MULTIMODAL VS MULTI-MODEL: A COMPARATIVE ANALYSIS OF METADATA UTILIZATION IN SKIN LESION CLASSIFICATION

FRANK RITCHI

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2007402

COMMITTEE

dr. J.S. Olier Jauregui
dr. G. Chrupala

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2021

Acknowledgements

Before you lies the Master thesis “Multimodal vs Multi-Model: A Comparative Analysis of Metadata Utilization in Skin Lesion Classification”, in which the performance of different approaches to incorporating patients’ metadata into the classification of pigmented skin lesion are compared. While writing a thesis can be a solitary undertaking at times, especially during a pandemic, I could not have done it without guidance and support.

I would like to thank my supervisor, Dr. J.S. Olier Jauregui for his valuable feedback and excellent guidance, in particular during the conceptualization stage of this research. This guidance and feedback has helped me set my research on track and has provided me confidence during the remainder of the process. I also greatly appreciate the feedback and suggestions of the other students in my thesis group. Furthermore, I would like to thank Dr. Philipp Tschandl from the Department of Dermatology of the Medical University of Vienna. First of all for the creation of the groundbreaking dataset without which this research would not have been possible but in particular for his invaluable feedback and extensive answers to my many questions regarding the dataset and research topic.

As much as the scientific guidance has been helpful and is greatly appreciated, I would most of all like to thank my wonderful parents, Marij and Ton, and my lovely girlfriend, Adriana. Without the continuous and unconditional support of my parents throughout the years I would not be where I am today, let alone this thesis. I hope this thesis would have made you proud, mom, and makes you proud as well, dad. Adriana, thank you for your continuous support and motivation, in particular during the writing of this thesis. Your loving care and support has enabled me to focus on this thesis and I am tremendously grateful for that.

I hope you enjoy your reading.

Frank Ritchi

Sittard, The Netherlands

June 20, 2021

Link to full modeling procedure on GitHub:

https://github.com/FrankRitchi85/Master-Thesis

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Research Rationale and Scientific Relevance

Pigmented skin lesions can range from completely harmless to life threatening. The fact that benign and malignant can be visually very similar complicates diagnoses and hampers early detection, which is especially detrimental as early detection and treatment is crucial for patient's chance of survival (Claeson, Gillstedt, Whiteman & Paoli, 2017). The introduction of dermatoscopy, the examination of the skin using skin surface microscopy, has significantly improved early detection of (potentially) pigmented malignant carcinomas (Sinz et al, 2017). Another advantage of the introduction of dermatoscopy is the introduction of high-quality dermatoscopic images. In 1994 already, Binder et al. used dermatoscopic images in a neural network with the aim of distinguishing melanoma from nevi (e.g. birthmarks). Classification performance has continued to rise since then, but has been significantly accelerated by the collection, and distribution of large annotated datasets of dermatoscopic and microscopic images and was accelerated further by the introduction of new, advanced machine learning. At this point, the most advanced models outperform human professionals both in binary as well as some forms of multiclass pigmented skin lesion classification. In 2000 already Binder et al. broadened the research field again as they included of seven clinical parameters in the classification of lesions and found significant, yet marginal results.

One way of incorporating metadata is in multimodal classification, where multiple modalities, here images and metadata, are fed into a single classifier to produce a single outcome. Recently, Yap, Yolland & Tschandl (2018) compared such a multimodal approach to a baseline classifier and found significant performance increases. They trained a Random Forest classifier as well on just the metadata and found that it indeed contains predictive power with respect to lesion classification. Other high-performing multimodal models, utilizing both image and metadata modalities have been created and have too shown increases in overall performance (see the ISIC leaderboard, that ranks pigmented skin lesion classification models that were trained on the HAM10000 dataset: https://challenge.isic-archive.com/leaderboards/2018). However, a drawback of the multimodal approach is that such an approach does not tell us which features are predictive as neural networks are black boxes in this respect. Furthermore, classification performance increases from adding metadata have in general be moderate, including Yap et al.' approach. Because of the availability of high quality data, the novelty and importance of skin lesion classification, the research field is very popular, including under non medically trained machine learning professionals. Incredibly complex algorithms have been developed, however an approach that has yet not been tried, or at least not been published, is utilizing the availability of the metadata to create different models for different subsets of patients and lesions. The most well-known and utilized dataset of pigmented skin lesions, which will also be used in this research, the HAM10000 dataset, contains the age and sex of the patient as well the localization of the lesion on the body. The forming of skin lesions are influenced by age and gender. Older men in particular have a significantly higher risk of forming deadly melanomas or other forms of skin cancer. Furthermore, most malignant lesions are UV-induced; implying that skin that receives more sunlight is more susceptible to malignancy, in particular white skin. As all three included forms of metadata affect the chances of developing melanoma it is not unreasonable to pose that utilizing this metadata in a non-black box approach might be worth investigating. The conduciveness of features in the metadata on the development of malignant lesions will be further discussed in the literature review. Even if a performance increase is not achieved, it might provide handles for further research in this field, for example the addition of more potentially relevant metadata as well as using distinct techniques per subset.

1.2 Research Questions

From the points discussed above, several problems and opportunities can the can be distilled:

- Improving classification performance in skin lesions detection has significant societal benefits, especially now that the most advance models outperform human experts.
- 'Traditional' inclusion of metadata, i.e. a multimodal approach, can improve classification performance in pigmented skin lesion classification. However, in such a 'black-box' approach it is not clear which features in the metadata are conducive to improved classification.
- Furthermore, it is not clear how the performance of a multi-model approach would compare to that of a multimodal approach. Despite being trained on less data, given the same dataset, models trained on relevant subsets might (partially) outperform a multimodal model in terms of relevant performance criteria.
- Significant performance differences between the sub-models in the multi-model approach would provide more insights into which metadata is more conducive and could as such provide a rationale for incorporating more metadata than the current features, e.g. race and ancestral information, or provide a rationale to investigate the benefits of creating entirely distinct models based on metadata features, rather than a single classifier for all lesions or subsetting by features.
- It is important to classify lesion types correctly, but in real-life application it is more relevant that (potentially) malignant lesions are distinguished from harmless benign lesions.

Considering the societal relevance and impact of improved skin lesion performance, it is worthwhile to address these problems and opportunities, and by doing so explore a gap in the current scientific framework, and perhaps widen it. Therefore, the following research question has been formulated:

- *How does the classification performance of a multi-model classification approach compare to that of a multimodal model in pigmented skin lesion classification?*

As there might be performance differences between multiclass and binary classification performance, this research question can be subdivided into two sub-questions:

- *How do the two approaches compare in a binary classification, distinguishing between all lesion classes?*
- *How do the two approaches compare in a multiclass classification, distinguishing malignant from benign lesions?*

Several issues need to be address to be able to provide answers to these questions. A question that needs to be answered before any meaningful performance comparison can be undertaken is which classification performance metrics are most appropriate in this comparison. Considering how crucial (early) detection is, in the case of, in particular, malignant lesions it is evident different misclassifications have different costs attached to them.

Relevant considerations will therefore be formulated to determine the most appropriate performance metrics. Furthermore, an appropriate dataset is needed. This research will utilize a readily available and widely used benchmark dataset, the HAM10000 dataset, created by dermatologists from the Medical University of Vienna., containing of 10015 annotated dermatoscopic images including metadata. To answer the research questions several appropriate models have to be created. A model without any utilization of metadata will act as a baseline for the performance comparison. As the baseline model will also be used as the foundation of the multimodal model as well as in the multi-model approach it will need a sophisticated network architecture and be properly constructed and tuned, yet be relatively straightforward. Next, a multimodal model will be build, based on the baseline however with the metadata modality fused into the model. Finally, several models will be created based on subsets in the metadata, that contains the gender and age of the patients as well as the anatomic sites of the lesions. These subsets will be determined based on relevant literature. All models will be made suitable to handle multiclass as well as binary classifications. With all the relevant performance metrics determined and all the models complete, the results of the procedure can be reviewed and discussed and provide answers to the defined research questions.

1.3 Thesis Structure

The remainder of this thesis is structured in the following manner. Section 2 (Theoretical Framework) discusses the relevant research fields and discusses the existing framework in which this thesis is set. The section start with background information about (pigmented) skin lesions and the contribution of automated classification (computer-aided diagnosis) in the detection of malignant skin lesions. The second part of this section discusses the utilization of patients' metadata to aid classification. The final part of this section discusses the concepts of multimodal and multi-model classification. Section 3 (Methodology) explains the methodology that will be used to provide answers to the research questions and how this methodology is suitable for this goal. Specifically, it discusses the underlying baseline architecture (ResNet50), the three applied models/approaches (baseline, multimodal, and multi-model approach) and the evaluation and comparison of these approaches. Section 4 (Experimental Setup) describes the dataset used in this research and the full experimental procedure. The reproducibility of this research is touched upon as well. After obtaining the results of the experimental procedure they are are presented and discussed in Section 5 (Results) as well as their validity. Section 6 (Discussion) reflects on the research goal and discusses the findings with regards to the earlier formulated research questions. The limitations of the research are then discussed as well as its contribution to and implications within the existing scientific framework. Section 7 (Conclusion) concludes this thesis.

2. Theoretical Framework

3. Methodology

This section describes the basic research setup and methods used to answer the formulated research questions and the motivation behind applying them, including the setup of the most appropriate models and evaluation metrics. To answer the research questions, a thorough model comparison has to be undertaken. First, a baseline neural network will be constructed to serve as a benchmark, to which a multimodal and multi-model approach will be compared. Both the multimodal and multi-model approach will be based on the baseline model; the multimodal will fuse the metadata into the baseline model while the multi-model approach will apply the baseline model to different subsets of the data.

3.1 Baseline

To be able to act as a baseline, the baseline model cannot utilize any metadata and will only use dermatoscopic images as input. Although maximal performance is explicitly not a goal in this comparative analysis, a comparison of poorly performing models would not be relevant and scientifically meaningful. The models should therefore be high performing and as the baseline model will form the backbone of the other models, this applies to the baseline model as well. Another consideration is that in order to draw meaningful conclusions, the models should be relatively simple and straightforward. This is in contrast with the most intricate and complex models top-performing models applied to the HAM10000 dataset. These models apply advanced ensemble techniques (and sometimes even ensembles of ensembles) and often feed the models with external data as well. Applying such techniques in this research would hamper the ability to draw any meaningful conclusions. As such, it is deemed appropriate to apply a simple, yet sophisticated and proven neural network architecture.

ResNet50 architecture

The network architecture that will used in the baseline model, and in extension the multimodal and multi-model models, is the ResNet50 DCNN architecture. ResNet50 is a 50-layer residual neural network and has proven to be very suitable for image classification tasks, including skin lesion classification. Deep learning is thought of as learning a hierarchical set of representations, such that it learns low-, mid- and high-level features. DCNN-models are therefore very well suited for image classification, as such models can learn shapes, edges, colors etc. In principle neural networks should perform better, or at least not worse, as more layers are added to the network. When deeper layers are not learning any new information, the network could at least preserve the information learned in previous layers by simply copying the shallow network with identity mappings, to preserve the information gained in the earlier layers. This is however computationally hard in plain network structures and is therefore not applied. The deeper a neural network gets, the harder it becomes for the optimizer to determine the most optimal parameter. As a result of this, and without identity mapping, at some point adding extra layers leads to performance degradation. In 2015, a new type of model architecture was introduced by He, Zhang, Ren & Sun (2016); the deep residual neural network. The authors demonstrated the described constraints of traditional deep networks in their paper, by empirically showing there is a maximal depth threshold in traditional convolution neural networks. Residual neural networks alleviate this issue by introducing so-called *residual blocks* and *skip-connections*, or *identity connection*. Skip-connections add identity mapping to a forward point in the network, from the beginning of the residual block to the end, shortcutting the residual block. When a residual block gained no extra information, only the identity mapping is pushed forward into the network, instead of the residual mapping of that block. The output of a residual block *i* can be mathematically expressed as:

$$y = F(x, \{W_i\}) + W_s x \tag{1}$$

Where $F(x, \{W_i\})$ is the residual mapping with W_i the block's weight layers and x the identity mapping, with W_s a linear projection to match the output size of the residual layer. A visualization of a residual block and skin connection can be seen in Figure 1. This characteristic of residual networks make very deep models possible without the usual problems of degradation and vanishing gradients. As the identity connection does not introduce any new parameters the computational complexity is almost identical to that of a 'plain' deep convolutional neural network.

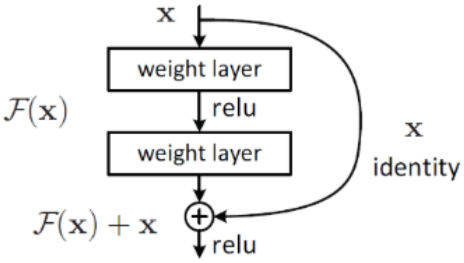

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 1. Schematic overview of a residual block (He et al, (2016)) ResNet-models have been created with various depths, up to a previously unheard of 152 layers. The most applied Resnet-model is ResNet50, with 50 layers, consisting of 16 residual blocks of three convolutional layers, between a initializing convolutional layer and a final Fully Connected layer. Its widespread use, relatively low computational complexity and high performance in image classification tasks are the main motivations of applying this architecture is this research. The full architecture can be found in Appendix A. Schematically, the baseline model looks, as intended, straightforward. The Resnet model takes the images as input and outputs predictions:

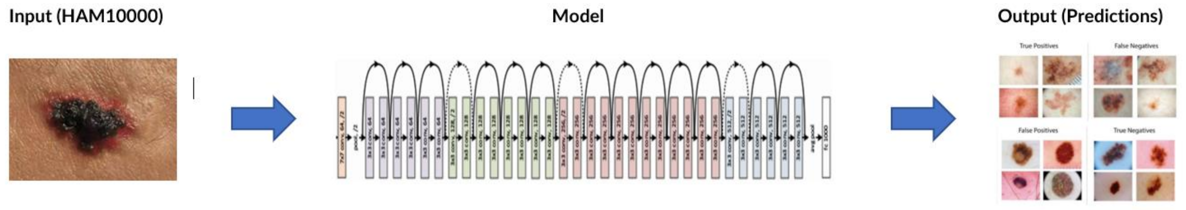


Figure 2. Schematic overview of the baseline procedure

3.2 Multimodal Approach

The multimodal approach will be an adaptation of the baseline model. In addition to the images, metadata will also be fused into the model. This will be accomplished by first training a separate artificial neural network on the metadata and then merge the ResNet model with this newly trained model. As the used metadata is rather straightforward, consisting of two binary and one one-hot encoded features a simple Multilayer Perceptron is deemed sufficient for the training of the metadata. In accordance with an earlier highly performing multimodal modal that was trained on the HAM10000 dataset (Yap et al., 2018) a late fusion technique will be applied, where the last (flattened) layers of the ResNet model are merged with the last (flattened) layer of the MLP. After merging, several fully connected layers will be added after the results are outputted via a softmax (multiclass) or sigmoid (binary model) activation function. Schematically, the multimodal model looks as follows:

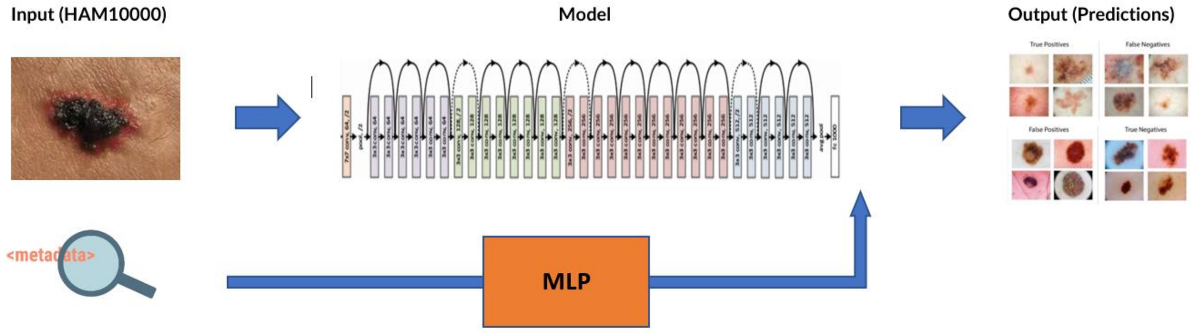


Figure 3. Schematic overview of the multimodal procedure

3.3 Multi-model Approach

The multi-model procedure basically applies the baseline procedure to several subsets of lesions. First, relevant subsets will be created on all of the available features in the metadata; age, sex and the anatomic site of the lesion. These subsets will be based on reviewed literature. The original dataset will be divided in two groups per feature, i.e. a total of six multiclass models and six binary models. The classification outputs will be evaluated separately, but will also be aggregated. Schematically, the multi-model approach looks as follows:

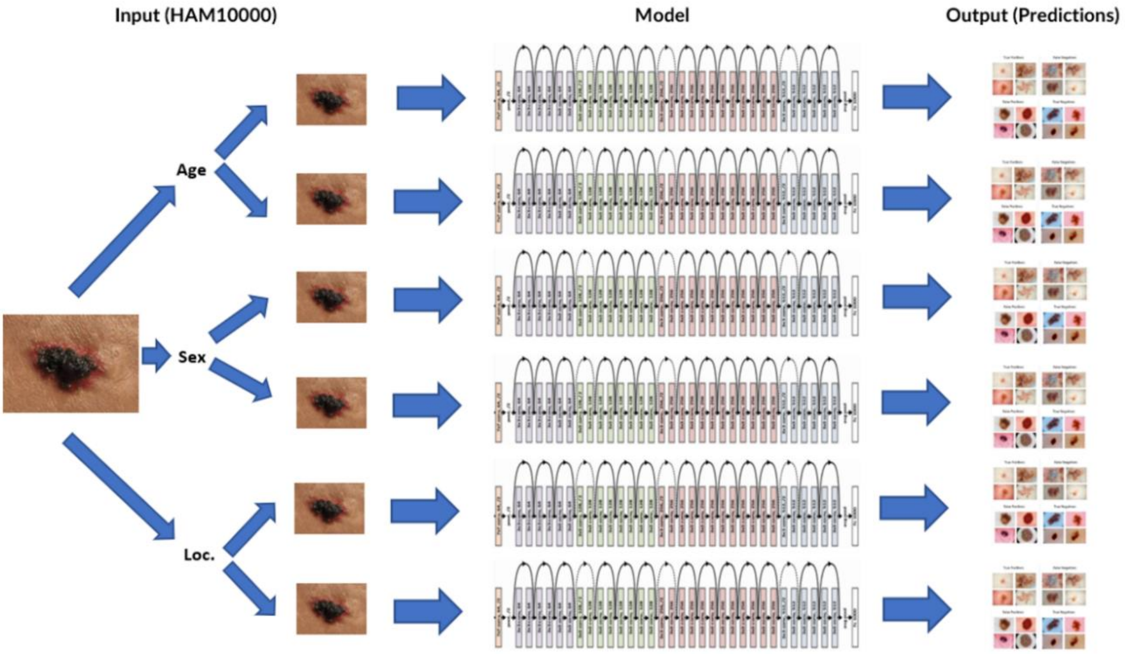


Figure 4. Schematic overview of the multi-model procedure

3.4 Evaluation Metrics

The goal of this research is to compare the performance of the multimodal and multi-model approaches, in multiclass as well as binary classification. In order to formulate the most appropriate metrics, it is imperative to understand on what basis the misclassifications should be weighed. In binary classification, two type of misclassifications can occur; a malignant lesion can be classified as a benign lesion and vice versa. These misclassifications are however not equally harmful from a medical perspective. As discussed earlier, for the malignant lesions present in the dataset early detection and treatment can greatly influence patients' health outcome. This applies to melanomas in particular, as they are by far the deadliest form of skin cancer while they can often be simply removed surgically if detected in an early stage. The same applies to basal cell carcinomas and to a lesser extent actinic keratoses and intraepithelial carcinoma. As such, the ability of a model to differentiate malignant lesions, in particular melanoma, from harmless benign lesions should be the most crucial performance indicator. On the other hand, misclassifying a benign lesion as a malignant one has much less severe consequences. Such misclassifications might lead to otherwise unnecessary extra examinations that lead to higher workloads and costs, as well as induce unnecessary stress in patients, but such misclassifications do not have any (potential) detrimental health effects. Albeit still undesirable, these negative consequences pale in comparison to the devastating effect of untreated malignant lesions.

In multiclass classification, more types of misclassifications can occur, as benign/malignant lesions can also be misclassified as another type of benign/malignant lesion. From a practical medical viewpoint, such misclassifications are again undesirable, but would not lead to significant harm. In case of mistaking a benign lesions for another, no different actions would take place, nor would the health outcome change. Furthermore, as different types of malignant lesions need different treatment, practically such a misclassification would not matter as in the case of suspected malignancy a lesion would be thoroughly examined and confirmed via histopathology. Any of such mistakes would therefore not be harmful, as long as all cases labeled malignant would be examined within a similar timeframe.

In both types of classification, finding as many malignant lesions as possible is most important, even if this leads to more false positives. In the multiclass method this can be split out per malignant lesion. The most appropriate measure to report is therefore recall, that reports how many of the relevant items (here malignant lesions) are classified as such. The formula for recall is:

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{2}$$

On the other hand too many false positives will also have a detrimental effect on the quality and usefulness of the model. As such, it should be determined how many of the identified positives are true positives. A low ratio would indicate that many benign lesions are classified as malignant. The formula for this ratio, precision, is:

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{3}$$

Both recall and precision are useful metrics but both only tell part of the tale. Accuracy is often used as an ‘overall’ measure, however the number of true negatives is not relevant in skin lesions classification. A more appropriate measure to address both paradigms in this context is the F1-score, that is the harmonic mean of precision and recall. It is calculated as follows:

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

The ROC, or receiver operating characteristic curve, illustrates the diagnostic ability of a binary classifier at different thresholds. The size of the area under this curve (AUC) indicates the performance of the classifier at distinguishing between a positive and negative class, where a perfect classifier’s AUC would be 1, that of a random classifier 0.5. As such, the AUC is another relevant metric and will be included in the performance comparison. Together, the recall, precision, F1 and AUC of malignant lesions in the binary model (malignant lesions combined) and multiclass model (for each of three malignant lesions) form a comprehensive performance indication which will enable us to make the relevant model comparisons. How the results of the classifications will be visually presented will be discussed in the next section.

Both the baseline and the multimodal models produce two classification prediction outputs, one for the multiclass and one for the binary classification. The multi-model however will produce two outputs per feature per classification method, a total of twelve (2*3*2) model outputs. The performance of the individual models are relevant as we want to compare them to both the multimodal performance as well as to the other model for that feature (e.g., male vs female). As these multi-model models only train on about half of the dataset, as their counterparts train on the other half, to compare the performance on the entire dataset with the multimodal model the results for both will also be aggregated. This can simply be done by summing both confusion matrices and calculate the defined performance metrics derived from it. For the AUC this is not possible. Instead the weighted average of the AUC's of both sub-models will be used.

4. Experimental Setup

This section describes the experimental setup, including a detailed description of the dataset and of the experimental procedure, including preprocessing, exploratory data analysis, the creation, and implementation of the models, The output and reproducibility of the procedure are discussed in section 4.3 and 4.4 respectively.

4.1 Dataset

The HAM10000 Dataset

This research utilizes the HAM10000 dataset, created by the ViDiR Group of the Department of Dermatology of the Medical University of Vienna, Austria in collaboration with the School of Medicine of the University of Queensland, Australia (Tschandl Rosendahl & Kittler, 2018). The total dataset consists of 11788 dermatoscopic images, of which only 10015 have been publicly released for academic purposes (the remainder served as test data for the ISIC challenge 2018), and contains seven types of lesions, both benign and malignant. The images were collected over twenty years in Australia and Austria. Several preprocessing techniques have been applied to make the collected data uniform and suitable for classification. This includes extraction from digital dermatoscopy systems and PowerPoint files, digitalization of diapositives, image filtering, unifying pathologic diagnoses and manual quality review. Several preprocessing techniques, like hair removal and removal of other noise were deliberately not applied, as the presence of such features "reflects the situation in clinical practice" (Tschandl et al, 2018). Some lesions appear multiple times in the dataset. In such cases, images are different with regards to zoom, angle of camera used. The creators of the dataset consider the presence of multiple images per lesion a form of "natural data augmentation" (Tschandl et al, 2018). The 10015 images in the dataset represent 7470 unique lesions. The images are annotated with diagnosis and patient metadata; the diagnosis and type of ground truth for the diagnosis, the age and sex of the patient and the localization of the lesion on the body (anatomic site).

The accompanying documentation (referenced above), with an in-depth description of the applied preprocessing methods has been published in Nature the code used for these methods is available on GitHub[1] . The dataset has been made publicly available through the archive of the International Skin Imaging Collaboration (ISIC), and can be retrieved via the Harvard Dataverse[2] or via Kaggle[3]. In 2018, the dataset was used in the MICCAI ISIC classification challenge. The leaderboard with the best performing models is publicly accessible[4]. Because of the unprecedented size, the combination of high quality images, manual quality review, presence of metadata and public availability this dataset has been serving as a benchmark dataset for pigmented skin lesion classification as is widely used and known in the research field (288 citations as of 10th of June 2021). It are these characteristics that make this dataset also suitable for this research, in particular the presence of metadata and its size, that make it possible to create sufficiently large subsets for the multi-model approach. The benchmark status of the dataset facilitates the incorporation of this research into the existing scientific framework, as the approach in this research can be applied to and compared with other utilizations of the dataset.

Skin Lesions in Dataset

Seven generic classes of skin lesions are present in the dataset. These account for 95% of all pigmented skin lesions encountered in clinical practice (at least in Austria and Australia) (source: HAM10000), making the dataset representative. This thesis will not discuss the lesions in-depth, but lists the information relevant for this research, which was retrieved from the HAM10000 documentation:

Actinic Keratoses and intraepithelial carcinoma: A common and non-invasive lesion class. These types of lesions can usually be treated locally without surgery, but may progress to invasive squamous cell carcinoma. Early detection is therefore important.

Basal cell carcinoma: Malignant skin cancer. Basal cell carcinomas rarely metastasize, but if untreated they can grow destructively. Early detection and treatment is therefore crucial.

Benign keratosis-like lesions: A benign, harmless type of lesion. Some of these lesions however share morphologic features with melanoma and are therefore often biopsied.

Dermatofibroma: A benign, harmless type of lesion.

Melanocytic nevi: A benign, harmless type of lesion. Includes typical birthmarks (nauvus naevocellularis). Because of their frequent occurrence and morphological similarity to melanoma many images of nevi are available and included in the dataset.

Melanoma (mel): The most dangerous form of skin cancer, that can be invasive or non-invasive. However, if detected and treated in an early stage might be completely removed by surgical excision.

Vascular skin lesions (vasc): In contrast to other classes not pigmented by melanin, but by hemoglobin. Vascular skin lesions are non-malignant.

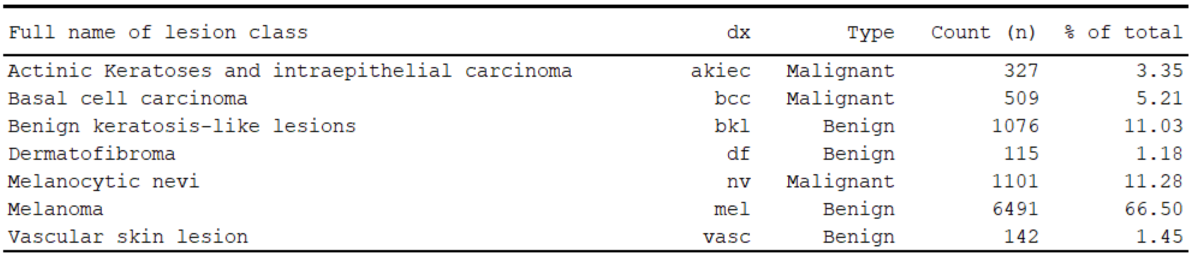
Table 1 gives an overview of the classes present in the dataset as well as their count and whether they are considered benign of malignant in the binary models. Actinic Keratoses and intraepithelial carcinoma are not malignant, but as they sometimes can progress into a malignant carcinoma early detection is still important. As such, for the purposes of this research, it has been categorized as malignant.

Retrieved examples of each of the classes can be seen in Figure 5. Inspection shows the challenge of pigmented skin lesion classification; difference within the classes in form of shape, color and borders can be quite significant, while there is also a large overlap between classes regarding these features, at least to an untrained eye.

[1] https://github.com/ptschandl/HAM10000_dataset
[2] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T
[3] https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000
[4] https://challenge.isic-archive.com/leaderboards/2018

Table 1

Description of skin lesion classes present in the dataset.



Full name of lesion class	dx	Type	Count (n)	% of total
Actinic Keratoses and intraepithelial carcinoma	akiec	Malignant	327	3.35
Basal cell carcinoma	bcc	Malignant	509	5.21
Benign keratosis-like lesions	bkl	Benign	1076	11.03
Dermatofibroma	df	Benign	115	1.18
Melanocytic nevi	nv	Malignant	1101	11.28
Melanoma	mel	Benign	6491	66.50
Vascular skin lesion	vasc	Benign	142	1.45

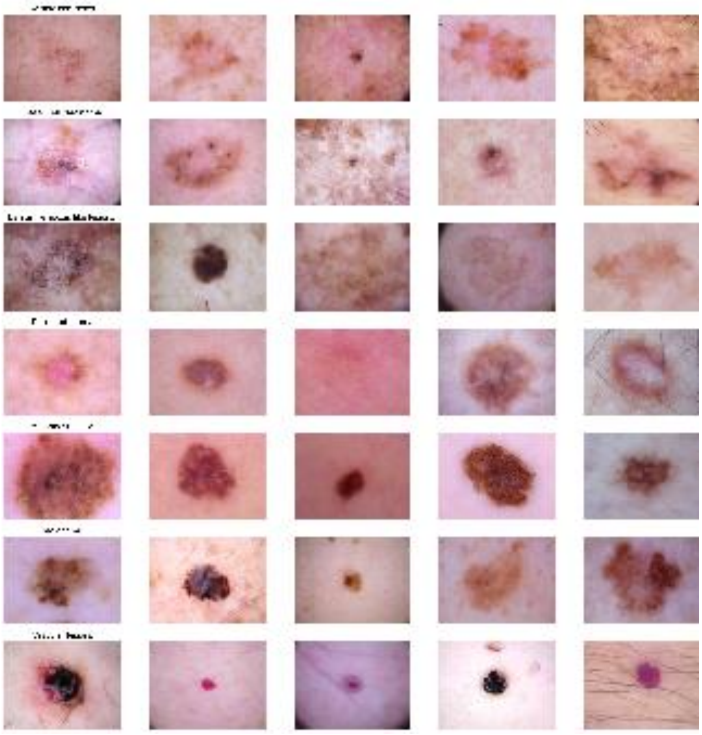


Figure 5. Examples of skin lesions present in dataset.

Technical Validation

Four types of ground truths have been used to confirm the diagnoses of the lesions in the dataset, namely histopathology (images analyzed by dermatopathologist), confocal microscopy (with near cell level resolution), follow-up (monitoring changes in lesions over time) and consensus (consensus between (trained) dataset creators; only used in benign cases). Images with any form of ambiguity in their diagnosis were not included in the dataset. The lesions and ground truths are not independent of each other; histopathology generally occurs only when there is a perceived plausibility the lesion could be malignant. All malignant lesions were confirmed with histopathology. Confocal microscopy, consensus and follow-ups were only used to confirm benign lesions.

Dataset characteristics:

The publicly available dataset consists of 10015 .jpg files, randomly divided over two folders, in 450*600 resolution, and a .csv file containing metadata. The image names correspond with the image ID's in the metadata file, forming a bridge between the images and metadata. As discussed, some lesions appear several times in the dataset. This is not an issue per se, however having pictures of the same lesions in the training and test has to be avoided. Column *dx* contains the seven classes of lesions, *dx_type* the form of ground truth, *age* the age of the patient (aggregated in seventeen bins of five year), *sex* the gender of the patient and *localization* the anatomic site of the lesions.

4.2 Experimental Procedure

4.2.1 Preprocessing and Exploratory Data Analysis

Data preprocessing commenced with outlier and missing value detection. No illegal values were identified, neither as legal but extreme outliers. However, several missing values were detected (57 NA's in *age*, 57 'unknown' in *sex* and 234 'unknown' in *localization*). 47 instances lacking *all* metadata were removed. Categorical imputation with mode or random sampling was considered but not undertaken, as it was deemed inappropriate as the dataset would be split up in the multi-model approach. With seven target classes in the multiclass models the risk of imputed data in relatively small datasets skewing the results was deemed too high. As the percentage of complete cases would still be 97.5% after removal of all incomplete instances it was decided to remove all incomplete instances. 9761 instances remain in the final dataset.

As the multi-model approach requires subsetting on age, gender and localization new columns were created, binning the variables *age* and *localization* in two bins; *young* and *old* and *sun-exposed* and *non-exposed* respectively. In accordance with the earlier discussed literature the cut-off for *age* was made at 50 years (with 50 being part of the 'older' category). The fourteen localizations were divided as follows: ear, face, hand, lower extremity (legs), upper extremity (arms), neck and scalp were categorized as sun-exposed, whereas abdomen, acral, back, chest, foot, genital and trunk were categorized as non-exposed. This distinction was made based on sun-exposure in typical summer wear and is in line with literature that identifies the localizations in the former group as having the most incidence of sunlight-induced lesions (Arisi et al., 2018). The type of diagnosis (benign/malignant) was also added, as well as the full diagnosis name, and a column '*unique*' to identify lesions with multiple images. The images were appended to the dataset, in the form of a column of arrays of pixel values. Appropriate columns were binarized/categorized.

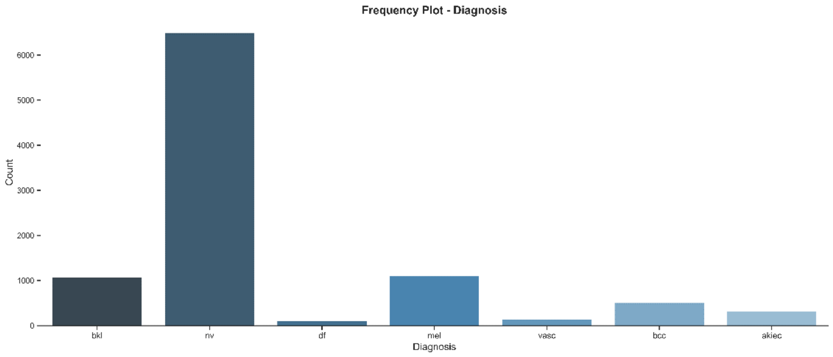


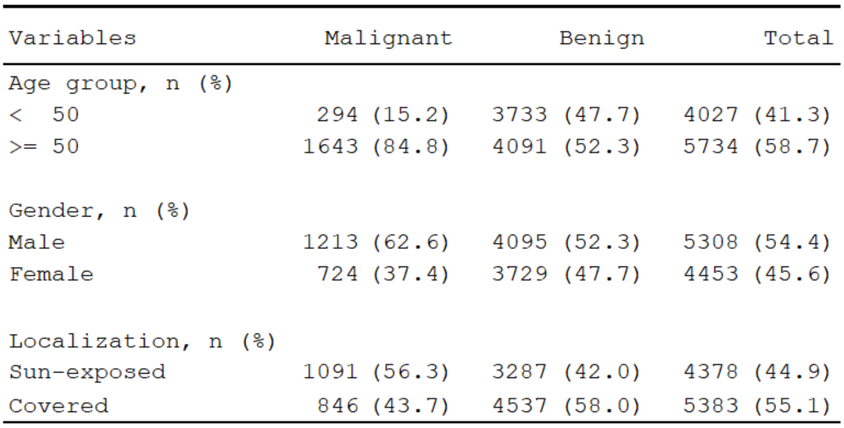
Figure 6. Distribution of the classes, showing significant class imbalance.

Extensive exploratory data analysis was performed to identify peculiarities and potential problems with respect to modeling. Although the HAM10000 dataset has been constructed with care and is in general very balanced and complete there is a large class imbalance, as there are many more melanocytic nevi present in the dataset than any other class, as can be seen in Figure 6. This is due to the fact that melanocytic nevi are the most common type of lesion recorded, as they are often examined for fear of being a melanoma. For the binary classification the class imbalance is somewhat offset by the other classes, but still very significant. The creators of the dataset have included these nevi in the dataset despite the class imbalance they cause as they still add predictive power to classification models. However, without class balancing any classification model is doomed to overfit.

Inspection of the variables via single-feature and multi-feature frequency plots did not identify any further potential issues. Although *age* and *localization* are not completely evenly distributed, as can be expected, these distributions are no reason for concern, in particular because the subsets in the multi-model approach seem still sufficiently large, as can be seen in Table 2. The last step in the exploratory data analysis was the creation of heatmaps, displaying the number of occurrences of malignant lesions per localization and age, for both genders. The heatmaps are grouped in Figure 7. These heatmaps provide indications that the distributions of the lesions differ over the ages, localizations as well as genders, which is in line with the literature and clinical experience.

Table 2

Distribution of skin lesion types over determined subsets.



Variables	Malignant	Benign	Total
Age group, n (%)			
< 50	294 (15.2)	3733 (47.7)	4027 (41.3)
>= 50	1643 (84.8)	4091 (52.3)	5734 (58.7)
Gender, n (%)			
Male	1213 (62.6)	4095 (52.3)	5308 (54.4)
Female	724 (37.4)	3729 (47.7)	4453 (45.6)
Localization, n (%)			
Sun-exposed	1091 (56.3)	3287 (42.0)	4378 (44.9)
Covered	846 (43.7)	4537 (58.0)	5383 (55.1)

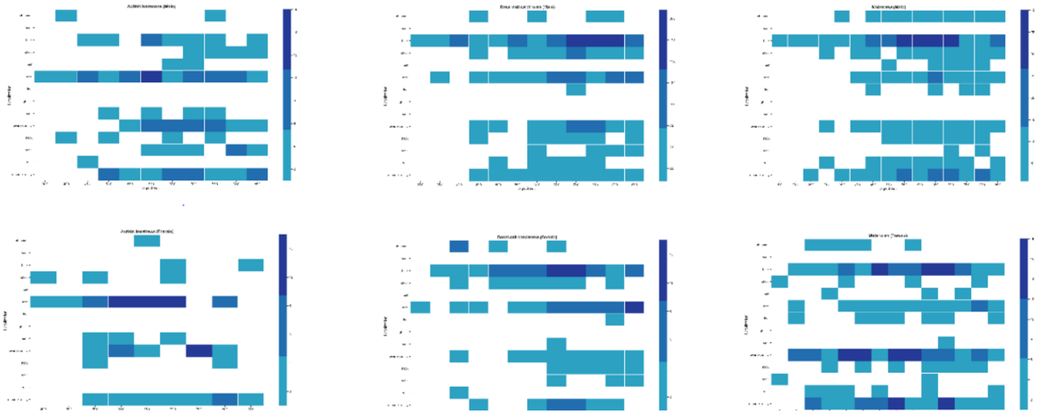


Figure 7. Heatmaps of distribution of lesions over ages, genders and localizations

4.2.2 Implementation

Several techniques have been applied to strengthen the performance and validity of the models. To counter the class imbalance, random oversampling of the minority classes was performed to match the majority class (nevi). This approach does not introduce new information, but at least helps to retain the information in the largest class. It comes with an increased risk of overfitting to the smaller classes though. A rather controversial decision was taken by also balancing the test set. While such a decision in certainly no standard practice, it was taken nonetheless as the class imbalance in the set is not representative for the population. While the classes in the dataset are indeed the most common lesions seen in clinical practice, the distribution is not as because of (financial, technical and time) barriers not all lesions are processed well enough to be suitable for use in benchmark dataset, which is a common challenge in the medical field. Next, the data went through stratified train/validation/test split. A division of 70%, 10%, 20% was deemed appropriate. Lesions with multiple images were put in the same set, not in both the training and test set. Data augmentation was performed to the training set to artificially inflate the amount of training data. Only simple techniques, like zooming and rotating, were used to keep the procedure relatively simple. Examples of this augmentation can be seen in Figure 8. The last step before initializing the model was setting up the parameters. Previously trained weights, trained on ImageNet, were used as that drastically lowers running times while still achieving high performance. The learning rate was made variable to balance running time and performance. All models used the Adam optimizer and ran (for a maximum of) 50 epochs.

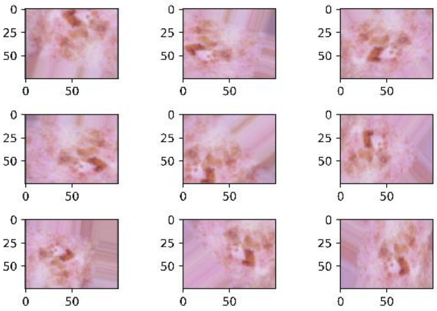


Figure 8. Applied data augmentations

4.3 Evaluation

The performance metrics that will be evaluated have been determined in the previous section; Precision, recall, F1 and the AUC score will be calculated for all binary models (with malignancy as the positive class). For the multiclass classifications the precision, recall and F1-scores will be obtained for all three malignant classes individually. All metrics will be displayed in two well-structured tables. As all these metrics are derived from the number of true and false positive and negatives, confusion matrices of all models, binary and multiclass, will be created as they convey all classifications in a clear and visually appeasing manner, supporting identification of differences in classification performance. The same rationale is applied to the inclusion of ROC-curves of the binary models and the three malignant classes in the multiclass classification.

4.4 Reproducibility

In order for scientific research to be a relevant scientific contribution to the research field, it is crucial that its findings can be validated, challenged and scrutinized by scientific peers. Reproducibility and replicability should therefore be integral to any research process. As such, all steps to ensure full reproducibly have been documented in a GitHub repository: *https://github.com/FrankRitchi85/Master-Thesis*. This repository contains the full code generated for this research, which is extensively commented, as well as all output and a requirements.txt file containing all software (Python 3.7.10) and packages and versions used in this research. All code was written and run on Google Colab, utilizing its GPU kernel. Running the code is however feasible on any moderately adequate personal computer. As Google Colab is a specialized version of Jupyter Notebook, the code is in the form of .ipynb files. A total of 19 notebooks have been created; one for preprocessing, sixteen for each of the models, one for evaluation and one for producing the output. The preprocessing, model and evaluation notebooks store data, variables and models in pickles that can then be loaded by subsequent notebooks. These pickles are not included in the GitHub repository due to file size limits.

5. Results

5.1 Model Comparison

With the models trained, the performances of the procedures can be compared. First, the binary classifiers shall be discussed, then the multiclass classifiers

Binary Classification

A first glance on the binary confusion matrices in Figure 9, shows little difference in the classification between the baseline, multimodal model and the three aggregated models for age, sex and localization. In fact, they all seem to very, and roughly equally good in distinguishing between both two classes, although the age, sex and localization models seem to have misclassified more malignant lesions as benign, which is the most undesirable type of misclassification. The insight the confusion matrices offer is valuable, however with such similar performances the confusion matrices are not suited for a more thorough comparison. Considering that the multi-model models trained on only half the data (remember that the three matrices for age, sex and localization are simply aggregates of their

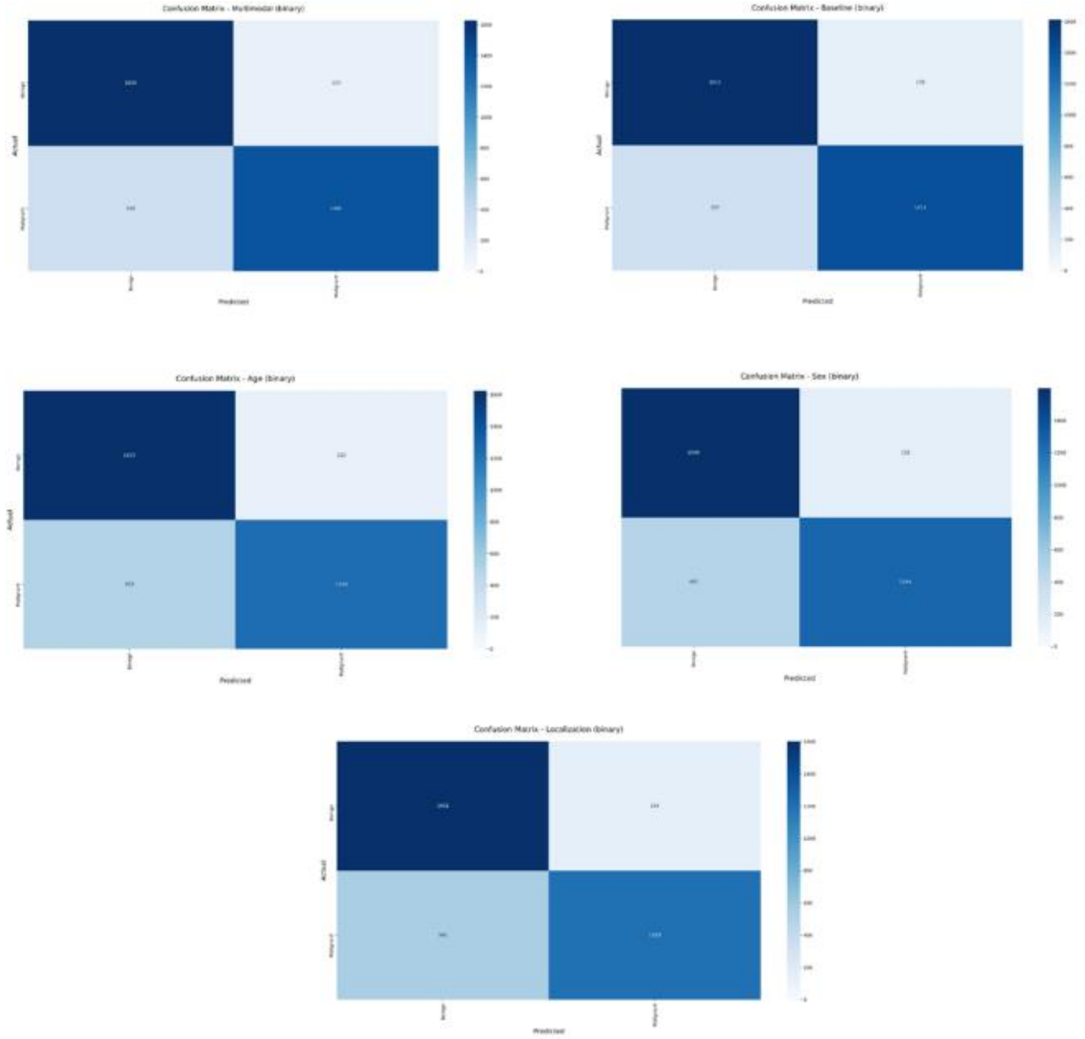


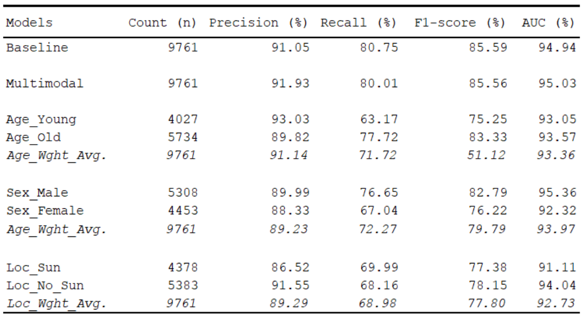
Figure 9. Confusion matrices for the binary classification for the baseline, the multimodal model and for the aggregated age, sex and localization models.

sub-models) the relatively small performance increases is quite surprising. Table 3 lists the earlier defined and calculated performance metrics for all developed models and provides a more detailed performance overview.

The extra misclassifications did indeed lead to a lower recall score for all aggregated and sub-models. The AUC score's however are more on par with the baseline and multimodal model, with the model for the male gender scoring the highest AUC of all models. Performance on precision is again quite similar between the models, implying they can almost equally well define the negative (benign) class as such. As precision is similar and recall lower, it is no surprise that the sub-models cannot match the F1-score of the baseline and multimodal model. Interesting is the performance difference within the sub-models, most notably the fourteen percent point difference in recall between the model for males and females. The fact that the male model contained a higher percentage of malignant lesion undoubtedly plays a role, but the difference is still remarkable. Furthermore, as men, especially older men, more often get malignant lesions, such a discrepancy is justified. Remarkable is the very minimal performance difference between the baseline and the multimodal model on all of the four performance metrics. The fusions of the metadata into the model did not seem to have had any significant impact. It is interesting to see whether this remains the case in the more difficult multiclass classification.

Table 3

Model performances in binary classification



Models	Count (n)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
Baseline	9761	91.05	80.75	85.59	94.94
Multimodal	9761	91.93	80.01	85.56	95.03
Age_Young	4027	93.03	63.17	75.25	93.05
Age_Old	5734	89.82	77.72	83.33	93.57
Age_Wght_Avg.	*9761*	*91.14*	*71.72*	*51.12*	*93.36*
Sex_Male	5308	89.99	76.65	82.79	95.36
Sex_Female	4453	88.33	67.04	76.22	92.32
Age_Wght_Avg.	*9761*	*89.23*	*72.27*	*79.79*	*93.97*
Loc_Sun	4378	86.52	69.99	77.38	91.11
Loc_No_Sun	5383	91.55	68.16	78.15	94.04
Loc_Wght_Avg.	*9761*	*89.29*	*68.98*	*77.80*	*92.73*

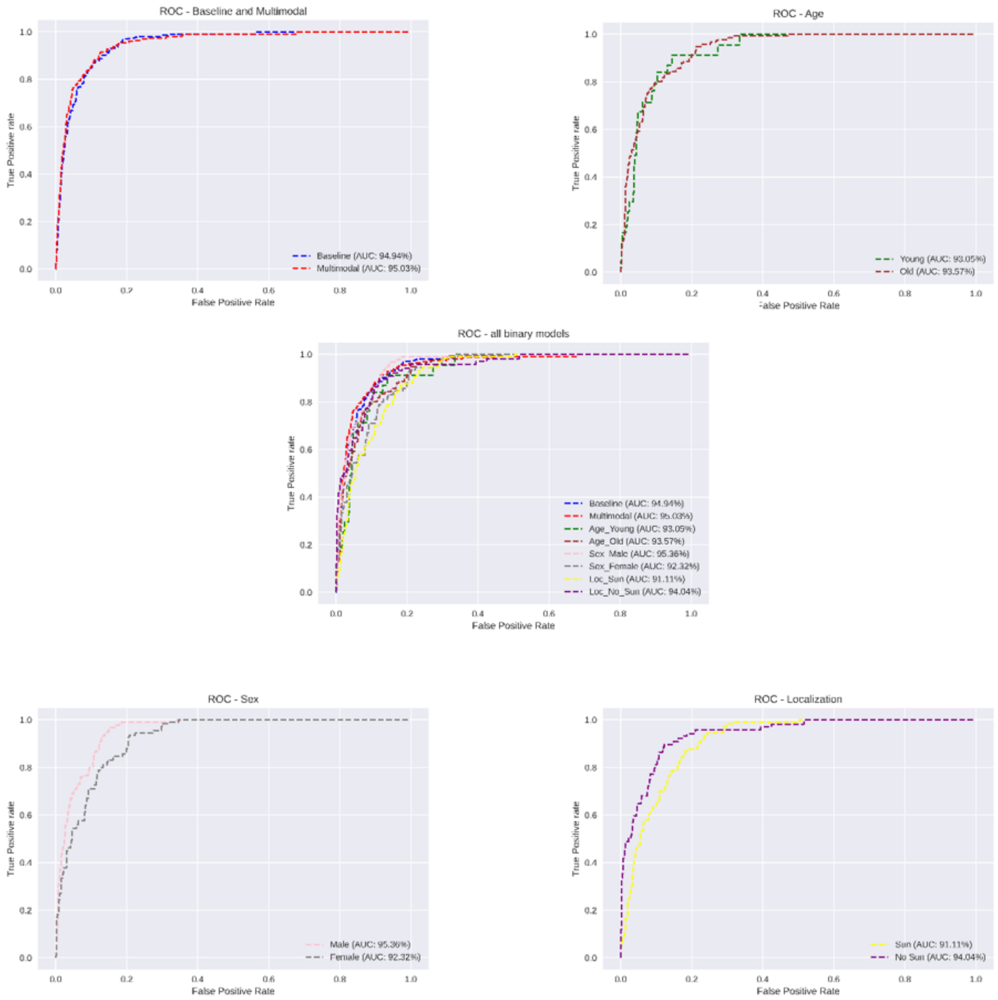


Figure 10. ROC-curves for the binary classification. The middle figures display all models in one plot. The other plots plot the baseline and the multi-model, both age, both sex and both localization models respectively.

The ROC-curves of the models in Figure 10 are in line with the information in the table, the baseline and multimodal model perform almost exactly the same. The most interesting conclusion is the performance difference between the sub-models. The models for females and covered localizations perform significantly worse than their counterparts. This might be because lesions for females and covered localization are harder to classify or that the constructed model was more fit for males and sun-exposed localizations. On the other hand, considering the role age plays in inducing malignancy, it is remarkable the models for younger and older patients perform so equally. This might be due to the fact that a neural network might learn to differentiate between older and younger skin without the metadata, but solely based on the skin condition.

Multiclass Classification

For the multiclass classification, no large performance differences can be determined based on the confusion matrices (Figure 11). It is clear though that all models have a much harder time differentiating between all the classes.

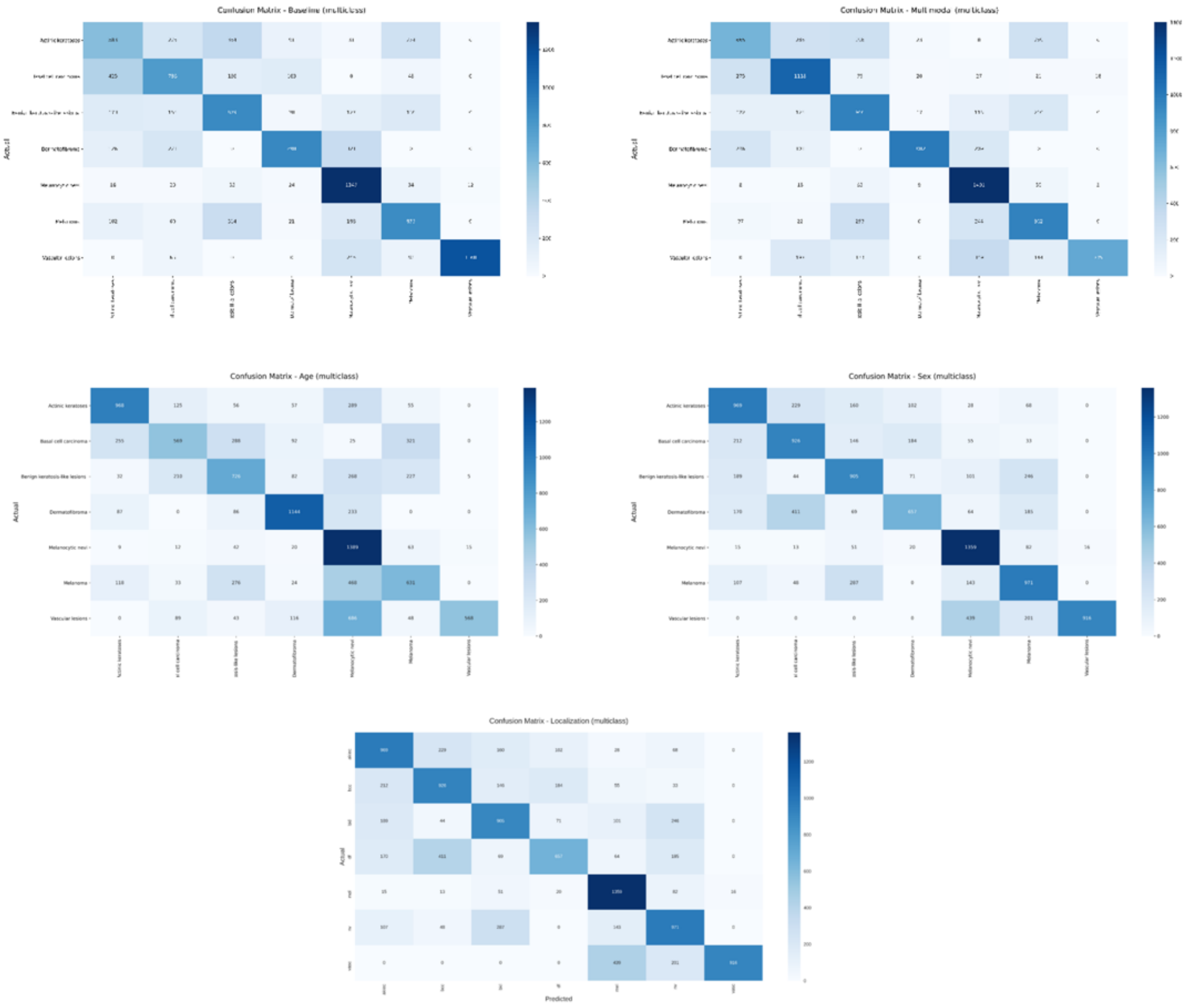


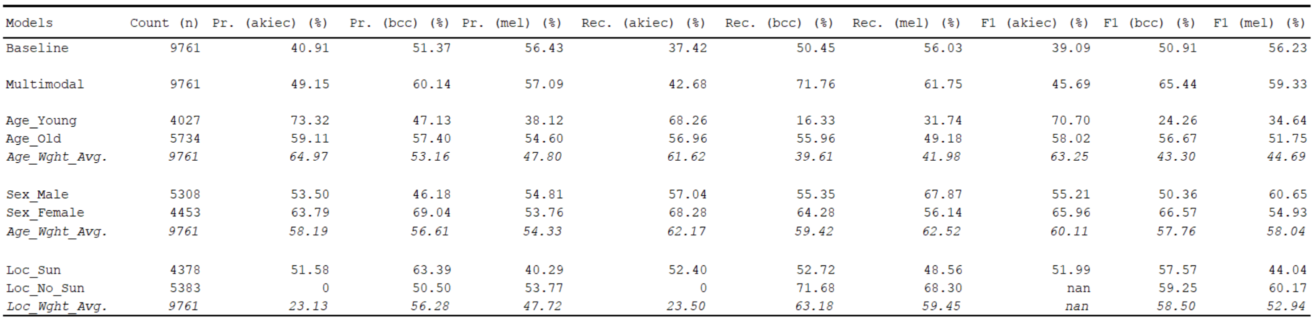
Figure 11. Confusion matrices for the multiclass classification for the baseline, the multimodal model and for the aggregated age, sex and localization models.

Table 4 gives a detailed overview of the performance on the predetermined metrics. In contrast to the binary classification, here the performance per malignant lesion will be examined. In contrast with the binary classification, here the multimodal classifier clearly outperforms the baseline model, on every metric for every class. Apparently, the binary classification was not hard enough for the ResNet baseline to have a detrimental effect of lacking the information contained in the metadata. Now the multimodal model clearly shows the benefit of utilizing the metadata it is very interesting to see how the sub-models compare. Inspection shows the results are rather ambiguous; the scores on all *akiec* metrics shatter that of the baseline and multimodal model, while the performance on the other metrics is more fickle. This behavior can also be seen in the ROC-curves for the three malignant classes (Figure 12). A probable cause for this is that there are not enough instances of each class to cater for stable performance. The sun-covered model (no_sun) even lacks one class, resulting in the inability to calculate some metrics. Data augmentation and oversampling can only do so much. Although the sum of all these observations lead to a somewhat ambiguous conclusion as the capriciousness of the classifiers do not provide stable results, one similar observation can be done as in the binary classifier; the performances between the two age, sex and localization models is again significant.

The capricious results of the multi-model differentiate it from the much more stable multimodal model. Undoubtedly, the limited data, as compared to the multimodal, plays a role, but it should be noticed that this is inherent to model based on subsets. This issue might be partially mitigated by including more data, however that would benefit the multimodal model as well.

Table 4

Model performance in multiclass classification



Models	Count (n)	Pr. (akiec) (%)	Pr. (bcc) (%)	Pr. (mel) (%)	Rec. (akiec) (%)	Rec. (bcc) (%)	Rec. (mel) (%)	F1 (akiec) (%)	F1 (bcc) (%)	F1 (mel) (%)
Baseline	9761	40.91	51.37	56.43	37.42	50.45	56.03	39.09	50.91	56.23
Multimodal	9761	49.15	60.14	57.09	42.68	71.76	61.75	45.69	65.44	59.33
Age_Young	4027	73.32	47.13	38.12	68.26	16.33	31.74	70.70	24.26	34.64
Age_Old	5734	59.11	57.40	54.60	56.96	55.96	49.18	58.02	56.67	51.75
Age_Wght_Avg.	*9761*	*64.97*	*53.16*	*47.80*	*61.62*	*39.61*	*41.98*	*63.25*	*43.30*	*44.69*
Sex_Male	5308	53.50	46.18	54.81	57.04	55.35	67.87	55.21	50.36	60.65
Sex_Female	4453	63.79	69.04	53.76	68.28	64.28	56.14	65.96	66.57	54.93
Age_Wght_Avg.	*9761*	*58.19*	*56.61*	*54.33*	*62.17*	*59.42*	*62.52*	*60.11*	*57.76*	*58.04*
Loc_Sun	4378	51.58	63.39	40.29	52.40	52.72	48.56	51.99	57.57	44.04
Loc_No_Sun	5383	0	50.50	53.77	0	71.68	68.30	nan	59.25	60.17
Loc_Wght_Avg.	*9761*	*23.13*	*56.28*	*47.72*	*23.50*	*63.18*	*59.45*	*nan*	*58.50*	*52.94*

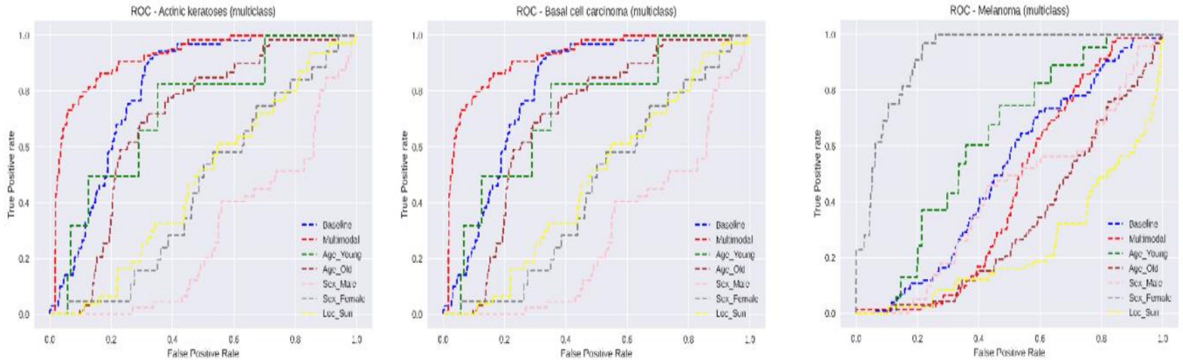


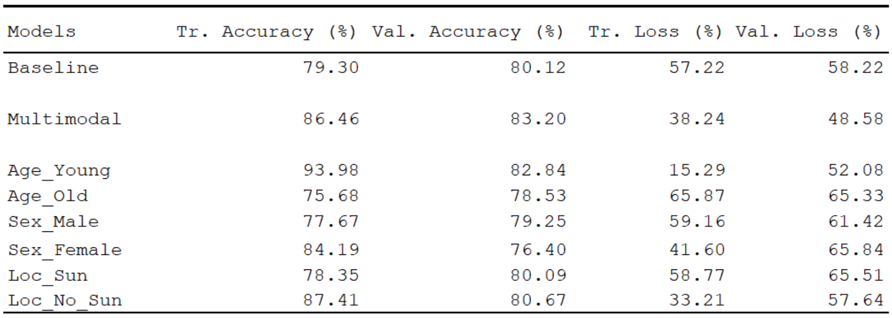
Figure 12. ROC-curves for the multiclass classification, showing the performance of all models per malignant class

5.2 Validation

All models were over 50 epochs, or less when the model stopped learning. Adam is widely considered the best performing optimizer for image classifications. Inspection of the development of the training accuracy and loss over the epochs (examples in Figure 13) Show that training losses steadily dropped, and accuracies increased. Most models did not run over the full 50 epoch, but were terminated earlier because the learning rate dropped too low. Final training accuracies and losses were in general on par with the validation accuracies and losses (see Table 5 and 6). which indicate the models trained well.

Table 5

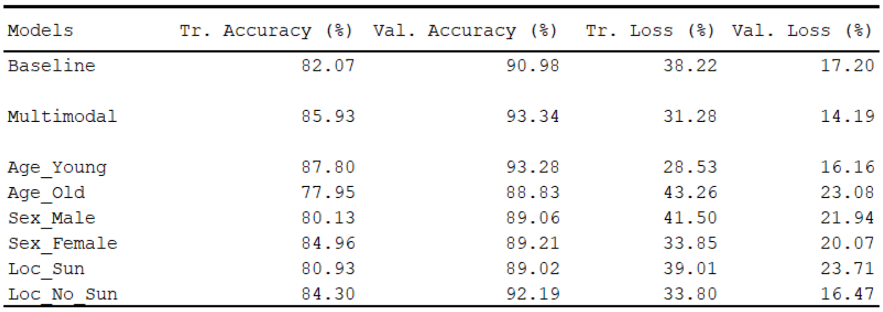
Training and validation accuracy and loss comparison in binary classification.



Models	Tr. Accuracy (%)	Val. Accuracy (%)	Tr. Loss (%)	Val. Loss (%)
Baseline	79.30	80.12	57.22	58.22
Multimodal	86.46	83.20	38.24	48.58
Age_Young	93.98	82.84	15.29	52.08
Age_Old	75.68	78.53	65.87	65.33
Sex_Male	77.67	79.25	59.16	61.42
Sex_Female	84.19	76.40	41.60	65.84
Loc_Sun	78.35	80.09	58.77	65.51
Loc_No_Sun	87.41	80.67	33.21	57.64

Table 6

Training and validation accuracy and loss comparison in multiclass classification.



Models	Tr. Accuracy (%)	Val. Accuracy (%)	Tr. Loss (%)	Val. Loss (%)
Baseline	82.07	90.98	38.22	17.20
Multimodal	85.93	93.34	31.28	14.19
Age_Young	87.80	93.28	28.53	16.16
Age_Old	77.95	88.83	43.26	23.08
Sex_Male	80.13	89.06	41.50	21.94
Sex_Female	84.96	89.21	33.85	20.07
Loc_Sun	80.93	89.02	39.01	23.71
Loc_No_Sun	84.30	92.19	33.80	16.47

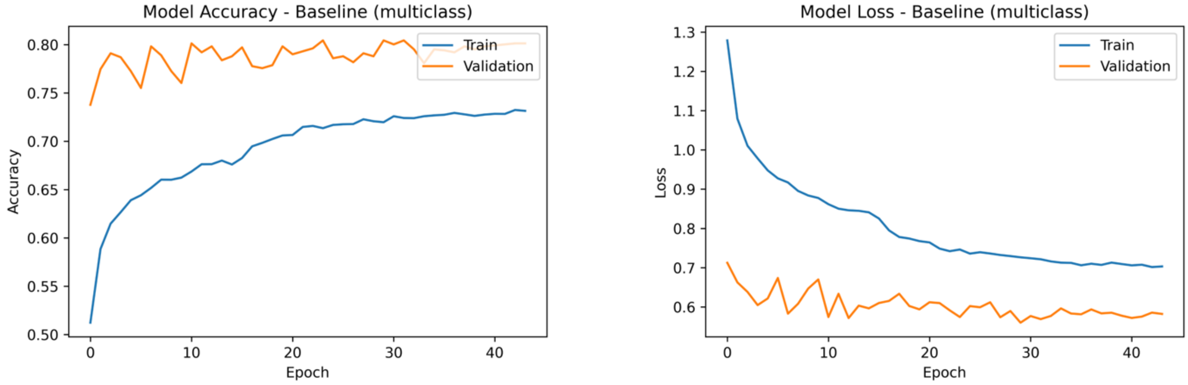


*Figure 13.*Training and validation accuracy and loss for the baseline multiclass model

6. Discussion and Conclusion

6.1 Goal and Findings

Investigate performance difference between multimodal and multi-model approaches. The analysis has indeed showed several differences; First of all the multimodal performance is much more stable, for better or for worse. It beats the performance of the baseline when the classification task gets more difficult. The multi-model approach's performances are less stable, but sometimes significantly outperform the other models. It is however very interesting to see how these models behave if trained on more data (at least concerning the smaller classes). The performance differences within the sub-models also beg for more understanding. This research does not aim to understand these differences, but the performance differences are worth investigating deeper, although it should also be checked first whether this behavior holds when larger training sets are used.

6.2 Limitations

The most significant limitations are the class imbalance and the very limited size of the smaller classes. The latter is especially of concern as it makes the results less stable, making drawing conclusion harder. Larger and more balanced datasets would mitigate these issues.

6.3 Scientific Contribution and Implications

This research has shows that different utilizations of metadata can lead to different performances, for better or worse. The fact that some classifiers vastly outperformed the baseline classifier is reason to dig deeper into how multi-models can increase performance and understanding; the latter of which could lead to even more sophisticated methods, for example distinct models per subset and the including of more forms of metadata. In any case it is recommended to use (even) larger datasets for these tasks.

References

Arisi M, Zane C, Caravello S, Rovati C, Zanca A, Venturini M, Calzavara-Pinton P. Sun Exposure and Melanoma, Certainties and Weaknesses of the Present Knowledge. Front Med (Lausanne). 2018 Aug 30;5:235. doi: 10.3389/fmed.2018.00235. PMID: 30214901; PMCID: PMC6126418.

Binder M, Steiner A, Schwarz M, Knollmayer S, Wolff K, Pehamberger H. Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study. Br J Dermatol. 1994 Apr;130(4):460-5. doi: 10.1111/j.1365-2133.1994.tb03378.x. PMID: 8186110.

Binder M, Kittler H, Dreiseitl S, Ganster H, Wolff K, Pehamberger H. Computer-aided epiluminescence microscopy of pigmented skin lesions: the value of clinical data for the classification process. Melanoma Res. 2000 Dec;10(6):556-61. doi: 10.1097/00008390-200012000-00007. PMID: 11198477.

M. Calderisi, G. Galatolo, I. Ceppa, T. Motta and F. Vergentini, "Improve Image Classification Tasks Using Simple Convolutional Architectures with Processed Metadata Injection," *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Sardinia, Italy, 2019, pp. 223-230, doi: 10.1109/AIKE.2019.00046.

Claeson, Magdalena & Gillstedt, Martin & Whiteman, David & Paoli, John. (2017). Lethal Melanomas: A Population-based Registry Study in Western Sweden from 1990 to 2014. Acta Dermato Venereologica. 97. 10.2340/00015555-2758.

K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

J. Kawahara, S. Daneshvar, G. Argenziano and G. Hamarneh, "Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 538-546, March 2019, doi: 10.1109/JBHI.2018.2824327.

Maron, Roman & Weichenthal, Michael & Utikal, Jochen & Hekler, Achim & Berking, Carola & Hauschild, Axel & Enk, Alexander & Haferkamp, Sebastian & Klode, Joachim & Schadendorf, Dirk & Jansen, Philipp & Holland-Letz, Tim & Schilling, Bastian & Kalle, Christof & Hling, Stefan & Gaiser, Maria & Hartmann, Daniela & Gesierich, Anja & Kähler, Katharina & Brinker, Titus. (2019). Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. European Journal of Cancer. 119. 57-65. 10.1016/j.ejca.2019.06.013.

Morgese F, Sampaolesi C, Torniai M, Conti A, Ranallo N, Giacchetti A, Serresi S, Onofri A, Burattini M, Ricotti G, Berardi R. Gender Differences and Outcomes in Melanoma Patients. Oncol Ther. 2020 Jun;8(1):103-114. doi: 10.1007/s40487-020-00109-1. Epub 2020 Feb 4. PMID: 32700073; PMCID: PMC7359998.

Ngiam, Jiquan & Khosla, Aditya & Kim, Mingyu & Nam, Juhan & Lee, Honglak & Ng, Andrew. (2011). Multimodal Deep Learning. Proceedings of the 28th International Conference on Machine Learning, ICML 2011. 689-696.

Sinz C, Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, Braun RP, Cabo H, Gourhant JY, Kreusch J, Lallas A, Lapins J, Marghoob AA, Menzies SW, Paoli J, Rabinovitz HS, Rinner C, Scope A, Soyer HP, Thomas L, Zalaudek I, Kittler H. Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin.

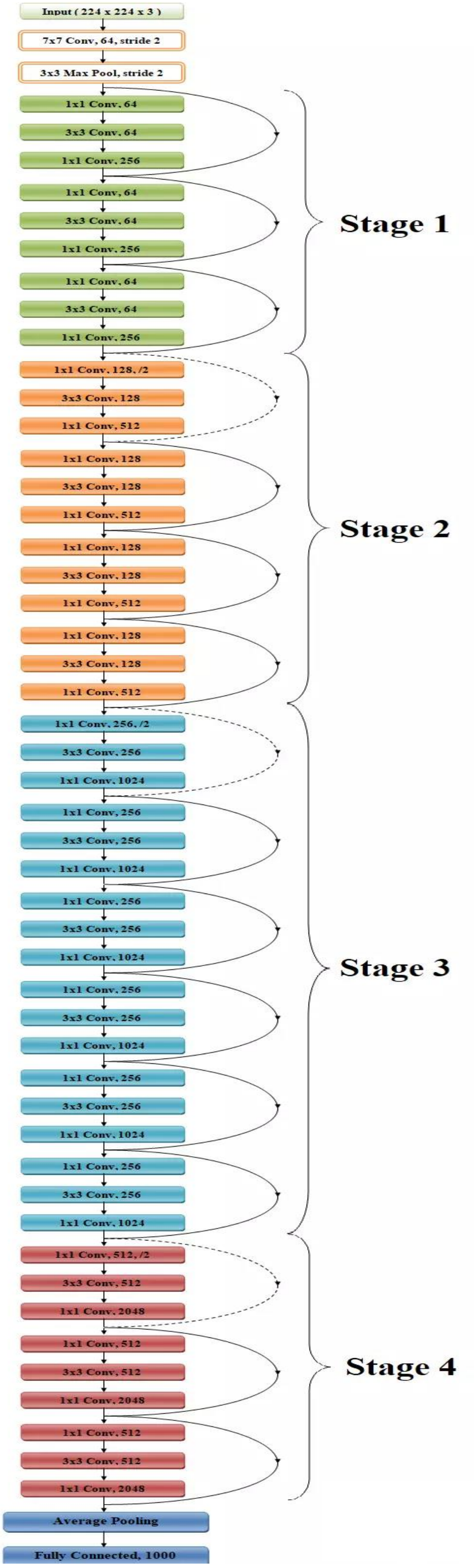
J Am Acad Dermatol. 2017 Dec;77(6):1100-1109. doi: 10.1016/j.jaad.2017.07.022. Epub 2017 Sep 20. PMID: 28941871.

Tschandl, P., Rosendahl, C. & Kittler, H. 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

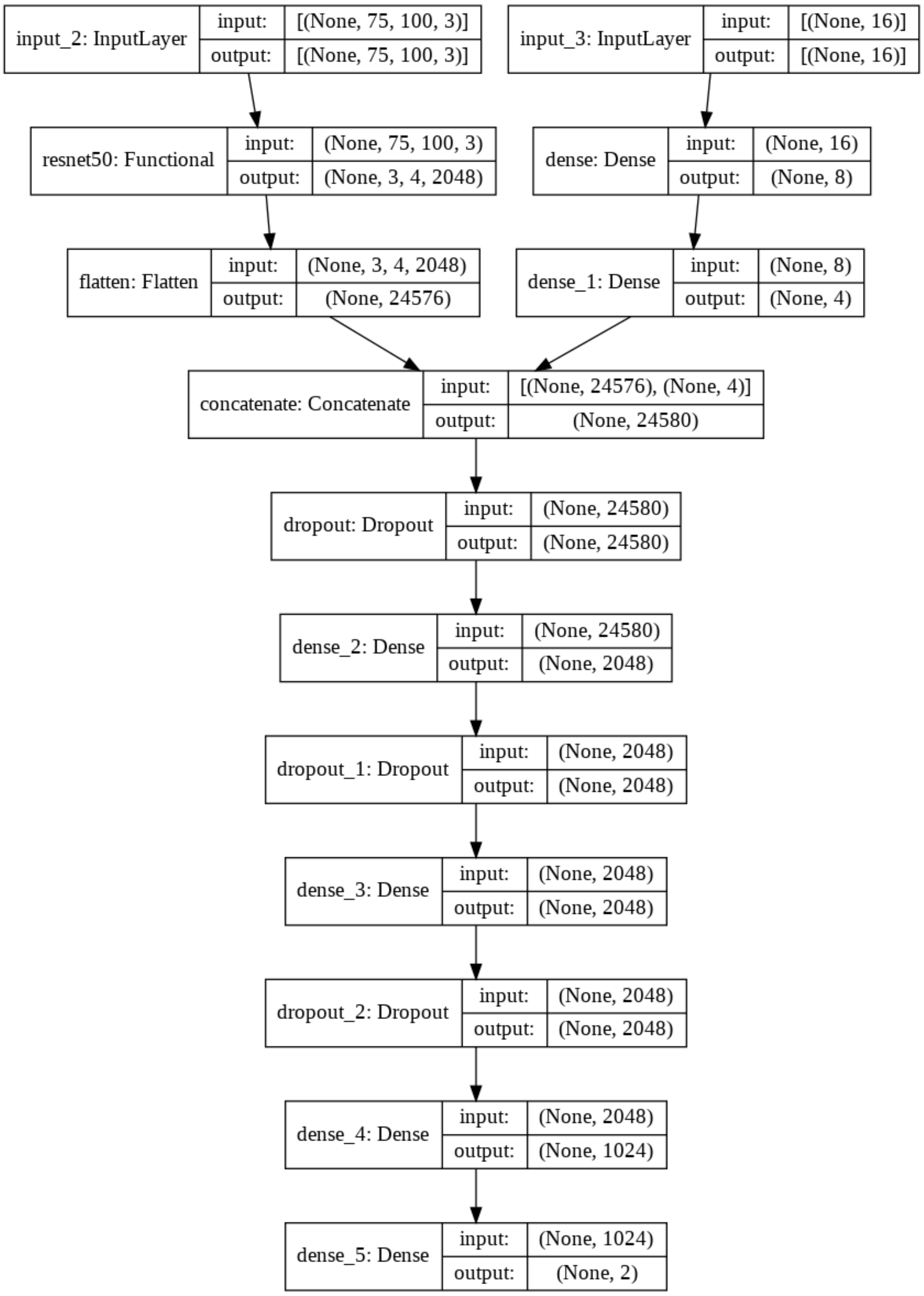
Yap, Jordan & Yolland, William & Tschandl, Philipp. (2018). Multimodal Skin Lesion Classification using Deep Learning. Experimental Dermatology. 27. 10.1111/exd.13777.

Appendix

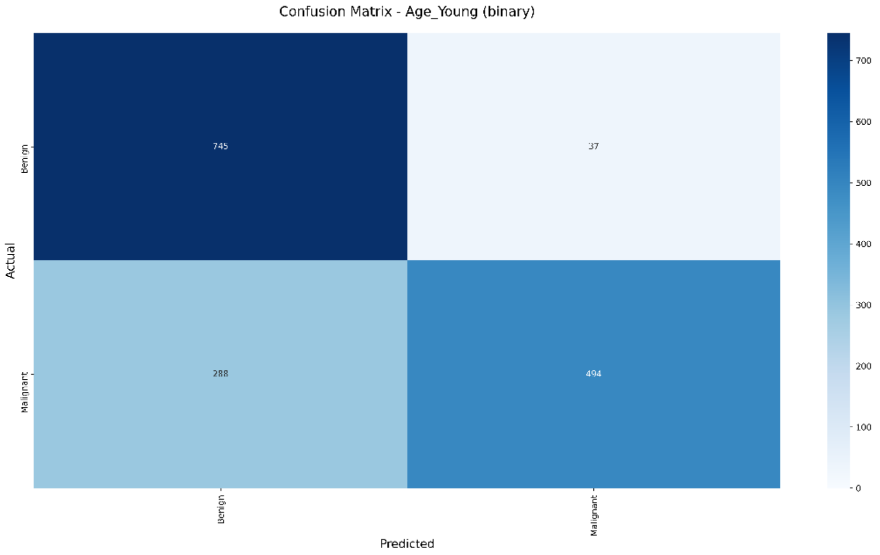
Appendix A: ResNet50 Architecture

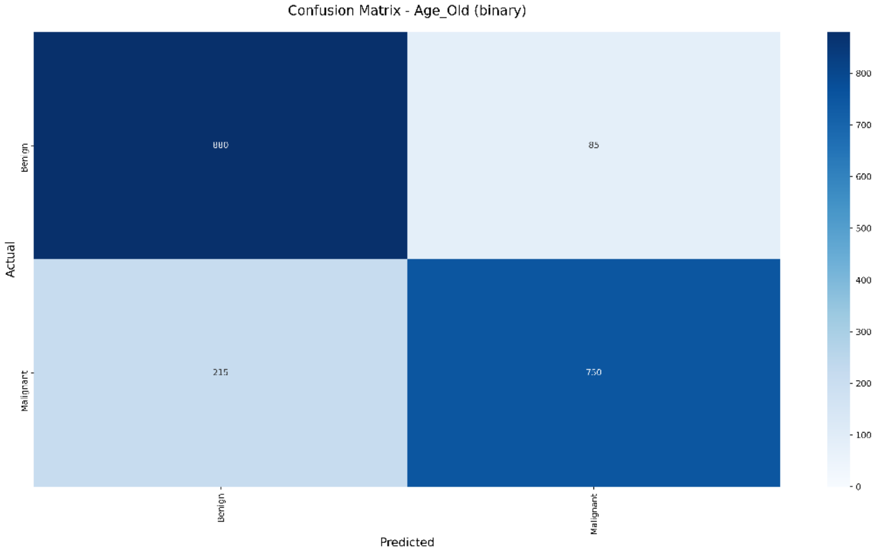


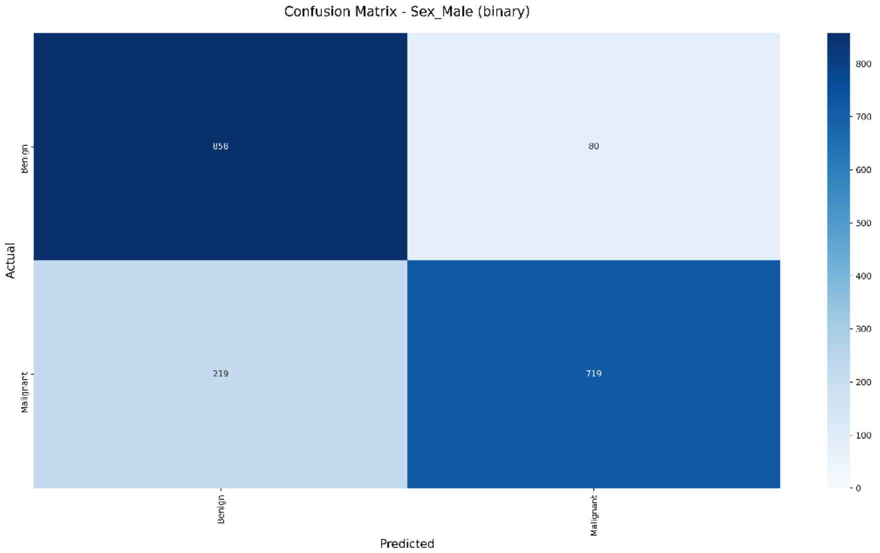
Appendix B: Full Multimodal Network Architecture

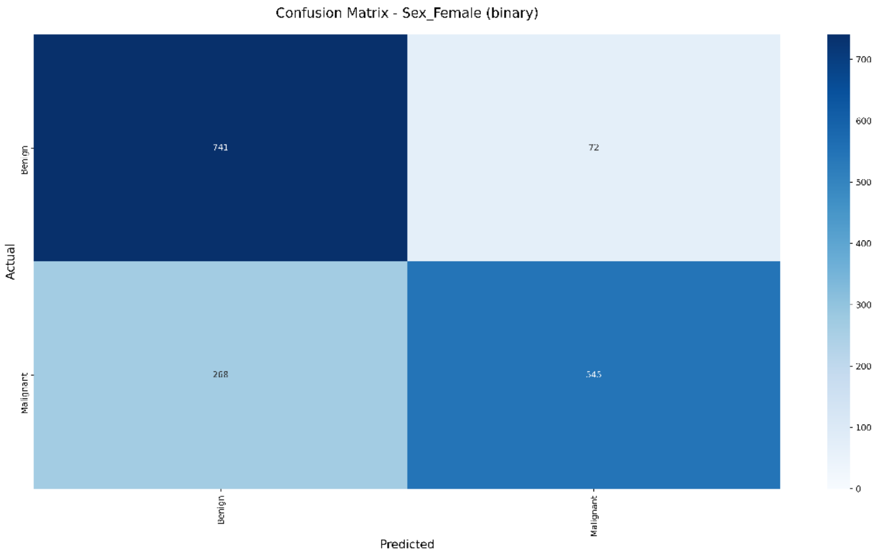


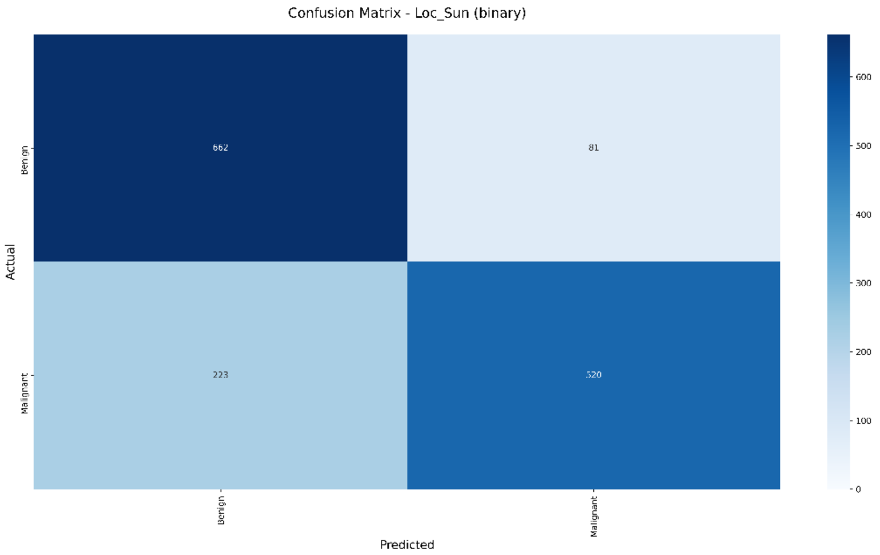
Appendix C: Confusion Matrices for all Sub-Models in the Multi-Model Approach

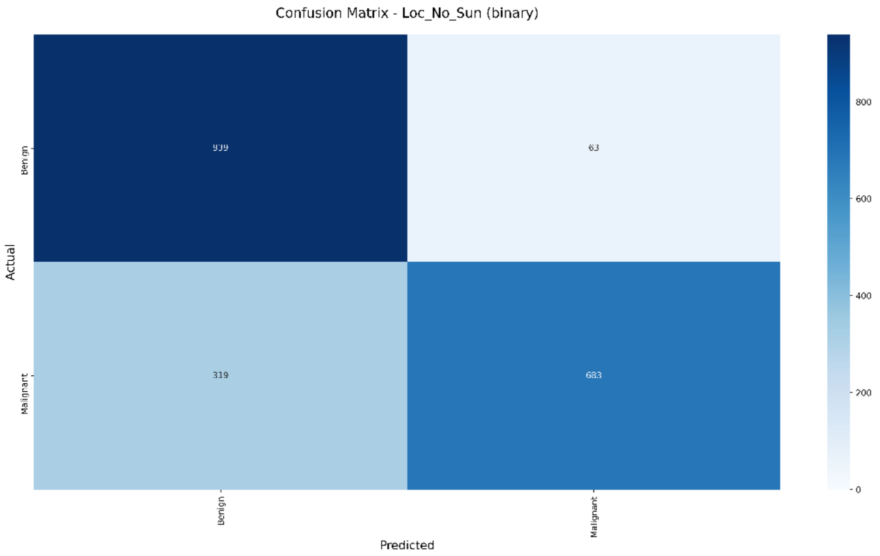


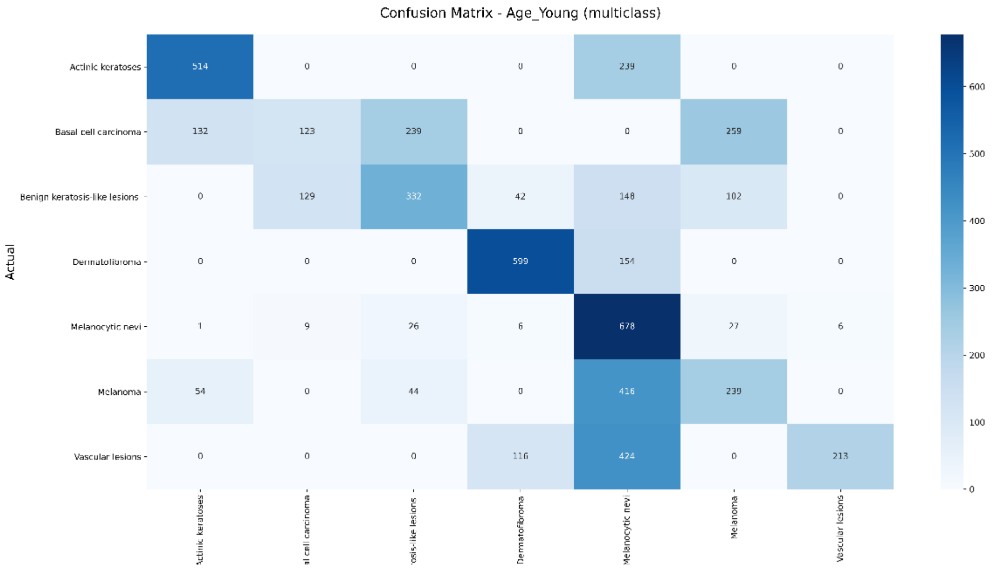


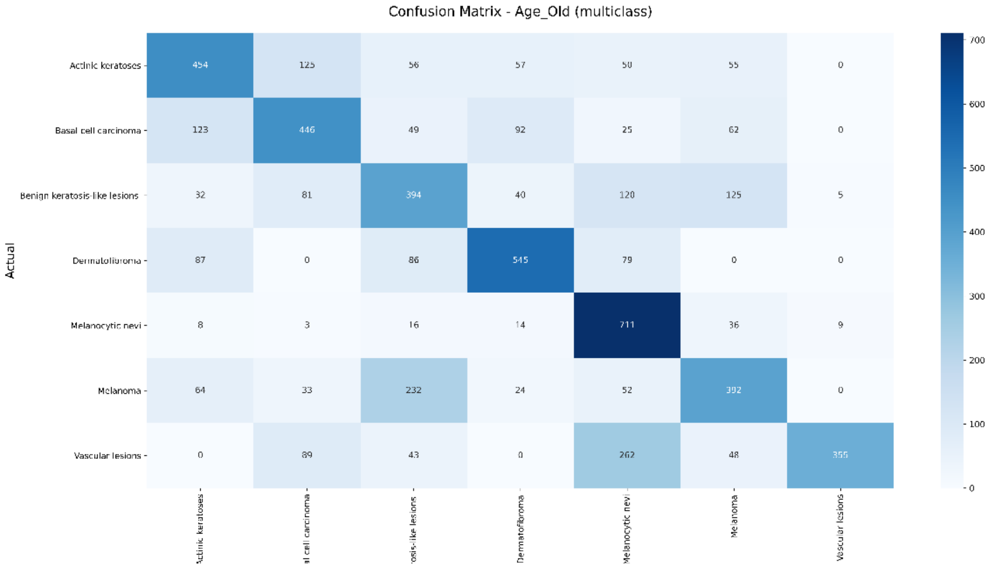


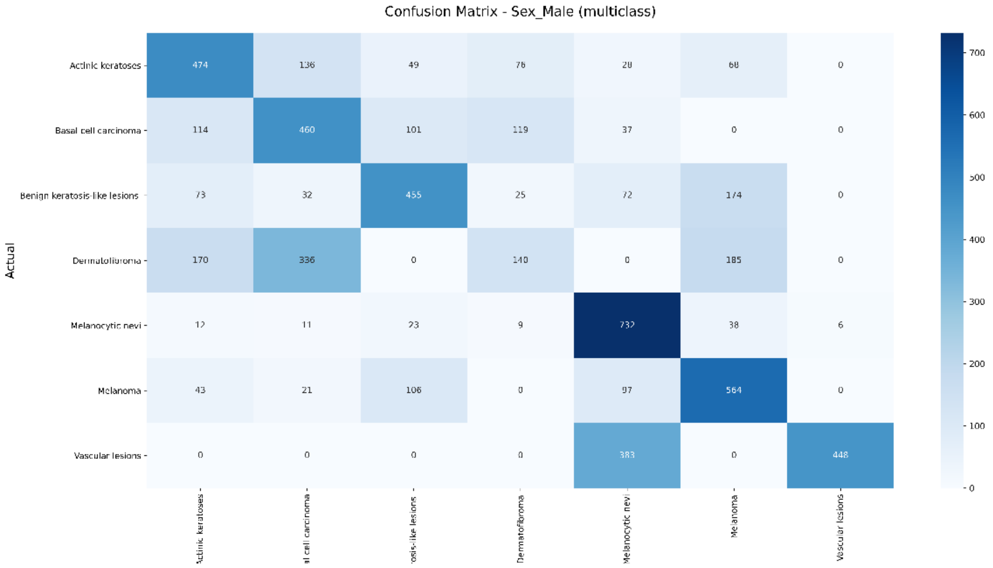

Confusion Matrix - Sex_Female (binary)
Actual
Benign
Malignant
741
72
268
545
Benign
Malignant
Predicted
700
600
500
400
300
200
100
0


Confusion Matrix - Loc_Sun (binary)
Actual
Benign
Malignant
662
81
223
520
Benign
Malignant
Predicted
600
500
400
300
200
100
0

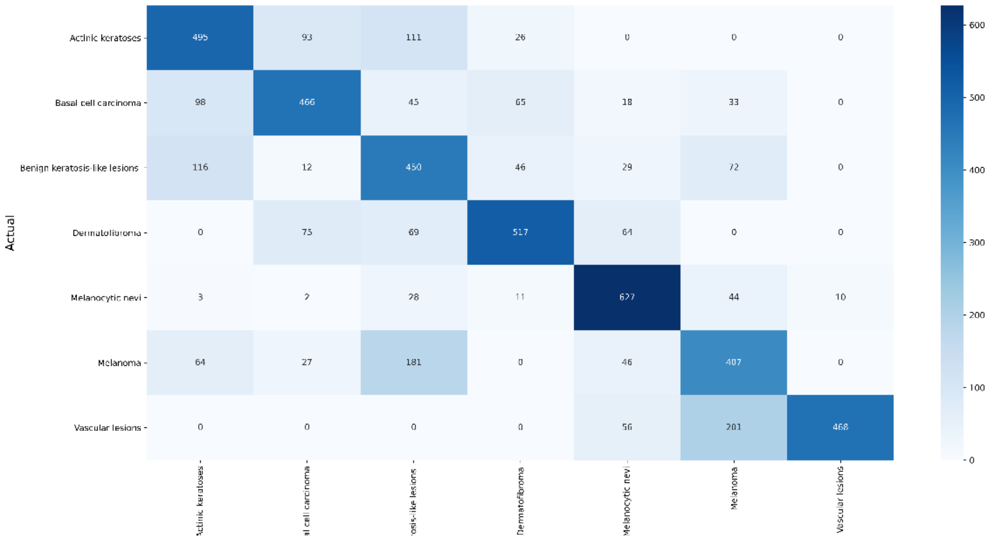

Confusion Matrix - Loc_No_Sun (binary)
Actual
Benign
Malignant
939
63
319
683
Benign
Malignant
Predicted
800
600
400
200
0


Confusion Matrix - Age_Young (multiclass)
Actual
Actinic keratoses
514 0 0 0 239 0 0
Basal cell carcinoma
132 123 239 0 0 259 0
Benign keratosis-like lesions
0 129 332 42 148 102 0
Dermatofibroma
0 0 0 599 154 0 0
Melanocytic nevi
1 9 26 6 678 27 6
Melanoma
54 0 44 0 416 239 0
Vascular lesions
0 0 0 116 424 0 213
ctinic keratoses
al cell carcinoma
osis-like lesions
Dermatofibroma
Melanocytic nevi
Melanoma
Vascular lesions
600
500
400
300
200
100
0

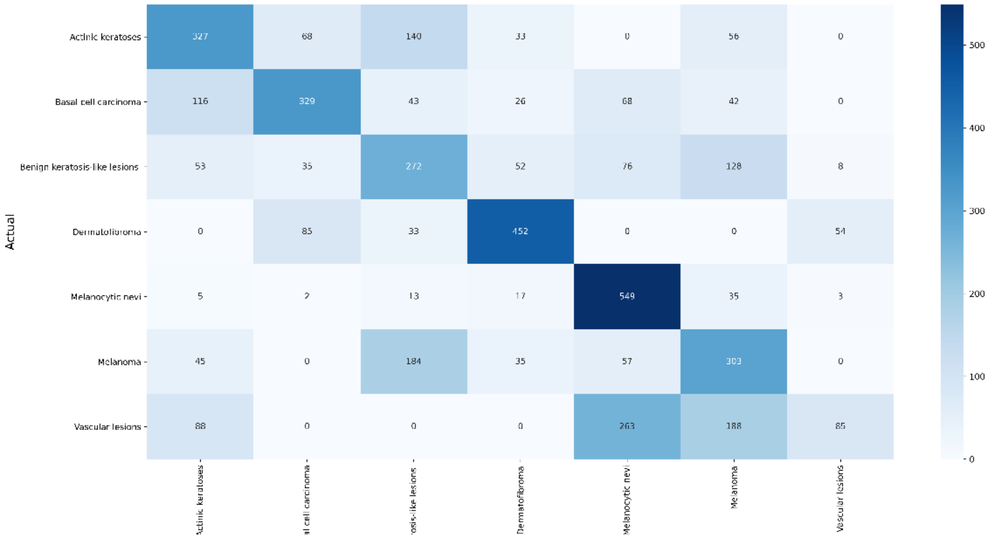

Confusion Matrix - Age_Old (multiclass)
Actual
Actinic keratoses
454 125 56 57 50 55 0
Basal cell carcinoma
123 446 49 92 25 62 0
Benign keratosis-like lesions
32 81 394 40 120 125 5
Dermatofibroma
87 0 86 545 79 0 0
Melanocytic nevi
8 3 16 14 711 36 9
Melanoma
64 33 232 24 52 392 0
Vascular lesions
0 89 43 0 262 48 355
ctinic keratoses
al cell carcinoma
osis-like lesions
Dermatofibroma
Melanocytic nevi
Melanoma
Vascular lesions
700
600
500
400
300
200
100
0


Confusion Matrix - Sex_Male (multiclass)
Actual
Actinic keratoses
474 136 49 76 20 68 0
Basal cell carcinoma
114 460 101 119 37 0 0
Benign keratosis-like lesions
73 32 455 25 72 174 0
Dermatofibroma
170 336 0 140 0 185 0
Melanocytic nevi
12 11 23 9 732 38 6
Melanoma
43 21 106 0 97 564 0
Vascular lesions
0 0 0 0 383 0 448
ctinic keratoses
al cell carcinoma
osis-like lesions
Dermatofibroma
Melanocytic nevi
Melanoma
Vascular lesions
700
600
500
400
300
200
100
0

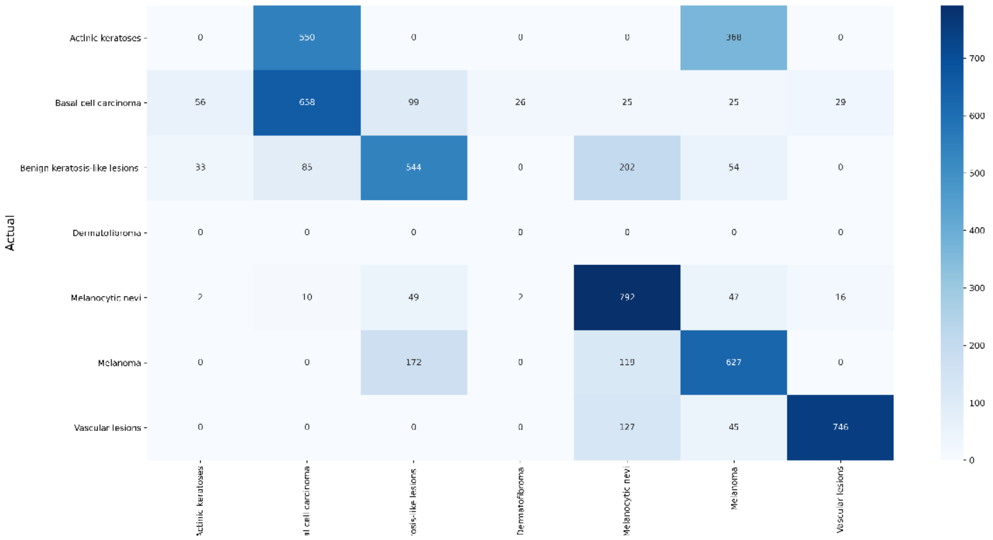
Confusion Matrix - Sex_Female (multiclass)



Confusion Matrix - Loc_Sun (multiclass)



Confusion Matrix - Loc_No_Sun (multiclass)


Actual
Actinic keratoses 0 550 0 0 0 368 0
Basal cell carcinoma 56 658 99 26 25 25 29
Benign keratosis-like lesions 33 85 544 0 202 54 0
Dermatofibroma 0 0 0 0 0 0 0
Melanocytic nevi 2 10 49 2 792 47 16
Melanoma 0 0 172 0 119 627 0
Vascular lesions 0 0 0 0 127 45 746
ctinic keratoses
al cell carcinoma
osis-like lesions
Dermatofibroma
Melanocytic nevi
Melanoma
Vascular lesions