Enhancing Melanoma Detection: A Comparative Study of Machine Learning Techniques Using the Melanoma Skin Cancer Dataset

Allan Zhang [*1] **David Huang** [*1] **Minseok Son** [*1] **Aditya Shrey** [*1]

Abstract

Early detection of melanoma skin cancer is critical for effective treatment. In this study, we compare machine learning techniques, including CNNs, SVMs, Logistic Regression, and Transfer Learning with ResNet-18, using the Melanoma Skin Cancer Dataset. Through preprocessing and hyperparameter optimization, we assess model performance in terms of accuracy, precision, recall, and F1-score. Our findings highlight CNNs' efficacy in achieving high accuracy, with SVM and Transfer Learning also showing promise. This study contributes to advancing machine learning applications in medical imaging for melanoma detection.

Figure 1. On the left, we have an example of a benign image. On the right, an example of a malignant detection is shown. From an initial glance, these images look similar, emphasizing how misdiagnosis can potentially occur.

1. Introduction

Melanoma skin cancer is one of the most deadly forms of skin cancer (Javid, 2022). Even though melanoma accounts for only 1% of all skin cancer, it accounts for a majority of the skin cancer related deaths according to the American Cancer Society (2024). Often, if a correct diagnosis is given early, melanoma can be treated properly with surgery and no other intervention. The misdiagnosis of melanoma can be costly for the patient as new drugs that have been created are not effective in treatment (Ther, 2019). In practice, various diagnosis techniques are used in the medical setting including immunohistochemistry, where markers, such as clinical or biomarkers, are used on a biopsy of the lesion (Ther, 2019). Given how melanoma and other skin cancers can often be indistinguishable from non-cancerous skin, enabling individuals to determine whether the suspecting area is benign or malignant is essential both before consulting a doctor and also while under the care of a healthcare professional.

Recently, there has been a growing trend of making use of deep learning techniques in classifying skin cancer related images as being either benign or malignant during the early stages. Convolutional neural networks have proven to give better results than other neural networks in the classification of skin cancer images (Dildar et al., 2021). Other machine learning techniques, such as support vector machines and logistic regression, could prove to be more effective and also create simpler models. Success of these simpler models could also come from the heuristic filtering of data that is deemed to be most relevant through principal component analysis. The aim of this paper is to conclude which of the machine learning techniques had superior results in melanoma skin cancer classification tasks. We will make use of the Melanoma skin Cancer Dataset of 10000 Images from Kaggle created by Javid[2]. The data has been labeled benign or malignant, which corresponds to our binary classification problem. This dataset enables us to utilize the various machine learning techniques on representative data for the melanoma classification problem.

2. Methods

2.1. Data Exploration

In this section, we delve into the dataset used for developing machine learning models targeting melanoma skin cancer.

[1]CS 4262 Foundations of Machine Learning (Dr. Sprinkle), Department of Computer Science, Vanderbilt University, Nashville, Tennessee.

[2]The dataset is linked in the references section of this paper.

[*]Equal contribution.

We discuss the steps taken in preprocessing the data and the features extracted.

Inspecting the Dataset: Key to the robustness and applicability of our models is understanding the fundamental properties of our dataset. It comprises a balanced distribution of 5,000 benign and 4,605 malignant images. This equilibrium is advantageous as it prevents model bias towards the more frequently represented class—a common challenge in medical imaging datasets, which often have a predominance of benign examples. Moreover, the quality of the images is notably high; they are largely intact with minimal noise. This clarity is beneficial for precise feature extraction, enhancing the model's ability to discern subtle distinctions between benign and malignant lesions. The high-quality, well-balanced nature of our dataset provides a solid foundation for the development of reliable and effective machine learning models for melanoma classification.

Preprocessing Techniques: To prepare the image for model training, we standardized the input sizes by resizing all images to a consistent resolution of 100 x 100 pixels. Additionally, we normalized the pixel values to fall within the [0, 1] range. This normalization is essential for promoting faster and more stable convergence during model training. The final format of each image is a 100 x 100 x 3 array with values from [0,1] which represent the RGB value of each pixel.

Data Augmentation: To bolster the models' robustness and enhance its ability to generalize across varying inputs, we incorporated data augmentation techniques such as random horizontal flips and rotations. These techniques introduce realistic variations into the training data, simulating potential real-world scenarios not represented in the initial dataset. Such augmentation helps mitigate overfitting, where a model learns the specific details and noise of the training data at the expense of its generalizability. This strategic introduction of variability ensures that the model can recognize and learn from a broader array of image features, significantly improving its ability to accurately predict new, unseen images (Perez and Wang, 2017).

Dimensionality Reduction: Facing the challenge of high dimensionality in the image data, we employed Principal Component Analysis (PCA) on the flattened and normalized images. PCA was instrumental in reducing the original feature set from 9,605 features to a more manageable 66 features, while retaining 95% of the original variance. The significant reduction in dimensionality not only makes the model more efficient to train, but it also helps preserve the critical information of the dataset, thereby supporting the detection of nuanced patterns associated with melanoma skin cancer.

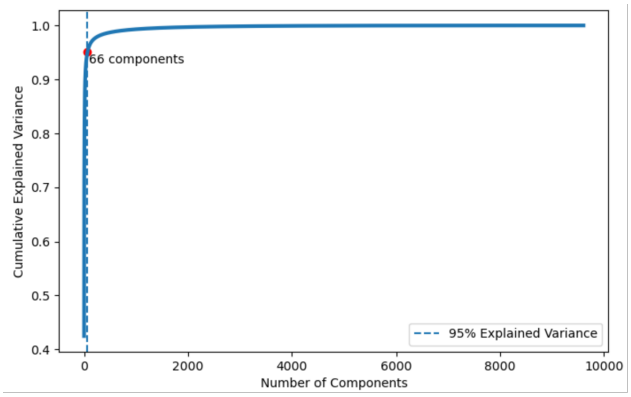

Figure 2. The plot depicts the cumulative explained variance against the number of components, with a threshold line at 95% indicating 66 features.

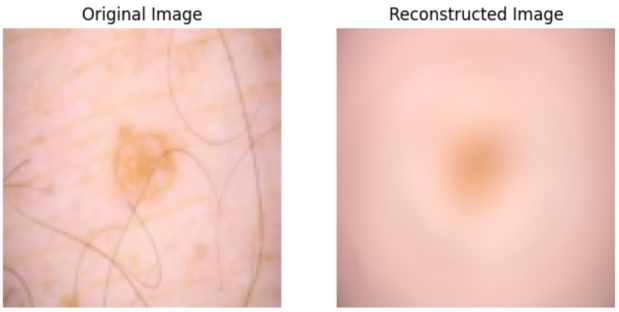

Figure 3. The plot compares the original image with the image that was reconstructed with 66 principal components.

2.2. CNN

The first model we tried to use to solve this problem was a convolutional neural network (CNN). Convolutional neural networks have historically been good at solving image classification problems due to their strength in finding and learning patterns in spatial hierarchies within images. Using our preprocessed images, which are in the form 100 x 100 x 3, we built a CNN model around each image that predicted whether the cancer shown was malignant or benign. We chose 20% of the images randomly for testing and trained on the rest. Our CNN is designed with the following outline: the input image is sent to a series of 2D-convolution, rectified linear unit (ReLU) activition, and max pooling layers, then the data is flattened and sent through two feed forward neural network layers where a prediction is made at the end. The model is pictured similar to the one shown in Figure 4.

We then performed grid search hyperparameter optimization on the number or layers and training epochs. The layer possibilities we chose for were [16, 32] (two layers), [16,

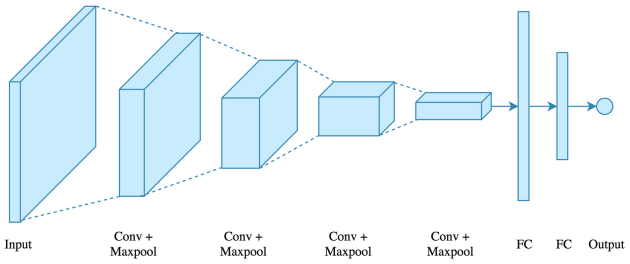

Figure 4. The figure shows the structure of a basic CNN model

32, 64] (three layers), [16, 32, 64, 128, 256] (five layers) where each number in the array corresponds to the number of filters in each CNN layers or in other words the size of the out channel. The epochs we tested were 3, 5, and 10 epochs. The loss function we chose was cross-entropy loss, and the optimizer we used was ADAM with a learning rate of 0.001. After training all the possible combinations of parameters using this setup, we found that the parameters with the best testing accuracy was [16, 32] (two layers) trained over 10 epochs.

2.3. Support Vector Machines

Compared to more basic classification models such as Naive Bayes or Logistic Regression, SVM is good at handling more complex and high-dimensional data because of its ability to apply kernel functions to map input features to a higher dimension (Redshift). Although nowadays CNN and other deep learning methods have a more prominent place in image classification, usages of SVM have been popular in past medical image classification research like the work on liver lesion by Virmani, et al. and the work on breast tumors by Huang, et al. SVMs have also shown superiority to deep learning even in situations with limited training samples availability like in the field of hyperspectral image classification (Kaul, 2022). Therefore, in our paper, we also decided to explore SVM as a comparison to our deep learning models.

For the training process, to be consistent with the way we trained CNN, we preprocessed images in the form of 100 x 100 x 3 and trained test split of 0.8, 0.2 again. Due to the high number of features of image data, we then flattened the data and scaled it using Standard Scaler and applied PCA to reduce the number of features to 66, which explains 95% of the variance. The PCA process allowed us to speed up the training and avoid over-fitting. The scaled data and fitted PCA model are then saved to files to avoid redundant works between sessions.

Then, we assessed the performance of the Support Vector Machine (SVM) with both linear and radial basis function (RBF) kernels under the default hyperparameters of the `Sklearn` library. The RBF kernel demonstrated significant performance on the validation set, achieving an accuracy of 0.85, compared to 0.78 for the linear kernel. Using the RBF kernel, we proceeded to tune the hyperparameters with grid search. Our grid search included 10 values for C and 10 numerical values for γ. To cover a larger range, they were chosen to be logarithmically spaced between 10^{-2} and 10^{10} for C and between 10^{-9} and 10^{3} for γ. Taken together, the search grid produced 100 different combinations of hyperparameters. We used 5-fold cross-validation to evaluate the performance of each combination. Finally, with the best parameters, we trained the model on the whole training set and tested it against our test set. Below is a graph showing the predictions of the resulting model.

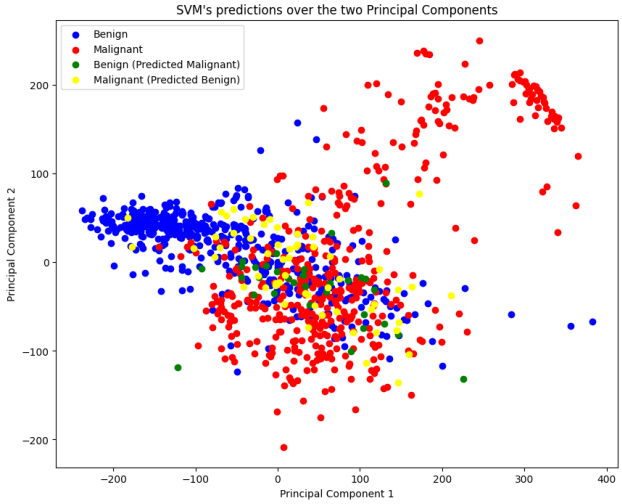

Figure 5. A scatterplot of the prediction of SVM over the first two principal components.

To better visualize the behavior of SVM, we also trained a model with only two principal components and plot its decision boundary in Figure 6.

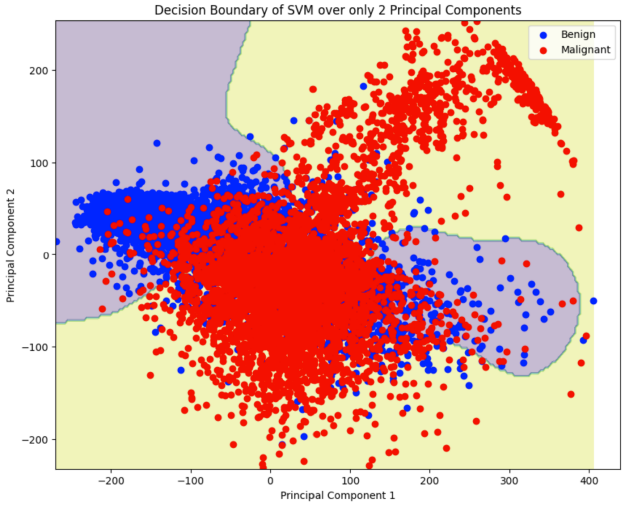

Figure 6. Decision boundary representations generated from our SVM models.

2.4. Logistic Regression

Besides SVM, logistic regression is another supervised learning model we applied to our dataset. As one of the most simple classification models, it can serve as a good baseline model for comparing all of our other models' performances. The training process for logistic regression was similar to that of SVM. We chose values for C to be logarithmically spaced between 10^{10} and 10^{20} , and then used 5-fold validation to evaluate the performance of each combination and trained the model over the full dataset with the best parameters.

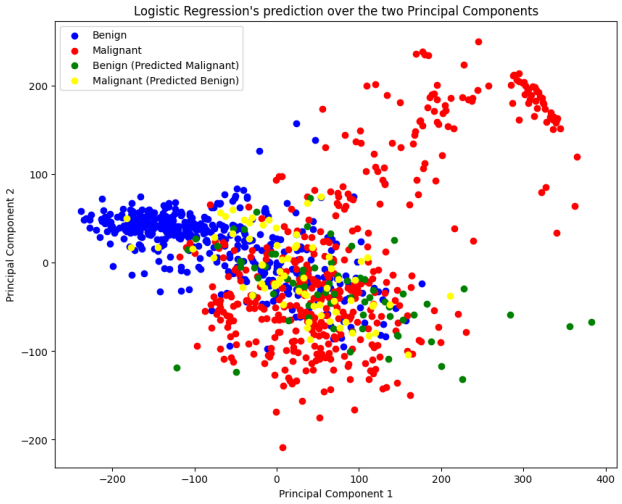

Figure 7. A scatterplot of our logistic regression model when plotted with the top 2 principal components.

2.5. Transfer Learning

Transfer learning marks another technique that we used in attempting to solve the classification problem of Melanoma skin cancer detection. The idea behind transfer learning is that a model that has been trained on a specific dataset can have the last output layer retrained on another dataset that is similar in nature. Occasionally, more layers can be added at the end of the model that is trained on the new data. This technique is often employed in situations where the data of both models are similar in nature. The idea of transferring knowledge from one domain to another originates from educational psychology theory of transfer which states that the ability to transfer knowledge comes from having experience that is generalizable (Zhuang et al., 2020). Applying this to our classification problem, we can utilize the learnings of a pre-trained model that was trained on images as our dataset also comprises of images.

The model that we used for our transfer learning models was based on the Residual Network models that were developed by He, Zhang, Ren, and Sun (2015). They recognized the difficulty in training deeper neural networks due to problems like the vanishing or exploding gradient. In theory, one would expect that as more layers are added to a network, the error of the model would decrease. However, in practice, this was not the case. A series of networks were trained by these researchers making use of residual learning blocks, as opposed to the typical learning layers, to make the models easier to train. The visualization of the residual learning block is depicted in the following figure (Figure 8).

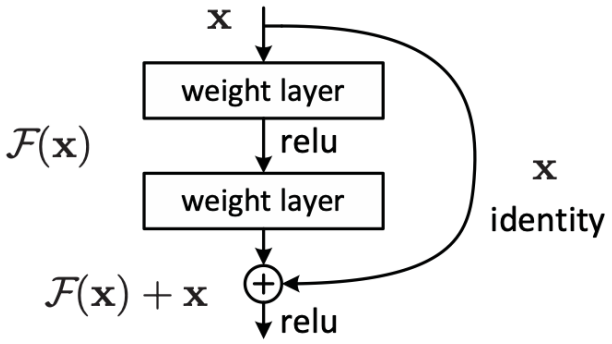

Figure 8. A residual learning block. Originally, we would have our inputs go through one weighted layer and then have the ReLU activation function applied. This would be applied one more time. With residual learning blocks, the final output is the activation function applied on the summation of the final layer values and the original input values. This is sometimes called the skip connection (He et al., 2015).

Of the various models that were trained by He and his peers, we decided to apply transfer learning to ResNet-18. This model was trained on the ImageNet dataset, which is highly relevant to our image classification problem (He et al., 2015). The original architecture of the model starts with a convolutional layer with a filter size of 7 by 7 with strides of 2. This is followed by a 3 by 3 max pooling layer with a stride of 2. This is the initial layer that is consistent with all ResNet-18 examples. We then move to the residual blocks where the ResNet-18 consists of 8 residual blocks that each have two convolutional layers. Each of these blocks utilizes a 3 by 3 filter. These blocks are divided equally into 4 sets. The number of channels in each set go from 64, 128, 256, and 512. Then, another layer is a global average pooling layer which averages all of the spatial dimensions and reduces the complexity by decreasing the number of parameters. This is then put into a fully connected layer which is used for the final classification tasks (He et al., 2015). An image of the original architecture is shown in Figure 9. In our transfer learning application, we removed the top layers of the pre-trained model and replaced it with a binary output layer: one to represent benign images and the other to represent malignant images. Further, we froze the weights of

Table 1. Performance Metrics of Different Models

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE
CNN	0.917	0.917	0.917	0.917
SVM	0.91	0.91	0.91	0.91
LOGISTIC REGRESSION	0.86	0.86	0.86	0.86
TRANSFER LEARNING (RESNET-18)	0.899	0.90	0.899	0.8986

the pre-trained model, meaning only the weights of the final layer would change during the training process. By freezing the weights, the information that was recognized from these weights would be used in the training of our model. We made use of binary cross entropy, to train the remainder of this model and furthermore utilized the Adam optimizer. We also trained the model on 10 epochs. To be consistent with the training of the other models, we preprocessed the images and used the train test split. Random flips and rotations were also added to the transformations of data for consistency.

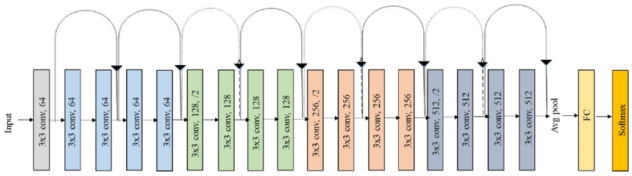

Figure 9. The original ResNet-18 Architecture. The skip connections are depicted as the curved arrows. (Ramzan et al., 2019).

3. Results

In this study, we employed various machine learning models to classify images of melanoma skin cancer, focusing on metrics such as F1-score, recall, and accuracy to evaluate performance.

Convolutional Neural Networks: We utilized a CNN, known for its efficacy in image classification, configured with layers of 2D-convolutions, ReLU activations, and max-pooling, followed by a feedforward network. After hyperparameter optimization via grid search, our best model used two layers (16 and 32 filters) trained over 10 epochs, achieving an accuracy, precision, and F1-score of 0.917.

Support Vector Machines: For SVM, we reduced the dataset's dimensionality to 66 principal components using PCA, which explained 95% of the variance. We chose the RBF kernel with a regularization parameter C of 100 after a grid search, resulting in an F1-score of 0.89.

Logistic Regression: Using a similar setup as the SVM model we created, we made use of the 66 principal components using PCA. The logistic regression model, primarily used as a baseline, yielded an F1-score of 0.86.

Transfer Learning: We adopted transfer learning with the ResNet-18 architecture, retraining the last output layer on our dataset. This approach leveraged the pre-trained network's capability to process image data efficiently, achieving an accuracy of 0.8990 and an F1-score of 0.8986.

Although our SVM has achieved a surprisingly good result, it might be due to the small size and simple nature of our dataset. In an article at the 2020 International Conference on Computer Science by S. Y. Chaganti et al., the researchers performed a similar study on image classification and found SVM initially performed well on their small dataset. However, as they increased their database size through data augmentation techniques, the performance of SVM decreased and was outperformed by the CNN model, which benefited from having more data and generalized well over noise. Our case might be similar. Although SVM's accuracy was close to our deep learning models, it is likely that it will not generalize well when data becomes more varied and includes more noise.

4. Discussion and Conclusion

In conclusion, our study thoroughly investigated various machine learning models using a robust dataset for melanoma skin cancer classification. The CNN emerged as the superior model, demonstrating outstanding accuracy and an ideal balance between precision and recall, underscoring its effectiveness in medical image diagnostics. While SVM and logistic regression showed competitive results, they fell short in handling more complex datasets—a common limitation for simpler models. Moreover, the promising results from applying transfer learning using the ResNet-18 architecture underscore the potential of using pre-trained models in specific medical imaging tasks. This project not only reaffirmed the effectiveness of CNNs in medical image classification but also opened avenues for future research such as image segmentation and the progressive monitoring of melanoma, potentially enhancing early diagnosis and treatment strategies. The collaborative efforts in various aspects of this study—from data preprocessing to model training—highlighted the interdisciplinary approach necessary for addressing complex health-related challenges through technology.

Over the course of this exploratory study, the challenge of weighing the crucial features of the model was important. While it is true that the CNN architecture proved to better perform our other models, one aspect to consider is the complexity of these models. The simpler models, such as those generated by our baseline model, would be significantly easier to explain and deduce from. Given the literature and past use cases of convolutional neural networks in the context of image classification, it is likely that going forward CNNs will continue to have better performance in identifying potential cancer occurrences in individuals. Especially in the context of cancer diagnosis, the improvement in accuracy and recall, even if not by much, is critical for patient safety, further proving that a CNN is the best model for this task.

As we took our dataset from Kaggle, we would like to discuss how our approach compares to the current models already implemented. Our study achieved similar results to the best performing models on Kaggle, but our approach brings originality in its exploration and comparison of multiple different models. In terms of overall accuracy, among the 31 existing notebooks on Kaggle, the best-performing code achieved 0.919 accuracy and 0.921 F1-score, from which our result of 0.917 accuracy and 0.917 is not far. Our code differs from all existing Kaggle notebooks by exploring supervised learning and transfer learning. The majority of notebooks only focused on deep learning techniques such CNN, GAN, and clustering. Our approach explored using PCA combined with supervised learning models such as SVM, logistic regression, and have achieved significant success with it. By doing this we showed that simpler models are able to solve this problem with a similar degree of success as more complex models; however, as discussed in the previous paragraph, for the task of cancer diagnosis, the more complex model is still the best option.

In the future, we could extend from what we have utilized in this report and use segmentation techniques of the images. This could make pre-scanning of individuals of cancer much easier. For instance, if there were occurrences of cancer in areas where individuals did not have a constant visual exposure to, such as their back, taking a photo of their entire back and running segmentation on potential cancer cites would make the model significantly more useful for individuals. Further, making note of the progression of the melanoma cancer occurrence and giving useful metrics and potential areas of concerns for the healthcare provider is another area of avenue that could stem from this work.

The risk of false negatives and false positives must also be evaluated when discussing whether machine learning techniques should be used in classification of cancer as explored in this study. As our best performing CNN model achieved an accuracy of 91.7%, inevitably, false positives and false negatives will occur if the models are used in practice. Reducing false negatives appears to be the most critical as they cause significant delay in the treatment of patients and can increase mortality rates (Burt et al., 2017). The effects of false positives are less clear as in some cases, such as breast cancer, these occurrences cause individuals to get more cancer screenings while they also do the opposite, as in the case of colorectal cancer (Taksler et al., 2018). All this to say, machine learning techniques do appear promising as an initial screen for patients. However, improving accuracy of future models and decreasing the instances of false negatives is essential for creating a more useful model. In the actual clinical setting, machine learning techniques can serve as an aid by healthcare providers, but other biopsy methods, as mentioned in the introduction, should still remain the primary method of confirming the cancer cases for patients.

References

American Cancer Society. (2024). Key Statistics for Melanoma Skin Cancer. https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html.

Burt T., Button K.S., Thom H., Noveck R.J., & Munafò M.R. (2017). The Burden of the "False-Negatives" in Clinical Development: Analyses of Current and Alternative Scenarios and Corrective Measures. Clin Transl Sci, 10(6):470-479. doi: 10.1111/cts.12478.

He, K., Zheng, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.48550/arXiv.1512.03385.

Huang, Y., Wang, K., & Chen, D. (2005). Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines. Neural Computing and Applications, 15(2), 164-169. https://doi.org/10.1007/s00521-005-0019-5.

Javid, M.H. (2022). Melanoma Skin Cancer Dataset of 10000 Images [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3376422.

Kaul, A. & Raina, S. Support vector machine versus convolutional neural network for hyperspectral image classification: A systematic review. Concurrency Computat Pract Exper. 2022; 34(15):e6945. doi:10.1002/cpe.6945.

Ramzan, F., Khan, M.U.G., Rehmat, A., Iqbal, S., Saba, T., Rehman, A., & Mehmood, Z. (2019). A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks. Journal of Medical Systems 44(2). DOI:10.1007/s10916-019-1475-2.

Taksler, G., Keating, N., & Rothberg, M. (2018). Impli-

cations of False-Positives for Future Cancer Screenings. Cancer, 124(11):2390-2398. doi: 10.1002/cncr.31271.

Verma, Mayank. "CNN Model." Medium, 8 May 2022, https://medium.com/@mayankverma05032001/binary-classification-using-convolution-neural-network-cnn-model-6e35cdf5bdbb.

Virmani, J., Kumar, V., Kalra, N., & Khandelwal, N. (2013). Characterization of primary and secondary malignant liver lesions from b-mode ultrasound. Journal of Digital Imaging, 26(6), 1058-1070. https://doi.org/10.1007/s10278-013-9578-7.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A Comprehensive Survey on Transfer Learning. Proceedings of IEEE. https://doi.org/10.48550/arXiv.1911.02685.

Perez, L. & Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. arXiv preprint arXiv:1712.04621.

Acknowledgements

All members contributed equally to the project. We all developed an understanding of the background to the Melanoma Skin cancer identification problem. Min primarily worked with the data preprocessing and identifying heuristics in the data from which we could train on. Allan worked primarily with training CNN models. David worked with training the SVM and Logistic Regression models. Aditya worked on training the Transfer Learning models. All members worked on the report equally.

A. Code Appendix

The attached code repository contains the source code and research materials for a project aimed at improving the detection of melanoma through advanced machine learning methods. It explores a variety of models such as Convolutional Neural Networks, Support Vector Machines, Logistic Regression, and Transfer Learning using the ResNet-18 architecture, applying these to the Melanoma Skin Cancer Dataset. Detailed Jupyter Notebooks are provided for each model, including notebooks for data exploration and analysis techniques like Principal Component Analysis (PCA). Users can clone the repository, navigate through the directories for each model, and follow detailed instructions in the notebooks to run experiments and evaluate the models. The project requires several dependencies, including Python, TensorFlow, Scikit-learn, and libraries for data handling and visualization like Matplotlib, NumPy, and Pandas. The code can be found here: https://github.com/allanwzhang/cs4262_project