Tarek El-Hajjaoui, Micah Fadrigo, Sheldon Gu, Cecilia Nguyen
Professor Shahbaba, Professor Li
MDS Capstone Projects (DATA 296P / 297P)
15 December 2023

Accelerating Juvenile Dermatomyositis Diagnosis with Artificial Intelligence using Nailfold Capillaroscopy

Abstract

Juvenile Dermatomyositis (JDM) is a rare, chronic autoimmune disease affecting individuals during childhood. To diagnose patients with JDM, Convolutional Neural Networks (CNNs) can use raw image data for classification of patients with or without JDM, and can be used as an aid to correctly diagnose and allow timely treatment. Although CNNs achieve state of the art performance, they tend to have poor explainability and generalizability compared to general classification ML models. We propose looking at simpler classification models of Logistic, Random Forest, and Support Vector Machine - to find an alternative to CNNs with similar performance but better explainability of features.

A total of 1,120 NFC images from 111 children with active JDM, diagnosed between 1990 and 2020, and 321 NFC images from 31 healthy controls were retrieved from the CureJM JDM Registry. Images were downscaled by interpolation techniques to reduce the computational cost. We vectorized the images and performed Histogram of Oriented Gradient transformation to all images with the intention to extract more useful information.

The SVM model on the HOG images achieved high performance in differentiating patients with JDM from controls, with an area under the ROC curve (AUC ROC) of 0.920. The good performance of SVM combined with HOG competes with the performance of NFC-Net, showcasing the potential of simple machine learning models. It also reinforces the idea that NFC images are sufficient for detecting often unrecognized JDM disease activity, providing a reliable indicator of disease status.

1 Introduction

Juvenile Dermatomyositis (JDM) is an autoimmune disease that only occurs in childhood and may last into adulthood. The average age the disease starts is seven years. JDM is very rare, affecting about two to four children for every one million children in the United States per year. Common symptoms for JDM include muscle inflammation (myositis), skin rash (dermato), and muscle weakness. There are currently no known causes of JDM disease, and no effective treatment is available. However, there are treatments to help patients alleviate or eliminate some symptoms, which controls muscle inflammation so patients with JDM can live more comfortably with the illness.

1.1 Motivation

Early diagnosis of JDM is essential for patients to start disease management at an early stage. The current diagnosis process of JDM involves biomarkers. They usually involve expensive lab tests and are difficult to identify. On the other hand, there has been research showing the effectiveness of Nailfold Capillaroscopy (NFC) for diagnosis of JDM. NFC is a sensitive, simple, safe, and noninvasive imaging technique. It is much

easier to obtain and affordable compared to labs and biomarkers. Our project researches on using Machine Learning techniques to perform JDM pre screening using NFC. This project may help facilitate the pre-screening process so potential patients can start disease management at an early stage.

1.2 Goals and Objectives

This project focuses on quick, simple, and affordable pre-screening of JDM. Specifically, this project tries to solve a classification problem: given an NFC picture, use machine learning techniques to classify it as a JDM patient or a healthy person.

2 Data

Dataset for this project is provided by Children's Hospital of Orange County (CHOC). Between 1990 and 2020, CHOC has collected in total 1441 NFC images, including 1120 images for the JDM group and 321 images for the healthy control group. These images are taken from 111 JDM patients and 31 healthy people respectively. The image files are named in the format of <PatientID + Finger Number>. The dataset is completely deidentified so we receive no demographic information.

2.1 Data Challenges

As introduced in the beginning, JDM is an extremely rare condition, making it difficult to collect a large amount of data. These images are also medical sensitive data, requiring extra protection and constraints and are not permitted to be uploaded to any external devices. This limitation restricts our group from acquiring additional computing resources. On top of that, the time span for collecting the NFC images goes across several decades. During this time, technology and medical equipment have evolved dramatically, making it difficult to stay consistent in the data collection process. Therefore, the collected images consists of many noises.

Figure 1 presents some example images from the original dataset. The first notable issue in the figure concerns the variation in finger orientation: some fingers face upward, while others face downward. Not all images have the nailfold area centered in the middle. Another challenge steaming from the collection process is the absence of standardized lighting conditions. Certain images appear notably darker than others. Skin tone variations introduce extra noise into the dataset.

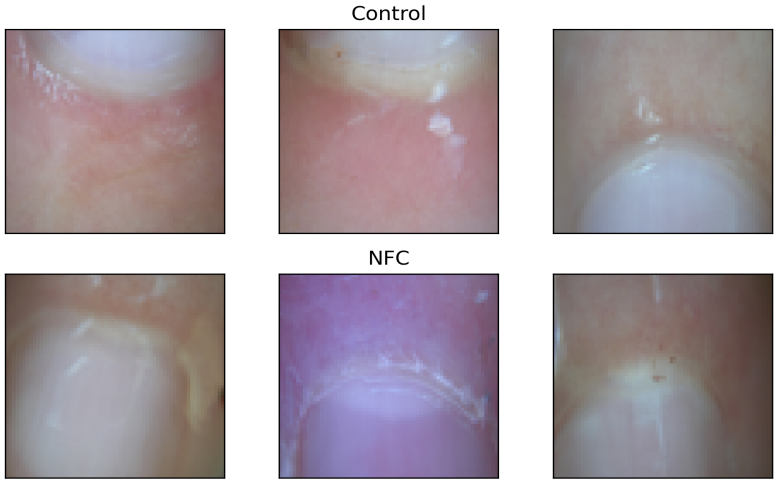

Figure 1: Example Nailfold Capillaroscopy images from the original dataset provided by CHOC.

Constrained by data protection requirements, we are not permitted to show further examples besides Figure 1. Here are several other issues we observed with the dataset. Some nailfold surfaces are obstructed, primarily

caused by nail polish decorations or uncleaned nail polish residuals. There is an inherent imbalance between two groups in the dataset, with the JDM group containing significantly more data than the healthy control group. Furthermore, some patients have more pictures taken than others, suggesting potential underlying correlations among the images. Lastly, the original image size poses a great challenge given our limited computing resources: it may take weeks to finish training and tunning a model.

2.2 Exploratory Data Analysis

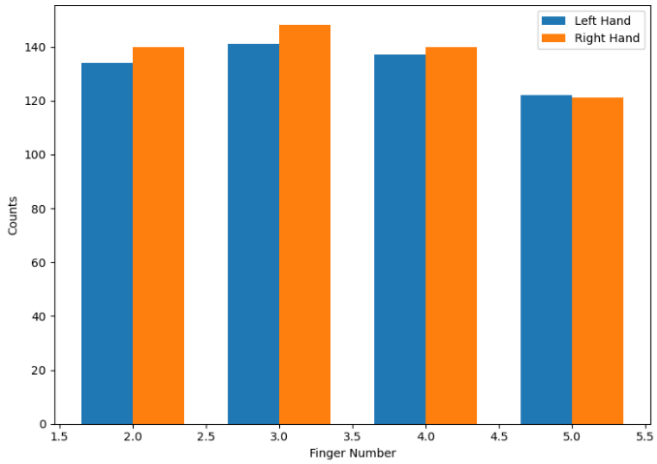

Figure 2: Distribution of all NFC images among all eight fingers, including both JDM group and healthy control group.

In order to build successful, well-performing models and comprehend any issues within the data, we first explored the distribution of all NFC images between the left and right hand, as well as the distribution of total images taken from each patient in the case and control groups. From Figure 2, we can observe the raw counts of finger images numbered from 2 to 5, starting from the pointer finger to pinky. From both groups, the images of each finger (excluding the thumb) shows a relatively equal and balanced proportion of finger images taken between the left hand and right hand of each patient, indicating no abnormalities in distribution.

Figure 3 depicts the number of NFC images taken per patient. The plot on the left indicates counts for the healthy control group and the right plot indicates counts for the JDM group. In both groups, an individual is most likely to take eight NFC images. However, there are considerably many individuals who take less than eight images or much more than eight images. In the most extreme case, an individual has taken more than 30 NFC images. An imbalance issue can be observed patient-level wise, in both our case and healthy control groups. The underlying correlation issue may cause bias in our model predictions toward some patients with much more images, possibly decreasing model performance if tested on NFC images outside of our dataset.

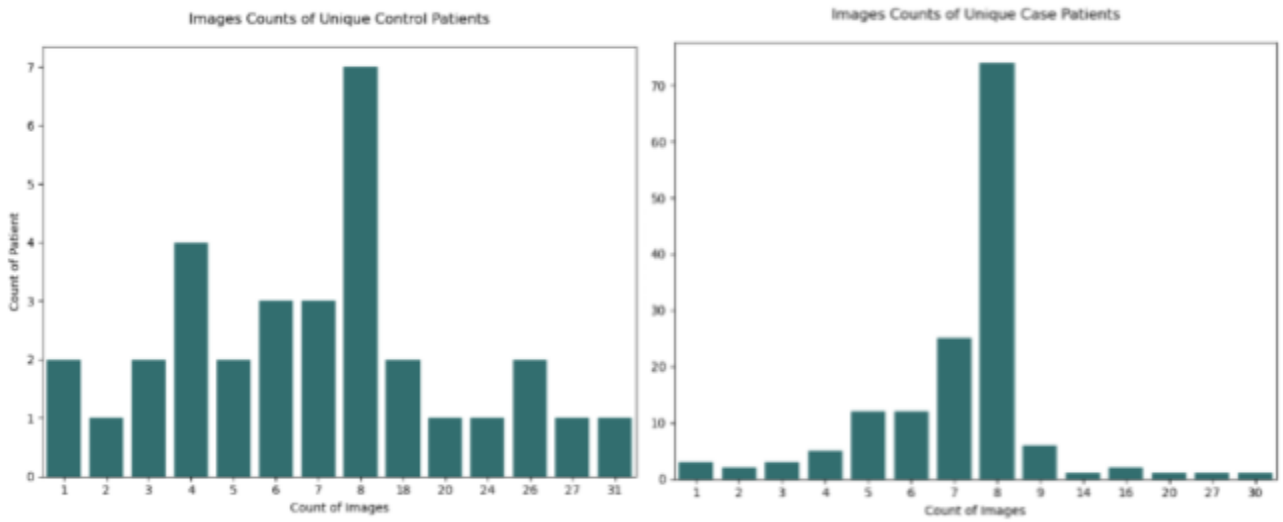

Figure 3: Distribution of NFC Images taken per person.

2.3 Data Preprocessing

Before we received the data, the CHOC data science research group examined the images and manually corrected the orientations so they all face the same upward direction. This standardization greatly reduces the inconsistency, making it much easier to look for a common pattern.

The first step we take in the preprocessing stage is to downscale the raw image into smaller sizes. We chose size 128x128, 64x64, and 32x32 for comparisons. The next preprocessing step is to scale the input pixel to (0,256) to (-1, 1). Then the 3D images are flatten into a 1D array so they can be input into machine learning models for training. Additionally, we perform Histogram of Oriented Gradient transformation to all images with the intention to extract more useful information, such as shapes, patterns, and structures.

3 Machine Learning Models

Most image classification tasks are conducted using standard deep neural network (DNN) architectures for computer vision. These standard architectures implement batch normalization - a method used to improve training speed and model performance. However, batch normalization is data dependent and can fail if there is a large variation in the data, resulting in unstable and poorly trained models. Due to the large variability in the data with regards to skin tone, skin color, use of nail polish, and devices used to capture the NFC images, standard architectures such as ResNet were abandoned. Instead, the CHOC team customized a CNN (referred to as NFC-Net) that excludes batch normalization, has a much simpler architecture, and consists of only 3 convolutional layers. NFC-Net achieved nearly excellent performance at discriminating between case and control images given its AUC ROC score of 0.93. However, NFC-Net may still be too complex and uninterpretable to clinicians who are unfamiliar with AI. Clinicians need insight into *how* the model is arriving at its predictions to ensure that it aligns with their research and knowledge. CHOC is continuously working towards explainable AI - being able to identify which regions of the image are driving the model's predictions.

In this age of artificial intelligence (AI) and deep learning, we tend to forget that simple algorithms can work well for a surprisingly large range of problems. With model explainability in mind, we decided to go in the direction of intrinsically simpler models that have a simple structure compared to CNNs. We also chose this

approach for several other reasons: to establish a baseline measurement, to improve computational efficiency, and to deal with high levels of noise in the data. We needed to test simpler models to see how they would measure up to NFC-Net. Simpler models are also more computationally efficient, making it more suitable to be wrapped within an edge-based device like a mobile application that is accessible to both clinicians and patients. Additionally, simpler models have better generalization performance due to their robustness to noise.

3.1 Machine Learning Training Methods

All of the models that are mentioned in this paper utilized Stratified 10-Fold Cross Validation (CV) for calculating model scores: ROC AUC Score, Class 0 Accuracy, Precision, Accuracy, Recall, and F1 Score. The most important scores for the JDM classification task are ROC Auc Score and Class 0 Accuracy because the goal is to create a model that can reliably distinguish against JDM versus control. Each fold has 80% assigned to the training set and 20% to the testing set; it is important to use cross validation to ensure that scores are not biased from only being based on one train/test set split. Stratified ensures that samples of each class are proportionally distributed among the train and test datasets. An improvement in this method that could be implemented is to perform a train test split in which images from a given patient are either all in the train set or all in the test set; this method would ensure there is no data leakage during training.

For the first set of models, a version of itself is created for each combination of image size - 32x32, 64x64, 128x128 and feature type - vectorized image or Histogram Oriented Gradients (HOG) feature descriptor. In the training of the second iteration of the Support Vector Machine (SVM) model, only 32x32 HOG feature descriptors are utilized due to the large hyperparameter training times. For this model, the GridSearchCV module from Scikit-Learn is used to find the optimal HOG feature descriptors and model parameters. A set of values for each parameter is given as a grid to the module, then it calculates the scores for each set of parameter values and returns the best set of values in the end. Grid search is configured in these experiments to use 3-Fold Stratified Cross Validation (80/20 splits per fold) and to refit based on ROC AUC Score.

3.2 Importance of Model Explainability

Model explainability is the ability to understand how a model is arriving at its predictions. It's necessary to build trust and user confidence in the model's predictions, especially in such a sensitive application like healthcare where equity and reliability are major pillars. Clinicians need to be able to trust that the model is looking at significant parts of the NFC image that align with their knowledge and research, prior to model deployment. Clinicians also need a way to debug model behavior and inform corrective actions, thus working towards a model that helps decrease the rate of misdiagnosis. Overall, interpretable model explanations contribute to end user trust and can act as a catalyst to adoption of the machine learning system.

We commonly see a tradeoff between model performance and explainability - as performance increases, explainability decreases. For example, Convolutional Neural Networks (CNNs) achieve state-of-the-art performance, but are considered black-box algorithms, in which it is unclear what is being learned due the activation functions. As a result, it can exhibit unreliable behavior and reinforce undesirable biases that result in poor outcomes for many stakeholders.

Methods based on gradient information have been developed to explain CNN predictions. Integrated Gradients (IG) is a post hoc explanatory method that aims to explain the relationship between a model's predictions and its

features. It uses gradient information as a measure of importance in the feature space and returns an attribution heat map that highlights regions of the original image that are important in the model's decision.

A goal for this project was to be able to create a similar translation interface visualization for one of the machine learning models, if it had comparable results to NFC-Net.

3.3 Histogram of Oriented Gradients

A feature engineering technique, Histogram Oriented Gradients (HOG), is implemented to compensate for the lack of features in the dataset. This technique counts occurrences of gradient orientation in localized portions, cells, of an image grid to create feature descriptors. These feature descriptors represent edge distributions based on their gradient magnitude (edge thickness) and gradient direction. The HOG feature descriptors reduce the noise in Computer Vision ML tasks by enabling the models to focus on the shapes and edges of objects in images instead of noisy patterns such as color, background, etc. that may distract from the task at hand. The skimage function hog is used to create HOG feature descriptors; the function has several parameters that were configured during development: orientations, pixels per cell, and cells per block. Orientations indicate how many slices of a cell will be considered to calculate the descriptors, pixels per cell specify how many pixels will be included per cell for calculations, and cells per block indicate how many cells will be used per each block of cells from the image grid. The initial sets of models including Logistic Regression, Random Forest, and Support Vector Machine (SVM) were all configured with one set of values: orientation = 9, pixels per cell (8,8), cells per block (2,2). The second iteration of the SVM model utilized different values of the parameters that were chosen with the help of the GridSearchCV module from the Scikit-Learn library (discussed in 3.1).

3.4 Logistic Regression and Lasso (L1) Regularization

The simplest statistical model for a binary classification problem is logistic regression, which predicts the likelihood of an event happening. Logistic regression was used in combination with lasso (L1) regularization to address multicollinearity in the image data. L1 regularization is a common feature selection technique used in high-dimensional data settings where there are redundant or highly correlated features. It identifies the most important features while setting others to zero, thereby simplifying the model and can lead to improved generalization performance and interpretability.

The model performed the best on 32x32 vectorized images with an ROC AUC score of 0.767 and specificity of 0.62. The ROC AUC score of 0.767 indicates that the model was better than random chance, however there is still room for improvement given the ROC AUC score of 0.93 from NFC-Net. The specificity of 0.62 means that only 62% of the control images (minority class) were classified correctly.

Aside from the subpar metrics, this model was not a promising model worth pursuing further due to its violation of assumptions. The linear relationship assumption between predictors and log odds of JDM was violated since image data is more likely to have non-linear relationships. The independence assumption was also violated in which images associated with the same individual would likely yield similar risk of JDM. If more data is collected in the future, highly correlated images for each patient could be removed so that each patient is only represented once in the dataset. Lastly, the absence of multicollinearity assumption was violated in which pixels within the NFC images are highly correlated. Multicollinearity makes it difficult to isolate the unique contribution of each feature to the model. More advanced methods are necessary to remove highly correlated features (pixels) in the image.

3.5 Random Forest

Another general classification algorithm chosen for the design of the simple model is Random Forest. It is a supervised learning technique involving an ensemble learning technique which combines classifiers and averages of all the predictions of each decision tree. Noting the trade-off between explainability and accuracy, compared to a CNN model without additional interpretation techniques, Random Forest is more explainable due to its simple tree structure which provides feature importance scores across trees, and hence was chosen as a simple model candidate. In addition, Random Forest is less sensitive to imbalanced data and noises in training data and has no assumptions regarding underlying distribution of data.

The Random Forest model's initial set of values were the default values from the Scikit-Learn library, aside from the hyperparameter of class weights which was {0:1.5 1:1}, to account for the influence of class imbalance. For performance, the model with vectorized features had slightly higher metric scores aside from recall, but overall metrics indicated subpar performance in the model for classification.

The metrics indicated poor performance, with the SVM model having the best results. The highest two Random Forest models with vectorization (32 x 32) and HOG features (64 x 64) had an AUC-ROC score of around 0.75 and 0.65 , and with specificity scores of 0.53 and 0.30 ; This indicates poor performance especially compared to NFC Net's. The low performances in the initial tests may be due to limitations with the Random Forest algorithm. Random Forest has difficulty with high-dimensional data and does not fully capture the spatial hierarchies and local relationships between pixels in images. Random Forest is not a viable simple model to further optimize compared to the other simple models.

3.6 Support Vector Machine

The idea of using SVM for the JDM classification tasks stems from the fact that SVM models are known to be effective for binary classification tasks, computationally efficient, and robust to overfitting. And perhaps most importantly, the combination of HOG feature descriptors and SVM is proven effective and widely used in Computer Vision applications such as face detection and vehicle detection tasks. Two types of SVM models were utilized, a Linear SVM and RBF SVM. Linear SVM models utilize a linear decision boundary, meaning it assumes that class features are linearly separable; in contrast, the RBF SVM applies a kernel function that increases the number of dimensions of the data and then creates non-linear decision boundaries. Aside from the kernel, the SVM has three other key hyperparameters: C, class weights, and gamma. Generally speaking, the higher the value of C the more tight the decision boundary will be based on the training data set; a too high C value results in overfitting and a too low C value results in underfitting. Class weights tell the SVM model to consider the number of class samples based on their weights when creating the decision boundary; this parameter is important for the JDM task given the imbalance dataset. The gamma parameter defines how far the influence of a single training example reaches. Only the default value of C, 1, and class weights, {0: 1, 1: 1}, are used in the initial set of SVM models, shown in Figure 4. The ROC AUC Score and Class 0 Accuracy scores of the Linear SVM and RBF SVM across the different image sizes (32, 64, 128) are shown below. Since these models were not tuned, the Linear SVM model utilizing the 32 x 32 vectorized images performed the best.

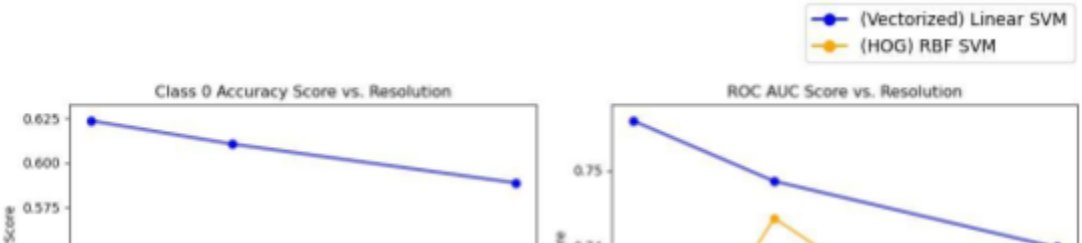

Figure 4: ROC AUC Score and Class 0 Accuracy of the original set of SVM models.

The C and class weights parameters are later tuned with the help of GridSearchCV (discussed in 3.1); the gamma parameter is never optimally tuned, but can be for future improvements. There is a significant improvement across all metric scores in the SVM model using optimal hyperparameters and optimal HOG feature descriptors, shown in Figure 5 below. This improvement highlights the importance of hyperparameter tuning as well as the capabilities of SVM with HOG feature descriptors in Computer Vision classification tasks.

Original Best - Linear SVM [32x32] (Vectorized)					
ROC AUC	Class 0 Accuracy	Precision	Accuracy	Recall	F1
0.756	0.624	0.892	0.830	0.890	0.890

New Best - RBF SVM [32x32] (HOG)					
ROC AUC	Class 0 Accuracy	Precision	Accuracy	Recall	F1
0.920	0.726	0.923	0.896	0.945	0.934

Figure 5: Final best performing SVM compared to original SVM before tuning.

A goal of this project mentioned in section 3.2 was to create a translation interface visualization that communicates important regions of the image that contribute to the model's decision. With the optimal SVM model achieving comparable results to NFC-Net, model explainability was of interest. As for model explainability for the optimal SVM model, the support vectors were visualized on top of the HOG image. In this context, the support vectors are the HOG feature descriptors that have the most influence on the position of the decision boundary. Areas with a high concentration of support vectors indicate regions where the model heavily relies on HOG features for making decisions. Two random examples were pulled from the dataset and it was found that support vectors were not as concentrated in the skin area and moreso concentrated near the bottom of the nail. Due to privacy concerns, the images and visualizations cannot be shown. It can be seen that analyzing support vectors in the HOG image can provide some insight into what the model is focusing on.

3.7 Convolution Neural Network

CHOC's NFC-Net consists of three hidden layers and one final classification layer. Based on NFC-Net, we constructed a customized Convolutional Neural Network with five layers. As described

in the data section, the dataset is limited in number for effective CNN training and it also suffers from imbalance. Therefore, image augmentation techniques are applied in order to deal with some of these issues. The first layer of the customized CNN is an image augmentation layer. The image augmentation techniques applied include horizontal random flipping and 10% random zooming out. Then, the augmented images are sent to three hidden layers, which are made of Convolution and Max Pooling. In the end, a classification layer produces the classification decision. The customized CNN model is able to achieve an overall accuracy score of 84%. One limitation of CNN is that its training uses massive computing power and resources and we are restricted to limited computational resources available. The image augmentation layer further magnifies this issue by adding many more images into the model training. Furthermore, we are restricted with a narrow timeline so we could not continue parameter tuning and experiment with other layer constructions.

4 Conclusion

To achieve our objective of using ML to create a quick, simple, and affordable pre-screening for the diagnosis of patients with JDM, we explored the concept of simple model algorithms as an effective and more explainable alternative to a CNN model, and then optimized our designed models further through hyperparameter tuning. With simple models having more advantages such as robustness to noise and computational efficiency, we noticed that a simple model can compare to a complex CNN in terms of performance in classification**.** Completing the steps of the data lifecycle, we performed EDA and then addressed issues within the image preprocessing stage, which were variations within the images (such as the orientation of the nails), and interpolated the images for faster computation due to downsizing. We then accounted for the imbalance of the JDM and healthy patient classes during the design of our simple classification models (Logistic Regression, Random Forest, and SVM), and found SVM as the most promising model. To determine the potential of the SVM model, we performed additional feature engineering with the HOG feature descriptors and optimized our model through hyperparameter tuning to achieve similar results to the NFC-Net's. Our model experimentations also included possibly optimizing the CHOC's current NFC-Net Model, by adding image augmentation and changing the architecture with the addition of more layers. However, there was no improvement in performance, and further experimentation is required.

Extending from our research, there is high potential in the real-world application of using ML models for diagnosis, not just for JDM. The possibility of deploying and building a mobile application integrated with ML algorithms (such as SVM), can be beneficial for a patients' care as they are able to receive a timely diagnosis–reducing the progression of disease with earlier treatment and intervention. The use of simple models can improve explainability in the features determining disease compared to more complex models like CNNs, and can positively improve efficiency in care of a patient's health as an additional aid for a physician's diagnosis.

4.1 Limitations

The main challenges that we encountered during our model experimentations were: the low quality of data and the lack of computing power, which decreased the efficiency in training, and the fine-tuning of the hyperparameters in our models. Decreasing our model performance, the data had a small sample of 1441 patients' nail images taken over a 30 year period, as well as variations in the nail images taken (brightness, skin color, position, etc.). The issue of computation would notably occur across the images of increasing interpolation sizes (with the sizes 64, 128), and also when tuning the HOG feature descriptors. During the training of our model with multiple stratified cross-validation, a method used to reduce bias and overfitting, we weren't able to obtain the performance metrics within a reasonable timeframe. The short timeframe to further develop and optimize our model's performance, including the NFC-Net model, was also a limitation.

4.2 Possible Future Improvements

Possible improvements in our research would be during the image preprocessing and model optimization stages. Our best simple model is currently the SVM model with the rbf kernel, HOG feature descriptors, and 32*32 size. Although we achieved higher recall and F1 scores compared to NFC-Net, we still need to optimize and finetune the hyperparameters for other important metrics such as AUC-ROC, compared to NFC-Net's (0.92 vs 0.93) and the specificity scores, which is lower than NFC-Net's (0.73 vs 0.90). For medical diagnoses, there is focus on reducing Type 1 error and improving the classification of actual healthy patients; this is because of the possible detrimental effects of incorrectly classifying healthy patients as ones with JDM. The unnecessary expenditures on treatments and resources can be costly, as well as damage patient-physician trust with a misdiagnosis. To improve our image data for the data pipeline, We would also perform better image preprocessing, such as including centering of images when fixing orientation.

For the explainability of our models, we would like to explore different interpretability techniques other than the integrated gradient maps, to have a comprehensive understanding of the model's predictions and gain a different perspective of important features, confirming the reliability of our interpretations.

To improve overall data quality, we would like a larger quantity of image data, as well as a standardized process for the collection of NFC images; this would reduce variability and noise, which streamlines the image preprocessing stage. In the future, utilizing simple models for diagnosis of JDM can be implemented from a proof-of-concept to the real world, in the form of a mobile application, which can be a convenient medium and tool for diagnosis by physicians.

5 Acknowledgements & Contributions

We would like to acknowledge and express our thanks toward our sponsor, the Children's Hospital of Orange County (CHOC), specifically the staff of the data science team for their collaboration with us for the project, with the involvement of Dr. Peyman Kassani, Nadine Afari, and Louis

Ehwerhemuepha who helped guide us throughout the project–giving us a strong, foundational understanding of the JDM CNN classification models developed in their research.

Tarek is responsible for team communication, SVM, HOG and parameter tuning. Micah contributed to the data preprocessing framework and SVM explainability and was responsible for the logistic regression & lasso regularization model. Sheldon is responsible for data challenges and the customized Convolutional Neural Network construction. Cecilia is responsible for the EDA, Random Forest, and Conclusion sections.

References

Kassani, Peyman Hosseinzadeh, et al. “Artificial Intelligence for Nailfold Capillaroscopy Analyses – a Proof of Concept Application in Juvenile Dermatomyositis.” Nature News, Nature Publishing Group, 22 Nov. 2023, www.nature.com/articles/s41390-023-02894-7.

Etehad Tavakol, Mahnaz, et al. “Nailfold Capillaroscopy in Rheumatic Diseases: Which Parameters Should Be Evaluated?” BioMed Research International, U.S. National Library of Medicine, 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4569783/.

Lambova S. N., Müller-Ladner U. Capillaroscopic pattern in systemic lupus erythematosus and undifferentiated connective tissue disease: what we still have to learn? Rheumatology International. 2013;33(3):689–695. doi: 10.1007/s00296-012-2434-0.

Mallick, Satya. “Histogram of Oriented Gradients Explained Using Opencv.” LearnOpenCV, 30 Nov. 2021, learnopencv.com/histogram-of-oriented-gradients/.