Hochschule Konstanz
Fakultät Informatik

INSTITUTE FOR OPTICAL SYSTEMS

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

Master Thesis MSI

Autor: Ivonne Kovylov
Supervisors: Prof. Dr. Oliver Dürr, Prof. Dr. Beate Sick

Motivation

- Malignant melanomas: Aggressive skin tumors
- Over 90% of all skin tumor deaths
- Risks:
    - UV-exposed leisure, skin color, age, …
- Use of dermoscopy for melanoma diagnosis
- Early detection is critical
    - Convolutional Neural Network (CNN) in dermoscopy images shows good performance

Cp. [15,17,22,23,24]

Motivation

- Neural Networks (NNs) are often ‘black box’
    - Additional interpretation is relevant for medical applications
    - ‘What effect does age have on the type of melanoma?’

    → Statistical regression models: Obtain interpretable parameters of structured data

    - Combination of structured (tabular) and unstructured (image) data

    → Semi-structured model: Combine benefits of statistical and deep learning community

Motivation

- Predictions of NNs are often overconfident
    - Uncertainty modeling is crucial in medical application
    - Often Bayesian Neural Networks
    - Posterior distribution hard to determine exactly
    - Approximation complex posterior distribution: Transformation model-based variational inference (TM-VI)

Cp. [7,9,11,16]

Agenda

1. Aim and objectives
2. Dataset and methods
    - ISIC melanoma dataset
    - Logistic regression
    - Convolutional Neural Network
    - Combining image and tabular data
    - Bayesian Neural Network
3. Results
    - Non-Bayesian models
    - Bayesian models
4. Conclusion and outlook

Aim and objectives

Aim: Predict the melanoma types by using semi-structured Bayesian models, while interpretable components and model uncertainty is quantified.

Objectives:

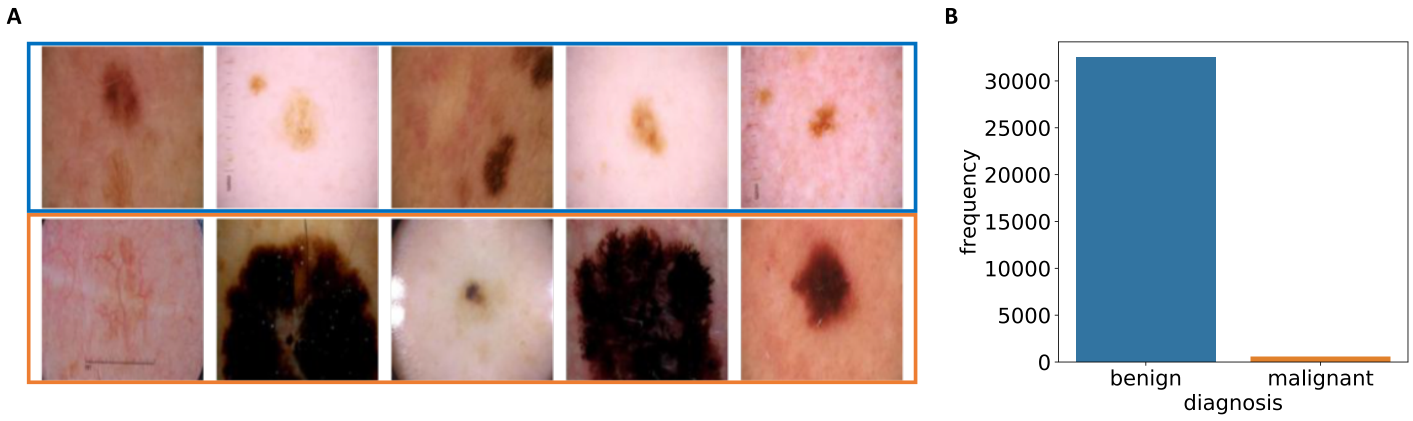
1. Develop models by using the ISIC dataset based on: lesion images, patient's age, and a combination of both
2. Interpret the effect of patient’s age with and without the impact of the image
3. Apply TM-VI to semi-structured Bayesian Models
4. Quantify uncertainty of the model parameter estimations
5. Evaluate and compare the prediction performance of all models
6. Evaluate the prediction uncertainties

2. Dataset and methods

- **ISIC melanoma dataset**

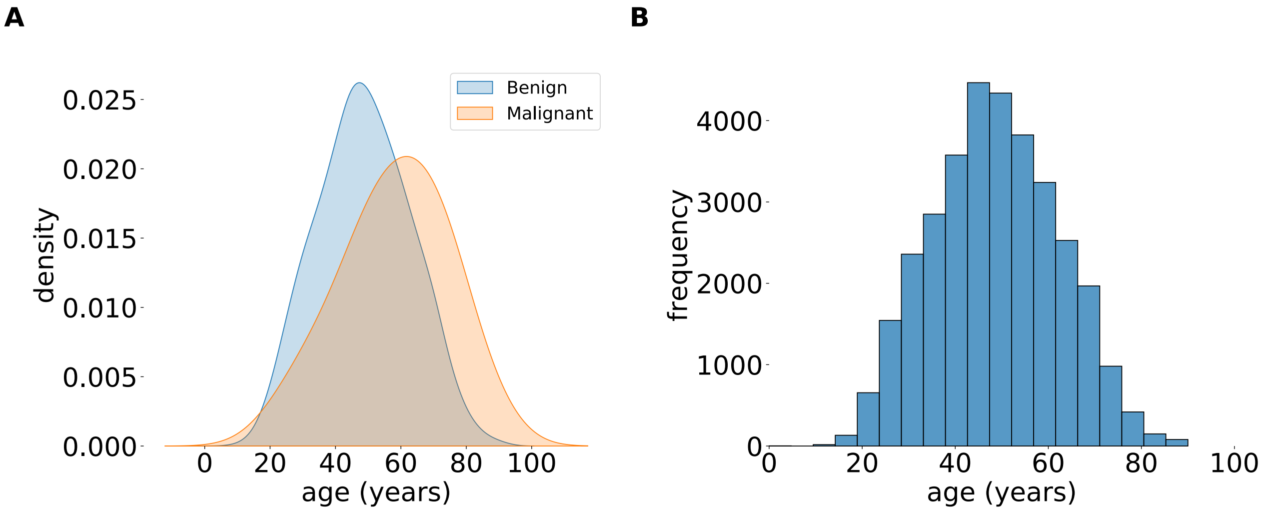
ISIC melanoma dataset

- International Skin Imaging Collaboration (ISIC): Skin lesion analysis towards melanoma detection 2020 https://challenge2020.isic-archive.com/
- Binary classification: Benign / malignant
- 32542 benign / 584 malignant
- Metadata available in addition to the images: Patient ID, gender, age, location site, diagnosis



Cp. [13,25]

Patient's age



2. Dataset and methods

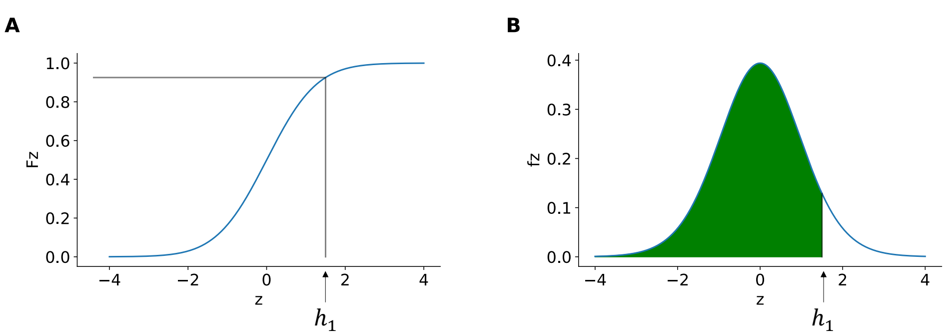
- **Logistic regression: Tabular data as input**

Logistic regression

- Binary outcome and tabular data as input for an interpretable model
- Conditional probability: $p_D = p(y = 1|D)$
- Conditional probability distribution (CPD): $(y|D) \sim \mathrm{Bern}(p_D)$
- Logit model:
  - $\mathrm{Odds}(\mathrm{y} = 1) = \frac{p_D}{1-p_D}$
  - $\mathrm{Log}(\mathrm{Odds}) = \mathrm{logit}(p_D) = z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

- Probability estimated by logistic regression:
  - $p(y = 1|D) = p_D = \frac{1}{1+e^{-z}}$

$p_{xi} = F_z(h(x_i))$

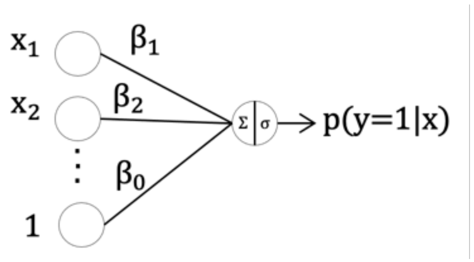


$h(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

Logistic regression as Neural Network

- Parameters are estimated by the maximum likelihood approach
- NN is trained by minimizing the negative log likelihood (NLL):

$$\mathrm{NLL} = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log\left(p_1(x_i)\right) + (1-y_i)\log\left(1-p_1(x_i)\right)\right)$$



Logistic regression: Interpretation

- Odds Ratio:
  - $\mathrm{OR}_{x\to x+1} = \frac{\mathrm{Odds}\,(x+1)}{\mathrm{Odds}(x)} = \frac{e^{\beta_0+\beta_k x_k+1}}{e^{\beta_0+\beta_k x_k}} = e^{\beta_k}$
  - How does the odds change with change in x by one unit, holding all other predictors constant
  - $\mathrm{OR}_{x\to x+1} > 1$: positive association, $\mathrm{OR}_{x\to x+1} < 1$: negative association, $\mathrm{OR}_{x\to x+1} = 1$: no association

- Interpretation difficult:
  - Consideration of confounders
  - Non-collapsibility

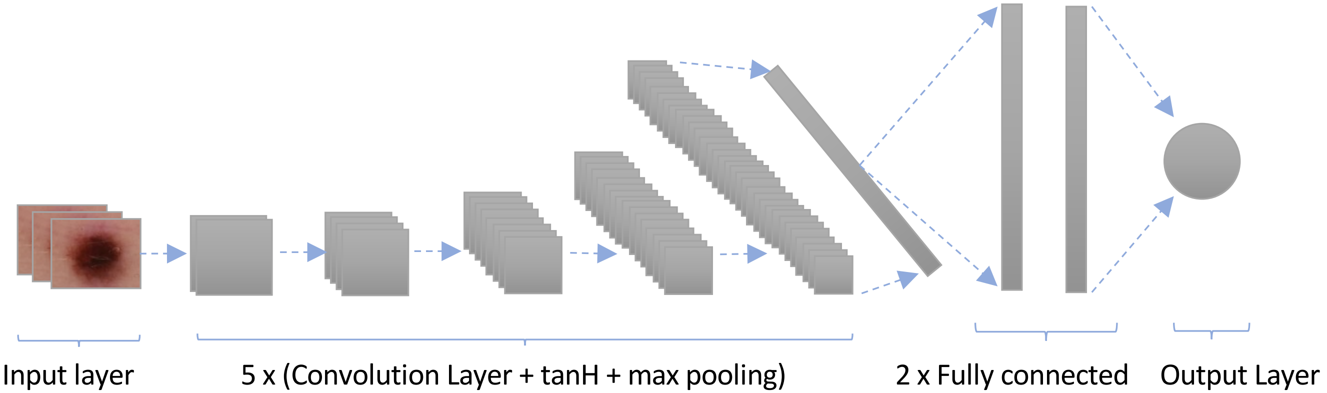
Cp. [5]

2. Dataset and methods

- **Convolutional Neural Network: Image data as input**

Convolutional Neural Network (CNN)

- Input: Skin lesions with size of 128x128x3 pixels
- 5 Convolutional Layer with max pooling (2x2 pixels)
- Batch Normalization
- Hyperbolic tangent activation function (tanH) as non-linear activation function
- 419,589 trainable parameters
- NLL as loss function



Cp. [1, 9, 28]

2. Dataset and methods

- **Combining image and tabular data**

Additive components

- Already been shown for ordinal regression → Ordinal Neural Network Transformation Models (ONTRAM)
- Jointly trained NNs: Additive components based on tabular and image data
- Transformation function $h(y|x)$: Transform outcome to cut-points of latent variable fz
- Logistic regression: 1 cut-point

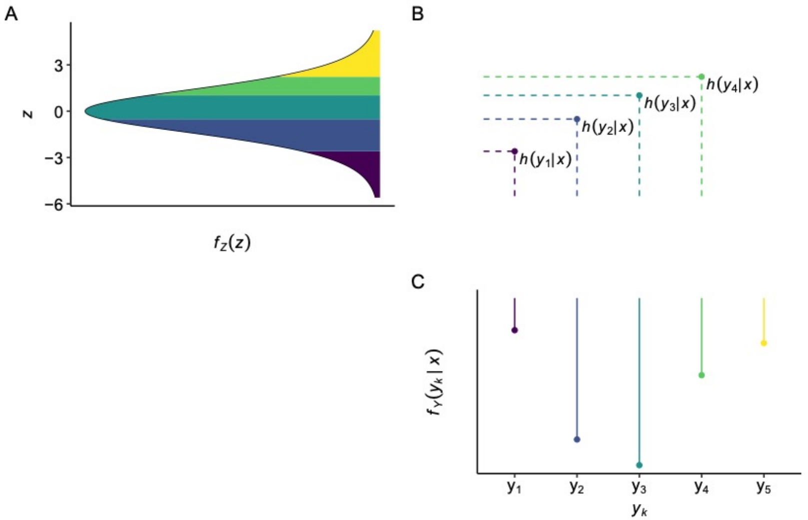
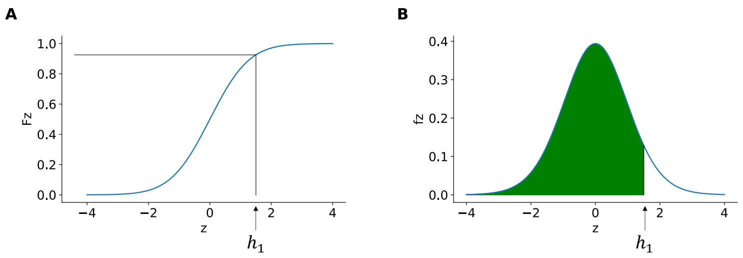
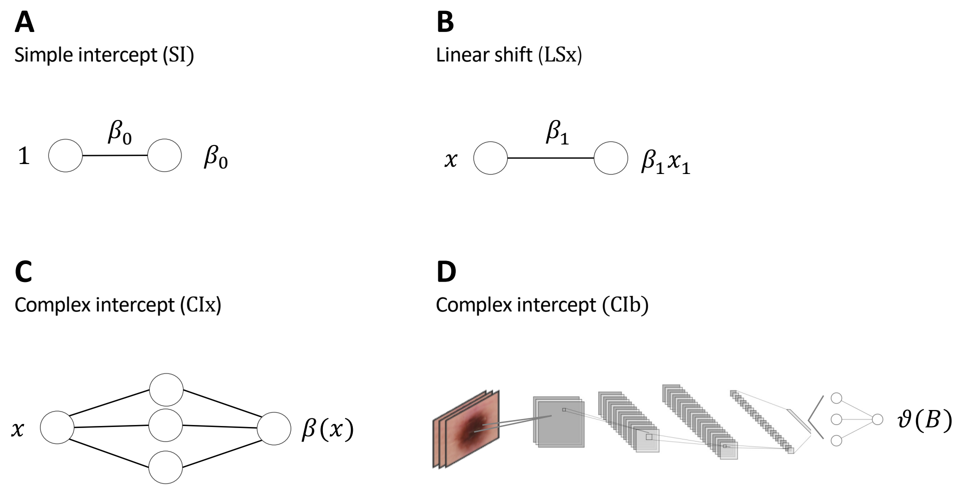


Illustration taken from Kook et al. [19]



Cp. [19]

Additive components



Model	$h(y\|x, B)$
M1: SI + LSx	$\beta_0 + \beta_1 x_1$
M2: CIx	$\beta(x)$
M3: CIb	$\vartheta(B)$
M4: CIb + LSx	$\vartheta(B) + \beta_1 x_1$

cut-point: $p = \sigma(h(y|x, B))$

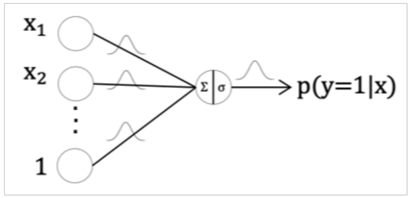
Own representation based on Kook et al. [19]

2. Dataset and methods

- **Bayesian neural network**

Bayesian Neural Network

- Capturing parameter and model uncertainty
- Bayesian approach:



likelihood

prior

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

posterior

normalization constant

- Posterior predictive distribution (PPD):
  - $$p(y|x,D) = \int_w p(y|x,w) \cdot p(w|D)dw$$
- Intractable problem: Need of approximation, e.g.: Markov-Chain-Monte-Carlo (MCMC), variational inference (VI), transformation model-based variational inference (TM-VI)

Cp. [2,3,7,14,26]

Variational inference as optimization problem

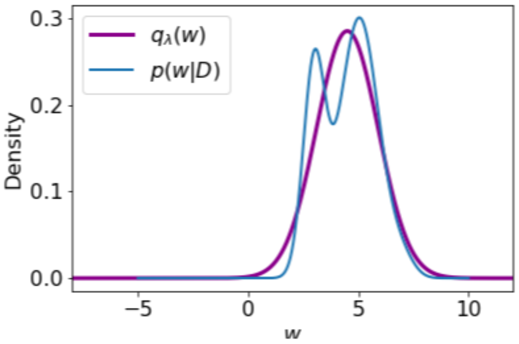
- Posterior distribution $p(w|D)$ can be approximated by a variational distribution (often Gaussians) $q_\lambda(w)$
- Minimizing Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(q_\lambda(w)||p(w|D)) = \int q_\lambda(w) \log\left(\frac{q_\lambda(w)}{p(w|D)}\right) dw$$

$$= \log(p(D)) - \underbrace{(\mathbb{E}_{w\sim q_\lambda}\big(\log\big(p(D|w)\big)\big) - \mathrm{KL}(q_\lambda(w)||p(w)))}_{\text{ELBO}(\lambda)}$$

$$\mathbb{E}_{w\sim q_\lambda}\big(\log\big(p(D|w)\big)\big) \approx \frac{1}{T}\sum_{t,i} \log(p(D_i|w_t)) \qquad \mathrm{KL}(q_\lambda(w)||p(w)) \approx \frac{1}{T}\sum_{t} \log\left(\frac{q_\lambda(w_t)}{p(w_t)}\right)$$

- Multiple parameter: often mean-field VI is used
- Disadvantage of Gaussian: Limited flexibility



Cp. [3,30]

Transformation models (TM)

- TMs allow the transformation of a simple distribution p(z) = N(0,1) to a potentially complex distribution $q_\lambda(w)$
- Main idea: Learn a bijective transformation function
    - $h(z) = f3 \circ f2 \circ f1$
- Core: Flexible Bernstein polynomial
    - Strict monotonous increase of parameters $\vartheta_1 \leq \vartheta_1 \leq \cdots \leq \vartheta_M$
    - Transform any function in the range [0,1]
    - Flexible by controlling the order of degree (M)
- $q_\lambda(w_t)$: Probability density by change of variable function

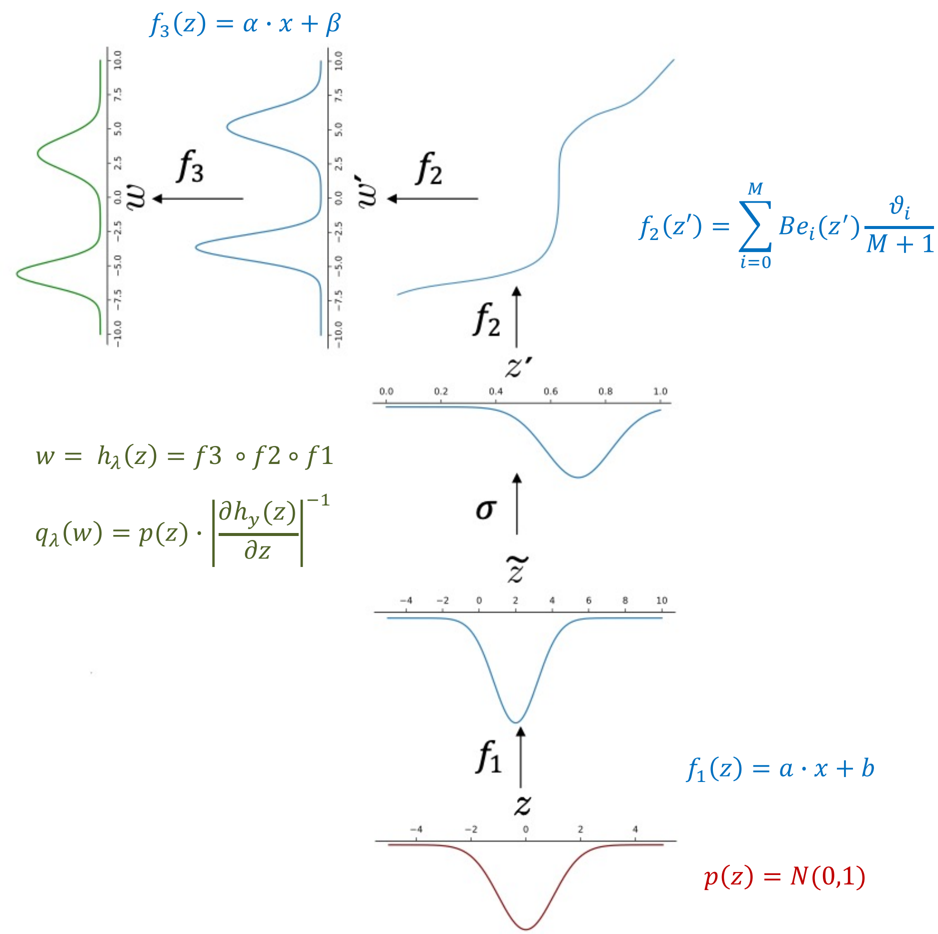


Illustration taken from Hörtling et al. [11]

Cp. [11,12,18,27]

Transformation model-based variational inference (TM-VI)

- Training:
  - Parameters $\lambda = a, b, \vartheta_0, \ldots \vartheta_M, \alpha, \beta$ are trained by minimizing the negative ELBO via SGD
  - Calculate expected log likelihood: $z_t \sim N(0,1), w_t = h(z_t)$
  - w_t: Calculate KL-Divergence between variational distribution and the prior

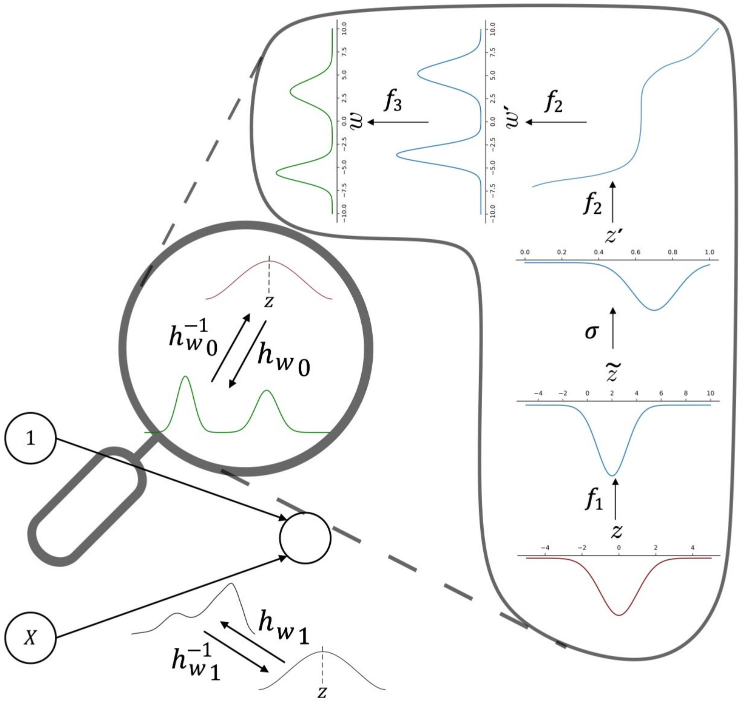
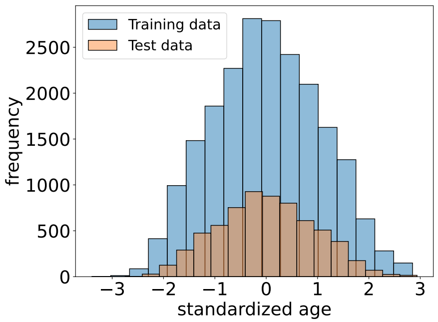


Illustration taken from Hörtling et al. [11]

Cp. [11]

Experiments Details

- Divided in: Non Bayesian and Bayesian models
- Patient's age standardized to mean 0 and variance 1
- Evaluation metrics:
  - Log-score:
    $$\frac{1}{N_{\text{test}}}\sum_{i=1}^{N_{\text{test}}}\log(p(y=y_i|x_i,D))$$
  - Area under the ROC curve (AUC)



Cp. [8,10]

3. Results

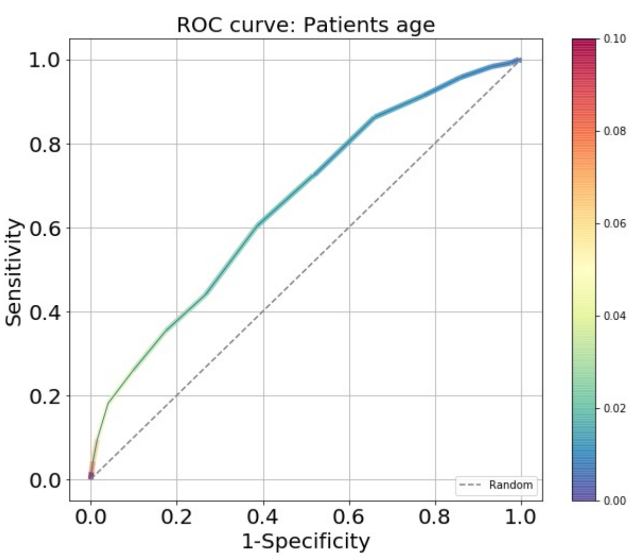
- **Non-Bayesian models**

Tabular data: SI LSx

- Modeling patient's age with simple Intercept (SI) and linear shift (LSx):

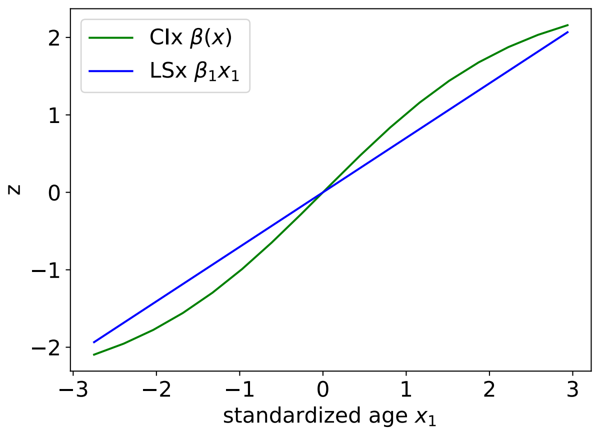
$$h = \beta_0 + \beta_1 x_1$$

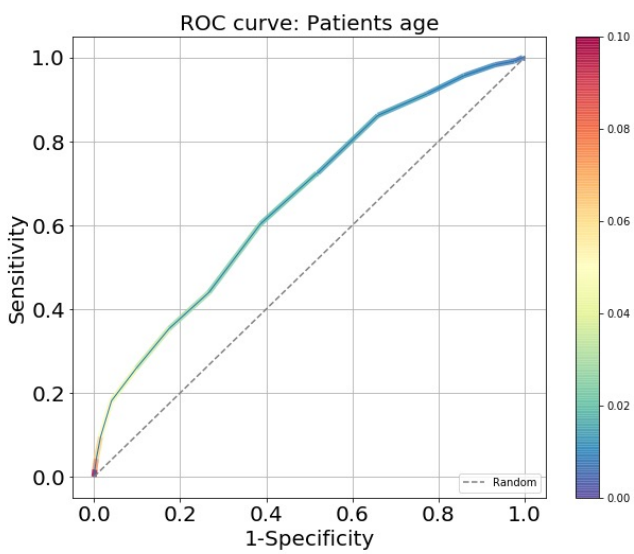
- Same as logistic regression
- Standardized coefficient: $OR_{Age} = e^{\beta_{Age}} = 2.01$



Model	Log-score	AUC (95% CI)	$\mathbf{OR_{Age}}$
M1: SI LSx $h = \beta_0 + \beta_1 x_1$	-0.085	0.66 [0.61-0.71]	2.01
Logistic regression	-0.085	0.66 [0.61-0.71]	2.01 [1.82-2.25]

Tabular data: CIx





- Modeling patient's age as complex intercept to make a nonlinear relationship:

$$h = \beta(x)$$

- Additional hidden layer with nonlinear activation function (tanH)
- Log-odds ratio function
- Comparison SI LSx $h = \beta_0 + \beta_1 x_1$ with CIx $h = \beta(x)$

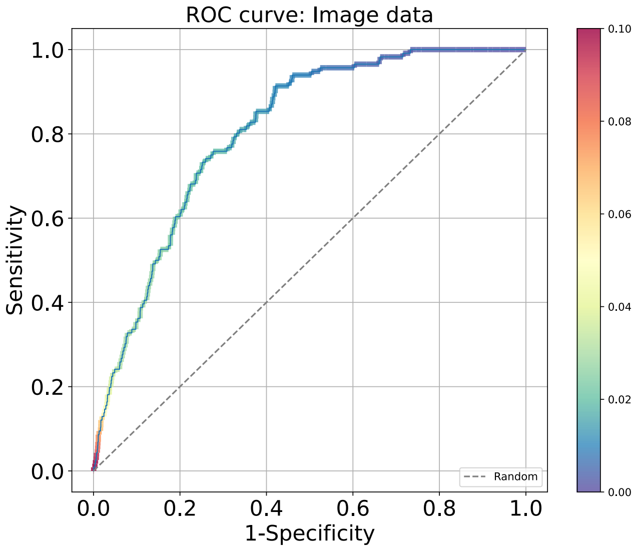
Model	Log-score	AUC (95% CI)	$\mathbf{OR_{Age}}$
M1: CIx $h = \beta(x)$	-0.085	0.66 [0.61-0.70]	-

Image data: CIb

- Modeling image data as complex intercept:

$$h = \vartheta(\mathrm{B})$$

- CNN with one output layer
- Outcome log-odds ratio function
- Interpretation not clear



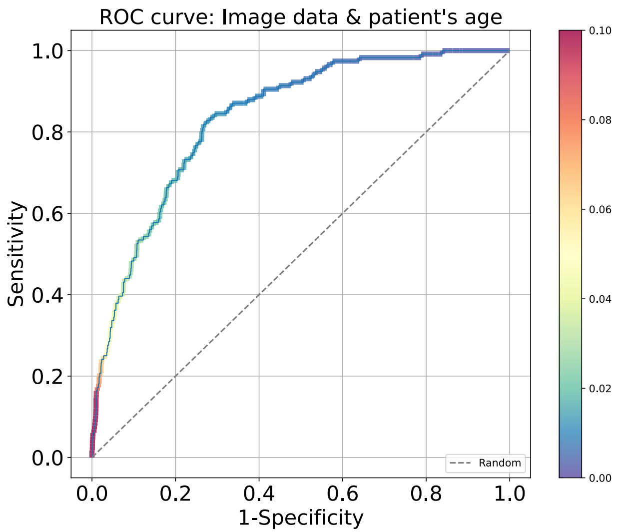
Model	Log-score	AUC (95% CI)	$\mathbf{OR_{Age}}$
M1: CIb $h = \vartheta(\mathrm{B})$	-0.078	0.81 [0.77-0.84]	-

Image and tabular data: CIb LSx

- Modeling image data as complex intercept and tabular data as linear shift term:

$$h = \vartheta(\mathrm{B}) + \beta_1 x_1$$

- Standardized coefficient: $OR_{Age} = e^{\beta_{Age}} = 1.83$ (holding image constant)
- Smaller effect of age after including the image (without $OR_{Age} = 2.01$)



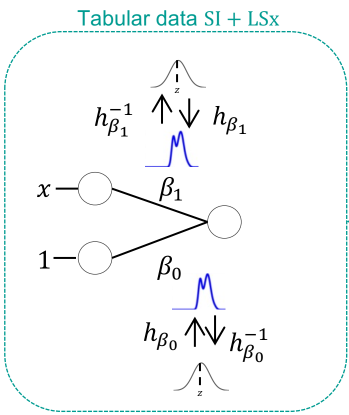
Model	Log-score	AUC (95% CI)	$\mathbf{OR_{Age}}$
M1: CIb + LSx $h = \vartheta(\mathrm{B}) + \beta_1 x_1$	-0.075	0.84 [0.80-0.87]	1.83

3. Results

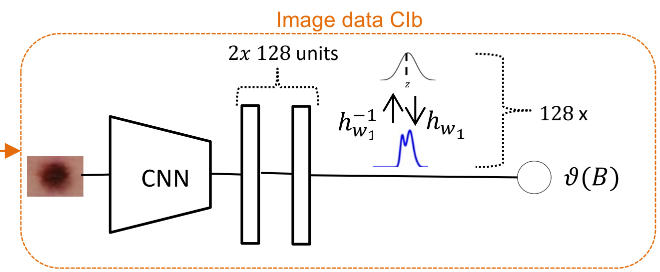
- **Bayesian models**

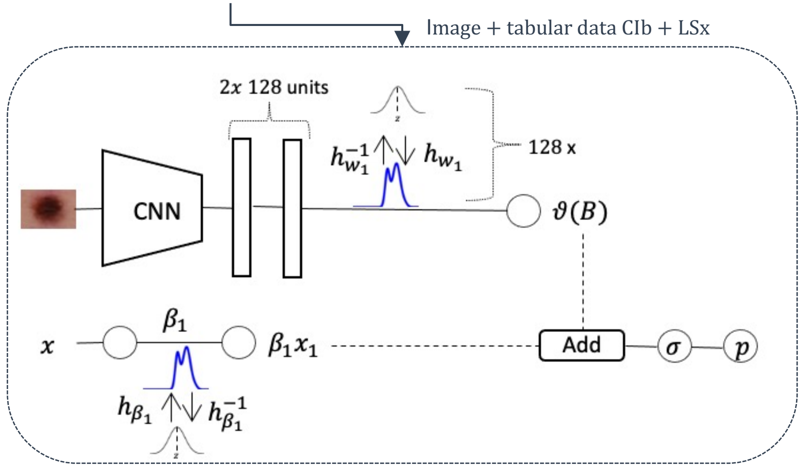
Model components

- Image $\vartheta(B)$: Determine posterior distributions only in the last layer
- Patient's age $\beta_1 x_1$: Determine slope posterior distribution



Model	$h(y\|x, B)$
M1: SI + LSx	$\beta_0 + \beta_1 x_1$
M3: CIb	$\vartheta(B)$
M4: CIb + LSx	$\vartheta(B) + \beta_1 x_1$

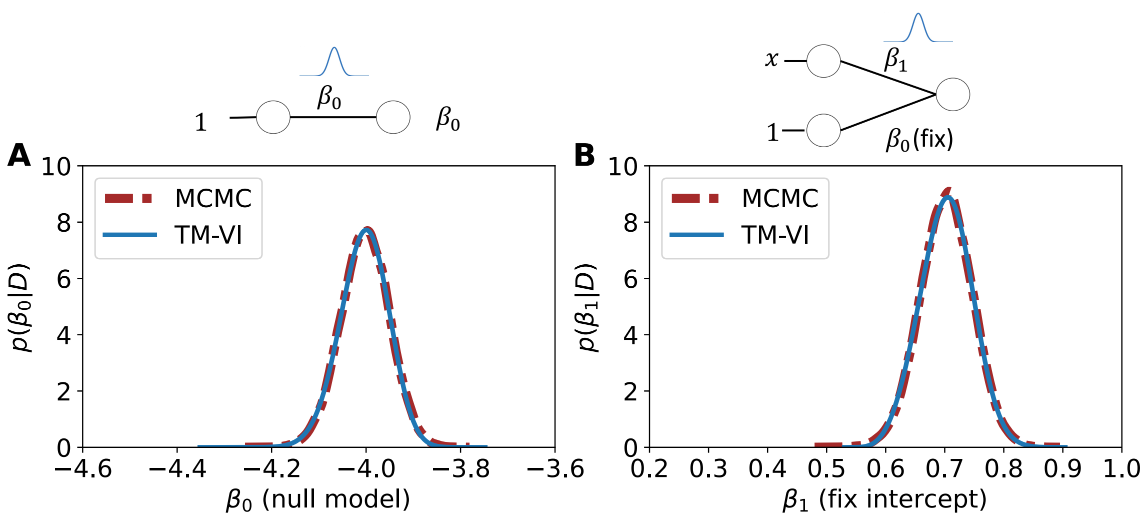




Cp. [4,20]

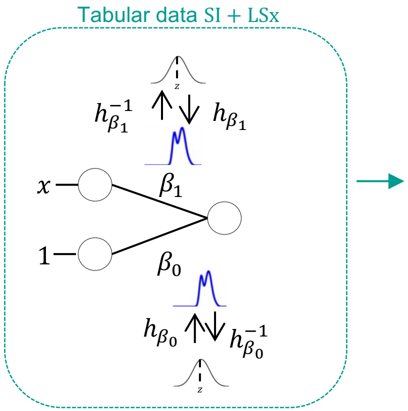
One-parameter models

- In single parameter model TM-VI yields accurate posterior approximations
  - Intercept parameter β_0 (null model)
  - Slope parameter β_1 (fix intercept)
- Compare with true posterior (MCMC)

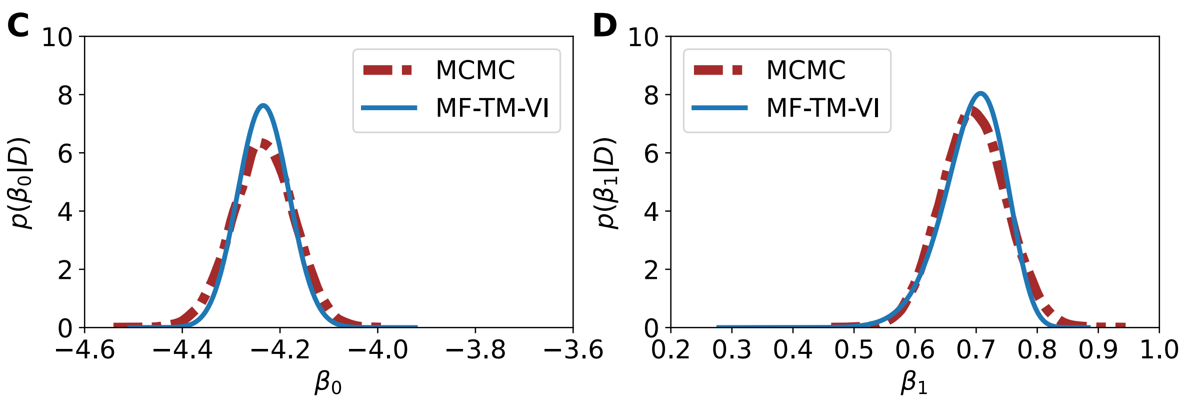


M1: Tabular data mean-field TM-VI

- Patient's age with parameter β_0 and β_1 compared to true posterior (MCMC)
- Performance:
  - Log-score: -0.085
  - AUC: 0.66 [0.61-0.71] 95% CI
  - OR_{Age} = e^{β_1}= 2.07 [1.84-2.20] 95% HDI



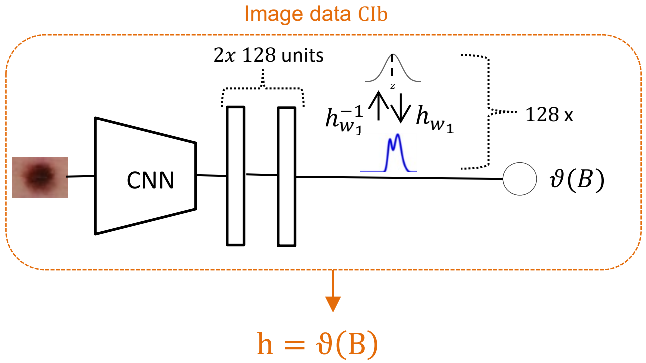
$h = \beta_0 + \beta_1 x_1$



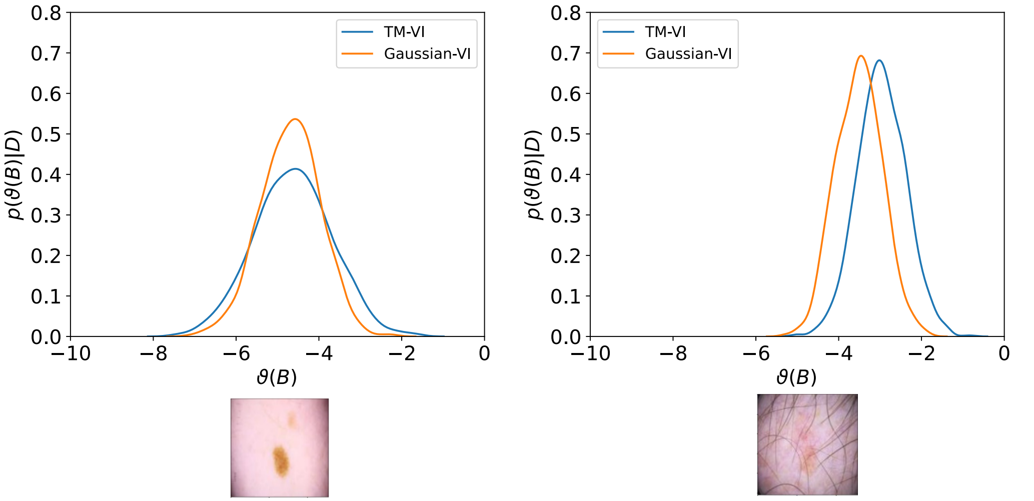
M3: Image data mean-field TM-VI

- MF-TM-VI in last layer of fully-connected part of CNN
- Compared to MF-Gaussian-VI
- Performance MF-TM-VI:
  - Log-score: −0.076
  - AUC: 0.83 [0.80−0.86]
- Performance MF-Gaussian-VI:
  - Log-score: -0.076
  - AUC: 0.83 [0.80-0.86]

→ Evaluation on posterior predictive distribution will be shown later



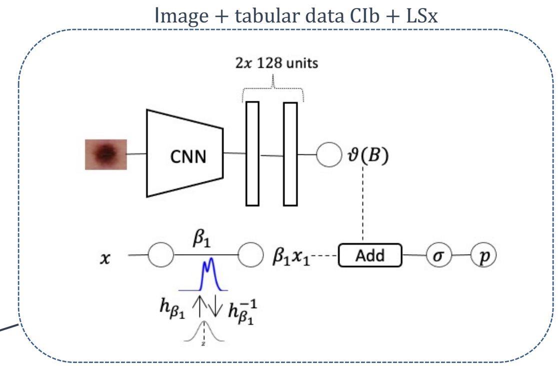
Posterior distribution $\vartheta(B)$ depends on image:



M4: Image and tabular data

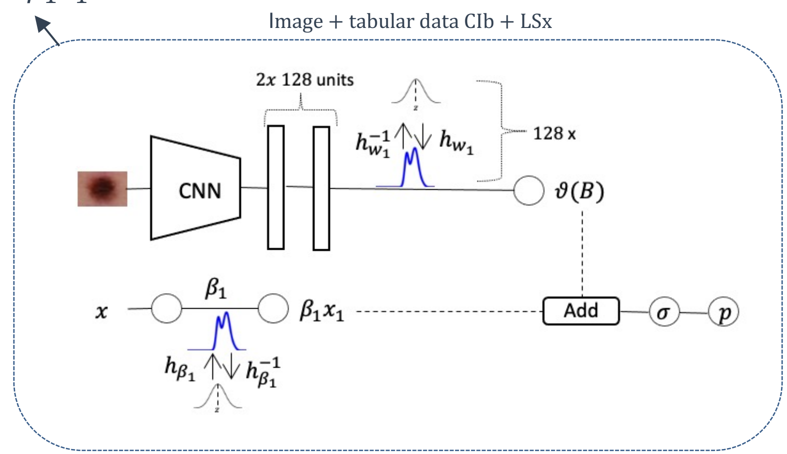
Different models:

1. Apply TM-VI only to LSx term $\beta_1 x_1$ & modeling image without uncertainty
    - Performance
        - Log-score: -0.075
        - AUC: 0.84 [0.80-0.87]



$$h = \vartheta(B) + \beta_1 x_1$$

2. Apply MF-TM-VI to image and tabular part
    - Performance
        - Log-score: -0.074
        - AUC: 0.85 [0.82-0.88]

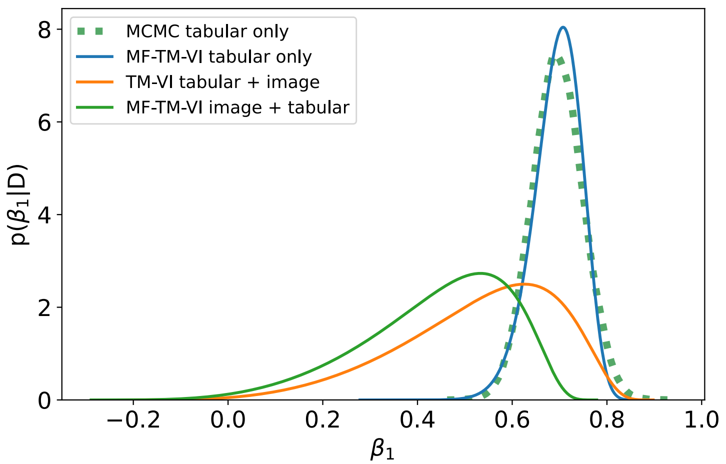


M4: Slope parameter with addition of image data

Models:

1. Apply TM-VI only to LSx term $\beta_1 x_1$ & modeling image without uncertainty
   - β_1: 0.59 [0.20-0.79] (95% HDI)
   - OR_{Age}= $e^{\beta_{Age}}$= 1.80 [1.21-2.20] (holding $\vartheta(\mathrm{B})$ constant)

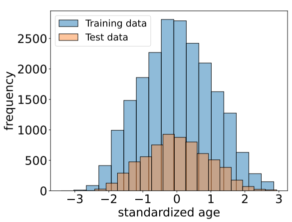
2. Apply MF-TM-VI to image and tabular part
   - β_1: 0.51 [0.14-0.69] (95% HDI)
   - OR_{Age}= $e^{\beta_{Age}}$= 1.67 [1.15-2.00] (holding $\vartheta(\mathrm{B})$ constant)

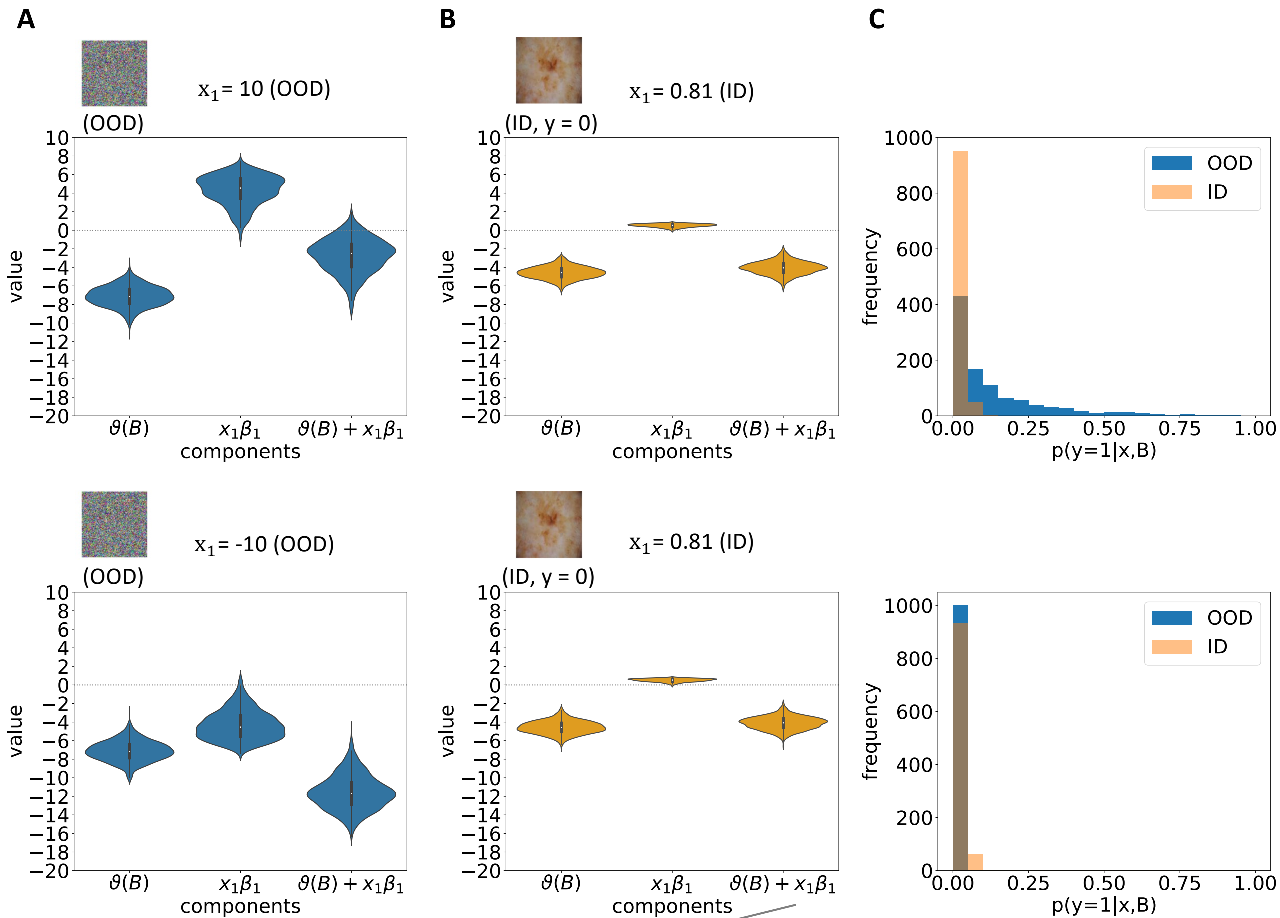


3. Results

- **Out-of-distribution (OOD) detection:**
  MF-TM-VI in image and tabular part

Experiments: Out-of-distribution (OOD) detection



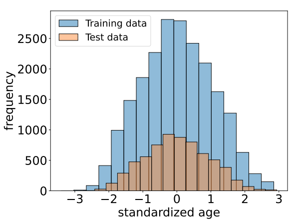


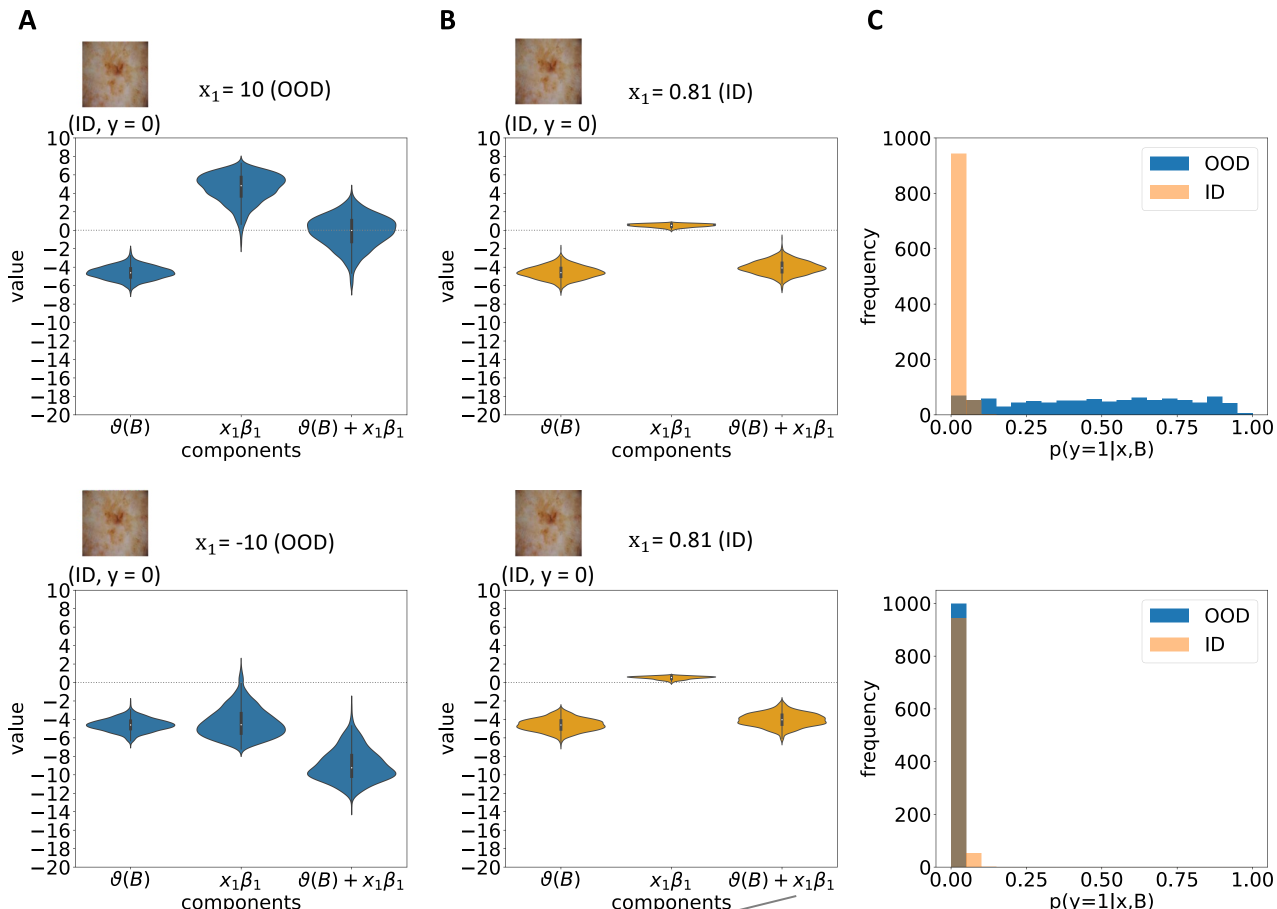
Outcome distribution z (log odds) = $\vartheta(B) + x_1\beta_1$

Posterior predictive distribution

$p(y = 1|x, B) = \sigma(\vartheta(B) + x_1\beta_1)$

Experiments: Out-of-distribution (OOD) detection



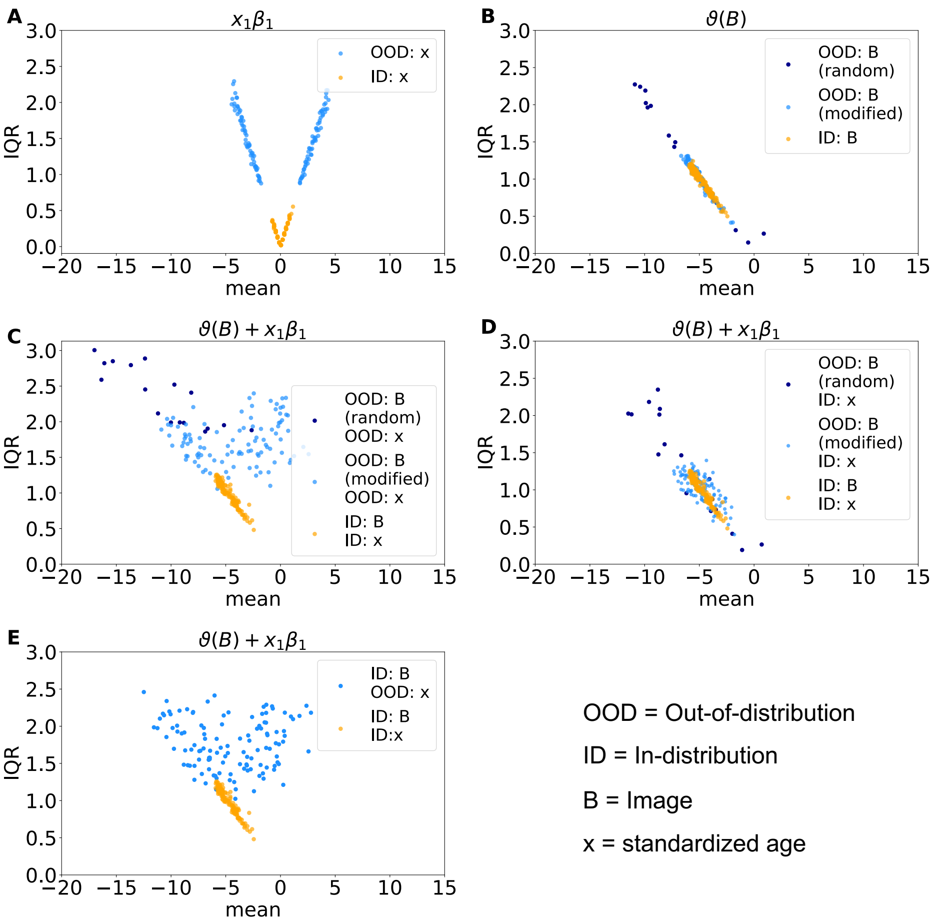


Outcome distribution z (log odds) = $\vartheta(B) + x_1\beta_1$

Posterior predictive distribution

$$p(y = 1|x, B) = \sigma(\vartheta(B) + x_1\beta_1)$$

Evaluation of multiple data



OOD = Out-of-distribution

ID = In-distribution

B = Image

x = standardized age

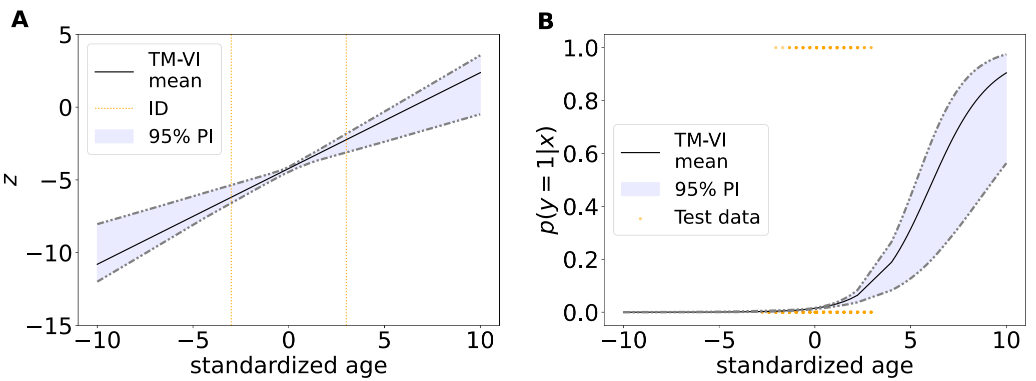
- Model components: $\beta_1 x_1$, $\vartheta(B)$, $\vartheta(B) + \beta_1 x_1$ before entering sigmoid
- 120 in-distribution test data:
  - 120 standardized age data from range [-3,3]
  - 120 images

- 120 out-of-distribution data:
  - 120 standardized age data from range [-10,4] & [4,10]
  - 17 random images

  - 103 slightly modified images

Evaluation MF-TM-VI of tabular data only

- In-distribution: 120 test data
- Out-of-distribution: 60 data from range [-10, -4] ; 60 data from range [4, 10]

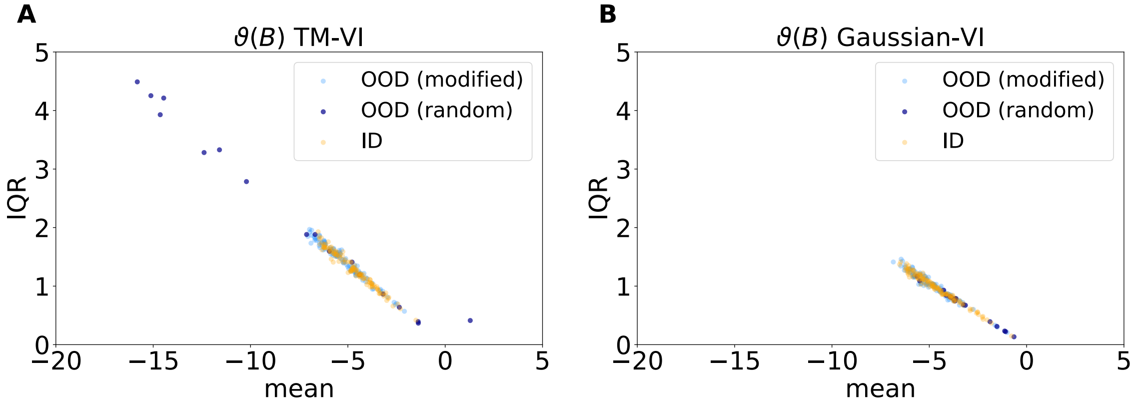


Outcome distribution z (log odds) = $\beta_0 + x_1\beta_1$

Posterior predictive distribution
$p(y=1|x,B) = \sigma(\beta_0 + x_1\beta_1)$

Evaluation MF-TM-VI of image data only

- Comparison of MF-TM-VI and MF-Gaussian-VI
- Log-Odds: $z = \vartheta(B)$
- In-distribution: 120 test data
- Out-of-distribution: 120 OOD images (random & image augmentation)



4. Conclusion and outlook

Conclusion

- Additional consideration of tabular data:
    - Interpretability effect of patient’s age
- Quantify uncertainty using TM-VI method:
    - Posterior distribution of the interpretable parameter of all models
    - Detection of OOD:
        - Uncertainties must be caught before using sigmoid
        - Image part: Not always reliable
- Performance
    - Performance increases when training with tabular and image data
    - Combination of both using the MF-TM-VI method: Best performance

Outlook

- Performance of CNN based models can be improved
- Bayes variant for the entire CNN and complex intercept based on tabular data
- Include other interpretable predictors like diagnosis, location of lesion
- First step to combine interpretable semi-structured models with the VI-procedure
    - Applied to other medical classification tasks

Thank you for your attention!

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. Hörtling, 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.

Bibliography

[15] M. A. Kassem, K. M. Hosny, R. Damaševicius, 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, 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.leitlinienprogrammonkologie.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. Malvehy, 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.

Bibliography

[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.

Additional material

Performance of all models

- Non-Bayesian models

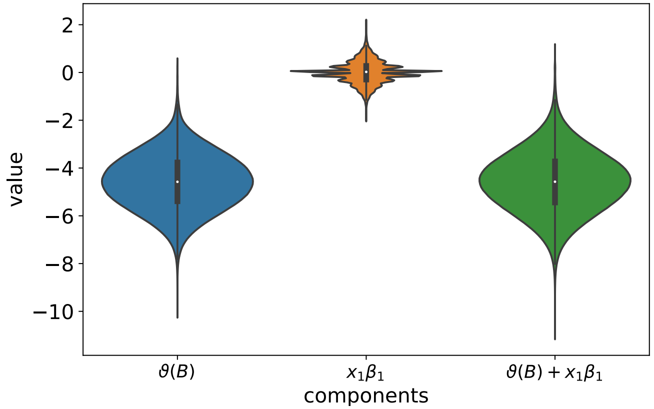
Model	Log-Score	AUC (95% -CI)	$\mathbf{OR_{Age}}$
Logistic regression	-0.085	0.66 $[0.61 - 0.71]$	2.01 $[1.82 - 2.25]$
M1 SI LSx $h = \beta_0 + \beta_1 x_1$	-0.085	0.66 $[0.61 - 0.71]$	2.01
M2 CIx $h = \beta(x)$	-0.085	0.66 $[0.61 - 0.70]$	-
M3 CIb $h = \vartheta(B)$	-0.078	0.81 $[0.77 - 0.84]$	-
M4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

- Bayesian models

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

Figures

Outcome distribution of components of combined model



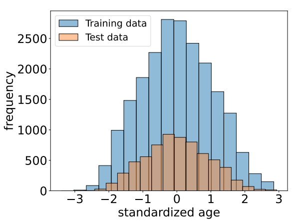
Example PyStan:
Distribution of slope parameter β_1

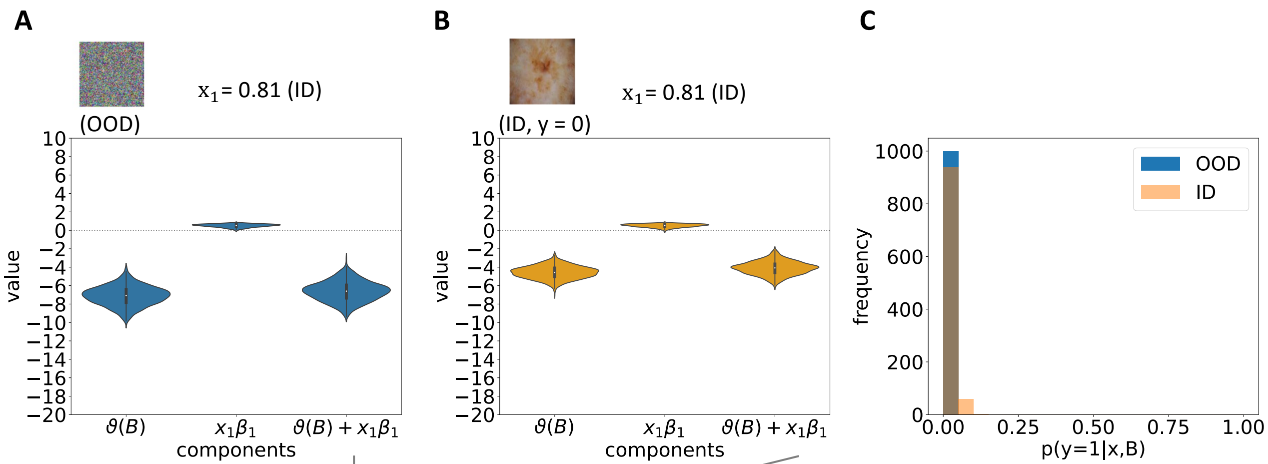
```
lr_code = """
  data {
    int N;
    int M;
    real X[N, M];
    int<lower=0, upper=1> y[N];
  }
  parameters {
    real intercept;
    real beta[M];

  }
  model {
    for (i in 1:N)
        y[i] ~ bernoulli(inv_logit (intercept+ dot_product(X[i] , beta)));
    beta[M] ~ normal(0, 1);


  }
"""
```

Experiments: Out-of-distribution (OOD) detection





Outcome distribution z (log odds) = $\vartheta(B) + x_1\beta_1$

Posterior predictive distribution

$p(y = 1|x, B) = \sigma(\vartheta(B) + x_1\beta_1)$