Project Document: Final Report
Classification of Skin Lesions
MISSION_VISSION

Lucas Wurtz
Jacob Abaare
Mohammad Fahim Shahriar
Ethan Gunderson

13 May, 2024

Contents

Abstract

Detection of skin lesions at early stages can be critical in tackling melanoma (a form of skin cancer). With the addition of more malignant cases to our dataset, our auto-encoder model demonstrated superior performance, achieving an overall accuracy of 93.3%, a recall of 91.7%, and a precision of 94.8%. These results align closely with the baseline accuracy range of 90% to 96%. However, they fall slightly short of the baseline precision and recall, which range from 93% to 95.1%, respectively. Future work could focus on training our models with datasets that are diverse across racial groups to enhance their robustness and accuracy when used in non-caucasian populations.

Chapter 1

Milestone 1: Project Ideas

1.1 Introduction

Our team is especially interested in problems related to computer vision and image processing. For idea generation, we found the most useful resources were paperswithcode, huggingface, the course canvas page, and faculty here at UNL.

We initially looked at problems that seemed interesting to work on. We looked for problems that, in the future, we may find stimulating and exciting. We also considered the real-life value of solving the given problem. We then narrowed down our search space by considering the feasibility and availability of the resource (primarily the dataset). Furthermore, the relevancy of current/recent work by UNL faculty members was prioritized in our project consideration. In particular, we see Project 2 as an avenue for future collaboration.

1.2 Project Idea 1: Classifying Skin Lesions

The goal is to create an architecture capable of classifying multiple classes of skin lesions given color image data into their respective classes. Such models have practical applications in facilitating early diagnosis of skin cancers and enhancing accessibility to and efficiency of diagnostic procedures [24]. Further, exploring the viability and efficacy of human-machine collaboration to improve outcomes in health and other settings.

Generally, approaches to this problem involve pre-processing images, extracting feature information via CNN, before dense layers or an ensemble of classifiers classify the image [3]. While other work focuses on effective feature extraction through the use of attention mechanisms and capsule networks to minimize information loss [**?**]. Data is readily available, we will be taking advantage of the ISIC2019 dataset, which is a collection of over 25,000 labelled images [4][5][25]. Though review of the cited literature would reveal that there is a class discrepancy in the data, and so augmentation is likely to be necessary.

A trained model will take an image as input and return either a one hot or affine vector with its class prediction.

1.3 Project Idea 2: Plant Phenotyping

The goal of plant phenotyping is to identify and classify the various growth stages of flower crop plants to determine overall flower health, which is a principle factor in crop growth and production. Hence, this is a multiclass classification problem. Practical applications include automated crop production detection, manual labor minimization through automation, and yield optimization [14]. By focusing on these issues, plant phenotyping seeks to improve crop yields and increase food security in the long run.

Approaches involve pre-processing images, feature extraction using CNNs, and classification through dense layers. The input can be a set of 2D images or a 3D image and the output will be a probability distribution of the model's confidence in various phenotypes. Data is available via UNL's Plant Vision Initiative as the FlowerPheno dataset [6]. The dataset comprises RGB image sequences of three flowering plant species: sunflower, canna, and coleus.

The total dataset contains 17,022 images, which would be divided into training and test datasets. Each image in the dataset has a resolution of 420 x 420 pixels. To enhance the size and quality of the training datasets, data manipulation techniques such as data augmentation will be employed. These techniques increase the variety of the dataset through various transformations, such as random horizontal flipping, scaling, and adjusting contrast. One challenge to this approach is identifying overlapping leaves. To mitigate this, approaches such as visual growth tracking and semantic segmentation can be used. Visual growth tracking monitors the edges of leaves to determine various types of overlaps [1]. Additionally, a method known as "Deep Coloring" can simplify instance segmentation to semantic segmentation, which will enable the identification of individual leaves while simplifying training [10]. A trained model will take an image and predict the type of flower and its growth stage with a confidence interval.

1.4 Conclusions

We consider both of these projects to be interesting as well as applicable in future endeavors. They also share some distinct similarities, being related to computer vision. Firstly, the importance of feature extraction in both problems. Identifying and extracting important information from the images will be essential in achieving comparable test performances; as a consequence, effective image pre-processing is also critical. We also have the issue of data. Computer vision tasks are often data-hungry, and any discrepancies in the data will need to be accounted for. As we mentioned, conscientious data processing and augmentation will be necessary to address these issues.

As a group, we consider whichever has more agreeable data to be the preferred candidate. Given the difficulties that potentially unlabeled data may present, it may be more feasible to select problem 1. However, if we can secure a higher-value dataset for problem 2, then it is also a strong candidate for the final project.

Table 1.1: Contributions by team member for Milestone 1.

Team Member	**Contribution**
Lucas Wurtz	Introduction; Conclusion; Project Idea 1: sources, data, summary
Ethan Gunderson	Project Idea 2: sources, data, summary
Jacob Abaare	Project Idea 2: summary, data
Mohammad Fahim Shahriar	Introduction; Project Idea 2: Proposal, Background Study

Chapter 2

Milestone 2: Project Selection

2.1 Introduction

After thorough deliberations and research, our team has opted for Project Idea 1: the development of a deep architecture designed for the classification of skin lesions. While Project Idea 2 also presented an interesting problem for us to explore, our evaluation of the time and resources required for its realization within a three-month time frame seemed not feasible. Furthermore, the significance of Project Idea 1, particularly in its potential to aid in the early diagnosis of cancers—a matter of great importance today—compelled our choice of the project. Our decision reflects our commitment to contributing meaningfully to critical health concerns, aligning with our interest in leveraging computer vision and image processing for impactful real-world applications.

2.2 Problem Specification

2.2.1 Project Statement

We have decided on classifying skin lesions. The goal is to train a model to take an image or images of a skin lesion and predict whether or not the growth is malignant or not. One of the biggest difficulties lies in the high variance of our data, and being able to reliably extract useful information will be of utmost importance.

2.2.2 Motivation

The motivation behind this project is to increase the detection rate of skin cancer. Skin cancer is a common illness that requires early detection for effective treatment. Many health experts rely on visual examinations and small skin

samples to find and treat skin cancer before it spreads throughout the body. However, these methods of detecting skin cancer may not always provide an accurate or confident answer. This model seeks to accurately detect malignant skin lesions in the early stages so that the model can be incorporated into the existing healthcare system to aid in the early treatment of skin cancer.

2.2.3 Auxiliary Resources

We will require the ISIC2020, PAD-UFES-20, and HAM10000 datasets. The ISIC2020 dataset is a benchmark for skin cancer images and includes testing data [17]. The PAD-UFES-20 is a collection of images taken by phones [15]. The HAM10000 is a collection of images focused on providing a diverse set of cases and includes over 1100 cases of melanoma [23]. For Python packages, we will of course be using TensorFlow for models. For image augmentation, we will use the Albumentations package and in some cases, we will use Lime to help us visualize what our networks are focusing on.

2.3 Related Work

There are many existing, successful approaches to this problem. One of which employs a large ensemble of capable feature extractors, with the inclusion of a secondary network that incorporates given metadata (such as patient age, location, etc.) [9]. This approach is a bit beyond our computing capabilities; however, we think that an ensemble would be one of our best options.

If we wanted to avoid training many different models we would want to focus on efficient feature extraction, other successful approaches have made this the forefront of their models. Such as the approach from Zhangli Lan et al. which implements attention inside its convolutions blocks [11]. In contrast to our task, the FixCaps approach was used for multi-class classification, whereas we are simply concerned with whether or not a case is melanoma or not. We believe that more discriminative models like these have the potential to perform well in the binary malignant task.

Further augmentation strategies were gleaned from successful implementations [9][22]. Potential pitfalls when working with dermoscopic images were explored by Pewton et al. While some images in the dataset contain a circular black artifact (due to the process of dermoscopic imaging), they didn't present strong conclusions as to the effect [16].

2.4 Proposed Method 1: Transfer Learning Ensemble

Method one involves transfer learning on the following backbones: Xception, ResNet50v2, EfficientNetB4, ConvNextTiny, and ResNet101. In the final, trained,

ensemble, each model would be fed the input image and would classify it as malignant or not, and the ensemble would predict the most common class. The backbones were chosen for their sizes, they are relatively memory efficient at around 100MB for the weights. This should allow us space for batches of our data, making training and testing more efficient. Further, pre-trained models are all publicly available through TensorFlow applications. Given the ease of transfer learning through the TensorFlow platform, we would like to test with larger models as well, if possible.

The datasets we will use for training and testing are splits of the ISIC2020 and PAD-UFES-20 datasets [17][15]. From each, we split 10% of the confirmed melanoma (positive classification target) and 10% of the non-melanoma cases to be reserved for the test set. Resulting in roughly 30000 train samples and 4000 test samples. The training data will be preprocessed in several ways using the Albumentations Python package. We will be applying the following commonly used augmentations to our melanoma-class images in the training set: random rotation and translation, vertical and horizontal flips, random brightness and contrast changes, cutout, and CLAHE. The image will also be resized to match the expected input size of the given backbone. These augmentation strategies have shown to be effective, cutout especially seems to be effective at slowing the rate of over-fitting, which is an exceptional risk when re-using a limited number of training instances as is the case with our positive melanoma samples.

To evaluate the models' and overall ensemble performance, we will use the recall and F1 scores as evaluation metrics. Given the nature of the problem, we are particularly interested in minimizing the false negative rate. To put it bluntly, if you have cancer, it is imperative you know about it. Monitoring recall gives us a sense of our false negative rate, the F1 score is there to evaluate the performance through the class discrepancy. Furthermore, the use of LIME will help us confirm or deny that the models are using reasonable information in their predictions (we don't want to focus on the surrounding skin or any visual artifacts). Mid-checks would include evaluating each backbone's individual performance on the task and deciding on the final selection of models included in the ensemble.

The costs and risks are minimal. The ensemble uses publicly available and relatively lightweight backbones, such that the training and inference for each are inexpensive. The most substantial consideration is the time required to fine-tune each of the ensemble models on the dataset. The applicable risks lie in the model either over or under-predicting malignant cases. However, given that this would be a tool used in conjunction with human health experts and that this is a screening tool only, the risks are quite low.

2.5 Proposed Method 2: Stacked Autoencoders with Classification Model

Method two involves stacking encoders and integrating a classifier to determine the presence of melanoma or not. It involves two critical stages. First, input images are fed into a stack of sparse autoencoders. Each autoencoder extracts characteristics of an input image, which is then reconstructed as output. The output is then given as input to the next autoencoder to extract a new set of features. This enables each autoencoder to extract a smaller, but ideally more important, set of features as it moves down the stack [18]. Once rich features have been extracted, the outputs are fed to a classification model. The classification model could be either a Support Vector Machine (SVM) or simply an activation layer [26]. By using both autoencoders and a classification layer, this approach seeks to overcome the data imbalance issue, but quality datasets and effective preprocessing techniques are necessary.

We intend to rely on the publicly available datasets used for skin lesion classification and analysis, such as the ISIC2020, HAM10000, and PAD-UFES-20 for the training and testing of the autoencoders [17][23][15]. In total, the three datasets contain 45439 training images. We begin by preprocessing the skin lesion images, which include normalization to ensure computational efficiency during model training and resizing to the dimensions of the images to match the encoder's architecture. To ensure the generalization ability of the trained encoder, we will consider data augmentation techniques such as rotation and flip to increase the diversity of the dataset. The dataset will then be partitioned into 80% for training, 10% for validation, and 10% for testing. This separation allows for the evaluation of the model's performance and its generalization capability. The autoencoder is then trained to reduce the reconstruction error through an iterative process. Features learned through the stacked encoders are then extracted and built into a classification model that classifies images as malignant or not.

To evaluate the performance throughout the project, several metrics will be used. During the training phase, reconstruction error metrics such as Mean Squared Error will be used to monitor the autoencoder's ability to extract necessary features. Once the stack of autoencoders has been incrementally trained, the classification model will be evaluated. The classification model will be evaluated with standard metrics such as accuracy, precision, recall, and F1-score. For final checks, generalization will be measured using AUC-ROC scores. In both approaches, minimizing false negatives is very important, so monitoring recall will be crucial.

The risks involved with this approach include a high chance of overfitting. With a high data imbalance and the convoluted nature of autoencoders, data augmentation is critical to ensuring a robust model. Techniques such as dropout, regularizers, and cross-validation can also be used to lower this risk. Additionally, the computational cost of incrementally training each layer of the stacked autoencoders and measuring each layer's reconstruction error is high. To miti-

gate this, training each layer can be done in an unsupervised manner, and then the entire stack can be fine-tuned via supervised training [18]. Overall, the risks and costs are relatively low as long as the model is used as an assisting tool by health experts and is periodically updated with essential feedback.

2.6 Conclusion

In conclusion, we explored two distinct methods for classifying skin lesions, which is crucial in melanoma detection. Method 1 focused on transfer learning and ensemble modeling, showcasing how backbone networks such as ResNet50v2, and EfficientNetB4 can be modified to tackle classifying skin lesions. Method 2, on the other hand, focused on stacked encoders integrated with a classifier to solve the same problem. While Method 1 showed the flexibility and scalability of transferred learning, where existing resources can be optimized to solve other problems, Method 2 highlighted how auto-encoders can be vital in handling data imbalances, which is a common hurdle in most labeled medical datasets. As a group, our final choice in the approach to explore moving forward will heavily depend on our preliminary investigations, where we will consider both model performances and ease of implementation to make a choice.

Table 2.1: Contributions by team member for Milestone 2.

Team Member	**Contribution**
Lucas Wurtz	Research, Problem statement, auxiliary resources, related work, all of method 1
Ethan Gunderson	Research, Motivation, Method 2
Jacob Abaare	Research, Introduction, Method 2, Conclusion
Mohammad Fahim Shahriar	Background Study, Editing, Abstract

Chapter 3

Milestone 3: Progress Report 1

3.1 Introduction

Our goal is to classify malignancy given an image of a skin lesion. We propose a stacked autoencoder approach. We first train an autoencoder to reproduce cropped images of skin lesions. In future reports, additional autoencoders will be incrementally trained, and their outputs will be fed to the next autoencoder to create a stack of autoencoders. After training, a classification module is to be attached and fine-tuned. As of yet, our results are promising. We have successfully trained two autoencoders to reproduce images from our population distribution.

3.2 Related Work

Autoencoders have found extensive applications across various domains for image classification tasks[2][13]. Depending on the type of problem being solved, autoencoders have found widespread uses for denoising, anomaly detection, data augmentation, as in the case of variational autoencoders, and segmentation. In this project, we mainly focused on gleaning ideas from successful implementations of autoencoders for feature extraction and classifying. One such method used a deep autoencoder (AE) classifier to distinguish between malignant and benign lung nodules using two input features: appearance and geometric features[20]. In our case, we are focused on the geometric features (shape) as the appearance features require modeling the lung nodule using the 7th-order Markov Gibbs random field (MGRF) model, which is complex and computationally intensive for less complex images like ours.

Moreover, Gogoi and Begum explored the classification accuracy of using a support vector machine (SVM) and a softmax regression function as a classifi-

cation layer to the encoder part[8]. Results showed that SVM showed superior performance in handling both binary and multiclass classification effectively, influencing our choice to use SVM as our classifier in future milestones.

Variational encoders have been reported to effectively perform data augmentation to compensate for class imbalances[19]. However, the potential pitfall of adding noise to the images during the augmentation and propagating the extracted noisy features for classification significantly impacts accuracy. To mitigate this issue, we strategically adjusted our approach by performing data augmentation after the autoencoder extracted features from the original images. This sequence ensured that the augmentation process did not compromise the integrity of the feature set used for classification, allowing us to enhance our dataset while preserving the quality of features essential for accurate model performance.

3.3 Experimental Setup

Our initial testing saw us training our autoencoder to reproduce images from the ISIC2020 dataset. The ISIC train and test data was split into 80% train and 20% validation sets. The autoencoders were trained for up to 15 epochs. The overall validation score is the metric of concern, given this is pretraining the model to extract useful features. No test set was used.

3.3.1 Architecture A

The first architectures are largely proofs of concept. Architecture A utilizes multiple residual convolutional blocks in its encoder, to aid in the passing of gradients downstream as well as using more available information. The latent space generated by the model is $95 \times 95 \times 8$, down from the input image size $380 \times 380 \times 3$. Which is a decrease of around 83%. Future efforts will go to further reducing the dimension of the latent space, balanced with reconstruction accuracy. Since decreasing the size of this space lends itself to the efficient usage of our limited memory.

In developing the model, we took inspiration from Zhou et al. [27]. In particular, the organization of the deconvolution block, in which the paper implements conv2DTranspose operations along with conv2D operations and achieved very good results in processing medical images. Implementing it in this architecture, at first glance, appears promising.

3.3.2 Architecture B

The second architecture was focused on reducing the number of parameters to quicken training time. Architecture B uses two convolutional layers in the encoder and three convolutional transpose layers for the decoder. The latent space generated is $95 \times 95 \times 64$. This allows the encoder to extract the most important features while maintaining a large variety of features. After encountering

memory issues using this architecture, further reduction of the latent space is necessary. Including pooling layers would provide the reduction without significant loss of data. Additionally, using dense layers described by Sabbaghi et al. may provide the reduction, but may also increase the training time [18].

3.4 Experimental Results

For each of the autoencoders, we iterated through training and report the final validation loss. Further, we present figures of recreated images to demonstrate the recreation ability of the models.

Table 3.1: Performance Evaluation

Architecture	Validation Loss (MSE)
A	0.0021
B	0.000083

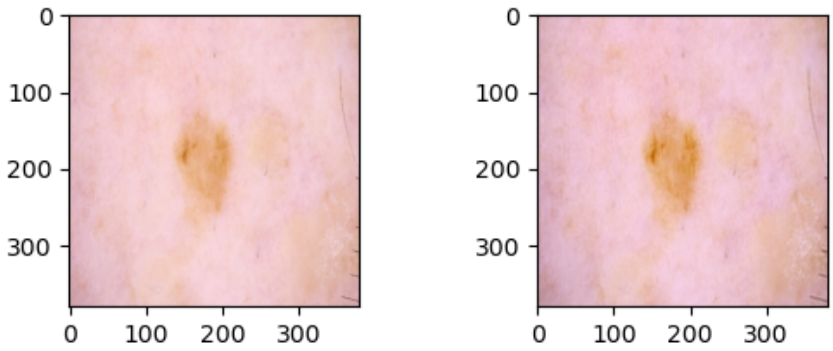

Figure 3.1: Architecture A Image 0 Recreation

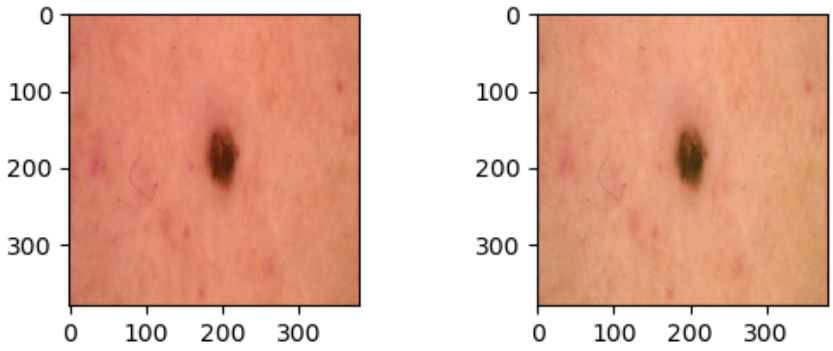

Figure 3.2: Architecture A Image 1 Recreation

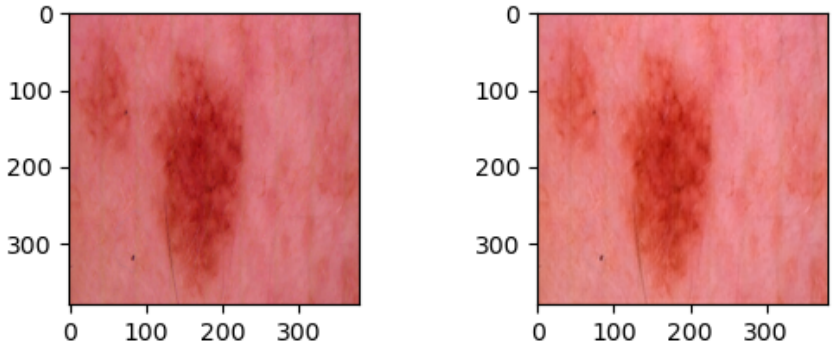

Figure 3.3: Architecture A Image 2 Recreation

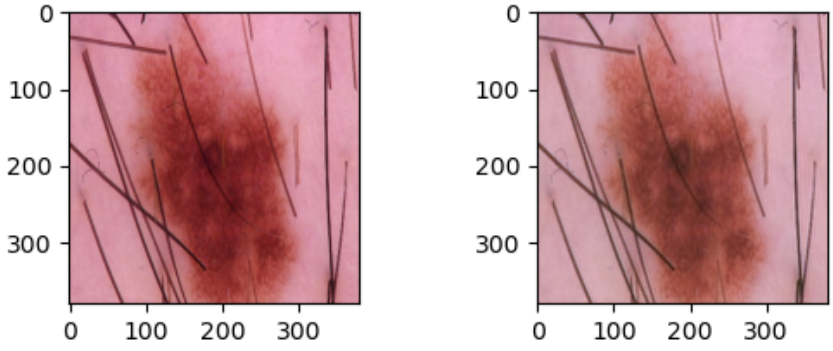

Figure 3.4: Architecture A Image 3 Recreation

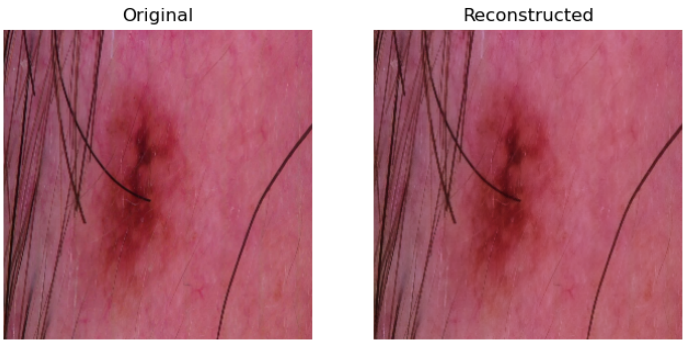

Figure 3.5: Architecture B Image Recreation

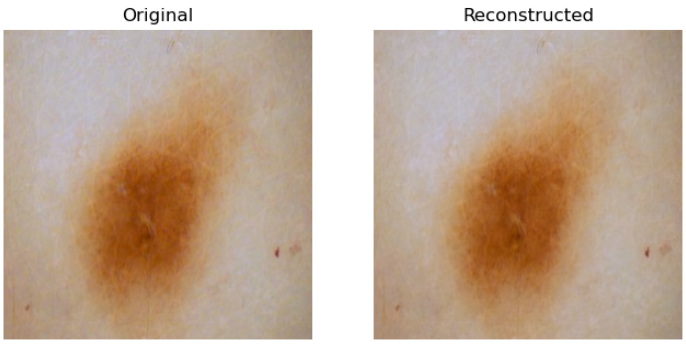

Figure 3.6: Architecture B Image Recreation

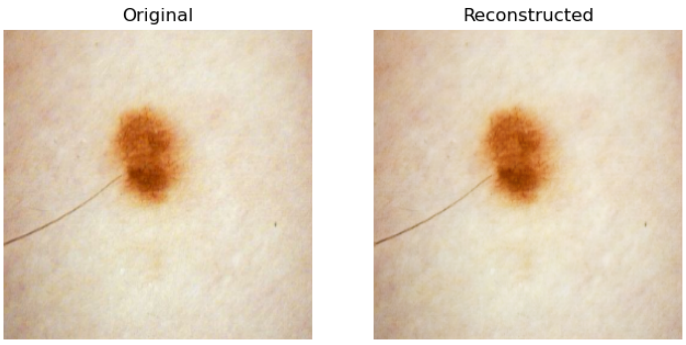

Figure 3.7: Architecture B Image Recreation

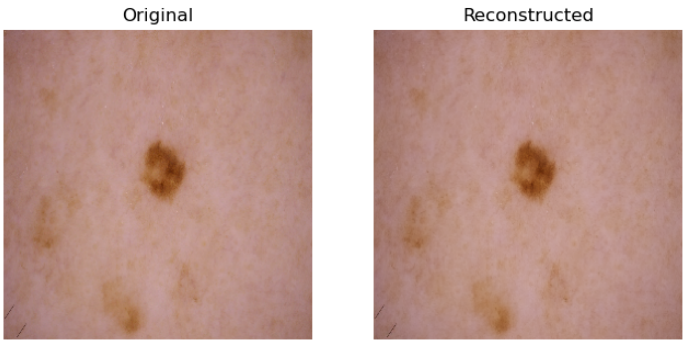

Figure 3.8: Architecture B Image Recreation

3.5 Discussion

From our experimental findings, we found significant differences in performance. Model A, with a latent space of $95 \times 95 \times 8$, had a reconstruction loss of 0.0021. Model B, with a latent space of $95 \times 95 \times 64$, had a reconstruction loss of 0.000083. The differences show the impact of latent space size on accuracy. Model B was able to retain much more information than Model A, but required much more memory. Because we plan to use an SVM classifier, efficient memory usage is crucial. This means implementing a more memory-efficient structure as well as increasing robustness to prevent potential memory loss. Experimenting with more pooling layers is a key consideration for building the next autoencoders.

In terms of image reconstruction, we can see a similar level of image recreation. However, redundant features not necessarily useful for the classification, such as hair and additional noise are apparent. More work can be done in denoising to reconstruct only essential target features.

3.6 Work Plan

Moving forward, we aim to further refine our architecture to achieve good image reproduction with as small a latent space as possible. Further, completing the end-to-end classifier. One possible continuation is to train two autoencoders on malignant/benign splits to compare the features they extract. Also an option is using these models in an ensemble mimicking a MoE strategy. We have created a tentative schedule outlining important milestones to complete in the next month.

Table 3.2: Task Schedule

Week of	Tasks to complete
April 1	Complete classification blocks for encoders & organize supervised datasets
April 8	Train and Evaluate autoencoder-classification models
April 15	Iterate Autoencoder design based on results/feedback
April 22	Finalize autoencoder model architecture/s and Evaluate
...	Determine further work (ensemble, transfer models)

3.7 Conclusions

The selected approaches have shown us promising results. The models are recreating the input images very accurately. This implies that the features we have extracted are good; however, we are eager to reduce the size of the latent space. We think we can decrease the dimension of the space without significant sacrifices to performance. Further, once the classifiers have been trained, exploring more avenues to extract performance will likely be necessary. This would in-

clude but not be limited to, changes to the architectures themselves (through additional layers, attention, etc.) as well as aggressive data augmentation.

Table 3.3: Contributions by team member for Milestone 3.

Team Member	**Contribution**
Lucas Wurtz	Architecture A, Results, Intro, Conclusions, Edits
Ethan Gunderson	Architecture B, Results, Discussion, Edits
Mohammad Fahim Shahriar	Background Study, Discussion, Edits
Jacob Abaare	Related Works, Edits

Chapter 4

Milestone 4: Progress Report 2

4.1 Introduction

Skin lesion classification is a challenging task. Our approach in milestone 3 was to use stacked autoencoders with an SVM classification layer. In our last report, we mentioned out-of-memory issues when classifying, which have been addressed by using the HAM10000 dataset [25]. The HAM10000 dataset contains about 20% malignant cases, which decreases the amount of data points necessary for SVM classification. However, we found that this method wasn't producing meaningful results as it would either predict largely malignant or largely benign no matter which classifier was used. The classifiers used were linear support vector classification, a random forest classifier, and a dense neural network. We have since made the decision to explore the transfer learning ensemble, which has provided promising results. We have trained a classifier based on our auto-encoders and we have transferred multiple models to the task via transfer learning. Lastly, these models were tested both separately and as an ensemble.

To evaluate the performance of the 5 models, we relied on accuracy, precision, and recall as our evaluation metrics since the need for accurate diagnosis against the risks of incorrect or missed diagnoses is critical for deploying models such as ours. Precision measures the accuracy of positive predictions, which is vital in determining whether malignancy is present. On the other hand, recall measures our model's ability to accurately find all malignancy conditions within the ISIC2020 dataset. Employing both precision and recall helps affirm the accuracy and reliability of the results of our model, which is critical for making medical decisions. The results are promising, shown in full in Table 4.1.

Notably, the overall performance of our ensemble currently is around 80.5% accuracy, with slightly lower precision at 75.6%, notably our recall benefited the most relative to individual performances, with a recall of 90% on our limited test set.

4.2 Related Work

Both transfer learning and an autoencoder integrated with a classifier have proven to be effective methods for classifying medical images accurately. For example, Showkat and Qureshi, suggested enhancing the classification accuracy of chest X-ray images by dynamically adjusting the weight ratios between ResNet-101 and ResNet-152 blocks [21]. Despite the proven successes of ResNet-101 and ResNet-152, we chose to implement two variations of ResNet-50, specifically version 1 and version 2, due to their lower computational demands.

Moreover, the successful applications by Lu and Zadeh, as well as Geetha and Prakash, utilizing XceptionNet and EfficientNetB4 blocks for classifying medical images—including conditions such as melanoma and glaucoma with high accuracy inspired us to adopt and utilize them for determining the malignancy of our images [12],[7]. Leveraging their key benefits, such as efficient parameter usage and scalability, we investigated how they performed relative to the other models.

Furthermore, we gleaned ideas from successful implementations of autoencoders for feature extraction and classifying. One such method used a deep autoencoder (AE) classifier to distinguish between malignant and benign lung nodules using two input features: appearance and geometric features [20]. In our case, we are focused on the geometric features (shape) as the appearance features require modeling the lung nodule using the 7th-order markov gibbs random field (MGRF) model, which is complex and computationally intensive for less complex images like ours.

4.3 Experimental Setup

The models used in the ensemble were; Xception, EfficientNetB4, ResNet50 (both V1 and V2), and for our fifth model we used an autoencoder that we made from scratch, which is a bottlenecked version of Autoencoder A (henceforth referred to as AE) as shown in Figure 5.1. Where the latent space was tightened from $95 \times 95 \times 8$ to $24 \times 24 \times 3$, achieving a similar loss as previously achieved. All backbones were connected to a relatively simple classification network.

The data used in this test came solely from the ISIC2020 dataset [17]. First, 150 malignant and 150 benign (300 total), were split from the available labeled training data. Then a validation set of 50/50 and a test set of 100/100 malignant/benign photos was established. This validation/test set was constant across all trained models. The remaining data was used for training, with the remaining malignant images, we applied random augmentations and injected them into the training set according to our previous plans. Further, we used TensorFlow's built in feature to sample from datasets to change the frequency of seeing malignant images in training. The probabilities for seeing true malignant, benign, and augmented malignant photos were 0.05, 0.85, and 0.1, respectively. This was done to reduce the intensity of our class imbalance. However, it was done at the cost of oversampling from our target class, which we conclude af-

fected the ability to train the models for longer periods of time. We are still left with a class disparity, however; so we also applied a weighted loss based on the function:

$$weight_i = N/(N_i * 2) \tag{4.1}$$

Where N is the size of the dataset and N_i is the number of instances of class i, and this value was adjusted based on training performance (i.e, it would sometimes encourage blanket positive predictions, we would decrease the weight of the positive class).

All models were trained in the same manner, the autoencoder and backbones had their weights frozen and an identical classification layer was used for each. They trained on 500 steps per epoch with each step consisting of a batch of 32 images before running on the validation set, the validation loss was the observed metric for early-stopping, weight-saving, etc. The metrics recorded were accuracy, precision, recall, and both the PR and ROC AUC. After training each model, they were tested individually on the test set before being tested as an ensemble.

To get the ensemble's prediction we took the sum of all predictions and compared it to the value:

$$0.5 \times n_m - gamma \tag{4.2}$$

Where n_m is the number of models in the ensemble and *gamma* is a hyper-parameter to tune the sensitivity of positive predictions. We also tried simply taking the most common prediction, however that approach yielded less performance.

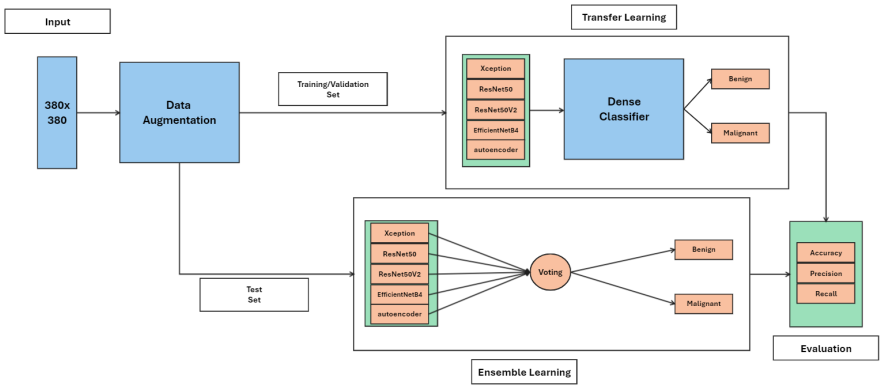

Figure 4.1: Transfer Learning Ensemble

4.4 Experimental Results

The final test results of the models are given in Figure 4.1. We see that the transferred models all performed similarly; however, the characteristics of the precision and recall of the models differentiate them and demonstrate the validity of the ensemble approach. Some of the models have a higher recall with lower precision, while others reverse that, or try to balance the two. This leads to the ensemble having a generally higher accuracy than any single model achieved while taking a strong combination of recall and precision. With the ensemble achieving 80% accuracy, with 90% recall and 75.6% precision, the results can certainly be improved.

Table 4.1: Model Test Results

Model	Accuracy	Recall	Precision
ResNet50	.775	.840	.743
ResNet50V2	.750	.82	.7193
Xception	.775	.93	.709
EfficientNetB4	.755	.660	.815
AE	.770	.880	.721
Ensemble	.805	.90	.756

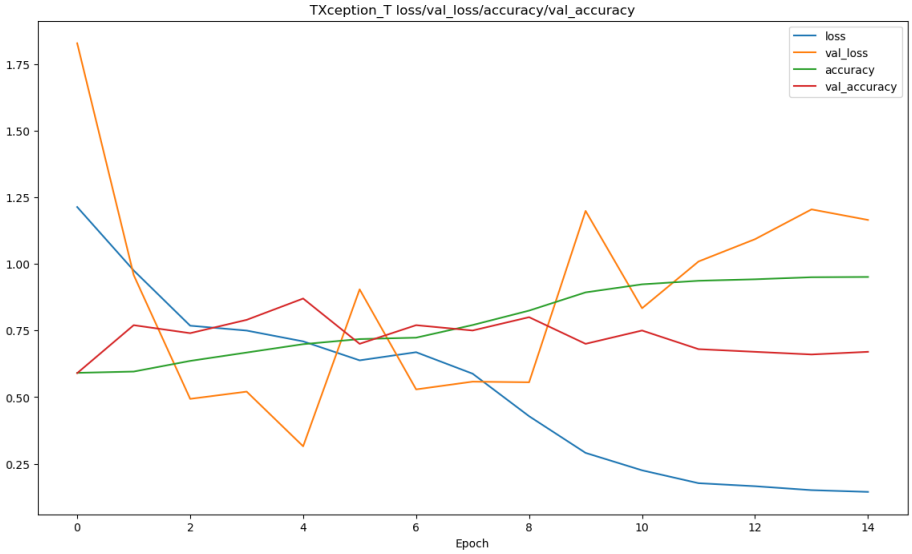

Figure 4.2: Xception Train/Val Loss and Accuracy

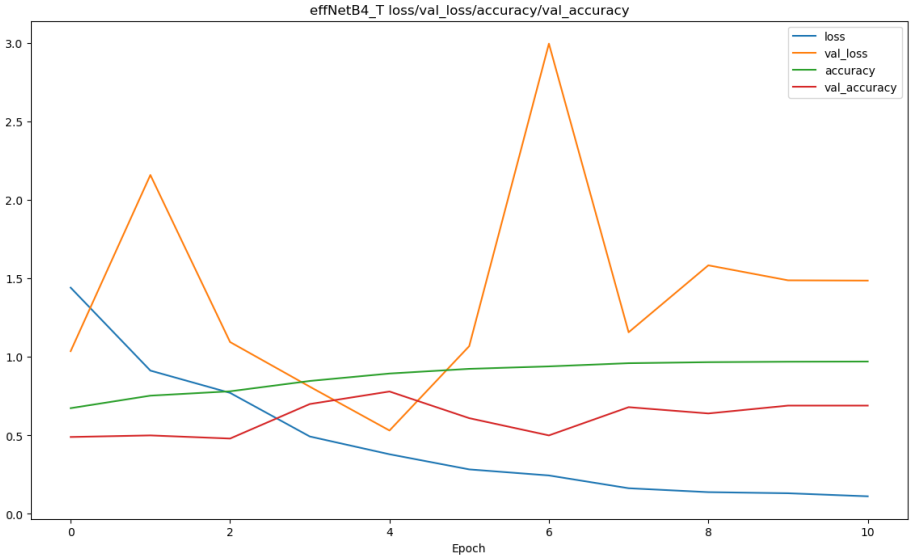

Figure 4.3: EfficientNetB4 Train/Val Loss and Accuracy

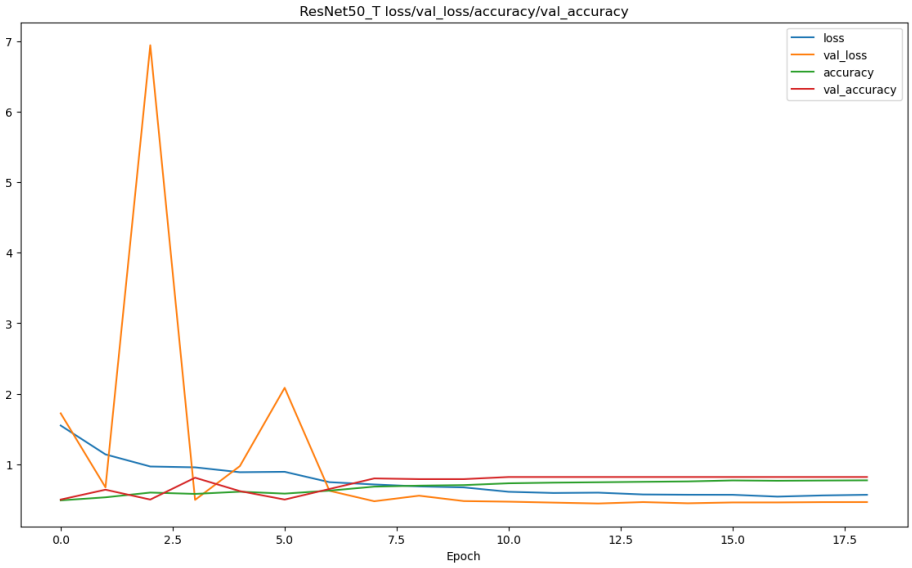

Figure 4.4: ResNet50 Train/Val Loss and Accuracy

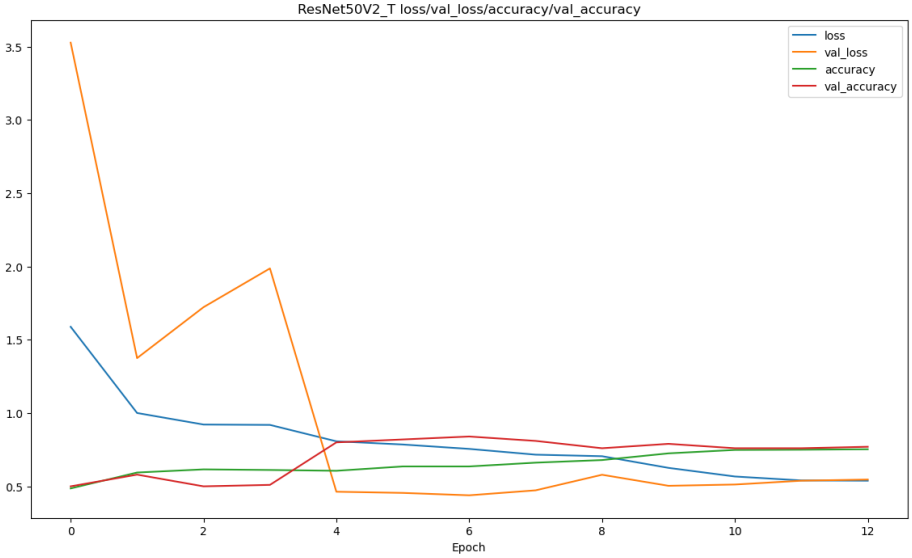

Figure 4.5: ResNet50V2 Train/Val Loss and Accuracy

4.5 Discussion

While our results are promising, we acknowledge there is much room to improve. Firstly, as we mentioned in the 4.3, a few of the models struggled as their training went on. Particularly the transferred Xception model, as Figure 4.2 shows, the model experienced very unstable validation loss scores and the validation accuracy stagnated while the train loss and accuracy improved. This was similar in the EfficientNetB4 model as well, Figure 4.3. The ResNet models behaved a bit differently and seemed to be more well-behaved in training - Figures 4.4 4.5, with the validation and training loss seeming to converge, much more usual training.

We see that the individual models clustered around 76-77% accuracy but have their own distinct behaviors, with there always being the trade-off between recall and precision - Table 4.1. This behavior may be something we can take advantage of by means of weighting predictions based on the model's specific proclivities during training (for instance, if a model is very precise but has low recall and predicts positively, we should give it more weight than a low precision, high recall positive guess).

A major point that we need to recover from is the very high likelihood that we are not representing the target population very well. This stems from the issue of the minuscule number of positive instances we began with that limited us to such small validation and testing splits. We simply need more information to have any confidence in the results. We think this can be easily remedied by the introduction of the datasets proposed earlier in the paper. We would simply introduce this data into the train/validation/testing splits and resume training from the current state of the models, only using more accurate validation and testing.

4.6 Work Plan

As the discussion may suggest, the introduction of more data is first and foremost what we want to accomplish along with the accompanying data augmentation. After that, further training of our classification layers before another round of testing. Then we will probably explore the fine-tuning of layers in our backbones.

Table 4.2: Task Schedule

Week of	Tasks to complete
April 28	Increase datasets used and augmentation strategies
May 5	Finalize reports and presentation

4.7 Conclusions

From this milestone, we have determined that the transfer learning ensemble approach is what will produce the best results. By combining multiple models, we hope to achieve high overall accuracy, which is highly important in cancer classification. Our results from this milestone are promising and by increasing the amount of data, we hope it will further improve overall accuracy and prevent overfitting. Specifically, introducing more malignant cases will hopefully improve precision and recall, which are the most relevant metrics in cancer classification.

Table 4.3: Contributions by team member for Milestone 3.

Team Member	**Contribution**
Lucas Wurtz	Transfer Models and Ensemble, write-up, tables/figs
Ethan Gunderson	Autoencoder training, intro, conclusion
Mohammad Fahim Shahriar	Background Study, Introduction, Abstract, Figures
Jacob Abaare	Introduction, Related Works, General Edits

Chapter 5

Milestone 5: Final Report

5.1 Introduction

The primary aim of this project is the classification of skin lesions, a notably complex task. The ultimate goal is to determine the probability of malignancy from images of skin lesions suspected of skin cancer. Continuing from Milestone 4, we utilized transfer learning models and a classifier integrated with our auto-encoder. As noted previously, each model was tested individually and then in combination as an ensemble. A significant checkpoint in this milestone was addressing the class imbalance by augmenting our training dataset with additional malignant cases. This involved adding malignant images from the HAM10000 and BCN20000 datasets [25] into the existing ISIC2020 dataset [17]. Overall, 7 distinct models, including transfer learning models, auto-encoders, and ensemble model, were developed and evaluated.

Our evaluation metrics did not change as we still relied on accuracy, precision, and recall to evaluate the performance of the 7 models. This is necessary to ensure that models are robust against the risks of incorrect or missed diagnoses which is critical for deploying models such as ours. Precision measures the accuracy of positive predictions, which is vital in determining whether malignancy is present. On the other hand, recall measures our model's ability to accurately find all malignancy conditions within the combined dataset. Employing both precision and recall helps affirm the accuracy and reliability of the results of our model, which is critical for making medical decisions.

The results, detailed fully in Table 5.1, are promising. The overall performance of our ensemble model achieved an accuracy of approximately 80.5%, with a slightly lower precision of 75.6%. Notably, the recall saw the most significant improvement, reaching 90% on our combined test set. It is also important to highlight the exceptional performance of our auto-encoder, which achieved an accuracy of 93.3%, a recall of 91.7%, and a precision of 94.8%, all of which was in range in terms of the baseline accuracy of 90% to 96% [21] but a little short of the baseline precision and recall which were both between 93% and

95.1% respectively.

5.2 Related Work

Throughout the literature review in this project, our focus has been on identifying the most effective methods for classifying medical images and discovering the best augmentation strategies, as evidenced by the works cited below. For instance, Showkat and Qureshi recommended enhancing the classification accuracy of chest X-ray images by dynamically adjusting the weight ratios between ResNet-101 and ResNet-152 blocks [21]. The successful applications by Lu and Zadeh, as well as Geetha and Prakash, who utilized XceptionNet and EfficientNetB4 blocks for classifying medical conditions such as melanoma and glaucoma with high accuracy, inspired us to adopt these models for determining the malignancy of our images [12], [7].

Additionally, we explored the successful implementation of autoencoders for feature extraction and classification. One notable method involved using a deep auto-encoder (AE) classifier to differentiate between malignant and benign lung nodules based on two input features: appearance and geometric features [20]. In our approach, we focused on geometric features (shape), as the appearance features require modeling using the 7th-order Markov Gibbs random field (MGRF) model, which is complex and computationally intensive for less complex images like ours.

With insights from successful augmentation strategies and best practices [9], [22], our efforts in this milestone concentrated on how the above models performed relative to our developed models. From all the above models, it was realized that the models had an accuracy around 90% to 96% and a precision and recall ranging around 93% to 95.1%. We used these metrics as benchmarks to evaluate the effectiveness of our models.

5.3 Experimental Setup

To address the class imbalance issue, we tried adding the malignant images from the HAM10000 and BCN20000 datasets to the ISIC2020 dataset. The ISIC2020 dataset only contains 584 malignant images, which is only 1.76% of the total dataset. After adding 6094 malignant images from the BCN20000 dataset and 1954 from the HAM10000 dataset, the total amount of malignant images in the combined ISIC2020 dataset was 8632, or 20.96% of the total dataset. With the added images, we increased the test size to 2048 and validation size to 1024 images with a split of 50/50 for malignant/bengin.

The combined dataset was subject to the same weighting and augmentation as the previous milestone. The augmentation strategies used were horizontal/vertical flips, random brightness, and random rotations. During training, the frequency of each class was modified to sample malignant, benign, and augmented malignant with probabilities of 0.05, 0.85, and 0.1, respectively. A

weighted loss was also applied based on the function:

$$weight_i = N/(N_i * 2) \tag{5.1}$$

Where N is the size of the dataset and N_i is the number of instances of class i. This value was adjusted based on training performance (i.e., it would sometimes encourage blanket positive predictions, so we would decrease the weight of the positive class).

All models were trained in the same manner, the auto-encoder and backbones had their weights frozen and an identical classification layer was used for each. They trained on 500 steps per epoch with each step consisting of a batch of 32 images before running on the validation set, the validation loss was the observed metric for early-stopping, weight-saving, etc. The metrics recorded were accuracy, precision, recall, and both the PR and ROC AUC. After training each model, they were tested individually on the test set before being tested as an ensemble.

To get the ensemble's prediction we took the sum of all predictions and compared it to the value:

$$0.5 \times n_m - \gamma \tag{5.2}$$

Where n_m is the number of models in the ensemble and gamma is a hyper-parameter to tune the sensitivity of positive predictions. We also tried simply taking the most common prediction, however that approach yielded less performance.

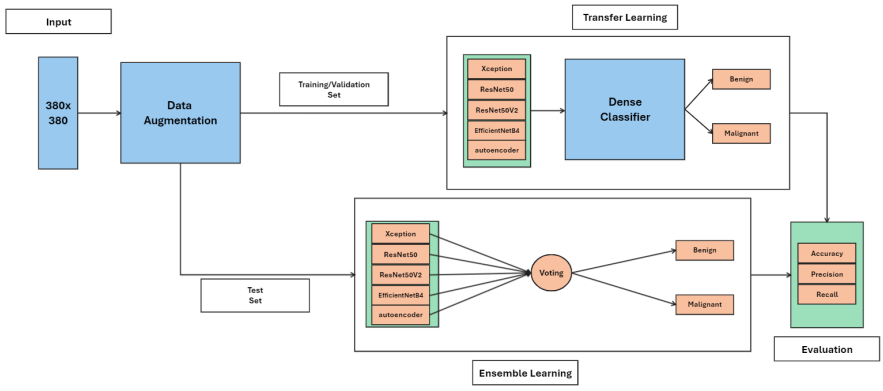

Figure 5.1: Transfer Learning Ensemble

5.4 Experimental Results

The final test results are given in Figure 5.1. We see that the auto-encoder on the combined dataset far exceeded any model, including the ensemble. This

is worrying, especially considering that only 584 out of 8632 of the malignant images come from the same dataset as the benign. The Discussion section contains more detail on the possible bias this may have introduced.

Table 5.1: Model Test Results

Model	Accuracy	Recall	Precision
ResNet50	.775	.840	.743
ResNet50V2	.750	.82	.7193
Xception	.775	.93	.709
EfficientNetB4	.755	.660	.815
AE	.770	.880	.721
Ensemble	.805	.90	.756
AE (w/combined dataset)	.933	.917	.948

5.5 Discussion

While the auto-encoder results are exceptional, we are concerned that additional bias may have been introduced. With only, 584 of the 8632 malignant images in the combined dataset, the ISIC2020 portion only represents 6.77% of malignant cases. The proportion of BCN20000 and HAM10000 malignant images are 70.60% and 22.64%, respectively. If there exists a commonality in either dataset, it will represent a large portion of the malignant images seen during training. For example, if a dermatoscope that produced a unique imaging artifact was used, the artifact would be present in every image. We think that this may be the reason the auto-encoder is achieving better results. One approach that might avoid bias would be to add an equal amount of benign images when combining the datasets. If an imaging artifact was present in all of the images, the model would have benign instances that could be trained against and the artifact would be ignored.

Moreover, a critical observation worth mentioning throughout our work with the various datasets mentioned above is the prevalent biases in data collected which consisted mainly of caucasian individuals. We believe that this could potentially affect the robustness our models especially when the models are used in a population outside the caucasian race. To enhance the generalizability and accuracy of our models across diverse racial groups, it is essential to further train them with skin lesion images representing a broader range of ethnic backgrounds. This will help in mitigating racial biases and improve our models effectiveness in clinical settings worldwide.

Additionally, to increase scalability, we may need to introduce the premalignant class for classification in the future. Since premalignant instances show the possibility of future cancer, they can be similar to both benign and malignant instances. This introduces an increased classification challenge, that will require more robust autoencoder as well as ensemble models. Building APIs to

incorporate with apps can also be a fruitful endeavor, as it can reveal hidden challenges when used in a more practical environment. This will also make data collection easier, as data can be stored when patients or health professionals use it for classification. There should also be a framework to incorporate feedback from medical personnel who may use it for classification. The input will be used to fine-tune the model, and detect and fix critical problems.

5.6 Conclusion

In conclusion, we explored two distinct methods for classifying skin lesions, which is crucial in melanoma detection. Method 1 focused on transfer learning and ensemble modeling, showcasing how backbone networks such as ResNet50v2, and EfficientNetB4 can be modified to tackle classifying skin lesions. Method 2, on the other hand, focused on an auto-encoders integrated with a classifier to solve the same problem. Despite all the models performing appreciably well, only the auto - encoder model trained on the combined dataset performed well comparable to the established benchmark metrics of 90% to 96% for accuracy and 93% to 95.1% for both precision and recall. Future works could include training our models with a dataset that is diverse across racial groups to make them more robust when used outside the caucasian race. Additionally, more robust model building can be pursued to tackle multi-class classification involving premalignant instances.

Table 5.2: Contributions by team member for Milestone 5.

Team Member	**Contribution**
Lucas Wurtz	Experimental setup
Ethan Gunderson	Combined datasets, experimental setup, results, discussion
Jacob Abaare	Introduction, Related Works, Discussion, Conclusion
Mohammad Fahim Shahriar	Background Study, Discussion, Figure

Bibliography

[1] Srinidhi Bashyam, Sruti Das Choudhury, Ashok Samal, and Tala Awada. Visual growth tracking for automated leaf stage monitoring based on image sequence analysis. *Remote Sensing*, 13:961, 2021. https://doi.org/10.3390/rs13050961 doi:10.3390/rs13050961.

[2] Ramzi Ben Ali, Ridha Ejbali, and Mourad Zaied. Classification of medical images based on deep stacked patched auto-encoders. *Multimedia Tools and Applications*, 79(35):25237–25257, 2020.

[3] Dusa Charan, Hemanth Nadipineni, Subin Sahayam, and Umarani Jayaraman. Method to classify skin lesions using dermascopic images, 2020. Available online at: https://arxiv.org/pdf/2008.09418v2.pdf.

[4] Noel C. F. Codella et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1710.05006*, 2017.

[5] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Allan Reiter, Allan Halpern, Susana Puig, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288, 2019.

[6] Sruti Das Choudhury, Samarpan Guha, Aankit Das, Amit Kumar Das, Ashok Samal, and Tala Awada. Flowerphenonet: Automated flower detection from multi-view image sequences using deep neural networks for temporal plant phenotyping analysis. *Remote Sensing*, 14(6252), 2022. https://doi.org/10.3390/rs14246252 doi:10.3390/rs14246252.

[7] A Geetha and NB Prakash. Classification of glaucoma in retinal images using efficientnetb4 deep learning model. *Comput. Syst. Sci. Eng.*, 43(3):1041–1055, 2022.

[8] Munmi Gogoi and Shahin Ara Begum. Image classification using deep autoencoders. In *2017 IEEE international conference on computational intelligence and computing research (ICCIC)*, pages 1–5. IEEE, 2017.

[9] Qishen Ha, Bo Liu, and Fuxu Liu. Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge, 2020. https://arxiv.org/abs/2010.05351 arXiv:2010.05351.

[10] Victor Kulikov, Victor Yurchenko, and Victor Lempitsky. Instance segmentation by deep coloring. *arXiv preprint arXiv:1807.10007*, Jul 2018. URL: https://arxiv.org/abs/1807.10007.

[11] Zhangli Lan, Songbai Cai, Xu He, and Xinpeng Wen. Fixcaps: An improved capsules network for diagnosis of skin cancer. *IEEE Access*, 10:76261–76267, 2022. https://doi.org/10.1109/ACCESS.2022.3181225 doi:10.1109/ACCESS.2022.3181225.

[12] Xinrong Lu and YA Firoozeh Abolhasani Zadeh. Deep learning-based classification for melanoma detection using xceptionnet. *Journal of Healthcare Engineering*, 2022, 2022.

[13] Pradeep Kumar Mallick, Seuc Ho Ryu, Sandeep Kumar Satapathy, Shruti Mishra, Gia Nhu Nguyen, and Prayag Tiwari. Brain mri image classification for cancer detection using deep wavelet autoencoder-based deep neural network. *IEEE Access*, 7:46278–46287, 2019.

[14] Sakib Mostafa, Debajyoti Mondal, Karim Panjvani, Leon Kochian, and Ian Stavness. Explainable deep learning in plant phenotyping. *Frontiers in Artificial Intelligence*, 6:1203546, 2023. https://doi.org/10.3389/frai.2023.1203546 doi:10.3389/frai.2023.1203546.

[15] Andre G C Pacheco, Gustavo R Lima, Amanda S Salomão, et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, 2020. https://doi.org/10.1016/j.dib.2020.106221 doi:10.1016/j.dib.2020.106221.

[16] Samuel William Pewton, Bill Cassidy, Connah Kendrick, and Moi Hoon Yap. Dermoscopic dark corner artifacts removal: Friend or foe? *Computer Methods and Programs in Biomedicine*, 244:107986, February 2024. URL: http://dx.doi.org/10.1016/j.cmpb.2023.107986, https://doi.org/10.1016/j.cmpb.2023.107986 doi:10.1016/j.cmpb.2023.107986.

[17] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, Allan Halpern, Brian Helba, Harald Kittler, Kivanc Kose, Steve Langer, Konstantinos Lioprys, Josep Malvehy, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander Stratigos, Philipp Tschandl, Jochen Weber, and H. Peter Soyer. A patient-centric dataset of images and metadata

for identifying melanomas using clinical context. *Scientific Data*, 8(1):34, 2021. https://doi.org/10.1038/s41597-021-00815-z doi:10.1038/s41597-021-00815-z.

[18] S. Sabbaghi, M. Aldeen, and R. Garnavi. A deep bag-of-features model for the classification of melanomas in dermoscopy images. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1369–1372, 2016. https://doi.org/10.1109/EMBC.2016.7590962 doi:10.1109/EMBC.2016.7590962.

[19] Jane Saldanha, Shaunak Chakraborty, Shruti Patil, Ketan Kotecha, Satish Kumar, and Anand Nayyar. Data augmentation using variational autoencoders for improvement of respiratory disease classification. *Plos one*, 17(8):e0266467, 2022.

[20] Ahmed Shaffie, Ahmed Soliman, Mohammed Ghazal, Fatma Taher, Neal Dunlap, Brian Wang, Victor Van Berkel, Georgy Gimelfarb, Adel Elmaghraby, and Ayman El-Baz. A novel autoencoder-based diagnostic system for early assessment of lung cancer. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 1393–1397. IEEE, 2018.

[21] Sadia Showkat and Shaima Qureshi. Efficacy of transfer learning-based resnet models in chest x-ray image classification for detecting covid-19 pneumonia. *Chemometrics and Intelligent Laboratory Systems*, 224:104534, 2022.

[22] Josef Steppan and Sten Hanke. Analysis of skin lesion images with deep learning, 2021. https://arxiv.org/abs/2101.03814 arXiv:2101.03814.

[23] Philipp Tschandl. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, 2018. https://doi.org/10.7910/DVN/DBW86T doi:10.7910/DVN/DBW86T.

[24] Philipp Tschandl, Christoph Rinner, Zoe Apalla, et al. Human-computer collaboration for skin cancer recognition. *Nature Medicine*, 26:1229–1234, 2020. https://doi.org/s41591-020-0942-0 doi:s41591-020-0942-0.

[25] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*, 5:180161, 2018. https://doi.org/10.1038/sdata.2018.161 doi:10.1038/sdata.2018.161.

[26] Nadia Smaoui Zghal and Imene Khanfir Kallel. An effective approach for the diagnosis of melanoma using the sparse auto-encoder for features detection and the svm for classification. In *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–6, 2020. https://doi.org/10.1109/ATSIP49331.2020.9231611 doi:10.1109/ATSIP49331.2020.9231611.

[27] Zhiwei Zhou, Junnan Wu, Zhengxia Wang, and Zhen-Li Huang. Deep learning using a residual deconvolutional network enables real-time high-density single-molecule localization microscopy. *Biomed. Opt. Express*, 14(4):1833–1847, Apr 2023. URL: `https://opg.optica.org/boe/abstract.cfm?URI=boe-14-4-1833`, https://doi.org/10.1364/BOE.484540 `doi:10.1364/BOE.484540`.