

# Prediction of melanoma types using semi-structured Bayesian deep learning models

Ivonne Kovylov

Konstanz, 30.11.2021

Master Thesis

**Thesis submitted in fulfillment of the requirements  
for the degree of**

**Master of Science (M. Sc.)**


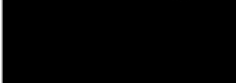
**Hochschule Konstanz**

University of Applied Science

**Department of Computer Science**

Program Master of Science Informatik

Title: **Prediction of melanoma types using semi-structured Bayesian deep learning models**

Masters candidate: Ivonne Kovylov   


1. Supervisor: Prof. Dr. Oliver Dürr  
2. Supervisor: Prof. Dr. Beate Sick

Date of issue: 05.05.2021  
Submission date: 30.11.2021

# Ehrenwörtliche Erklärung

Hiermit erkläre ich *Ivonne Kovylov*, 

- (1) dass ich meine Masterarbeit mit dem Titel

**Prediction of melanoma types using semi-structured Bayesian deep learning models**

am Institut für Optische Systeme unter Anleitung von Prof. Dr. Oliver Dürr und Prof. Dr. Beate Sick selbständig und ohne fremde Hilfe angefertigt habe und keine anderen als die angeführten Hilfen benutzt habe;

- (2) dass ich die Übernahme wörtlicher Zitate, von Tabellen, Zeichnungen, Bildern und Programmen aus der Literatur oder anderen Quellen (Internet) sowie die Verwendung der Gedanken anderer Autoren an den entsprechenden Stellen innerhalb der Arbeit gekennzeichnet habe.
- (3) dass die eingereichten Abgabe-Exemplare in Papierform und im PDF-Format vollständig übereinstimmen.

Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Konstanz, 30.11.2021

---

(Unterschrift)

# Abstract

Title:	Prediction of melanoma types using semi-structured Bayesian deep learning models
Masters candidate:	Ivonne Kovylov
Supervisor:	HTWG Konstanz, Institut für Optische Systeme Prof. Dr. Oliver Dürr Prof. Dr. Beate Sick
Submission date:	30.11.2021
Keywords:	deep learning, probabilistic modeling, interpretability, uncertainty, statistics

Interpretability and uncertainty modeling are important key factors for medical applications. Moreover, data in medicine are often available as a combination of unstructured data like images and structured predictors like patient's metadata. While deep learning models are state-of-the-art for image classification, the models are often referred to as 'black-box', caused by the lack of interpretability. Moreover, DL models are often yielding point predictions and are too confident about the parameter estimation and outcome predictions. On the other side with statistical regression models, it is possible to obtain interpretable predictor effects and capture parameter and model uncertainty based on the Bayesian approach. In this thesis, a publicly available melanoma dataset, consisting of skin lesions and patient's age, is used to predict the melanoma types by using a semi-structured model, while interpretable components and model uncertainty is quantified. For Bayesian models, transformation model-based variational inference (TM-VI) method is used to determine the posterior distribution of the parameter. Several model constellations consisting of patient's age and/or skin lesion were implemented and evaluated. Predictive performance was shown to be best by using a combined model of image and patient's age, while providing the interpretable posterior distribution of the regression coefficient is possible. In addition, integrating uncertainty in image and tabular parts results in larger variability of the outputs corresponding to high uncertainty of the single model components.

# Contents

<b>Ehrenwörtliche Erklärung</b>	<b>I</b>
<b>Abstract</b>	<b>II</b>
<b>Nomenclature</b>	<b>IV</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Aim and Objectives . . . . .	2
1.2 Structure . . . . .	2
<b>2 Dataset</b>	<b>3</b>
<b>3 Methods</b>	<b>5</b>
3.1 Logistic Regression . . . . .	5
3.2 Modeling image data . . . . .	7
3.3 Combining image and tabular data . . . . .	8
3.4 Bayesian neural network . . . . .	9
<b>4 Experiments</b>	<b>14</b>
4.1 Models . . . . .	14
4.2 Preparation . . . . .	15
4.3 Evaluation . . . . .	16
<b>5 Results and Discussion</b>	<b>17</b>
5.1 Non-Bayesian models . . . . .	17
5.2 Bayesian models . . . . .	20
<b>6 Summary and Outlook</b>	<b>29</b>
6.1 Results of non-Bayesian models . . . . .	29
6.2 Results of Bayesian models . . . . .	30
6.3 Outlook . . . . .	30
<b>Bibliography</b>	<b>VI</b>
<b>Appendix</b>	<b>IX</b>

# Nomenclature

## Acronyms

ANN	Artificial neural network
AUC	Area under the curve
BNN	Bayesian neural network
CDF	Cumulative distribution function
CI	Confidence interval
CI <sub>b</sub>	Complex intercept for image data
CI <sub>x</sub>	Complex intercept for tabular data
CNN	Convolutional neural network
CPD	Conditional probability distribution
CRI	Credible interval
DL	Deep learning
ELBO	Evidence lower bound
HDI	High density interval
ID	In-distribution
IQR	Interquartile range
ISIC	International Skin Imaging Collaboration
KL	Kullback-Leibler
LS <sub>x</sub>	Linear shift for tabular data
MAP	Maximum a posteriori
MCMC	Markov-Chain-Monte-Carlo
MF	Mean-field
NLL	Negative log-likelihood
NN	Neural network

ONTRAM	Ordinal neural network transformation models
OOD	Out-of-distribution
OR	Odds ratio
PPD	Posterior predictive distribution
SI	Simple intercept
TM-VI	Transformation model-based variational inference
VI	Variational inference

### **Symbols**

Be	Beta-density
$\beta(x)$	Log odds ratio function for tabular data, defined as Complex intercept
$\beta_0$	Intercept parameter
$\beta_1$	Slope parameter, log odds ratio
$\lambda$	Initialized variational parameter
$\mathbb{E}$	Expected log likelihood
$\sigma$	Sigmoid function
$\vartheta(B)$	Log odds ratio function for image data, defined as Complex intercept
$\vartheta_M$	Coefficients of Bernstein polynomial
$B$	Image data
$D$	Data
$h(z)$	Transformation function
$h$	Linear predictor
$M$	Order of Bernstein polynomial
$N$	Normal distribution
$n$	Number of samples
$p_D$	Conditional probability
$q_\lambda(w)$	Variational distribution
$w$	Weights of a NN
$x$	Tabular data
$y$	Outcome

# Chapter 1

## Introduction

Malignant melanomas are aggressive skin tumors of melanocytic origin and responsible for over 90% of all skin tumor deaths. In recent decades, there has been a significant increase in the incidence rate. This is on the one hand due to UV-exposed leisure and vacation behavior, which is a well-known major risk [23]. In Germany, an incidence of 19 cases per 100,000 population is recorded, with an approximately equal number of men and women affected [6]. Additionally, as with other tumors, the risk of being diagnosed with melanoma increases with age [22, 24]. But on the other hand, there is also a relatively high incidence at young ages compared to other tumor entities [22]. The use of dermoscopy as a noninvasive skin imaging technique for melanoma diagnosis leads to improved accuracy in the diagnosis of pigmented lesions when used correctly. However, the degree of experience of dermoscopy methods is important for diagnostic accuracy. Even for experts, diagnosing melanoma can be time-consuming and based on subjective judgment [17]. Because early detection of melanoma is critical for efficient treatment, there is a need for computer-aided systems for automated skin lesion models.

In recent years, artificial neural networks (ANNs) and the associated deep learning (DL) methods have gained great success in modeling unstructured data like images with the ability of automatic feature engineering [9]. Previous works show [15] that utilizing ANNs, which are based on convolutional neural networks (CNNs) in dermoscopy images, can result in a good performance in the early detection of melanoma type and therefore ease the diagnosis of medical experts.

However, ANNs are often referred to 'black box' models because they are generally difficult to interpret. For medical applications, additional interpretation possibilities focusing on different risk factors may be relevant, e.g. 'what effect does age have on the type of melanoma?' Within the usage of statistical regression models, it is possible to obtain interpretable parameters. These models take structured data as input, which are tables. The combination of both, structured and unstructured data, yielding to a semi-structured model, where the benefits of the statistical and deep learning community are combined. Another drawback of traditional deep learning models is that they are often overconfident within their predictions and only yield point estimates of parameters and the resulting predictions [9]. However, uncertainty modeling, with the underlying probabilistic view, is crucial especially in high-risk domains like medical applications. For example, what happens if a melanoma image appears with a structure that has not been observed by the model before?

In this case, Bayesian modeling helps to reveal an *out-of-distribution* situation, captured as *epistemic uncertainty* [16], manifesting in a wide uncertainty distribution. *Aleatoric uncertainty* [16] on the other hand captures data uncertainty and thus it can't be reduced by adding more data. Bayesian neural networks (BNN) [7] allow to capture both types of



uncertainties by means of posterior distributions over the weights and the outcome predictions of a NN. The determination of the model posteriors is computationally impossible due to a large number of parameters of NNs and therefore BNNs need an approximation. The posterior distribution is hard to determine exactly but can be approximated by variational inference (VI) [3]. In the case of a complex posterior distribution needs to be fitted, a recently developed method called *transformation model-based variational inference (TM-VI)* can be used [11].

## 1.1 Aim and Objectives

This master thesis aims to predict the melanoma type by using semi-structured Bayesian models, while interpretable components and model uncertainty is quantified.

This results in the following objectives:

1. Develop a CNN based on lesion images, a NN based on patient's age, and a combination of both by using a publicly available dataset consisting of skin lesions and patient's age.
2. Interpret the effect of patient's age with and without the impact of the image.
3. Apply TM-VI to semi-structured Bayesian models.
4. Quantifying uncertainty of the model parameter estimations of the different model components by using the TM-VI method.
5. Evaluate and compare the prediction performance of all models.
6. Evaluate the parameter and model uncertainty.

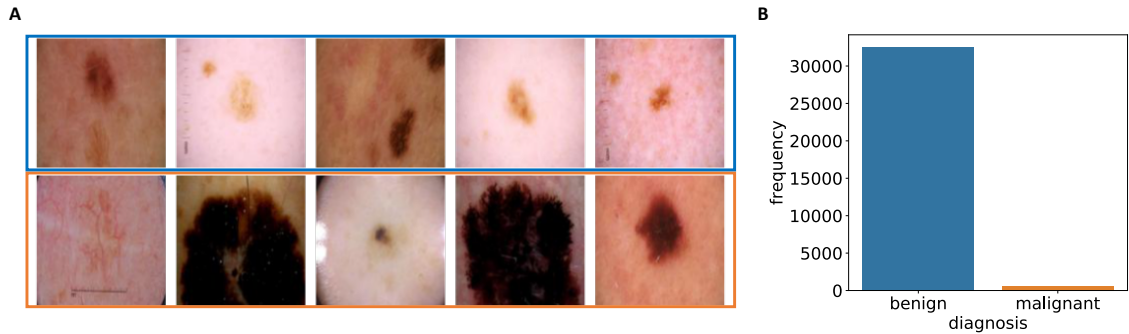
## 1.2 Structure

First, the dataset is described, which is used for implementing the models. Afterward, the methods are presented, which provide a brief introduction of the fundamentals of logistic regression, deep learning models to model image and tabular data, and Bayesian neural networks. Chapter 4 contains the experimental setup. This is followed by a discussion of the results. The findings are summarized in a conclusion in the last chapter. In addition, a brief exploration of future research points is given. Additional material is provided in appendix A-C.

## Chapter 2

# Dataset

Melanoma is a disease with a worldwide increasing incidence. One of the most extensive collections of melanoma datasets can be found in the International Skin Imaging Collaboration (ISIC) repository [13] including confirmed diagnosis of patient's skin lesions labeled by expert dermatologists. ISIC makes new public datasets available every year and has thus grown significantly over the years. For this thesis, the ISIC 2020 Challenge dataset 'Skin Lesion Analysis Towards Melanoma Detection' was used, which contains 33126 skin lesion images from over 2000 patients, collected from six different institutions [25]. The goal of the ISIC 2020 Challenge is to classify benign and malignant lesions. Figure 2.1 shows example lesions from the ISIC archive divided into benign and malignant melanoma types. As it is shown in figure 2.1, the class distribution is quite imbalanced with  $\approx 98\%$  benign and  $\approx 2\%$  malignant skin lesions.

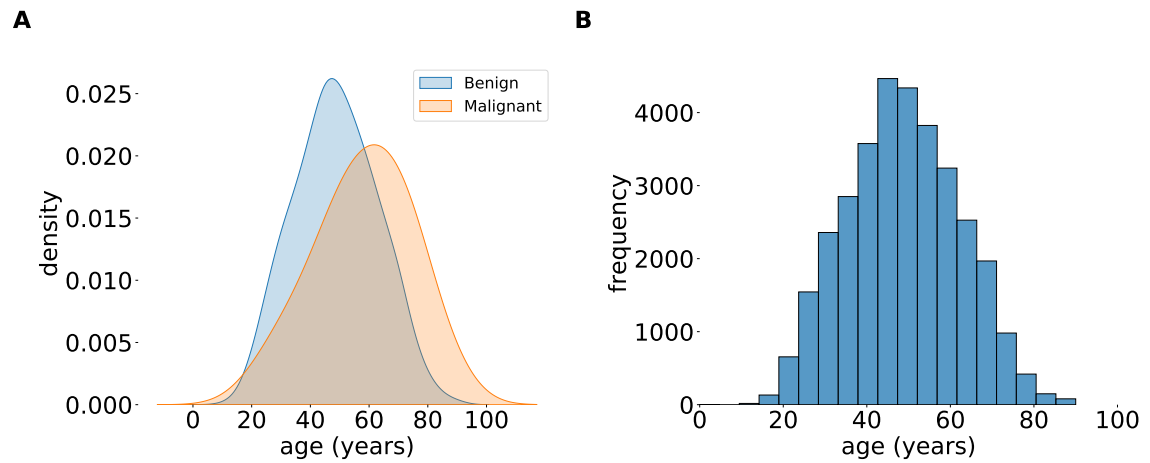


**Figure 2.1:** **A:** Example skin lesions from the ISIC 2020 Challenge Dataset with diagnosis 'benign' (top row) and 'malignant' (bottom row). **B:** Distribution of binary outcome variable of ISIC 2020 Challenge Dataset. The outcome is quite imbalanced with  $\approx 98\%$  'benign' ( $y = 0$ ) and  $\approx 2\%$  'malignant' ( $y = 1$ ) diagnosis.

In addition to the labeled dermoscopy imaging, the respective metadata, consisting of the patient's age, gender, location of lesion, diagnosis, and the patient's identification number is available. To demonstrate the use of interpretable Bayesian models, we restrict here to the age of the patient in combination with the lesion images.<sup>1</sup> The approximate age of the patient is almost equally distributed in both diagnosis groups with an average age of about 50 years and a standard deviation of about 14 years (see panel B figure 2.2). There were 68 missing values of the patient's age, which have been supplemented with the median age. Concerning the outcome variable, age seems to have an impact on the type of

<sup>1</sup>Statistical significance for the patient's age is shown in table 5.1.

melanoma, such that older patients are more likely to be affected by malignant melanoma than younger ones (see panel A figure 2.2).



**Figure 2.2:** Distribution of patient's age. **A:** Mass distribution of patient's age related to the outcome variable 'benign' and 'malignant'. **B:** Distribution of age with an average value of about 50 years.

# Chapter 3

## Methods

In this chapter the used methods are described that are relevant for this thesis to model Bayesian semi-structured models with the possibility of interpretation.

### 3.1 Logistic Regression

For a binary outcome and tabular data as input, logistic regression is used, when an interpretable model is preferred. From a probabilistic perspective, a conditional probability  $p_D = p(y = 1|D)$  is modeled, which indicates the probability for an event occurs ( $y = 1$ ), depending on the data  $D$ . The resulting outcome is a conditional probability distribution (CPD) which follows a Bernoulli distribution ( $y|D \sim \text{Ber}(p_D)$ ) where  $p_D$  is the probability that an event occurs ( $y = 1$ ). In logistic regression, instead of probabilities, odds are considered for interpretation and are defined as:

$$\text{odds}(y = 1) = \frac{p}{1 - p} \quad (1)$$

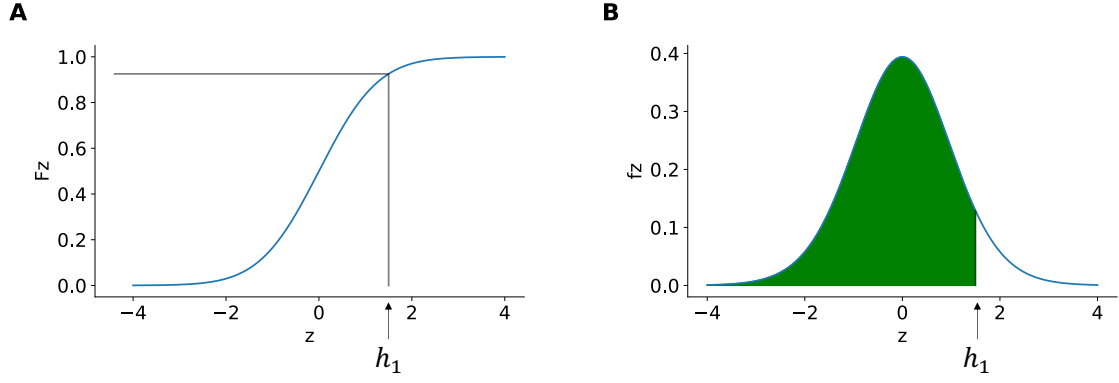
where  $p$  represents the probability that an event occurs ( $y = 1$ ). The range of odds is between 0 and  $\infty$ . To obtain the range of values between  $-\infty$  and  $\infty$ , the odds are logarithmized. Thus a linear predictor can be used to model the log-odds. The resulting logit scale is defined by:

$$z = \log\left(\frac{p}{1 - p}\right) = \text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (2)$$

The probabilities, estimated by a logistic regression, are obtained by solving equation 2 for  $p$ , also known as the sigmoid function  $\sigma(z)$ :

$$p(y = 1|x) = p(x) = \frac{1}{1 + e^{-z}} \quad (3)$$

As defined in equation 2, logistic regression can thus be represented as a continuous latent variable model (illustrated in figure 3.1), in which the continuous latent variable  $z$  (see panel B in figure 3.1) is modeled by a linear predictor. Therefore, rather than modeling the probability directly, one cutpoint  $h_1$  in the latent variable  $z$  is modeled. At this cutpoint, the cumulative distribution function (CDF) of the latent variable  $z$  is evaluated, which yields the probability  $p(y = 1|x)$ , modeled by logistic regression (see panel A in figure 3.1).



**Figure 3.1:** Logistic regression as latent variable model. Panel **A** depicts the cumulative distribution function (CDF) of the latent variable  $z$ , which is known as logistic regression. The probability density function (PDF) modeling  $z$  as linear predictor (panel **B**). There is one cutpoint  $h_1$ , which yields the probability for an outcome (**B**) and is evaluated by the logistic regression (**A**).

### 3.1.1 Interpretation of regression coefficients

For interpretation, the log-odds (see equation 2) are exponentiated ( $e^{\beta_k}$ ), since the link to the probabilities is nonlinear. This is known as odds ratio:

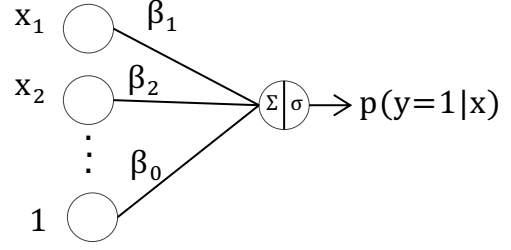
$$\text{OR}_{x \rightarrow x+1} = \frac{\text{odds}(x+1)}{\text{odds}(x)} = \frac{e^{\beta_0 + \beta_k x_k + 1}}{e^{\beta_0 + \beta_k x_k}} = e^{\beta_k} \quad (4)$$

Therefore  $e^{\beta_k}$  can be interpreted as the factor by which the odds for class 1 changes, when the predictor  $x_i$  is increased by one unit while keeping all other predictors constant. If  $\text{OR}_{x \rightarrow x+1} > 1$  there is a positive association, if  $\text{OR}_{x \rightarrow x+1} < 1$  it is negative and  $\text{OR}_{x \rightarrow x+1} = 1$  resumes no association between predictor and the outcome. It should be noted that the interpretation as OR's should be done with caution due to two different phenomena, namely, confounding bias and non-collapsibility [5]. Even if confounders can systematically be adjusted, there is still the effect of non-collapsibility, which means that the conditional OR is not equal to the marginal OR. For a more detail explanation refer to Burgess [5].

### 3.1.2 Logistic regression as neural network

Logistic regression can be modeled as a neural network (NN) without a hidden layer with a sigmoid as an activation function. The weights of the NN correspond to the regression coefficients (see equation 2).

**Figure 3.2:** Logistic regression as a one layer neural network (NN) without hidden layer. The weights correspond to the regression coefficients. The sigmoid activation function is used to model the probability for an event  $p(y = 1|x) = \sigma(z)$ .



The parameters are estimated by the maximum likelihood approach. Instead of maximizing the likelihood, the NN is trained by minimizing the negative log-likelihood (NLL) over all samples [9], which is defined as:

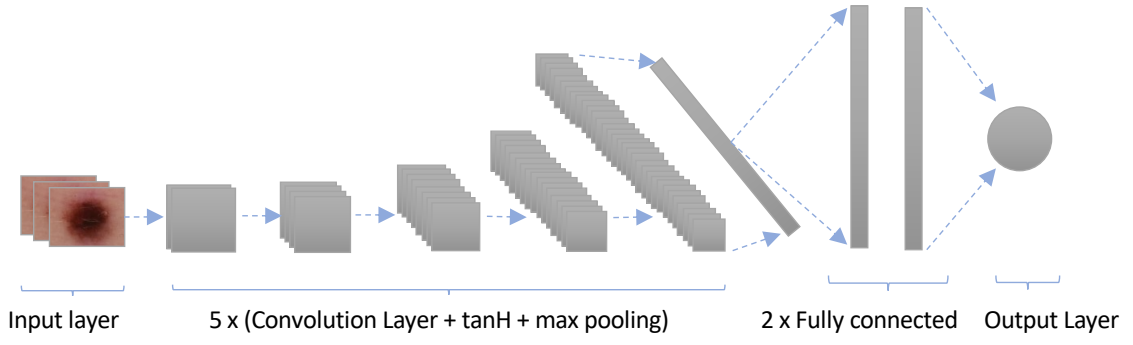
$$\text{NLL} = -\frac{1}{n} \sum_{i=1}^n (y_i \log(p_1(x_i)) + (1 - y_i) \log(1 - p_1(x_i))) \quad (5)$$

where  $n$  represents the number of samples,  $y_i$  the true label of a binary event based on data  $x_i$ , and  $p_1(x_i)$  the predicted probability for an event.

### 3.2 Modeling image data

The image data are processed by a deep convolutional neural network (CNN) since it is known for its good performance in the field of image classification. CNN is a type of NN that contains stacked convolutional layers followed by fully connected layers. While convolution is an operation process that detects local characteristics in certain regions of the image and is thus used for feature extraction, the fully connected part is used for classification. For more detailed information refer to Goodfellow et al. [9].

The architecture used in this thesis is inspired by the work of Abbas et al. [1] with a few adoptions. In their work, they used a self-proposed CNN architecture for dermoscopic images, consisting of five convolutional layers with max pooling (window size 2x2 pixels) followed by two fully-connected layers. The same layer composition is used for this thesis, however, the input of the 2D CNN are skin lesion images with the size of 128x128x3 pixels. Furthermore, batch normalization is used in each layer to normalize the inputs. In each convolution filter, a size of 3x3 pixels is used with 32,32,64,64,128 filters per layer. The fully connected part consists of two 128-unit fully connected layers. RMSprop [28] was used as optimizer and tanH (*hyperbolic tangent activation function*) as nonlinear activation function for all layers. The network contained 419,589 trainable parameters. The NLL is used as loss function (see equation 5).

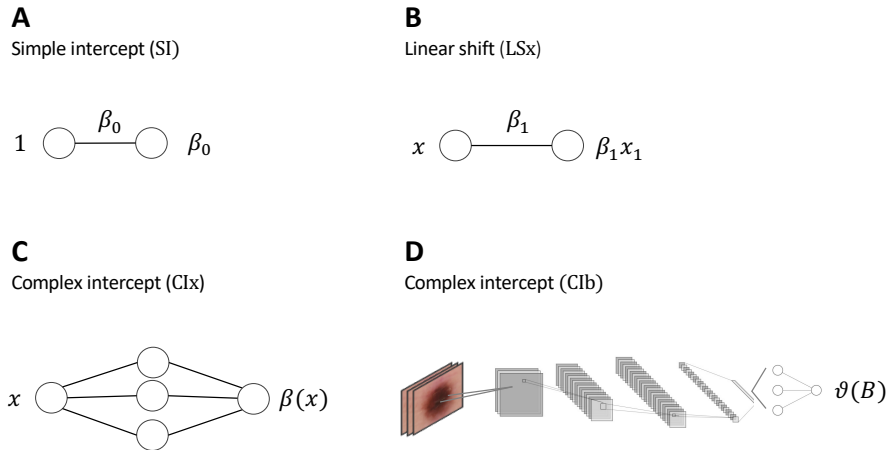


**Figure 3.3:** Schematic visualization of the CNN to model image data with binary outcome. TanH: Hyperbolic tangent activation function.

### 3.3 Combining image and tabular data

How to integrate tabular and unstructured data like images has already been shown for ordinal regression [19]. Ordinal neural network transformation models (ONTRAM) uses jointly trained neural networks, which compose of a transformation function  $h(y|x)$ . This transforms the ordinal outcome to cutpoints of a latent variable  $fz$ , instead of considering the probabilities of the outcomes.

Since the logistic regression is used to model an outcome variable, which has a binary outcome, the method simplifies to estimating a single cutpoint  $h_1$ , which is known from logistic regression (described in section 3.1). Through joint training of image and table data, it is possible to get the interpretation of the regression coefficients as log odds and at the same time have the prediction power of a CNN. Figure 3.4 provides an overview of the individual components for different NN models.



**Figure 3.4:** Architecture of the used model components and NN models. **A:** Simple intercept (SI) for a null model. **B:** Linear shift (LSx) term for tabular data as a dense NN without hidden layer with the possibility of interpretation as log odds ratio. **C:** Complex intercept (CIx) for tabular data as a dense NN with one hidden layer to model non-linear dependencies. **D:** Complex intercept (CIb) for image data by using a CNN.

In the following, the resulting models based on the model components of figure 3.4 are presented, which are necessary for this thesis.

1. Simple intercept + linear shift (**SI LSx**):  $h = \beta_0 + \beta_1 x_1$  is used for tabular data based on patients age as the only predictor variable  $x_1$  (see figure 3.4 components A & B). It is modeled by a single layer NN using a linear activation function. To get the probability for the outcome, a sigmoid function is used as a link function  $p(y = 1|D) = \sigma(h)$ . This corresponds to a logistic regression as a NN, which allows to interpret  $e^{\beta_1}$  from the LSx term as the OR (described in section 3.1.1). So far, there is no advantage to use a NN.
2. Complex intercept (**CIx**):  $h = \beta(x)$  is used to model patient's age in a more complex manner by a flexible dense NN, for example with one hidden layer and non-linear activation function (see figure 3.4 component C). It should be noted that the single output is used with linear activation where  $\beta(x)$  is a log odds ratio function, when the sigmoid function is used as link function  $p(y = 1|D) = \sigma(h)$ . Thus  $\beta(x)$  is different for each age  $x$  in contrast to the LSx term and can model non-linear dependencies.
3. Complex intercept (**CIb**):  $h = \vartheta(B)$  depends on the image data (see figure 3.4 component D) constructed by a 2D CNN (described in section 3.2). Similar to the complex intercept and the linear shift term for tabular data, the output  $\vartheta(B)$  can be interpreted as the log odds ratio.
4. Complex intercept + linear shift (**CIb LSx**):  $h = \vartheta(B) + \beta_1 x_1$  integrate image and tabular data into a single model (see figure 3.4 components D & B). It is necessary to model the output with a linear activation in both model components. Note that there is no bias term in the LSx term, which allows interpret  $e^{\beta_1}$  as OR using sigmoid function as link function  $p(y = 1|D) = \sigma(h)$ . Here, the advantage of modeling the tabular part with a NN becomes apparent with the possibility to combine structured and unstructured data.

### 3.4 Bayesian neural network

Capturing parameter and model uncertainty are often tackled by statistical Bayesian approaches, which can be transferred to neural networks, called *Bayesian neural network* (BNN). Instead of point estimates for the parameters, posterior distributions are learned, which represent the parameter uncertainty [7]. In BNNs, a prior distribution  $p(w)$  is chosen over the weights, which is known as prior belief before any data is observed. Furthermore, the definition of the likelihood distribution  $p(D|w)$  of the data  $D$  given by the parameters is needed. After observing the data  $D$ , the posterior distribution  $p(w|D)$  of the parameters are defined by the Bayesian theorem as follows:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} = \frac{p(D|w)p(w)}{\sum p(D|w)p(w)} \sim p(D|w)p(w) \quad (6)$$

The term  $p(D)$  is a normalization constant, which is usually an intractable problem due to the fact of a high dimensional integral. Hence an approximation is needed, for example Markov-Chain-Monte-Carlo (MCMC) [2] or variational inference (VI) [3]. With MCMC it is possible to approximate the true posterior well. However, since it is a sampling method, it is associated with high computational costs when sampling from a high dimensional



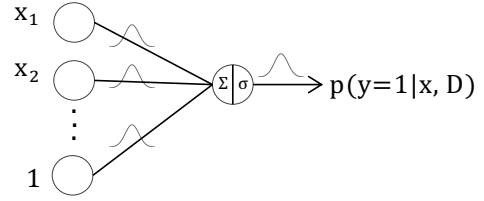
posterior [26]. Therefore VI is more suitable for larger NN. In section 3.4.1 the VI method is described in more detail.

The posterior predictive distribution (PPD) of the model on a test example  $x$  are then given by:

$$p(y|x, D) = \int_w p(y|x, w) \cdot p(w|D)dw \quad (7)$$

where  $p(y|x, w)$  is the outcome predictive distribution given the weights  $w$  and data  $x$ , that are weighted with the posterior probability  $p(w|D)$ . Thus, it is possible to capture the outcome uncertainty by an overall distribution of probabilities. Figure 3.5 gives a visual representation of a Bayesian logistic regression using a BNN. For detailed information on Bayesian deep learning, refer to e.g. [7, 14].

**Figure 3.5:** Logistic regression using a Bayesian neural network (BNN). Weights of the BNN have distributions. Posterior predictive distribution  $p(y|x, D)$  captures the outcome uncertainty.



### 3.4.1 Variational inference

Variational inference (VI) is an optimization problem [3], in which a variational distribution  $q_\lambda(w)$  can be assumed to approximate the posterior distribution. The approximation of the posterior is done by minimizing the Kullback-Leibler (KL) divergence between the VI distribution  $q_\lambda(w)$  and the posterior  $p(w|D)$ :

$$\begin{aligned} \text{KL}(q_\lambda(w) \parallel p(w|D)) &= \int q_\lambda(w) \log \left( \frac{q_\lambda(w)}{p(w|D)} \right) dw \\ &= \log(D) - \underbrace{(\mathbb{E}_{w \sim q_\lambda}(\log(p(D|w))) - \text{KL}(q_\lambda(w) \parallel p(w)))}_{\text{ELBO}(\lambda)} \end{aligned} \quad (8)$$

Since  $\log(D)$  is a constant, the KL is minimized, if the negative evidence lower bound (ELBO) is minimized via gradient descent instead. ELBO consists of two terms. The first term is an expected log-likelihood, which can be approximated by averaging over  $T$  weight samples:

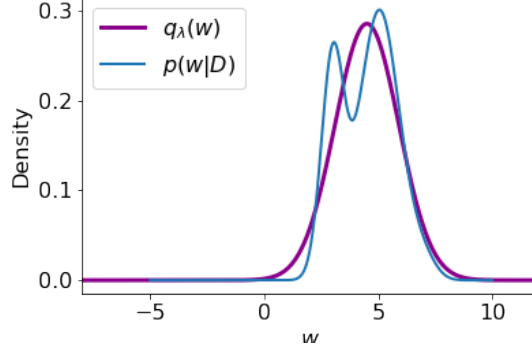
$$\mathbb{E}_{w \sim q_\lambda}(\log(p(D|w))) \approx \frac{1}{T} \sum_{t,i} \log(p(D_i|w_t)) \quad (9)$$

The second term is the negative KL-Divergence between the variational distribution  $q_\lambda(w)$  and the prior  $p(w)$ , which can be approximated by:

$$\text{KL}(q_\lambda(w) \parallel p(w)) \approx \frac{1}{T} \sum_t \log \left( \frac{q_\lambda(w_t)}{p(w_t)} \right) \quad (10)$$

As soon as several parameters have to be approximated, a mean-field VI is often used, which assumes that all parameters are independent. This assumption leads to the fact that posterior approximations are not determined as accurately, especially if the parameters have potential dependencies [3, 30].

Mostly Gaussian is taken as variational distribution  $q_\lambda(w)$ , but there is the disadvantage of limited flexibility in approximating a potentially complex distribution, as can be seen in figure 3.6. In this case the TM-VI method [11] is suitable, which is described in the following section.



**Figure 3.6:** Illustration of variational approximation. Variational Gaussian-VI approximation  $q_\lambda(w)$  compared to the true posterior  $p(w|D)$ .

### 3.4.2 Transformation model-based variational inference

Within the TM-VI method, it is possible to approximate a flexible posterior distribution [11]. This method consists of the concept of transformation model and VI. In figure 3.7 a visual representation of the method is given.

#### Transformation model

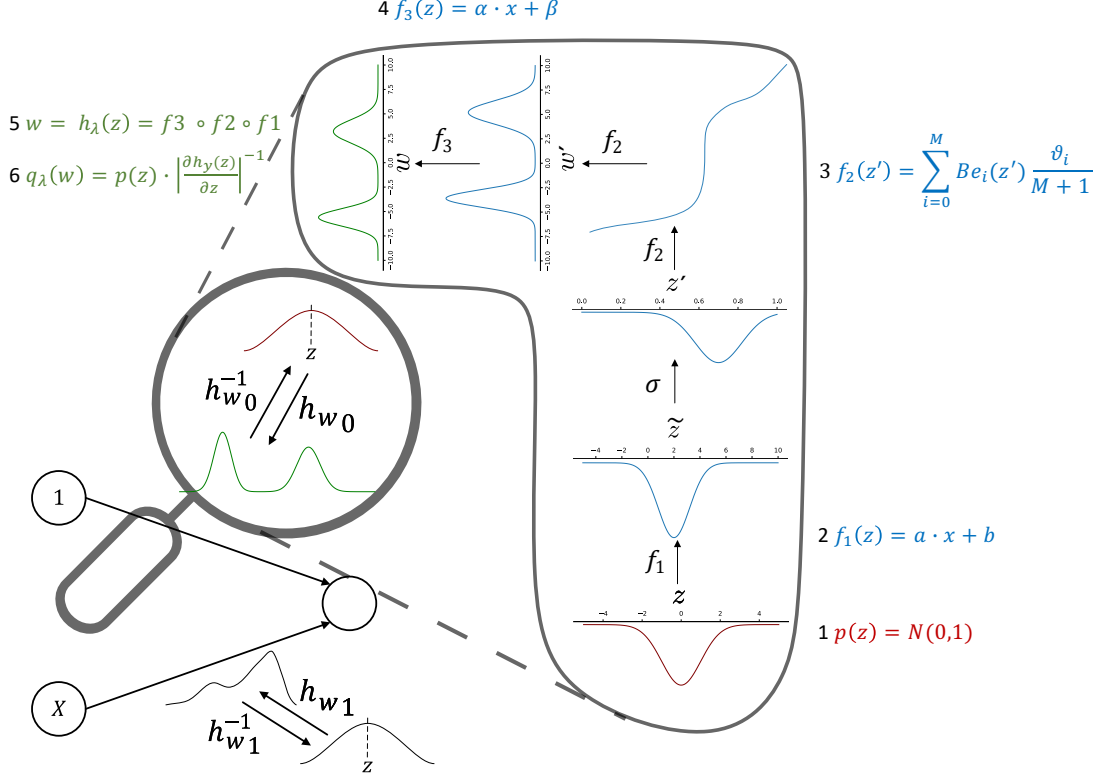
Transformation models (TM) in general allow the transformation of a simple distribution, such as Gaussian, to a potentially complex distribution and is a quite recently developed method within the statistical community [12]. Sick et al. [27] described a method to model complex regression distribution by joining ideas from statistical TMs and Normalizing Flows [18], which are known from the deep learning community. The main idea is to learn a bijective transformation function  $h$  that consists of a chain of transformation  $h(z) = f_3 \circ f_2 \circ f_1$  (demonstrated in the right part of figure 3.7).

The first transformation is a scale and shift transformation  $f_1(z) = a \cdot x + b$  (point 2 in figure 3.7), followed by a sigmoid function, which transforms  $z$ , that comes from a standard normal distribution, into the range of  $z' \in [0, 1]$ . For  $f_2(z')$ , a flexible Bernstein polynomial is used (point 3 in figure 3.7) with the properties of a strict monotonous increase of  $f_2(z')$ , which can be achieved by enforcing the Bernstein coefficients  $(\vartheta_0 \dots \vartheta_M)$  to increase. Furthermore, the Bernstein polynomial can transform any function in the range  $[0, 1]$ .

The Bernstein polynomial, consisting of  $M + 1$  parameters  $\vartheta_0 \dots \vartheta_M$ , is defined as follows:

$$f_2(z') = \sum_{i=0}^M \text{Be}_i(z') \frac{\vartheta_i}{M+1} \quad (11)$$

where the polynomials of order  $M$  are generated by beta-densities  $\text{Be}_i(z')$ .



**Figure 3.7:** Representation of the TM-VI procedure by a single layer BNN with one input  $x$  and bias term. The weights of the NN follow a posterior distribution, which is approximated by a variational distribution trained from a simple distribution  $N(0,1)$  by a bijective transformation function  $h(z)$ . This transformation function consists of a chain of transformation functions  $h(z) = f_3 \circ f_2 \circ f_1$  (point 5), which transforms the simple distribution (red distribution,  $N(0,1)$ ) to a potential complex distribution  $q_\lambda(w)$  (green distribution). After transforming  $N(0,1)$  into the range  $[0,1]$  with  $f_1$  (point 2) followed by a sigmoid function, the flexible Bernstein polynomial  $f_2$  (point 3) is used for transforming into a complex distribution. After the scale and shift function  $f_3$  (point 4), the probability density of  $q_\lambda(w_t)$  can be calculated (point 6). Illustration taken from Hörtling et al. [11].

The third transformation is again a scale and shift transformation  $f_3(w') = \alpha \cdot w' + \beta$  (point 4 in figure 3.7). To ensure that all transformations  $f_i$  increase monotonously to guarantee a bijective transformation, the softplus activation function  $f(x) = \log(1 + e^x)$  is used for the slope parameters  $a$  and  $\alpha$ . Moreover, the Bernstein coefficients are restricted as follows:  $\vartheta_0 = \vartheta'_0, \vartheta_i = \vartheta_i - 1 + \text{softplus}(\vartheta'_i)$  for  $i = 1, \dots, M$ . The transformation function  $h(z)$  has therefore  $5 + M$  parameters  $\lambda = a, b, \vartheta_0, \dots, \vartheta_M, \alpha, \beta$  which are controlled by the NN.

### Training TM-VI

For the TM-VI method, the initialized parameters  $\lambda = a, b, \vartheta_0, \dots, \vartheta_M, \alpha, \beta$  are trained by minimizing the negative ELBO via gradient descent as described in section 3.4.1. However,

for the expected log likelihood (see equation 9),  $T$  samples are drawn at first from the basis distribution  $z$  ( $z_t \sim N(0, 1)$ ). With these samples, the corresponding  $w$ -samples are calculated using the transformation function  $w_t = h(z_t)$ . Then,  $w_t$  is used to determine the KL divergence between variational distribution and the prior (see equation 10). The probability density  $q_\lambda(w_t)$  can be achieved by applying the *change of variable* function (point 6 in figure 3.7) from the samples  $z_t$ :

$$q_\lambda(w_t) = p(z_t) \cdot \left| \frac{\partial h_\lambda(z_t)}{\partial z} \right|^{-1} \quad (12)$$

# Chapter 4

## Experiments

Different models are implemented and evaluated by using the melanoma dataset (see chapter 2), which are described in the following sections.

All models are implemented in Python (version 3.9.7) using Keras based on Tensorflow backend (version 2.4.1) and trained on a GPU (see appendix B for implementation details.) For reproducibility, the full code is available on Github:

[https://github.com/IvonneKo/Master\\_Thesis](https://github.com/IvonneKo/Master_Thesis)

### 4.1 Models

For all models, a conditional outcome distribution  $(y|D) \sim \text{Ber}(p_D)$  is fitted where  $p_D$  is the probability of a melanoma being malignant. The data consists of image data (labeled as  $B$ ) and standardized patient's age (labels as  $x$ ). First, four non-Bayesian models (M1-M4) are fitted by using image data, tabular data, or the combination of both, as described in section 3.3. In the next step, the TM-VI method is used to model Bayesian models. Table 4.1 summarizes all components, which are used in this thesis.

Model	Bayesian variant
<b>M1</b> SI LSx $h = \beta_0 + \beta_1 x_1$	a) $\beta_0$ : TM-VI null model (SI); $\beta_1$ : TM-VI fix intercept (SI LSx) b) $\beta_0 + \beta_1 x_1$ : MF-TM-VI method (SI LSx)
<b>M2</b> CIx $h = \beta(x)$	-
<b>M3</b> CIb $h = \vartheta(B)$	a) Last layer MF-TM-VI b) Last layer MF-Gaussian-VI
<b>M4</b> CIb LSx $h = \vartheta(B) + \beta_1 x_1$	a) $\vartheta(B)$ : CNN ; $\beta_1 x_1$ : TM-VI b) $\vartheta(B)$ : Last layer MF-TM-VI; $\beta_1 x_1$ : MF-TM-VI

**Table 4.1:** Overview of the models implemented and evaluated in this thesis. In the left part of the table the models are listed, which are implemented at first without considering uncertainty (M1-M4). Models: SI LSx: Simple intercept, linear shift for tabular data, CIx: Complex intercept for tabular data, CIb: Complex intercept for image data, and CIb LSx: Complex intercept for image data, linear shift for tabular data. The TM-VI method is only added in the models M1, M3 and M4 (right part of the table), using different model setups. If more than one parameter is approximated, the meanfield TM-VI (MF-TM-VI) method is used (M1b, M3a&b, M4b).

### 4.1.1 Non-Bayesian models

Since no uncertainty is adopted in the first step, deep ensembling approach [21] is used for models with image data. This results in better prediction results by training the mentioned models three times using a different weight initialization. For model evaluation, the result of the predicted outcome is averaged over three runs.

To reduce computational costs for the combined model M4 CIb LSx, the weight of the linear shift model is initialized with the resulting coefficient from logistic regression. For this, additional Gaussian noise with mean 0 and scale 0.1 is added.

Learning rate, number of epochs, and batch size were adjusted for each model. The models are trained with NLL as loss function (defined in equation 5).

### 4.1.2 Bayesian models

Since the goal is an overall model that combines tabular and image data, which allows for both model uncertainty and interpretable regression coefficients, certain steps are taken to compare different models with different model setups.

The flexibility of the TM-VI method is tested on tabular data (M1 SI LSx) with 'one parameter models', where  $\beta_0$  is modeled at first without data as a null model. Furthermore, a fixed intercept term is taken for the parameter  $\beta_1$ , which is estimated from the maximum likelihood point estimation of the logistic regression. Next, the two parameters are modeled combined with the mean-field TM-VI method (MF-TM-VI). All four parameters are compared with the true MCMC posterior distributions modeled by PyStan (see appendix B.2 for implementation details).

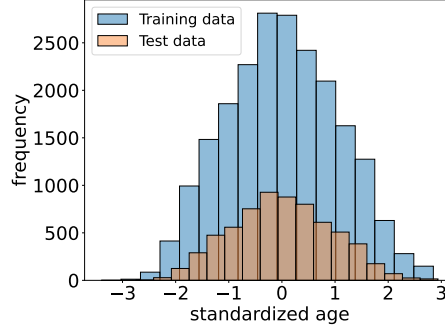
For image data (M3 CIb a)), the MF-TM-VI is only applied in the last layer of the fully connected part of the CNN. The reason is the reduction of the computational cost associated with the approximated inference [4], where reasonable results are obtained even when modeling only a certain area of Bayesian approximation [20]. To have a comparison, a model with MF-Gaussian-VI in the last layer (M3 CIb b)) is added.

For the combined model based on tabular and image data, the CNN is once modeled without (M4 CIb LSx a)), once with TM-VI (M4 CIb LSx b)) in the last layer. In the last case, the mean-field TM-VI is used. For numerical stability, the combined model M4b is initialized with pre-trained weights for the model components  $\vartheta(B)$  and  $\beta_1 x_1$ .

Learning rate, number of epochs, and batch size were adjusted for each model. The models are trained by minimizing the ELBO (defined in equation 8). For simplicity, all models assume a vague Gaussian prior  $N(0, 1)$ , which is often used in Bayesian neural networks [29].

## 4.2 Preparation

For evaluation purposes, the entire dataset was split, with 80% of the data going into the training set and 20% in the test set. 20% of the training set is used as validation data. The image data was resized to 128x128x3 pixels and normalized to have values between 0 and 1. The data of the patient's age are standardized with a mean 0 and variance 1. Figure 4.1 shows the distribution of the standardized age.



**Figure 4.1:** Distribution of the standardized age of training and test data with a mean of 0 and variance 1.

### 4.3 Evaluation

The prediction performance of all models are evaluated with the log-score, as this corresponds to a proper scoring rule [8]. It is thus well suited for the comparison of probabilistic models. The log-score indicates how well a model fits the data and can be used to determine the uncertainty quality. The higher the score the better:

$$\text{log-score} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \log(p(y = y_i | x_i, D)) \quad (13)$$

In addition, the Area under Curve (AUC) is determined, since the usage is very common especially in the medical field [10]. The higher the score, the better. A detailed description can be found in appendix A.

# Chapter 5

## Results and Discussion

This chapter presents the results and discussion divided into non-Bayesian and Bayesian models, using different combinations of model components. All models are evaluated and compared by using the log-score and the AUC value. In addition, we take a look at the interpretable parameter and at the PPD to detect OOD examples.

### 5.1 Non-Bayesian models

As described in section 4.1, four models are used in which uncertainty is not taken into account.

#### 5.1.1 Tabular data: Simple intercept, linear shift (M1 SI LSx)

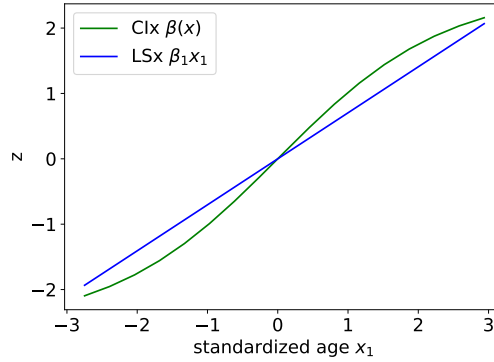
Model  $h = \beta_0 + \beta_1 x_1$  (SI LSx) can be compared to the solution from a logistic regression. Table 5.1 shows the prediction performance from both models and, as expected, the same log-score and AUC values are achieved. In addition, figure 5.2 gives a visual representation of the ROC curves. Furthermore, table 5.1 shows that the estimated odds ratio (OR) of the coefficient  $e^{\beta_{\text{Age}}}$  is the same. Here, the uncertainty of the coefficient estimates isn't addressed, since this is done with the addition of the TM-VI-method (results described in section 5.2). However, the solution of the logistic regression model can specify the 95% confidence interval (CI) of the coefficients, to have later the comparison with the credible intervals (CRI) of the Bayesian models. It should be noted that these intervals are not directly comparable, but give us an idea of the uncertainty distribution.

To interpret the effect of patient's age, the estimated coefficient  $\beta_1$  can be considered as a log odds ratio. It must be done with caution since the data is standardized to the mean 0 and variance 1 (see figure 4.1). Thus, the standardized coefficient is measured in units of standard deviation. Therefore, the interpretation of the standardized coefficient for the patient's age differs slightly as described in section 3.1. It would be interpreted as follows: The odds to have a malignant melanoma, when increasing the standardized predictor age  $x$  by one standard derivation, is  $\text{OR}_{\text{Age}} = e^{\beta_{\text{Age}}} = 2.01$  times higher, holding all other predictors constant. Since the odds are larger than one, the risk of a melanoma to become malignant increases with age. This result was expected since age can be considered as a risk factor of having a malignant melanoma.



### 5.1.2 Tabular data: Complex intercept (M2 CIx)

The assumption that the patient's age, related to the target variable malignant melanoma, has a linear effect, as specified in model M1 SI LSx, is not assured. For example, one possible assumption is that the effect of age becomes linear only after a certain age, while it remains almost the same at younger ages. To check this, the effect of age is modeled by a smooth function visualized by a NN with an additional hidden layer using *tangens hyperbolicus* ( $\tanh$ ) as a nonlinear activation function. This corresponds to the complex intercept model  $h = \beta(x)$  (CIx), in which  $\beta$  differs for each standardized age. In figure 5.1 it is obvious that the coefficient could be modeled linearly and therefore it is valid to use the simpler model SI LSx for further progression. Accordingly, when comparing the performance of CIx model with the linear shift model SI LSx, the same log-score is achieved, as can be seen in table 5.1. However, it must be emphasized that with the addition of the image or other tabular predictors, the relationship between age and the target variable does not have to be linear. This needs to be further investigated and is not covered in this thesis.



**Figure 5.1:** Estimated non-linear effect compared to estimated linear effect of standardized age. CIx: Complex intercept  $\beta(x)$  for tabular data to model standardized age in a complex manner as a log odds ratio function. LSx: Linear Shift of tabular data  $\beta_1 x_1$  to model linear dependencies from standardized age, where  $\beta_1$  is the log odds ratio. The outcome of the two models is  $z$ , which is defined as log odds.

### 5.1.3 Image data: Complex intercept (M3 CIb)

The prediction performance of the complex intercept  $h = \vartheta(B)$  (CIb) is significantly better compared to the model containing only tabular data, as can be seen in table 5.1 and figure 5.2. The outcome as  $\vartheta(B)$  can be interpreted as log-odds-ratio. However, it is not clear which features in the image are responsible for the prediction.

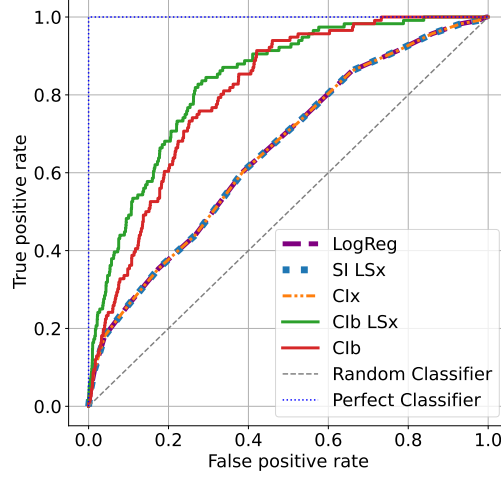
### 5.1.4 Image and tabular data: Complex intercept, linear shift (M4 CIb LSx)

Adding tabular data to the image  $h = \vartheta(B) + \beta_1 x_1$  (CIb LSx) yields a better test performance, compared to the models where both components are modeled alone (see table 5.1). Additionally, the main advantage of this common model is the possibility of interpreting the coefficients, which is especially crucial in the medical field. The estimated

regression coefficient of patient's age  $\beta_1$  shifts from the original value 0.70 with the addition of the image to 0.60, so the OR is with 1.83 lower than before, where the OR was 2.01. This indicates a smaller effect of age after inclusion of the image. The interpretation would be as follows: The odds to have a malignant melanoma, when increasing by one standard derivation  $x$ , is  $\text{OR}_{\text{Age}} = e^{\beta_{\text{Age}}} = 1.83$  higher when holding the image  $\vartheta(B)$  constant.

Model	Log-Score	AUC (95% -CI)	$\text{OR}_{\text{Age}}$
Logistic regression	-0.085	0.66 [0.61 – 0.71]	2.01 [1.82 – 2.25]
<b>M1</b> SI LSx $h = \beta_0 + \beta_1 x_1$	-0.085	0.66 [0.61 – 0.71]	2.01
<b>M2</b> CIx $h = \beta(x)$	-0.085	0.66 [0.61 – 0.70]	-
<b>M3</b> CIb $h = \vartheta(B)$	-0.078	0.81 [0.77 – 0.84]	-
<b>M4</b> CIb + LSx $h = \vartheta(B) + \beta_1 x_1$	-0.075	0.84 [0.80 – 0.87]	1.83

**Table 5.1:** Summary of performance measures log-score, area under the ROC curve (AUC), and estimated odds ratio (OR) of tabular part with age as the only predictor. SI LSx: Simple intercept, linear shift for tabular data, CIx: Complex intercept for tabular data, CIb: Complex intercept for image data, and CIb LSx: Complex intercept for image data, linear shift for tabular data. Higher values for log-score and AUC indicate higher model performance. For AUC, the 95% confidence interval (CI) is calculated with bootstrapping; the 95% CI of the coefficient in the logistic regression part is calculated with the Wald-interval.



**Figure 5.2:** ROC curves of the different models. LogReg: Logistic regression, SI LSx: Simple intercept, linear shift for tabular data, CIx: Complex intercept for tabular data, CIb: Complex intercept for image data, CIb LSx: Complex intercept for image data, linear shift for tabular data. The closer the curve is to the left corner, indicating a perfect classifier, the better. The grey dashed line indicates a random classifier. ROC curves for models LogReg, SI LSx, and CIx based on the tabular data lie on top of each other (purple, blue and orange line), as they achieve the same performance.

## 5.2 Bayesian models

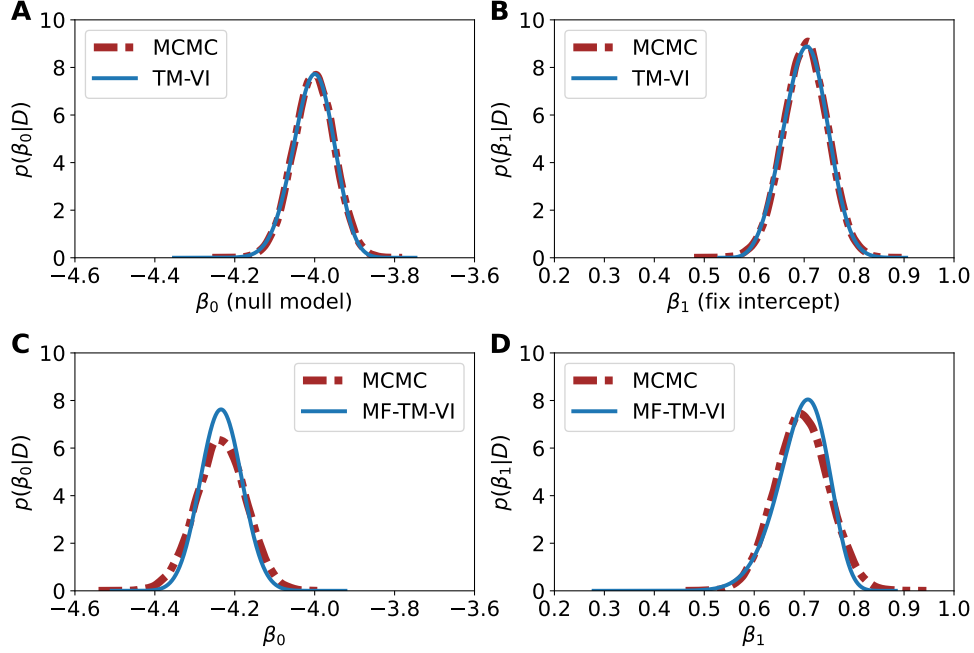
In this section we introduce Bayes to the three models: M1: SI LSx, M3: CIb and M4: CIb LSx by using the TM-VI method, as described in section 4.1.

### 5.2.1 Tabular data: Simple intercept, linear shift (M1 SI LSx)

Before using TM-VI to set up a Bayesian model  $h = \beta_0 + \beta_1 x_1$  (SI LSx), the method is evaluated on one-parameter models to avoid the mean-field assumption (see section 3.4.1). This is done once using a fix intercept and once as a null model (described in section 4.1.2). As can be seen in figure 5.3 panel A and B, an accurate posterior approximation can be achieved using the respective one-parameter models. This is in accordance with the results of Hörting et al. [11].

Using the mean-field TM-VI method to approximate two parameters of the model SI LSx (M1) shows that the variational approximation of  $\beta_0$  and  $\beta_1$  deviates slightly compared to the true posterior (figure 5.3 panel B), which can be caused by the mean-field assumption. Again, since the sigmoid function is used as a link function at the results,  $e^{\beta_1}$  can be interpreted as odds ratio. However, instead of a point estimation of the interpretable parameter, the posterior distribution represents the parameters uncertainty. To get the interval, which includes the most credible values, the 95% high density interval (HDI) is calculated. Furthermore the maximum a posteriori (MAP) is used to get the most likely value for the interpretable parameter. Calculating the MAP and the 95% HDI of  $e^{\beta_1}$ , resulting in a value and credible interval of 2.07 [1.84, 2.20]. Considering the prediction

performance (see table 5.2), the Bayesian modeling does not lead to any difference of the log score and AUC compared to the SI LSx model without uncertainty modeling (see table 5.1).



**Figure 5.3:** Posterior distribution of the parameter  $\beta_1$  and  $\beta_0$  based on tabular data compared to the true posterior resulting from Markov-Chain-Monte-Carlo (MCMC). The first row demonstrates the posterior approximation of a one parameter model modeled by the TM-VI method. **A:** Approximation of intercept parameter  $\beta_0$  as a null model without predictors. **B:** Approximation of slope parameter  $\beta_1$  with a fix intercept term from the maximum likelihood estimation. Second row demonstrates posterior approximation of the two parameters of the model SI LSx (simple intercept, linear shift for tabular data) by using the mean-field TM-VI approach. **C:** Posterior distribution of parameter  $\beta_0$ . **D:** Posterior distribution of slope parameter  $\beta_1$ .

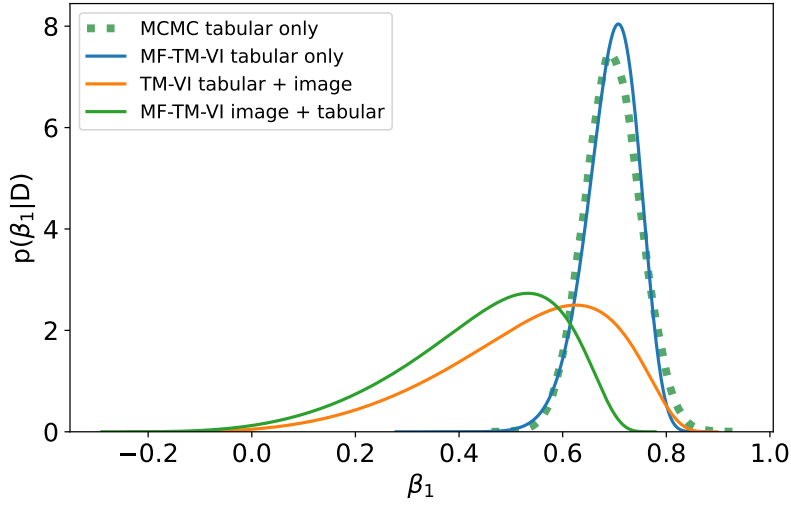
### 5.2.2 Image data: Complex intercept (M3: CIb)

Using the TM-VI method in the last layer of the fully connected part of the CNN  $h = \vartheta(B)$  (CIb) yields the same log-score for Gaussian-VI and TM-VI (see table 5.2). Both models provide a better prediction performance compared to the SI LSx model, which was to be expected due to the results of section 5.1.3, which shows a better prediction performance using image data. The uncertainty of  $\vartheta(B)$  can be evaluated using the posterior predictive distribution. The results are demonstrated in section 5.2.4.

### 5.2.3 Image and tabular data: Complex intercept, linear shift (M4 CIb LSx)

As mentioned in section 4.1.2, the combined model  $h = \vartheta(B) + \beta_1 x_1$  (CIb LSx) is compared with two model variants, where the image part  $\vartheta(B)$  is modeled once without (M4a) and once with Bayesian modeling (M4b). In both cases, the interpretable parameter  $\beta_1$  of the

LSx part is modeled in a Bayesian way. Considering the prediction performance, both models again have a higher log-score and AUC as the tabular and image model alone (see table 5.2). Modeling the mean-field TM-VI method in the tabular and image part (M4b) leads to a further increase of the log-score and AUC value. As shown in figure 5.4, with the addition of the image, the effect of patient's age related to the melanoma type becomes smaller than without the image. By using the MF-TM-VI method in the last layer of the fully connected part of the CNN, the position of the posterior distribution changes slightly (see figure 5.4). However, in both models, with the addition of the image, the odds for malignant melanoma, when age is increased by one standard deviation, is lower than without the lesion image (holding image  $\vartheta(B)$  constant). Calculating the MAP and the 95% HDI of  $\beta_1$  provides a value and credible interval of 0.59 [0.20, 0.79] for M4a and 0.51 [0.14, 0.79] for M4b (see table 5.2).



**Figure 5.4:** Posterior distribution of the slope parameter  $\beta_1$  compared for different models. MCMC: Markov-Chain-Monte-Carlo, MF: Mean-field, TM-VI: Transformation model-based variational inference. The dark green, dotted line is the true posterior of  $\beta_1$  modeled only with tabular data by MCMC. The blue line represents the MF-TM-VI method of  $\beta_1$  modeled only with tabular data. Adding the image part to the model is illustrated by the orange and green lines. Orange line: Tabular data is modeled by the TM-VI method while using a non-bayesian CNN. The green line: Image and tabular data are modeled with MF-TM-VI method.

Model	Log-Score	AUC (95% -CI)	OR <sub>Age</sub> (95%-HDI)
<b>M1</b> SI LSx (MF-TM-VI) $h = \beta_0 + \beta_1 x_1$	-0.085	0.66 [0.61, 0.71]	2.07 [1.84, 2.20]
<b>M3</b> CIb (MF-Gaussian-VI) $h = \vartheta(B)$	-0.076	0.83 [0.80, 0.86]	-
<b>M3</b> CIb (MF-TM-VI) $h = \vartheta(B)$	-0.076	0.83 [0.80, 0.86]	-
<b>M4a</b> CIb LSx (TM-VI) $h = \vartheta(B) + \beta_1 x_1$	-0.075	0.84 [0.80, 0.87]	1.80 [1.21, 2.20]
<b>M4b</b> CIb LSx (MF-TM-VI) $h = \vartheta(B) + \beta_1 x_1$	-0.074	0.85 [0.82, 0.88]	1.67 [1.15, 2.00]

**Table 5.2:** Summary of performance measures log-score, the area under the ROC curve (AUC), and estimated odds ratio (OR) of the tabular part. SI LSx: Simple intercept, linear shift for tabular data, CIb: Complex intercept for image data and CIb LSx: Complex intercept for image data, linear shift for tabular data. Higher values for log-score and AUC indicate higher model performance. For AUC the 95% confidence interval (CI) is calculated with bootstrapping. For models with tabular part, 95% high density interval (HDI) is specified.

#### 5.2.4 Uncertainty evaluation: Complex intercept, linear shift (M4 CIb LSx)

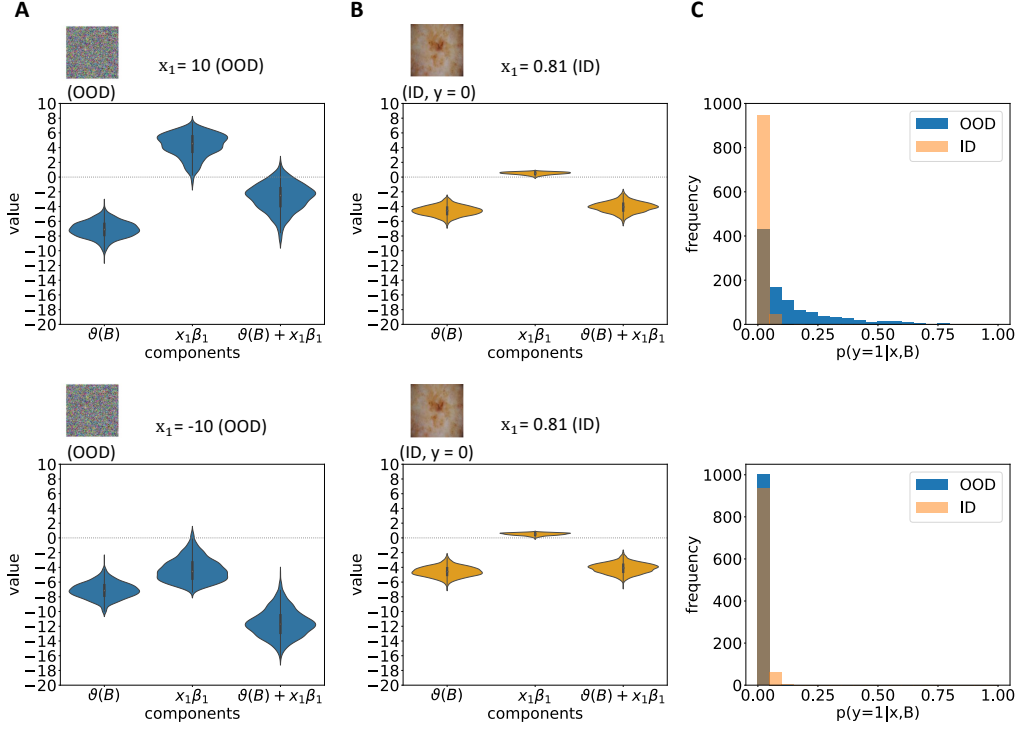
##### Example of out-of-distribution detection

To evaluate the modeled uncertainty, we look at the posterior predictive distribution (PPD) and check if it is possible to detect out-of-distribution (OOD) data. It is important to note that the evaluation of OOD behavior is challenging, since extreme values of the latent variable  $z = h(y|D)$  can be reached, which is not reflected in  $y = \sigma(z)$ . Therefore  $\sigma$  yields outside its working range around 0. Furthermore, the outputs from the image and tabular parts can influence each other through their combination. To better understand this, it is beneficial to demonstrate the uncertainty distribution of the three model components  $\beta_1$ ,  $\vartheta(B)$  as well as the combination of both  $\vartheta(B) + \beta_1 x_1$  by example inputs, before entering the sigmoid function for the PPD  $p(y = 1|x, B)$ .

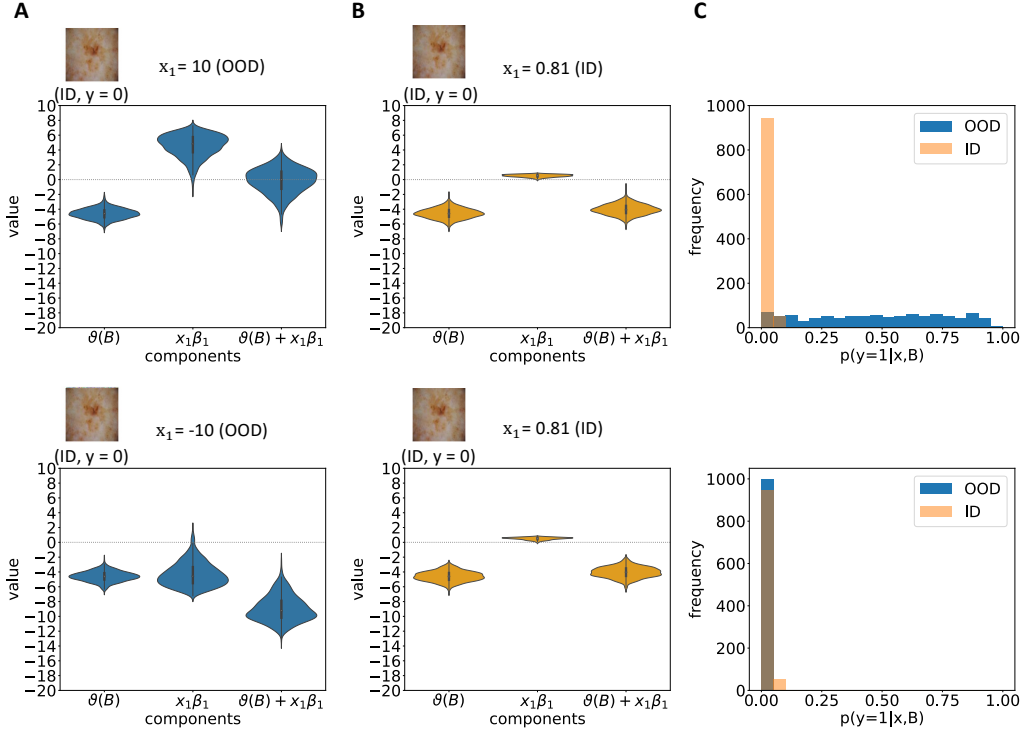
Figure 5.5 illustrates an OOD example once using a standardized age of  $x = 10$ , once a standardized age of  $x = -10$ , while the image is random in both cases<sup>1</sup>. As expected, both OOD cases show larger variability corresponding to high uncertainty (see panel A in figure 5.5). In contrast, the variabilities of in-distribution (ID) image and standardized age ( $x$ ) are smaller, indicating the model is more confident about the outcome (see panel B in figure 5.5). When taking a standardized age value of  $x = -10$ , the range of the combined model  $\vartheta(B) + \beta_1 x_1$  shifts to negative values, which lies outside the working range of the sigmoid function. This results to small variability corresponding to low uncertainty within the resulting PPD (see panel C in figure 5.5). This effect can also be seen in the next two examples illustrated in figures 5.6 and 5.7, where once only an OOD standardized age and once only a random image is taken.

<sup>1</sup>The range of the in-distribution data of standardized age is from  $-3$  to  $3$  (described in section 4.2).

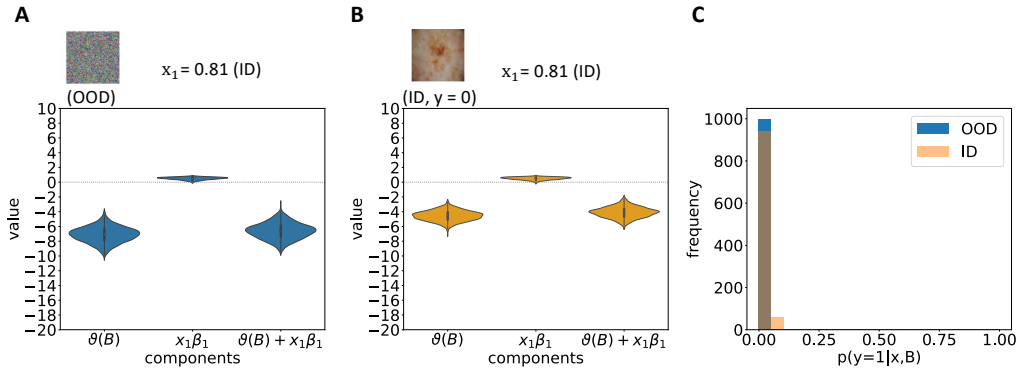
In summary, it can be said, that these examples clearly show the problems that can arise when extreme values before the sigmoid function are used for prediction. For this reason, it is important to consider the uncertainties before the sigmoid function is used for the prediction.



**Figure 5.5:** Impact and interaction of the individual model components  $\vartheta(B)$ ,  $\beta_1 x_1$ ,  $\vartheta(B) + \beta_1 x_1$  based on two example inputs for OOD and ID data and the corresponding posterior predictive distribution (PPD) of the combined model  $p(y = 1|x, B)$ . For the OOD example (column **A**, colored in blue), once using a standardized age of  $x_1 = 10$  (first row) and a standardized age of  $x_1 = -10$  (second row), while the image is random in both cases. For the in-distribution (ID) example (column **B**, colored in orange), a known image and standardized age  $x_1 = 0.81$  is taken in both examples. Column **C** represents the PPD of the combined log-odds model of image and tabular data after entering the sigmoid function  $p(y = 1|x, B) = \sigma(\vartheta(B) + \beta_1 x_1)$ .



**Figure 5.6:** Impact and interaction of the individual model components  $\vartheta(B)$ ,  $\beta_1 x_1$ ,  $\vartheta(B) + \beta_1 x_1$  based on two example inputs for OOD and ID data and the corresponding posterior predictive distribution (PPD) of the combined model  $p(y = 1|x, B)$ . For the OOD example (column **A**, colored in blue), once using a standardized age of  $x_1 = 10$  (first row) and a standardized age of  $x_1 = -10$  (second row), while the image is taken from the test data set. For the in-distribution (ID) example (column **B**, colored in orange) in both examples a known image and standardized age  $x_1 = 0.81$  is taken. Column **C** represent the PPD of the combined log-odds model of image and tabular data after entering the sigmoid function  $p(y = 1|x, B) = \sigma(h)$ .



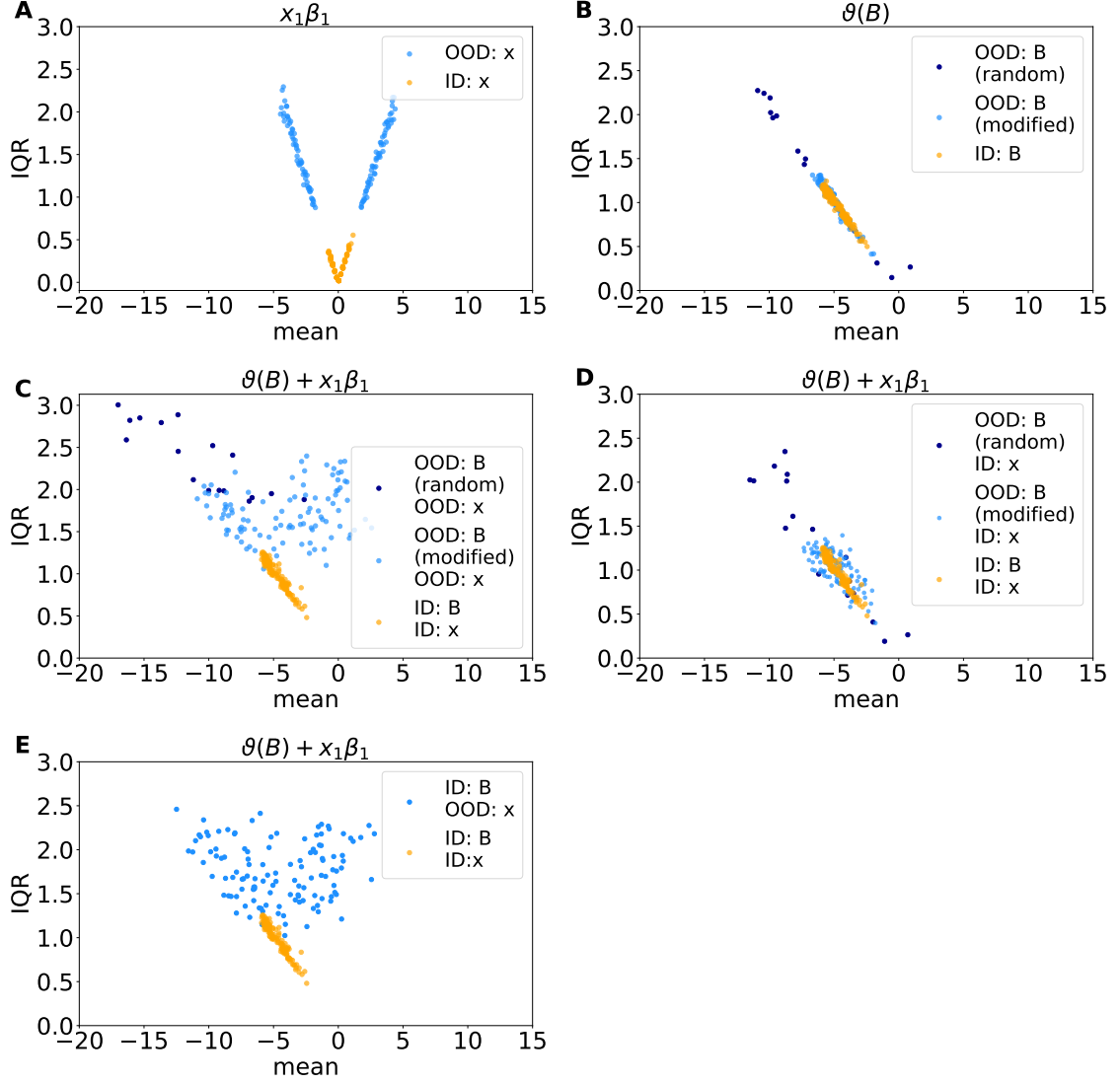
**Figure 5.7:** Impact and interaction of the individual model components  $\vartheta(B)$ ,  $\beta_1 x_1$ ,  $\vartheta(B) + \beta_1 x_1$  based on two example inputs for OOD and ID data and the corresponding posterior predictive distribution (PPD) of the combined model  $p(y = 1|x, B)$ . For the OOD example (panel **A**), a random image is used, while for the standardized age a value from the test data set  $x_1 = 0.81$  is taken. For the in-distribution (ID) example (panel **B**) in both examples a known image and standardized age  $x_1 = 0.81$  is taken. Panel **C** represent the PPD of the combined log-odds model of image and tabular data after entering the sigmoid function  $p(y = 1|x, B) = \sigma(h)$ .



### Uncertainty evaluation on multiple data

To enable a more systematic evaluation, multiple test data points, labeled as in-distribution data (ID), as well as OOD data, are considered with different input possibilities for image and tabular data, as shown in figure 5.8. For the OOD images, 17 random images and 103 slightly modified images of the test dataset are taken. Example OOD images with the used augmentation methods can be seen in appendix C. For OOD examples of standardized age, 120 data points from a range of  $[-10,4]$  and  $[4,10]$  are used.

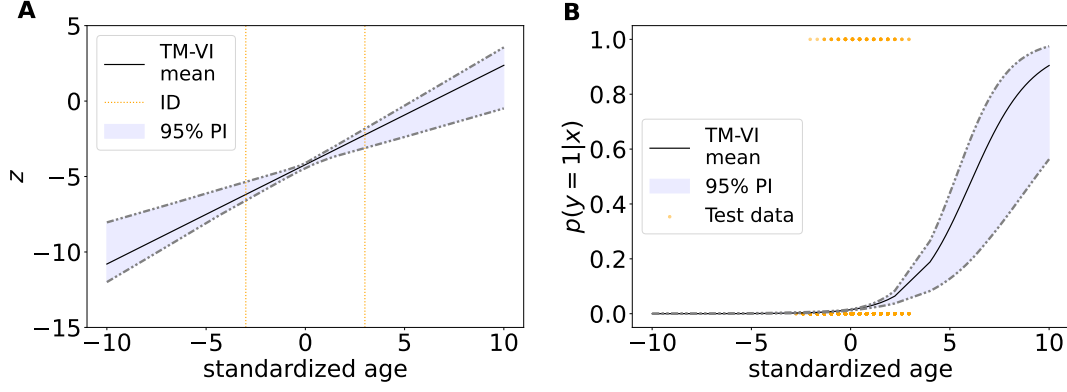
The following model components are considered:  $\beta_1 x_1$  (panel A in figure 5.8),  $\vartheta(B)$  (panel B in figure 5.8) as well as the combination of both  $\vartheta(B) + \beta_1 x_1$  (panels C-E in figure 5.8). For the combined model, the uncertainty is evaluated by three different model setups: First with OOD examples for image and tabular data (panel C in figure 5.8), second with OOD examples only for image data (panel D in figure 5.8), and last with OOD examples only for tabular data (panel E in figure 5.8). The uncertainty is quantified with the interquartile range (IQR) to summarize the dispersion in a single number. The higher the IQR is, the more uncertain the model. We would expect that in the cases with OOD examples the range is higher than for the ID data. As can be seen in figure 5.8, for most cases this assumption applies to be correct. However, for modified and random images the distinction between in-distribution and OOD data is not as clear as for the tabular data, so that there are some OOD examples where the model underestimates the uncertainties. This is shown by a similar IQR value of ID and OOD images. For some random images, the IQR is even very small, which indicates the model is certain for these images (panels B and D in figure 5.8). However, if the standardized age is taken from an OOD for the combined model, a good separation of ID and OOD is achieved in terms of IQR value (panels C and E in figure 5.8).



**Figure 5.8:** Uncertainty evaluation of multiple data with different input setups. IQR: Interquartile range of the model components indicated on the title of each plot (**A-E**), OOD: Out-of-distribution, ID: In-distribution, x: Tabular data, B: Image data,  $\beta_1 x_1$ : Linear shift term,  $\vartheta(B)$ : Complex intercept model. Light blue dots represent the OOD data of modified images and standardized age, orange dots are the ID test data and dark blue are random OOD images. **A:** Outcome of the linear shift term with standardized ID and OOD examples. **B:** Outcome of the complex intercept with ID, random and slightly modified images. **C:** Outcome of the combined model with ID and OOD random, slightly modified images and standardized age. **D:** Outcome of the combined model with OOD images only. **E:** Outcome of the combined model with only OOD standardized age.

### 5.2.5 Uncertainty evaluation: Simple intercept, linear shift (M1 SI LSx)

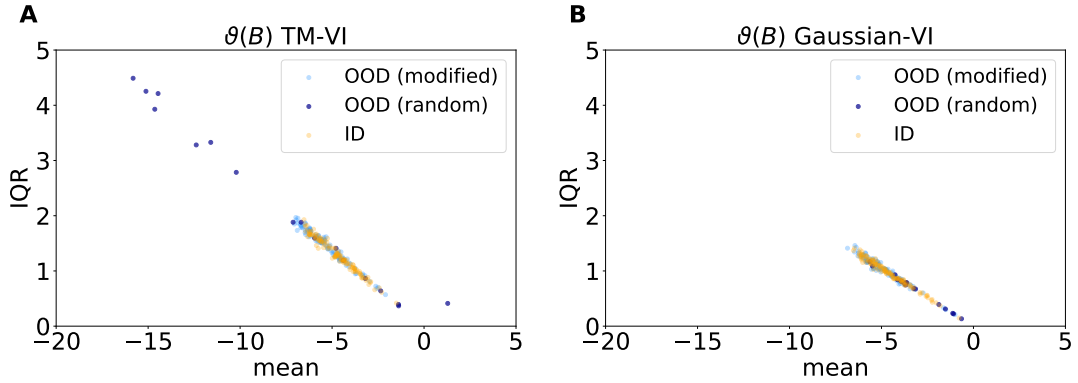
As it can be seen in figure 5.9 (panel A), the uncertainties on the logit scale  $h = \beta_0 + \beta_1 x_1$  for OOD examples become wider as soon as the known distribution is left. However, the resulting PPD  $\sigma(h)$  only has wide uncertainties for a positive extreme value of the standardized age (see figure 5.9 panel B). For negative extreme values for standardized age, the uncertainties are not visible caused by the extreme negative value for  $h$  before entering the sigmoid function. Again, for model uncertainty, it has to be captured before.



**Figure 5.9:** Uncertainty evaluation of tabular based model SI LSx. TM-VI: Transformation model-based variational inference, ID: In-distribution, PI: Prediction interval. **A:** Outcome of the log odds model  $z = h = \beta_0 + \beta_1$  with mean, 2.5 and 97.5 quantiles. Orange dotted line represent the range of the ID, which is taken from the test data set. **B:** Posterior predictive distribution as a CPD  $p(y = 1|x)$  with mean, 2.5 and 97.5 quantiles. Orange dots represent the test data.

### 5.2.6 Uncertainty evaluation: Complex intercept (M3 Clb)

For the image-based model, Gaussian-VI and TM-VI are compared. Figure 5.10 shows the range of uncertainties of the ID test data sets (120 data points) as well as 120 OOD images consisting of random and slightly modified images (example images can be found in appendix C). If the outcome of  $\vartheta(B)$  is taken into account, it is noticeable that TM-VI is more uncertainty aware in the case of OOD examples than Gaussian-VI.



**Figure 5.10:** Uncertainty evaluation of image-based model Clb: TM-VI (**A**) and Gaussian-VI (**B**). TM-VI: Transformation model-based variational inference, Gaussian-VI: Gaussian variational inference, IQR: Interquartile range, OOD: Out-of-distribution, ID: In-distribution. Light blue dots represent the OOD data of modified images, orange dots are the ID test data and dark blue are random OOD images. For the evaluation, the value of an IQR is taken, which indicates the higher the value the more uncertain the model.

## Chapter 6

# Summary and Outlook

In this thesis, the implementation of semi-structured Bayesian models is demonstrated, comprise of skin lesion images and the corresponding patient's age, to predict the melanoma types. In the analysis, we have seen that the additional consideration of tabular data in the classification task achieves interpretability of the effect of patient's age. Furthermore, using the TM-VI method leads to a model that gives the possibility to quantify the uncertainty. In order to achieve the goal and the resulting objectives, various models with different complexity are fitted. Therefore, the models are evaluated once without and once with the Bayesian variant. The evaluated models are SI LSx (simple intercept, linear shift for tabular data), CIx (complex intercept for tabular data) CIb (complex shift for image data) and CIb LSx (complex shift for image data, linear shift for tabular data). In the following, the results are described for the non-Bayesian and Bayesian models.

### 6.1 Results of non-Bayesian models

As baseline models, a 2D CNN based on patient's lesion images (CIb), corresponding to a binary DL model, and a single layer NN based on patient's age (SI LSx), corresponding to a logistic regression, were developed. The improved prediction power in terms of log-score and AUC of the CNN became clear in contrast to the simple NN model with age as the only predictor. These results were to be expected, as the images usually contain more information than just the age of the patient. The combination of both led, once again, to an improvement of the prediction power. In addition, there is the advantage of interpreting the patient's age to break down the characteristic of the 'black-box' DL model. The estimated odds ratio (OR) for the predictor age shows a positive association between patient's age and the outcome of having a malignant melanoma. This was to be expected, since age can be considered a risk factor. It could be observed that with the addition of the image, the effect of age becomes smaller. The assumption that the effect of age might also be nonlinear could not be confirmed by the addition of complexity in the NN. For this reason, it was valid to perform the further steps involving uncertainty models, based only on the three models: M1 SI LSx, M3 CIb, M4 Ib LSx.

## 6.2 Results of Bayesian models

We have confirmed that the TM-VI method approximates the posterior distribution accurately for one-parameter models. For the semi-structured model, which consists of multiple parameters, the TM-VI was used, except for one model, within a mean-field (MF) fashion. Modeling TM-VI only on the tabular linear shift part (LSx) within the combined model yields an approximation of the posterior distribution without the mean-field assumption.

In the case of tabular data (M1 SI LSx), modeling MF-TM-VI leads to the same predictive performance in terms of log score and AUC as the non-Bayesian variant of the model. Instead of a point estimation, a whole posterior distribution is determined for the interpretable coefficient. This distribution and the specification of credible intervals are especially important in the medical field to avoid misdiagnosis by having an overconfident model.

For the CNN-based models (CIb and CIb LSx), the MF-TM-VI method is only used in the last layer of the fully connected part. The reason is that it proved difficult to apply the method to multiple layers, recognizable by numerical instability. For this reason, it was also necessary to train the joint model (CIb LSx) consisting of tabular and CNN with initialized weights of both components. All models, except SI-LSx, achieve increasing log-score and AUC values in their Bayesian variant. Also in the Bayesian variant, by adding the image to the tabular part, the effect of age becomes smaller.

To evaluate the uncertainty, especially in the CNN-based models, we took a look at the predictive posterior distribution (PPD). The focus relied on the detection of out-of-distribution (OOD) examples, assuming large uncertainty ranges. This was demonstrated in more detail by using one example input for image and tabular part within the combined model CIb LSx. Five different combinations were used for tabular and image inputs, where the MF-TM-VI method was used in both parts, LSx and CIb. On the one hand, the results of the different experiments show how the model components influence each other. On the other hand, it becomes clear that the uncertainties must be caught before the actual PPD. The reason is that extreme values cause to leave the working range of the sigmoid function, which is used for the prediction of a binary outcome. Thus, extreme values (0 or 1) are reached, which leads to the fact that no uncertainties are recognized within the PPD. To quantify the uncertainty by using multiple data, the interquartile range (IQR) of the  $h$  components on the logit-scale was taken. It was observed that the OOD detection in the image part is not always reliable. This effect is also shown by comparing MF-Gaussian-VI and MF-TM-VI within the image-based models. However, TM-VI seems to be more aware in detecting OOD examples compared to MF-Gaussian-VI. Here, the expectation would be more that both methods would act similarly.

## 6.3 Outlook

While the primary interest of this thesis was to combine image and tabular data with the ability of interpretation and uncertainty modeling, the CNN-based models achieve a performance, which can be further improved by using different techniques, transfer learning, or fine tuning the network. For example, the implementation of preprocessing techniques, such as data augmentation, makes it possible to increase the diversity of the data. As a result, it can increase the performance by using the right combination of augmentation methods.

In the case of facing novel images, the implemented models underestimate the uncertainty for some types of images. It would be interesting to see how the results change when a Bayes variant is used for the entire CNN. To do that, a further adaptation of the TM-VI method is needed to provide results throughout a CNN. Furthermore, using the TM-VI method on the CIX model to model the predictor variable age in a complex manner was not covered in this thesis. What impact this might have on the performance and the uncertainty detection is something that can still be evaluated. In future work, other predictors such as gender or the location of the lesion could be included. For this, it would be advantageous to drop the mean-field assumption so that the dependencies of the different predictors are not ignored. Considering the interpretation task, the addition of several predictors can for example cause the linear effect of age to become nonlinear. In addition, confounding bias should be considered, which highlights the importance of adjusting the confounders in future analyses.

In conclusion, this work served as a first step to combine interpretable semi-structured models with the VI-procedure, which can be applied to other medical classification tasks.

# Bibliography

- [1] Q. Abbas, F. Ramzan, and M. U. Ghani. Acral melanoma detection using dermoscopic images and convolutional neural networks. *Visual Computing for Industry, Biomedicine, and Art* 2021 4:1, 4:1–12. <https://doi.org/10.1186/S42492-021-00091-Z>, 2021.
- [2] M. Betancourt. A conceptual introduction to hamiltonian monte carlo. arXiv:1701.02434, 2018.
- [3] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- [4] N. Brosse, C. Riquelme, A. Martin, S. Gelly, and Éric Moulines. On last-layer algorithms for classification: Decoupling representation from uncertainty estimation. arXiv:2001.08049, 2020.
- [5] S. Burgess. Estimating and contextualizing the attenuation of odds ratios due to non collapsibility. *Communications in Statistics - Theory and Methods*, 46:786–804. <https://doi.org/10.1080/03610926.2015.1006778>, 2017.
- [6] S. Friedrich and K. Kraywinkel. Faktenblatt: Epidemiologie des malignen melanoms in deutschland. *Der Onkologe*, pages 447–452. <https://doi.org/10.1007/s00761-018-0384-1>, 2018.
- [7] Y. Gal. Uncertainty in deep learning. *University of Cambridge*. PhD thesis. <https://mlg.eng.cam.ac.uk/yarin/thesis/thesis.pdf>, 2016.
- [8] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *American Statistical Association*. <https://doi.org/10.1198/016214506000001437>, 2007.
- [9] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>, 2016.
- [10] K. Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, 4:627, 2013.
- [11] S. Hörting, D. Dold, O. Dürr, and B. Sick. Transformation models for flexible posteriors in variational bayes. arXiv:2106.00528, 2021.
- [12] T. Hothorn, L. Möst, and P. Bühlmann. Most likely transformations. *Scandinavian Journal of Statistics*, 45(1):110–134. <https://doi.org/10.1111/sjos.12291>, 2017.
- [13] International Skin Imaging Collaboration. SIIM-ISIC 2020 Challenge Dataset. <https://doi.org/10.34970/2020-ds01>, 2020.

- [14] L. V. Jospin, W. Buntine, F. Boussaid, H. Laga, and M. Bennamoun. Hands-on bayesian neural networks – a tutorial for deep learning users. *arXiv:2007.06823*, 2020.
- [15] M. A. Kassem, K. M. Hosny, R. Damaševičius, and M. M. Eltoukhy. Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. *Diagnostics*, 11. ISSN 2075-4418. <https://doi.org/10.3390/diagnostics11081390>, 2021.
- [16] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv:1703.04977*, 2017.
- [17] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder. Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3:159–165. [https://doi.org/10.1016/S1470-2045\(02\)00679-4](https://doi.org/10.1016/S1470-2045(02)00679-4), 2002.
- [18] I. Kobyzev, S. J. Prince, and M. A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979. <https://doi.org/10.1109/TPAMI.2020.2992934>, 2021.
- [19] L. Kook, L. Herzog, T. Hothorn, O. Dürr, and B. Sick. Deep and interpretable regression models for ordinal outcomes. *arXiv:2010.08376*, 2021.
- [20] A. Kristiadi, M. Hein, and P. Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. *arXiv:2002.10118*, 2020.
- [21] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv:1612.01474*, 2016.
- [22] Leitlinienprogramm Onkologie (Deutsche Krebsgesellschaft, Deutsche Krebshilfe, AWMF). S3-Leitlinie Prävention von Hautkrebs, Langversion 1.0, AWMF Registernummer: 032/052OL, . [Online] <https://www.leitlinienprogramm-onkologie.de/leitlinien/hautkrebs-praevention/> (visited on 10/11/2021), 2014.
- [23] Leitlinienprogramm Onkologie (Deutsche Krebsgesellschaft, Deutsche Krebshilfe, AWMF). Diagnostik, Therapie und Nachsorge des Melanoms, Langversion 3.3, AWMF Registernummer: 032/024OL, . [Online] <https://www.leitlinienprogramm-onkologie.de/leitlinien/melanom/> (visited on 10/11/2021), 2020.
- [24] R. Marks. Epidemiology of melanoma. *Clinical and Experimental Dermatology*, 25: 459–463. <https://doi.org/10.1046/J.1365-2230.2000.00693.X>, 2000.
- [25] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvey, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, and H. P. Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8. <https://doi.org/10.1038/s41597-021-00815-z>, 2021.
- [26] T. Salimans, D. P. Kingma, and M. Welling. Markov chain monte carlo and variational inference: Bridging the gap. *arXiv:1410.6460*, 2015.
- [27] B. Sick, T. Hothorn, and O. Dürr. Deep transformation models: Tackling complex regression problems with neural network based transformation models. *arXiv:2004.00464*, 2020.



- [28] T. Tieleman, G. Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.
- [29] A. G. Wilson. The case for bayesian deep learning. arXiv:2001.10995, 2020.
- [30] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2008–2026, 2017.

# Appendix A

## Evaluation Metrics

### A.1 Log score

For the conditional outcome distribution  $p(y_i|x_i, D) \sim \text{Bernoulli}(y_i \in \{0, 1\})$ , the log score is used to compare models and is defined as:

$$\begin{aligned}\text{log-score} &= \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \log(p(y = y_i|x_i, D)) \\ &= \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i \log(p_1(x_i)) + (1 - y_i) \log(1 - p_1(x_i)))\end{aligned}$$

In Bayesian models the posterior predictive distribution is considered. For this, the integral is replaced with the mean samples  $S$  for one datapoint  $x_i$  from the posterior:

$$\begin{aligned}p(y_i|x_i, D) &= \int_w p(y_i|x_i, w) \cdot p(w|D)dw \\ p(y_i|x_i, D) &= \frac{1}{S} \sum_{i=1}^S p(y_i|x_i, w)\end{aligned}$$

### A.2 Area under the ROC Curve

A ROC curve (receiver operating characteristic curve) plots True Positive Rate (Sensitivity) against True Negative Rate (1-Specificity) at different classification thresholds:

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

To summarize the ROC curve to a single value, the area under the ROC curve is considered (AUC). The output of AUC is between 0 and 1. The closer the value to 1, the better.

# Appendix B

## Implementation details

GPU: NVIDIA GeForce RTX 2070 SUPER

Python: 3.9.7

### B.1 Important packages

#### Models

Tensorflow: 2.4.1, scikit-learn 1.0 (logistic regression)

#### Image Preprocessing

Keras-Preprocessing: 1.1.2

#### Analysis and Visualization

matplotlib 3.4.3, seaborn 0.11.2, arviz 0.11.2

#### MCMC simulation

PyStan: 2.19.1.1

### B.2 PyStan code for logistic regression

```
1 melanom_data = {'N': x_trainLR.shape[0], 'M': x_trainLR.shape[1],
2                  'X': x_trainLR, 'y': y_train}
3
4 lr_code = """
5
6     data {
7         int N;
8         int M;
9         real X[N, M];
10        int<lower=0, upper=1> y[N];
11    }
12
13    parameters {
14        real beta[M];
15        real beta0;
16    }
17
18
19    model {
20        for (i in 1:N)
21            y[i] ~ bernoulli(inv_logit(beta0 + dot_product(X[i], beta)));
22        beta[M] ~ normal(0, 1);
23        beta0 ~ normal(0, 1);
24    }
25    """
26
27 stm = pystan.StanModel(model_code=lr_code)
```

## Appendix C

# Out-of-distribution examples

**First row:** Slightly modified from original images: Random rotation, random translation, random flip, adjust brightness.

**Second row:** Random images.

