Concept-bottleneck modelling for interpretable melanoma detection

Xuelong An

ABSTRACT

Pure deep learning approaches achieve state-of-the-art melanoma detection accuracy, outperforming their human counterparts. However, their outputs are untrustworthy given that they are black-box models with which there is no principled way to understand their inner mechanisms underlying a decision. To mitigate such elusive behavior, we adopt a concept-bottleneck technique to make outputs from a existing high-performing ResNet interpretable for melanoma classification. Our results indicate that while the bottleneck technique hampers classification accuracy (achieving 43% compared to a pure ResNet which reaches 58% after finetuning for 5 epochs), this comes with the benefits of producing interpretable outputs by manipulating predicted concepts.

Keywords: concept-bottleneck model, melanoma classification, ResNet

1 INTRODUCTION

Melanoma is a type of malignant tumour with approximately 325000 cases estimated globally in 2020, corresponding to about 20% of cases of skin cancer worldwide. It has a mortality rate between 2% and 5% with variability owed to ethnicity, gender and age (Arnold et al., 2022).

This global health burden, similar to other cancers, can be mitigated with its early detection and subsequent treatment. The prognosis of melanoma is done through visual inspection suspicious pigmented lesions (SPLs), which is challenging owed to its similarity with other skin lesions (Jojoa Acosta et al., 2021). There exists non-invasive, imaging diagnostic methods using biomicroscopy, fundus photography, or ultrasonography (Nasr-Esfahani et al., 2016). However, these are often inflexible to scale-up due to the need of access to dermatologists with substantial expertise Patel et al. (2023), which may be infeasible for the high volume pigmented lesions available to process (Lewis, 2021). Thanks to the advent of artificial intelligence-based (AI) methods spearheaded by deep neural networks (DNNs), there is now extensive research exploring computerized approaches that provide a comfortable, less expensive, and speedy detection of melanoma (Dildar et al., 2021) that can aid practitioners in simplifying the workload of melanoma detection.

Deep learning (DL) is a technique that extracts task-dependent features from a given a dataset to solve a problem (Dildar et al., 2021). It has been extensively studied in the context of melanoma detection, often achieving high classification accuracy that outperforms human dermatologists (Patel et al., 2023). Despite constantly redefining state-of-the-art (SoTA) performance, DNN-based architectures, like convolutional neural networks (CNNs), are black-box models, and thus their outputs are untrustworthy. Suppose a dermatologist uses a DL model to identify melanoma given an image of a

SPL, and it gives a positive output. She might ask what is the model looking at to obtain such prediction, and if what the model is looking is irrelevant, can she tell it to change its prediction? This is important to know in a medical setting as it impacts what treatment a patient undergoes, or whether they can avoid unnecessary procedures and *why*. This is important to consider especially if we are interested in model deployment, where trustworthiness is paramount.

In this work, we explore a concept-bottleneck modelling technique by Pang et al. (2020), which consists of a plug-and-play module that can be added to existing high-performing DNNs to extract concepts from SPL images that help classify them in an interpretable way. We frame melanoma detection as a binary classification task that leverages contextual information on the image and run experiments to showcase advantages and limitations of our proposed method.

2 METHODS AND MATERIALS

2.1 Dataset

We use an open-source dataset from the Kaggle Competition platform provided by the Society for Imaging Informatics in Medicine (SIIM). It $\mathscr{D} = \{\mathbf{X}, \mathbf{c}, \mathbf{y}\}$ consists of $|\mathbf{X}| = 33126$ images downsampled to a 256×256 pixel resolution, of which values are normalized. Each is annotated with $y \in \{0, 1\}$ based on whether it is a malignant melanoma or benign lesion. There is contextual information $\mathbf{c}$ attached to each sample as metadata, measuring:

- the sex of the patient (when unknown, will be blank): either female or male.
- approximate patient age at time of imaging, which is either 0, 10, 15, 20, 25, 30, 35, 40, 45, 60, 65, 70, 75, 80, 85, 90 or unknown.
- anatomical site of the imaged lesion: head/neck, upper extremity, lower extremity, torso, palms/soles, oral/genital or unknown.
- detailed diagnosis of the lesion (train only): nevus, melanoma, seborrheic keratosis, lentigo NOS, lichenoid keratosis, solar lentigo, cafe-au-lait macule, atypical melanocytic proliferation, or unknown.

We treat each attribute as a categorical variable and one-hot them for modelling [1], resulting in 35 binary variables. We drop NaN values as a simple preprocessing step. Most SoTA DNNs ignore the above metadata, however they prove valuable to shed insight into black-box predictions. We denote the above 35 binary attributes as *concepts*.

We randomly partition the dataset into a training and validation split using an 80:20 ratio. The test set $\mathscr{D}_{test} = \{\mathbf{X}, \mathbf{c}\}$ only consists of images, of which pixel values we normalize, and concepts. Model performance is known after submitting its predictions to the platform.

This is a challenging task given the heavy class skewness towards benign samples ($|\mathbf{y}_0| = 32542$) (see Figure 1), as well as the presence of confounding objects per image: hairs, blood vessels, inflammation around the lesion, among others (Jojoa Acosta et al., 2021) that may influence a prediction. To partially mitigate this imbalance, we perform data transformation techniques (see Appendix 5.2) over the training images to increase their diversity.

[1] While age can be considered a continuous variable, we note that the original authors only annotated the age intervals, so it is reasonable to treat it as a categorical variable.

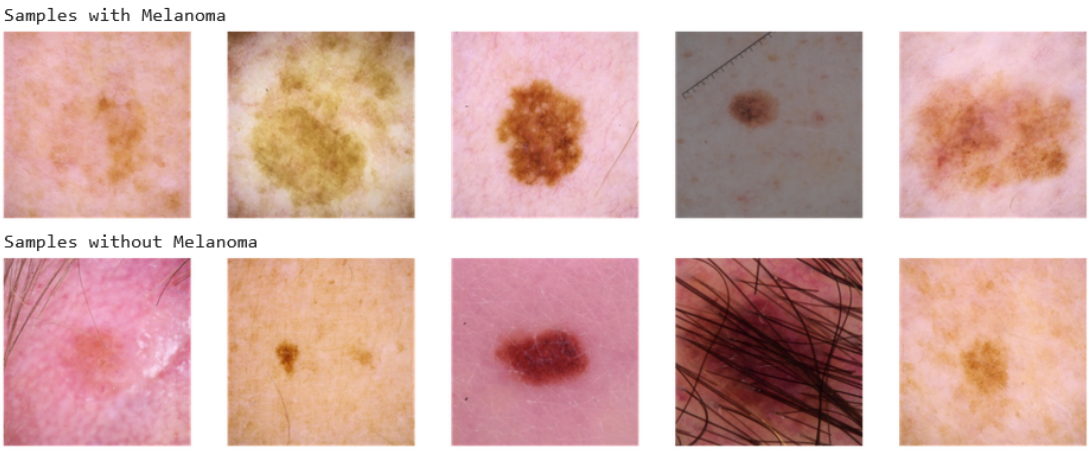

Figure 1. A depiction of training samples of SPLs, where top row consists of malignant samples and bottom row are benign

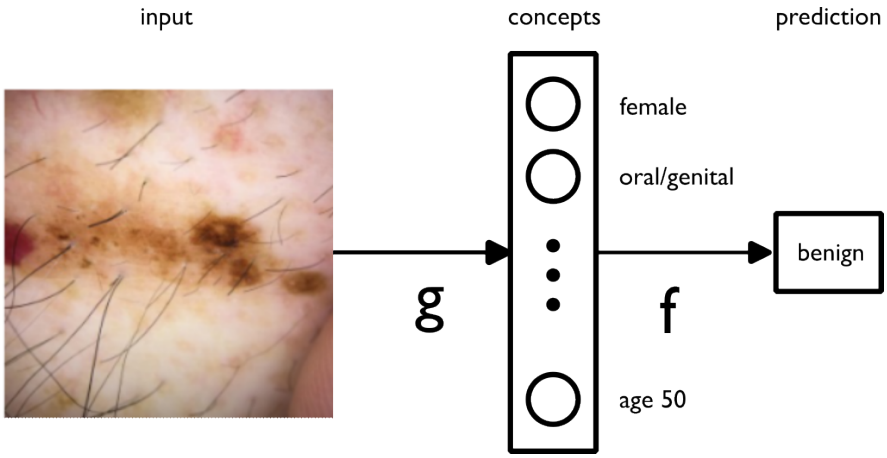

Figure 2. A depiction of the model architecture described by Pang et al. (2020) adapted to the task of melanoma detection.

2.2 Model

To mitigate black-box behavior, we resort to concept-bottleneck modelling[2]. We are given training points $\{(\mathbf{x}^{(i)}, \mathbf{c}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{n}$, where $\mathbf{x} \in \mathbb{R}^d$, and $\mathbf{c} \in \mathbb{R}^k$ is a vector of k concepts provided as metadata. The task of concept bottleneck models is to predict $\mathbf{c}$ from $\mathbf{x}$, and then predict $\mathbf{y}$. Their functional form is $\hat{y} = f_\phi(g_\theta(x))$, where $g : \mathbb{R}^d \to \mathbb{R}^k$ is parameterized by θ and maps an image $\mathbf{x}$ into the concept space (”palm lesion”, ”female”, etc.)[3], and $f : \mathbb{R}^k \to \mathbb{R}$ is parameterized by ϕ and maps concepts into a final prediction (”benign”)(Figure 2). In our setting, for $g_\theta(\cdot)$ we use a pretrained ResNet18 where we reshape the last layer to make a pointwise binary prediction on what attributes are present in the input image. We treat $f_\phi(\cdot)$ as a binary logistic regression model. For example, given an image, the overall pipeline first predicts the probabilities of concepts,

[2]We release our code at `https://www.kaggle.com/code/awxlong/concept-bottleneck-modelling-in-melanoma`

[3]They are described as ”bottleneck” because $k \ll d$, and the output depends on this low-dimensional vector.

e.g. with highest activations at being from a female, approximately 50, with a lesion in the upper extremity that is solar lentigo. From there, the model makes a decision on whether it is benign or not.

There are three ways to assemble f and g (see (Pang et al., 2020) for their details), either separately, sequentially or jointly. We pick the joint configuration where both components are trained simultaneously in an end-to-end manner by optimizing for θ, ϕ:

$$\hat{\theta}, \hat{\phi} = \arg\min_{\theta,\phi} \sum_{i}^{n} [\mathscr{L}_Y(f_\phi(g_\theta(x^{(i)})), y^{(i)}) + \sum_{j}^{|c|} \lambda \mathscr{L}_{C_j}(g_\theta(x^{(i)}), c^{(i)})]$$

, where $\mathscr{L}_{C_j} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is the binary cross-entropy (BCE) with logit loss that measures the discrepancy between the predicted and true *j-th* concept, $\mathscr{L}_Y : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is the BCE loss that measures the discrepancy between predicted and true targets, and $\lambda > 0$ is a hyperparameter that controls the influence of the loss of g_θ, where smaller value indicates that concepts contribute less to the final prediction. .

This is because in experiments done by Pang et al. (2020), the joint variant was the highest performing among the three, and achieved competitive performance or even surpassed that of pure DNNs approaches in their tasks of bird classification and bone spur severity prediction.

2.3 Experiment setup

We define a baseline to be a pretrained ResNet18 to illustrate the advantages of our proposed approach. We use the following hyperparameters taken from (Pang et al., 2020) for both models and do not optimize them: Adam optimizer with default settings, a learning rate of 0.01, a linear learning rate scheduler with a step size of 20, weight decay of 0.0004 to prevent overtraining, batch size of 64 and fine-tune for 5 epochs[4]. In the case of our bottleneck model, we define $\lambda = 0.001$ as borrowed from the original authors. We train for 2 seeds $\{42, 56\}$ on a P100 GPU on a Kaggle Notebook.

3 EVALUATION AND DISCUSSION

3.1 Results

We monitor and report several metrics during training and validation (Figure 3):

- Binary classification accuracy measured by the ratio of true positives and total amount of predictions, where the model's predictions are binarized via a cutoff of ≥ 0.5. Closer to 1 is better.
- The BCE Loss between target and predictions, where closer to 0 is better. For the bottleneck model, we choose not to report concept prediction accuracy (an average over each concept) as it is out of the scope of our work.
- For validation, we also report the area under the receiver operating characteristic (ROC) curve, or AUC for short, which is an aggregate measure of a model's performance across all possible classification thresholds, where closer to 1 means better performance.

[4]Preliminary experiments show that this is the amount feasible under constrained computational resources

- After training and validation, we also compute the F1-Score $= 2 \times \frac{precision \times recall}{precision + recall}$, a measure of the harmonic mean of precision[5] and recall[6], where closer to 1 indicates better performance. This is because the dataset is highly imbalanced and a high precision may be due to the model being biased towards identifying benign samples.

The results show high variability in the metric values, arguably because both models just started training and haven't converged. They achieve roughly similar training and validation accuracies ($\approx 98\%$). The F1-scores for training and validation splits are 0 for both models when the classification threshold is set at ≥ 0.5. In future work we can either augment the malignant samples in order to balance the dataset, or perform a grid search to choose a threshold catered to the class skewness.

We submitted the predictions and we obtained 58% for the baseline model, while 43% for our concept-bottleneck model. Despite achieving lower score, we note that we 1) haven't performed hyperparameter optimization and 2) we haven't trained for several epochs. Nonetheless, our concept-bottleneck model comes with a strong advantage of interpretability.

3.2 Diagnosis of learnt concepts

Test labels are not available due to concerns of test-set leakage, so if we resort to black-box deep learning, we wouldn't be able to gain any insight into why our ResNet achieved 58%. This is not the case for our concept-bottleneck model, mainly due to its logistic regression tail that has learnt $p(\mathbf{c}|\mathbf{x})$ a distribution of concepts given input image.

Consider the example where we diagnose Figure 4. The model predicts $y = 0$ it is a benign lesion. By taking the argmax of the concept-bottleneck model's tail, we extract that it predicts with high confidence it is from a female, from the oral/genital site, a lentigo NOS diagnosis (a harmless lesion), and approximately age 50 (see Equation 1). A dermatologist (or a user who took a photo of their lesion), by having access to these predictions, gains a deeper understanding of why the model gave a particular output. If the concepts are inaccurate, and the model's prediction is a false positive or negative, the user can *intervene* in the predicted concepts, and influence the model's predictions. This is because while the pipeline only needs an image input, the user can also correct predicted concepts as part of f (e.g. changing to male lesion misidentified as female), which can update the decision for the better. This showcases a form of human-computer interaction which we could explore in future work given that concepts are provided in $\mathscr{D}_{test}$, albeit it is out-of-scope here.

Moreover, we can also access the coefficients of f in order to understand what are the key concepts that trigger the label $p(y|\mathbf{c})$ (see Equation 1). We notice a negative bias, which is consistent with the class skewness towards class 0 images. The same analysis can be done for the other variables. A machine learning engineer without a dermatology background could also examine the coefficients for medical insight, i.e, those concepts with highest coefficients can inspire navigating the melanoma literature on discrepancies in incidence rates of benign/malignant lesions dependent on age group, gender or anatomical site of the lesion. These coefficients are also useful to diagnose

[5] ratio of true positives to true and false positives

[6] ratio of true positives to true positives and false negatives

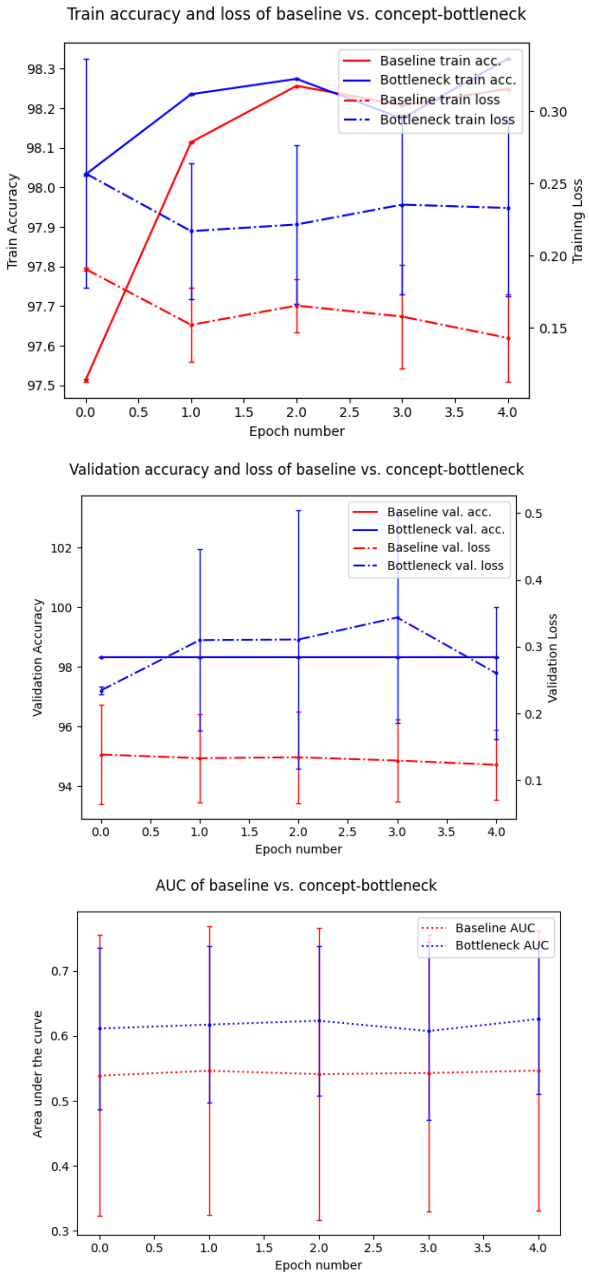

Figure 3. Results obtained after running for 2 seeds for the baseline and bottleneck models. We follow the convention that a blue line always corresponds to the bottleneck's metrics, a filled line to always report accuracy corresponding to the left axis, a dash-dotted line to report loss values corresponding to the right axis (note the different scales) and a dotted line for reporting AUC values. We plot error bars, where we note that their large margins are because we only fine-tune for 5 epochs. At the top and middle, we plot training and validations measurements respectively, where we stop when both models reach around 98% train and validation accuracy. At the middle plot, we notice that both baseline and bottleneck's performances converge, and as such lines overlap. At the bottom, we draw the AUC values, where we stop when the baseline achieves 0.55 ± 0.22 and the bottleneck achieves 0.63 ± 0.11.

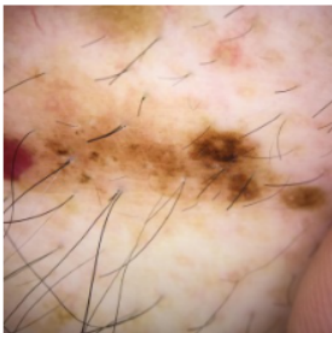

Figure 4. A depiction of test sample 42, where we can extract the predictions of concepts associated to the model, such as its gender, anatomical site, subtype of lesion and age.

potential biases in the collected data. For instance, had the learnt coefficient for females been higher w.r.t males, which is inconsistent with the literature (Bellenghi et al., 2020), we would argue for a scrutiny on the data collection process.

3.3 Limitations and improvements

Despite interpretable results, our concept-bottleneck model has lower test performance than a pure deep learning. We argue that this is mainly due to the class imbalance, as well as constrained computational resources. Also, despite being more interpretable, we note that the model is limited to the concepts available in the training data, i.e, it can't use ethnicity that helps detect melanoma, despite it being a very important factor that could also determine the survival rate of a potential patient (Lam et al., 2021). This is unless the original training data also measures these features and we retrain the model.

Other improvements include addressing the class skewness via artificial augmentation of malignant samples and retraining the model with different hyperparameters, such as increasing the number of epochs to 100 within tolerable computational demands. We are confident that the concept-bottleneck model can match or even surpass the pure ResNet's performance as reported by Pang et al. (2020) in other tasks[7].

Lastly, it is worth noting that thanks to the pipeline's modularity, we can replace either f or g with more powerful SoTA models, as long as they're end-end differentiable. In future work, for the feature extractor g_θ, we are interested in leveraging the self-attention mechanism of pretrained transformers, which has been shown useful for obtaining a disentangled representation of the input image (Faulkner and Zoran, 2022). For f_ϕ, we can use more complex models that explore the interaction of concepts (e.g $female * age\_50$) like symbolic regressors because it is well known that melanoma incidence rates are best studied via higher-order dependencies (Arnold et al., 2022; Bellenghi et al., 2020).

[7]Preliminary experiments show that by increasing epochs to 25 and augmenting class 1 images to build a balanced dataset, the concept-bottleneck model's test performance climbs up to 55.87%

4 CONCLUSION

In this work, we adopted a concept-bottleneck technique by Pang et al. (2020) for melanoma detection. It is a versatile method that can be combined with existing SoTA methods to achieve high performance. By leveraging concepts, predictions are not of black-box nature anymore, which is pivotal in a medical setting where model output must be trustworthy, examinable and modifiable given more contextual information.

Our model is in a better standing for deployment compared to black-box deep learning models. However, there is still a lot of work to do in order to make it a more powerful predictor, and thus widely accessible. This includes hyperparameter optimization for achieving higher test scores, a more intricately labeled meta-dataset to account for important concepts, and providing guidelines on how to perform concept intervention and interpretation.

REFERENCES

Arnold, M., Singh, D., Laversanne, M., Vignat, J., Vaccarella, S., Meheus, F., Cust, A. E., de Vries, E., Whiteman, D. C., and Bray, F. (2022). Global burden of cutaneous melanoma in 2020 and projections to 2040. *JAMA Dermatology*, 158:495–503.

Bellenghi, M., Puglisi, R., Pontecorvi, G., De Feo, A., Carè, A., and Mattia, G. (2020). Sex and gender disparities in melanoma. *Cancers*, 12.

Dildar, M., Akram, S., Irfan, M., Khan, H. U., Ramzan, M., Mahmood, A. R., Alsaiari, S. A., Saeed, A. H. M., Alraddadi, M. O., and Mahnashi, M. H. (2021). Skin cancer detection: A review using deep learning techniques. *International Journal of Environmental Research and Public Health*, 18:5479.

Faulkner, R. and Zoran, D. (2022). Solving reasoning tasks with a slot transformer.

Jojoa Acosta, M. F., Caballero Tovar, L. Y., Garcia-Zapirain, M. B., and Percybrooks, W. S. (2021). Melanoma diagnosis using deep learning techniques on dermatoscopic images. *BMC Medical Imaging*, 21.

Lam, M., Zhu, J. W., Hu, A., and Beecker, J. (2021). Racial differences in the prognosis and survival of cutaneous melanoma from 1990 to 2020 in north america: A systematic review and meta-analysis. *Journal of Cutaneous Medicine and Surgery*, page 120347542110528.

Lewis, M. (2021). An artificial intelligence tool that can help detect melanoma.

Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, S., Jafari, M., Ward, K., and Najarian, K. (2016). Melanoma detection by analysis of clinical images using convolutional neural network.

Pang, W. K., Nguyen, T., Tang, Y., Mussmann, S., Pierson, E., Kim, B., and Liang, P. (2020). Concept bottleneck models. *Proceedings of the 37 th International Conference on Machine Learning*.

Patel, R., Foltz, E. A., Witkowski, A., and Łudzik, J. (2023). Analysis of artificial intelligence-based approaches applied to non-invasive imaging for early detection of melanoma: A systematic review. *Cancers*, 15:4694–4694.

5 APPENDIX

5.1 Logistic regressor's functional form

$$
\begin{aligned}
\hat{y} = & -0.10 \times \text{female} - 0.10 \times \text{male} + 0.08 \times \text{site\_head/neck} \\
& - 0.00 \times \text{site\_low\_extremity} - 0.12 \times \text{site\_oral/genital} \\
& - 0.06 \times \text{site\_palms/soles} + 0.06 \times \text{site\_torso} \\
& - 0.03 \times \text{site\_up\_extremity} + 0.00 \times \text{diag\_atypical\_melanocytic\_proliferation} \\
& + 0.08 \times \text{diag\_cafe\_au\_lait\_macule} - 0.06 \times \text{diag\_lentigo\_NOS} \\
& + 0.04 \times \text{diag\_lichenoid\_keratosis} - 0.05 \times \text{diag\_melanoma} \\
& + 0.09 \times \text{diagnosis\_nevus} + 0.08 \times \text{diag\_seborrheic\_keratosis} \\
& - 0.00 \times \text{diag\_solar\_lentigo} - 0.08 \times \text{diag\_unknown} \\
& + 0.05 \times \text{age\_0} - 0.07 \times \text{age\_10} - 0.05 \times \text{age\_15} \\
& - 0.01 \times \text{age\_20} - 0.03 \times \text{age\_25} + 0.03 \times \text{age\_30} \\
& + 0.04 \times \text{age\_35} - 0.00 \times \text{age\_40} + 0.05 \times \text{age\_45} \\
& - 0.16 \times \text{age\_50} - 0.06 \times \text{age\_55} + 0.00 \times \text{age\_60} \\
& - 0.02 \times \text{age\_65} + 0.17 \times \text{age\_70} + 0.02 \times \text{age\_75} \\
& + 0.06 \times \text{age\_80} - 0.10 \times \text{age\_85} + 0.01 \times \text{age\_90} - 0.18
\end{aligned}
\tag{1}
$$

5.2 Data augmentations

With the existing training dataset, we performed with 50% probability the following image transformations:

- Cutout
- 90^o degrees rotation
- Flipping
- One of changing brightness, or hue saturation
- Adding Gaussian noise
- Blurring
- Optical distortion
- Shift scale rotation