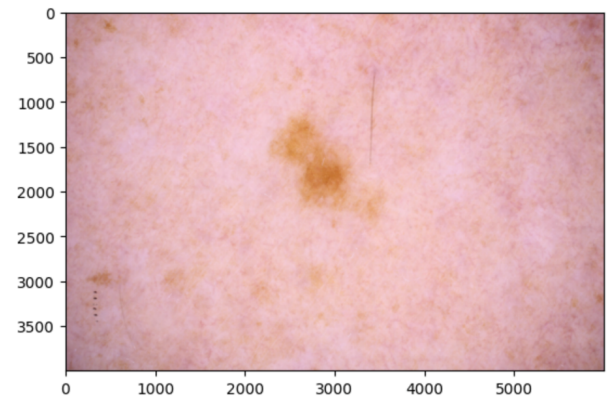
Fig. 1. Example Original Image

2 DATA

Numerous studies have been conducted producing thousands of images to aid in the accurate identification of melanoma. One such study had several iterations including a recent 2020 data set, which produced over 33,000 images with embedded metadata for training purposes and nearly 11,000 images with metadata for test data. [Rotemberg 2021] The imaging data is available in DICOM, which embeds metadata within the image itself or JPEG with separately attached metadata. Our group chose to utilize the JPEG format since this replicates more closely available formats on most if not all modern smartphones.

3 PROPOSED METHODS

The collected data will be preprocessed, segmented with OpenCV library, and evaluated for domain-specific features. Features will be extracted from the images to evaluate the ABCD (Asymmetry, Border irregularity, Color variation and Diameter) of the skin lesion. The extracted features will be passed to the machine learning model facilitating the classification of potentially cancerous lesions.

4 MELANOMA DETECTION

4.1 Preprocessing

The images from the data set were all sized differently so one of the first preprocessing tasked required the resizing of the images so that faster processing and more accurate feature extraction could be achieved. The images were also converted to a binary format to distinguish features and allow calculations to be made. The preprocessing revolved around making a suitable binary to pass to subsequent processing steps.

There were several images with excessive hair, skin markings, and other lesion obstructions that could have affected the image processing step's ability to distinguish the borders of the lesion. This necessitated the need to develop algorithms that could clean the images so processing could be accomplished. One algorithm our
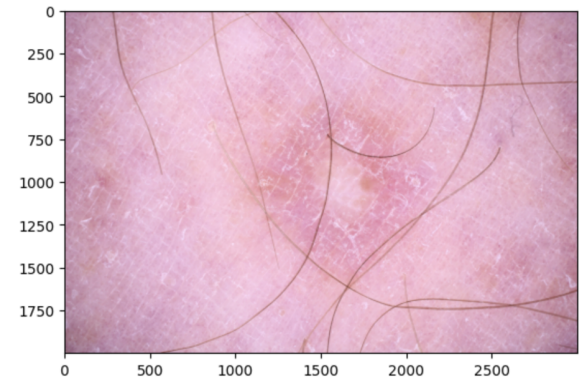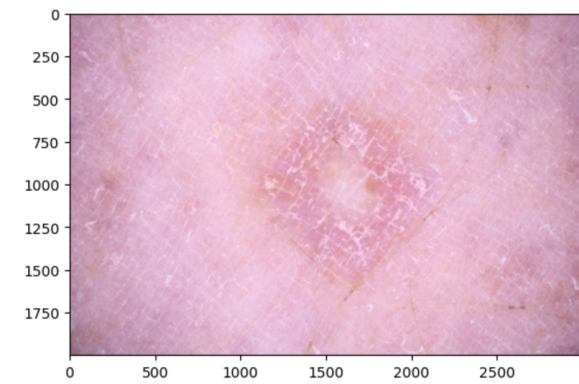
Fig. 2. Hair Removal Example (Before)

Fig. 4. Grayscale Blurred

Fig. 3. Hair Removal Example (After)

Fig. 5. Thresholded Sample Image

group utilized for processing was the hair removal algorithm, which attempted to remove as much hair as possible when applicable. That algorithm removed the hair effectively but also blurred the image to a point where edge detection and contour identification were too variable. The decision was made to drop that preprocessing step.

Figure 2, shown above is an example of one of our lesion image samples that contained hair. The figure illustrates how the hair causes a slight obstruction of the lesion, potentially obscuring features to be extracted. Figure 3 shows the same sample image after being processed by the hair removal algorithm. Figure 3 now shows a clear, unobstructed lesion image that is ready for additional pre-processing.

The next preprocessing step involved converting all the images to grayscale and then blurring with a median blur using a k-size of seven. The "Median Blur" offered excellent blurring and had a better performance than a "Gaussian Blur" with a similar k-size. "Gaussian Blur" with a k-size of less than seven did not blur the image enough for the thresholding step to distinguish the contours of interest.

The blurred images were then passed into a Thresholding Algorithm to help distinguish the lesion features from the surrounding tissues. The Thresholded images were then run through the Canny Edge Detection algorithm to find the edges of the lesion.

Fig. 6. Image Run Through Canny Edge Detection Algorithm

Once the edges were identified the "edged" image was passed to a function to find the contours in the image. The contour with the greatest area was preferred as the primary contour. That contour was then drawn on the image, highlighting the border of the lesion.

Once the contour was found, the additional contour features provided by OpenCV allowed us to calculate the center of the lesion,

Fig. 7. Sample Image Showing Distinguished Contours

Fig. 9. Sample Image Showing Fitted Ellipse

Fig. 8. Sample Image Showing Calculated Lesion Center and Extreme Edge Points
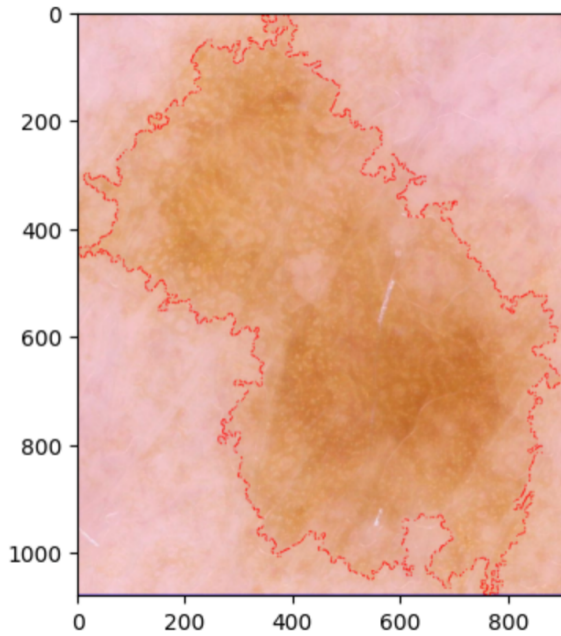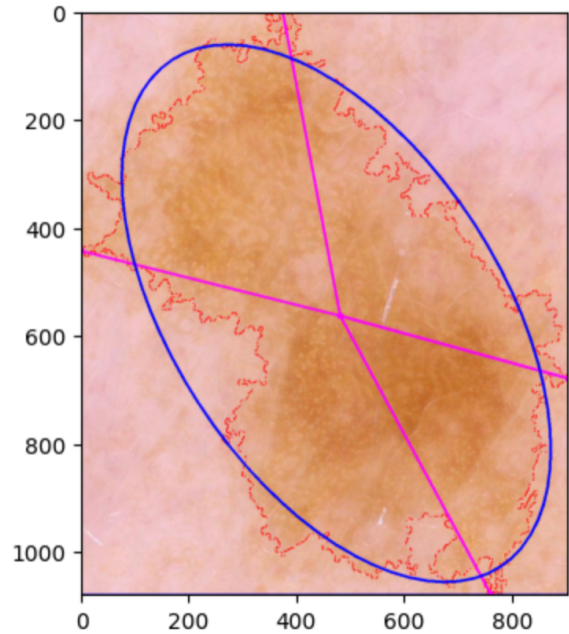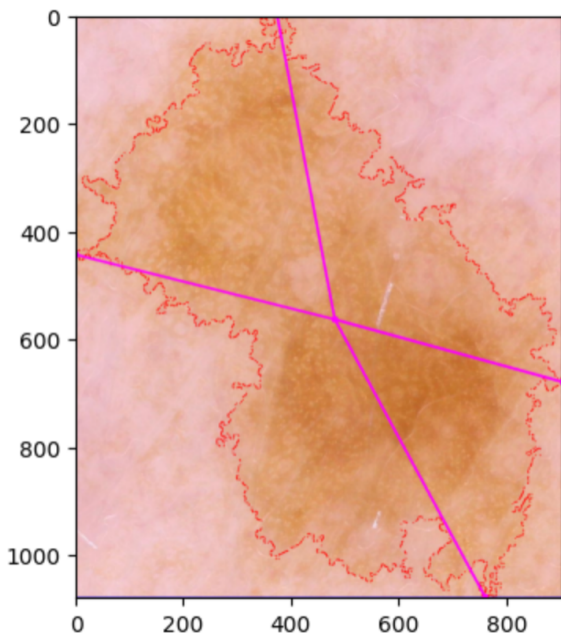
extreme edge points, fitting an ellipse to the lesion, and fitting a rectangle around the entire area of the lesion.

5 MAIN COMPONENTS

5.1 Model Development

The sample data set contained 33,126 images. Only 584 of those images were of malignant lesions. The imbalance of malignant samples to benign samples necessitated the use of undersampling of the data set, which allowed for more accurate features to be extracted. Our data also needed to be scaled due to the features being on vastly different magnitudes and scales. The sample data set included images that were as large as 3.5 MB and as small as 16 KB. This variation in file size caused issues with feature extraction so images were scaled so that consistency could be maintained across all image samples, allowing more efficient feature extraction.

Our group utilized the Random Forrest Model as it produced the best accuracy.

5.2 Feature Extraction

Our group extracted features from the data to represent the "ABCD" of melanoma diagnosis. "ABCD" is a common reference to the characteristics of Melanoma and refer to Asymmetry, Border Irregularity, Color Variations, and Diameter.

5.2.1 Asymmetry. Asymmetry was assessed by comparing the lesion's relative radius left versus right and top versus bottom portions. The lesion was isolated by Thresholding and the edges of it were found with Canny edge detection. The center of the contour in question was used as the origin for the 4 radii that attached the center point to the contour. The distances from the centroid to the contours were collected and used in the testing of asymmetry. The left radius was divided by the right radius to create a feature that represented relative asymmetry horizontally across the lesion. The
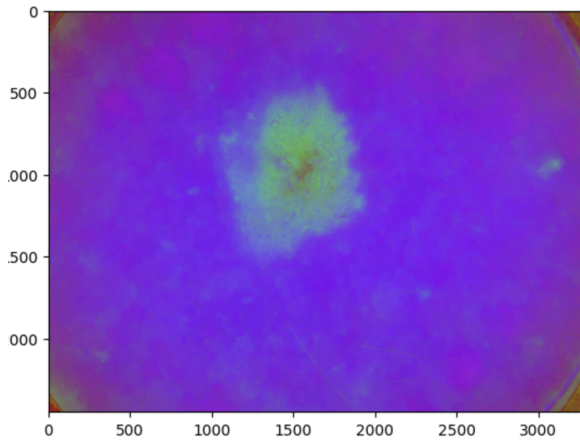
Fig. 10. Lesion Converted to "HSV" Color Space

Fig. 11. 3D Plot of "RGB" Color Space

same procedure was replicated for the top and bottom radius to create another feature to represent the symmetry top and bottom.

5.2.2 Border Irregularity. There were several proxies chosen to represent the border irregularity. One would expect a regular border to be consistent and geometrically accurate. The perimeter of the lesion, as measured on the contour, and the relative fit of the contour points to an ellipse. The length of the border contour would be higher in irregular border shapes that do not fit a circular or semi-circular path. The contour border was calculated to use as a feature to represent border irregularity. The ellipse is created by minimizing the mean least square error for fitting the ellipse. The extent that the lesion does not touch the ellipse would create a feature to help indicate the relative regularity of the border.

5.2.3 Color Variation. Color variations across a skin lesion (mole) have been linked to the development of melanoma. The image was in the red, green, and blue (RGB) space. It first must be converted into the hue, saturation, and value (HSV) color space. The HSV color space will help highlight the contrast between the colors present.

The "HSV" color space shows a clearer separation between colors than the "RGB" space.

A region of interest "ROI" was identified by adding a bounding rectangle around the image of the lesion. The image was then further cropped or zoomed into that region to calculate the features for color saturation across the region of the lesion. The bounding rectangle was used to crop the area of interest to prevent the surrounding tissues from affecting the color variation calculations. Color variation, represented by standard deviation, was calculated from the hue, saturation, and brightness of the region of interest.

5.2.4 Diameter. Lesion diameter has been identified as a clinical sign that correlates strongly with the prognosis. The equivalent diameter was calculated from the contour of the lesion image.

5.3 Steps For Pipeline

The images were converted to grayscale and blurred. Doing this allowed the sample images to be Thresholded. Once the images were
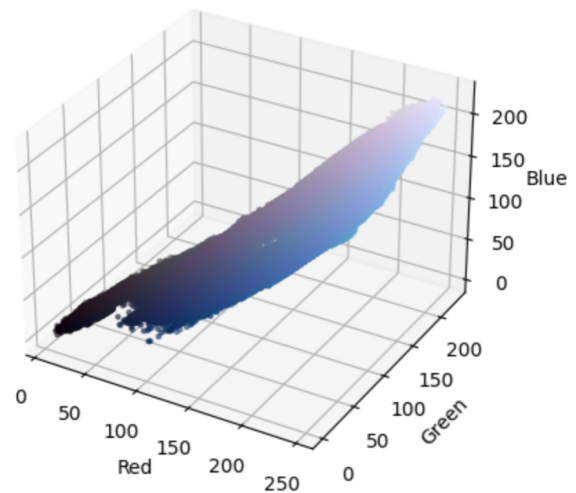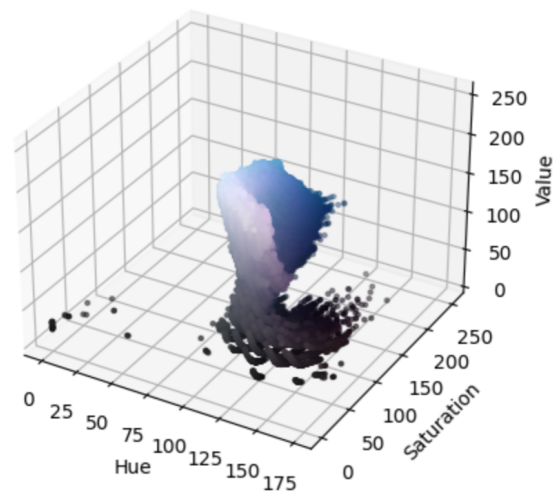
Fig. 12. 3D Plot of "HSV" Color Space

Thresholded, we obtained edges for the lesions. The lesion edges allowed for clearly distinguishable contours to be established. With the contours the image moments become accessible. The image moments allowed us to fit an ellipse to the sample image. The moments further provided a means to calculate the area of the contoured lesion. The center point, perimeter, and diameter of the contoured lesion were also obtained from the image moments. The difference in area between the contour of the lesion and the ellipse area was then calculated. The left, right, top, and bottom-most extreme contour points were established and lines were drawn connecting these points through the center of the lesion. Then we calculated the left, right, top, and bottom distances from the contour to the center of the contour to assess symmetry. Finally, we took the standard

Fig. 13. Confusion Matrix

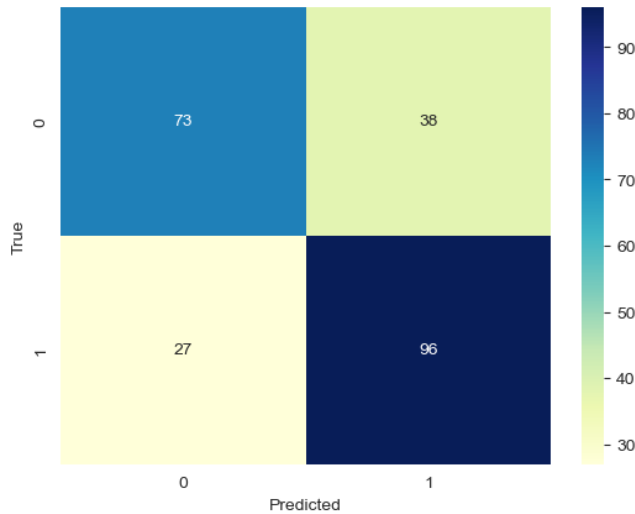Fig. 14. ROC Curve

```
Average Cross Validation Score from Training:  0.6830257029498016
Average Cross Validation Score STD from Training:  0.03298048399581561


Confusion Matrix:
Predicted    0    1  All
True
0           73   38  111
1           27   96  123
All        100  134  234


Test Statistics:
             precision   recall  f1-score   support

          0       0.73     0.66      0.69       111
          1       0.72     0.78      0.75       123

   accuracy                         0.72       234
  macro avg       0.72     0.72     0.72       234
weighted avg      0.72     0.72     0.72       234


Testing Accuracy: 0.7222222222222222
```

Fig. 15. Accuracy Results

```
Likely class Resembles benign lesions
Comparing characteristics to other benign lesions
        Asymmetry percentile 94.0,
        Border percentile 94.0,
        Color percentile 50.0,
        Diameter percentile 98.0
```

Fig. 16. Example of Comparative Percentiles

deviation of the "HSV" values for the region of interest (ROI) around the lesion to assess color variation across the ROI.

6 RESULTS

The model was trained on 1,168 images from the ISIC data set. Several different machine learning algorithms were applied to the training and test data sets. The random forest classifier produced the best results with an overall accuracy of 73.9 percent.

Additional information about the accuracy was obtained from prediction probabilities and the characteristics of the two classes, benign or malignant. Not only we were able to get a prediction of which class the new data belongs. This is accompanied by a percent likely related to one or both classes. The features extracted from any new image can be compared to the aggregate values within the class that it likely belongs. Percentile rank for each ABCD characteristic of the new image was calculated to give additional information about the image presented to the model for prediction.

7 DISCUSSION

The results obtained from this project show that there is promise in using OpenCV and a random forest classifier. The test accuracy was only in the low 70 percent range. There would need to be additional preprocessing or a different approach to the overall classification technique, possibly deep learning.

8 LIMITATIONS

The approach we employed relied on an accurate identification of the mole in the image. This was very complicated for several reasons. The collected images were not captured with standardized equipment or under a standardized protocol. As such some images contained artifacts which included hair, ink marks, and annotated markings on the image.

We attempted to employ a set of filters and thresholding to remove hair from the images with variable results. The hair removal steps
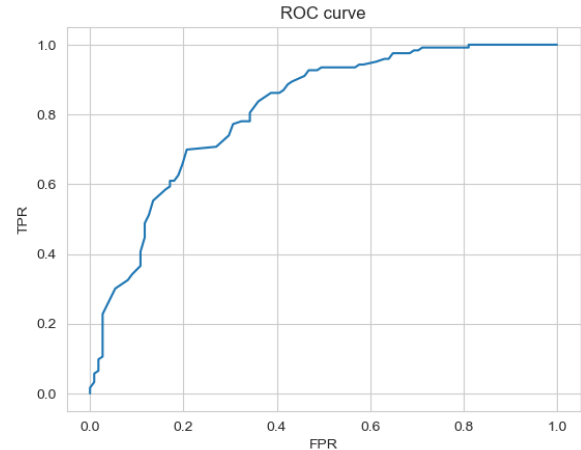
was mostly successful in removing the hair and other artifacts, but it also blurred the image to a point that the thresholding and contour identification was inconsistent.

The find contours methods from OpenCV identifies all possible contours in the converted binary image. To identify the most relevant contour, that likely represented the mole, we had to sort them by area, i.e. relative size.

Other techniques we tried included identifying the most relevant contour by creating a list of the all the available contours with additional descriptive features. Those features were contour area and

distance from the center of the image. The resulting list was sorted and the contour with the largest area closest to the center of the image was selected. Unfortunately this did not significantly increase the odds of identifying the correct contour or model accuracy.

9 FUTURE WORK

As discussed in the *Limitations* section identification of the lesion and the contours of the mole are key to the proper feature extraction. Other related work has demonstrated the ability of You Only Look Once (YOLO) algorithm to aid in mole identification or even classification.[Banerjee et al. 2020; Nie et al. 2019] Any future work would focus on the use of YOLO.

Most related work on discovered on melanoma detection involves deep learning.[Banerjee et al. 2020; Soenksen et al. 2021] The approach employed here did not use deep learning or deep learning techniques. It is logical to think the deep learning approach of progressively applying filters or layers would improve model detection and better segmentation the moles. Deep learning is another area to consider for future work.

10 CONCLUSION

Our group was able to show how publicly held data sets can be used to train a machine learning pipeline that can effectively predict melanoma. The hope is that this can be an effective tool in the early detection of melanoma and also demonstrate how machine learning can be used to make healthcare more universally obtainable. Melanoma when detected early enough is almost 100 percent curable and the use of our model could potentially cut down the guess time by allowing users to have a tool that suggests whether medical treatment is warranted, potentially increasing Melanoma survivability rates.

REFERENCES

2022. *Cancer Facts Figures 2022.* Retrieved November 1, 2022 from https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2022/2022-cancer-facts-and-figures.pdf

Shubhendu Banerjee, Sumit Kumar Singh, Avishek Chakraborty, Atanu Das, and Rajib Bag. 2020. Melanoma Diagnosis Using Deep Learning and Fuzzy Logic. *Diagnostics* 10, 8 (Aug. 2020), 577. https://doi.org/10.3390/diagnostics10080577

Yali Nie, Paolo Sommella, Mattias O'Nils, Consolatina Liguori, and Jan Lundgren. 2019. Automatic Detection of Melanoma with Yolo Deep Convolutional Neural Networks. In *2019 E-Health and Bioengineering Conference (EHB)*. 1–4. https://doi.org/10.1109/EHB47216.2019.8970033

Kurtansky N. Betz-Stablein B. Caffery L. Chousakos E. Codella N. Combalia M. Dusza S. Guitera P. Gutman D. Halpern A. Helba B. Kittler H. Kose K. Langer S. Lioprys K. Malvehy J. Musthaq S. Nanda J. Reiter O. Shih G. Stratigos A. Tschandl P. Weber J. Soyer P. Rotemberg, V. 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci Data 8, 34 (2021). https://doi.org/10.1038/s41597-021-00815-z

Luis R. Soenksen, Timothy Kassis, Susan T. Conover, Berta Marti-Fuster, Judith S. Birkenfeld, Jason Tucker-Schwartz, Asif Naseem, Robert R. Stavert, Caroline C. Kim, Maryanne M. Senna, José Avilés-Izquierdo, James J. Collins, Regina Barzilay, and Martha L. Gray. 2021. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine* 13, 581 (2021), eabb3652. https://doi.org/10.1126/scitranslmed.abb3652 _eprint: https://www.science.org/doi/pdf/10.1126/scitranslmed.abb3652.