Joshi, P., Dhar, R. EpICC: A Bayesian neural network model with uncertainty correction for a more accurate classification of cancer. *Sci Rep* 12, 14628 (2022).

Computational Cancer Research

Seminar Talk by Abhinav Mishra

December 7th, 2022

“It really is a nice theory. The only defect I think it has is probably common to all philosophical theories. It's wrong.”

Saul Kripke (1980, ‘Naming and Necessity’, pg. 64)

EpiCC

Epistemic Invariance in Cancer Classification*

Motivation

Goal

Background

Methods

Remarks

Motivation

Measure of confidence can improve decision-making ability

Goal

Classification of individual patient samples into cancer types, and subtypes with uncertainty with each prediction.

Background

Uncertainty

Aleatoric

Quality of data

- Labels
- Measurements

Epistemic

Model

Parameter selection

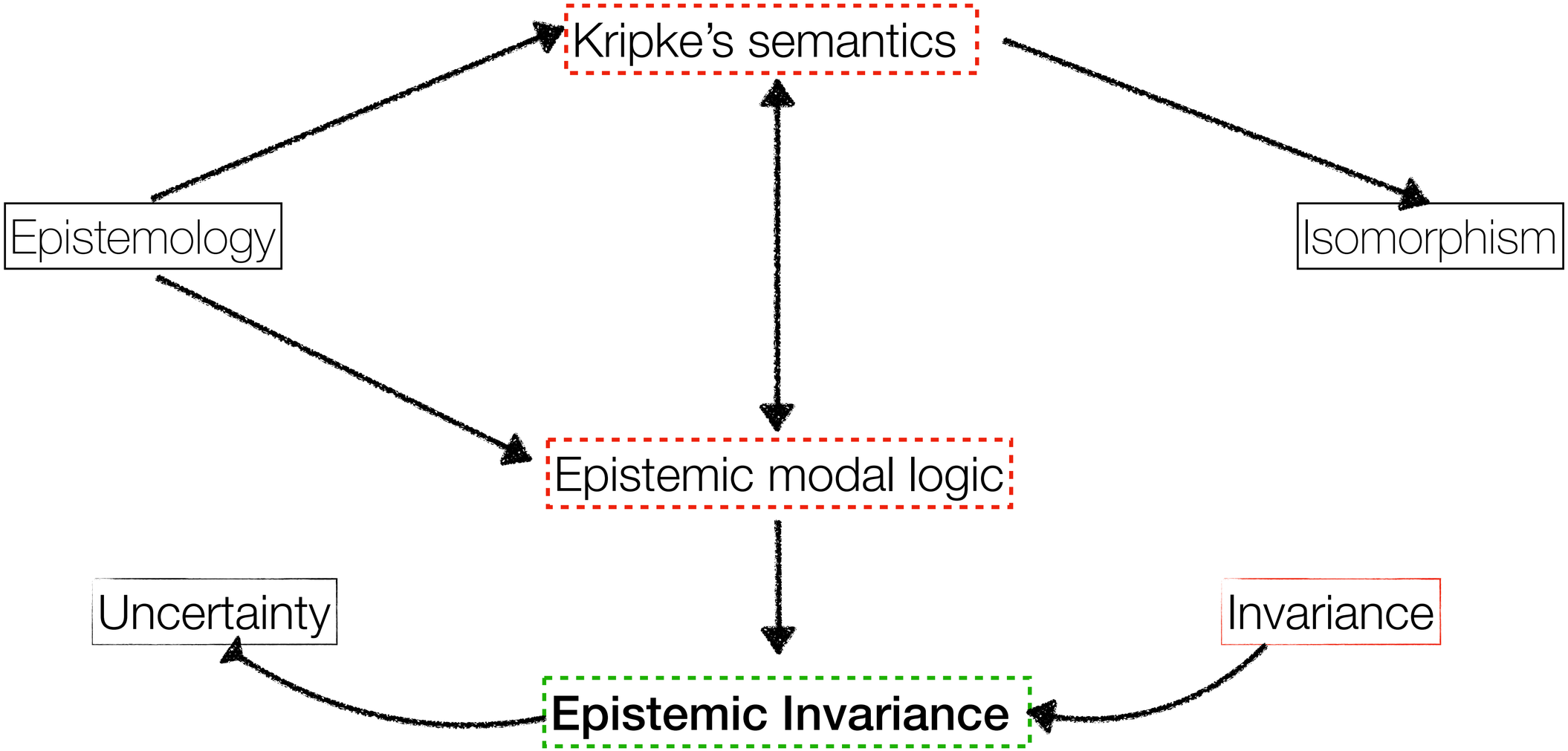
What is Epistemic Uncertainty ?

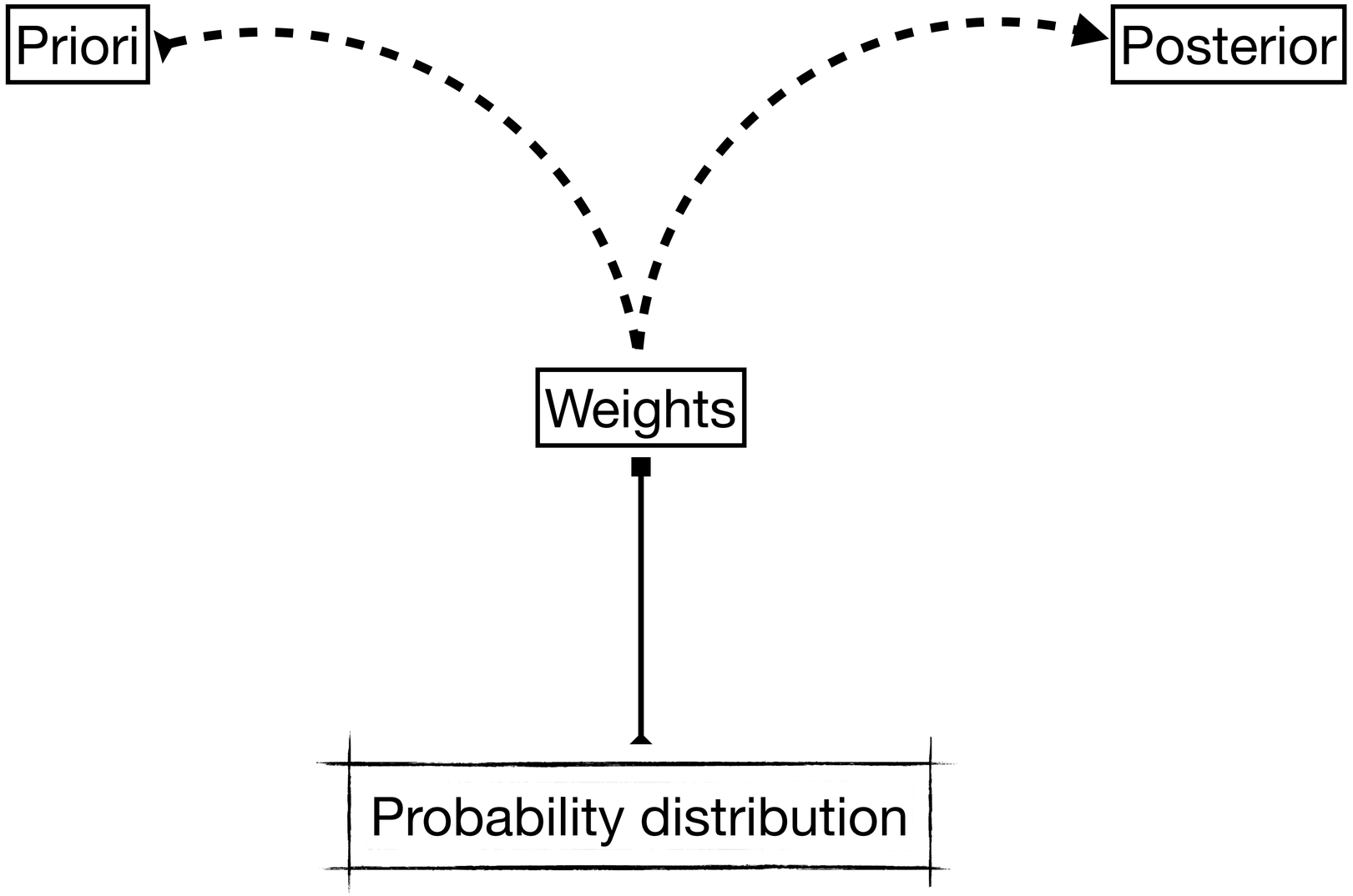
Variations in model fitting

Variations in classification

What is Epistemic Invariance ?

The property of being always equally accessible.

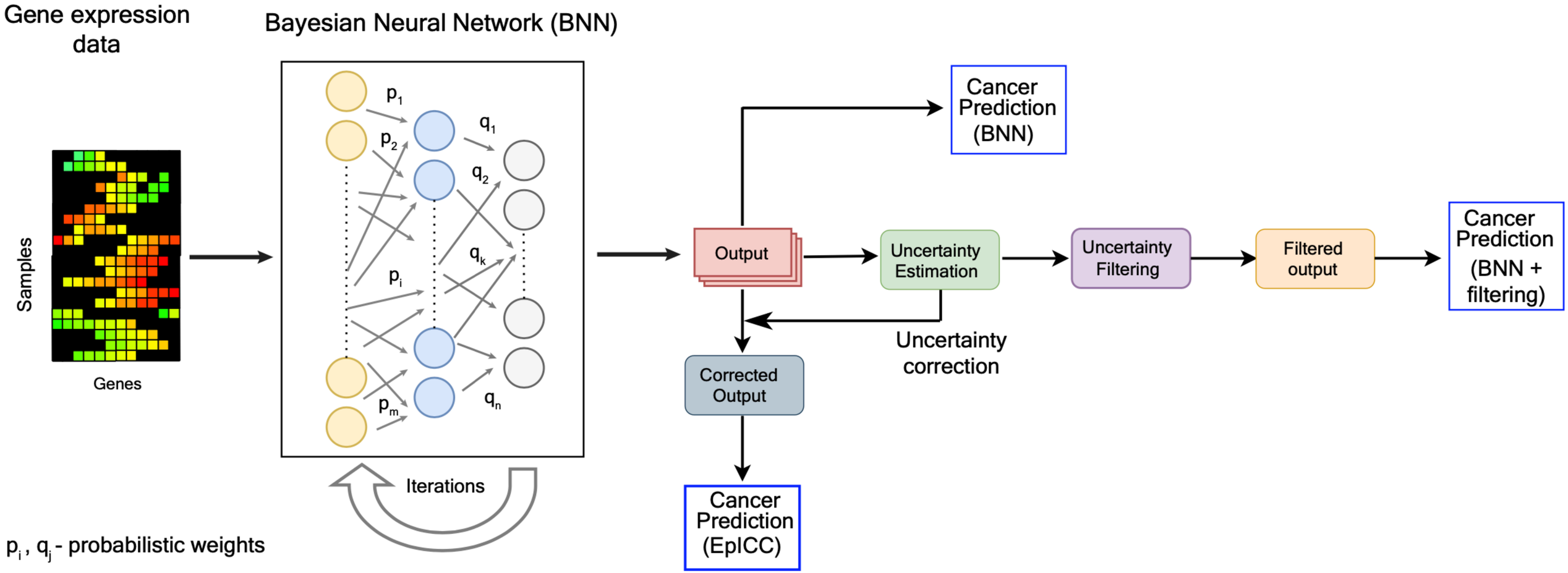

Kripke’s semantics
Epistemology
Isomorphism
Epistemic modal logic
Uncertainty
Invariance
Epistemic Invariance


Priori
Posterior
Weights
Probability distribution

Methods

3-layered BNN + uncertainty corrected

First Layer - 250 units, Hidden layer - 95 units, Output layer - 31 units



Source: Figure 1, pg. 3, https://doi.org/10.1038/s41598-022-18874-6

Data

Trancriptome (TCGA level 3)

Sequencing

Illumina HiSeq 200

Counts

$\log_2(x + 1)$ transformed RSEM normalized

Samples

10,013

Splitting

80% → training + feature selection | 20% → testing

Data

Trancriptome (TCGA level 3)

Cancer types

31

Cancer subtypes

4 out of 31

Source

UCSC XENA

Data

Binary classification: Cancer vs. Non-cancer

Values

Gene expression values of normal samples (GTEx)

Counts

$\log_2(x + 1)$ transformed RSEM normalized

Samples

7851

Splitting

80% → training + feature selection | 20% → testing

Data

Binary classification: Cancer vs. Non-cancer

Missing values

7

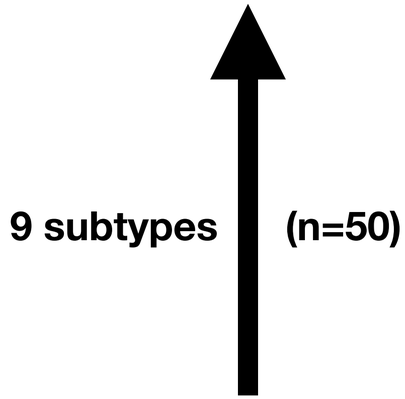
Model

L2-regularised logistic regression

Data

Cancer subtype

Expression values assigned to respective types



9 subtypes **(n=50)**

Phenotypic information

Feature Selection

Two-step PCA

Why ? Reducing the risk of **over-fitting**, and **redundancy**.

1. Selecting a set of genes (n=103) from the original RNA-seq data.

2. Selecting an ever smaller set from the step above

Feature Selection

Two-step PCA

First step Principal Component Analysis

For each component up to 10,

number of optimal genes ~ gene = **max** |Factor loading|

Second step Logistic Regression

Minimum number of genes required to achieve high accuracy

Bayesian Neural Network

BNN

Objective

Data points D: for i^{th} predictor variable and target variable y_i

$$D = (x_i, y_i) \forall i \in 1,2,3,...,N,$$

$N =$ number of sample points

Bayesian Neural Network

BNN

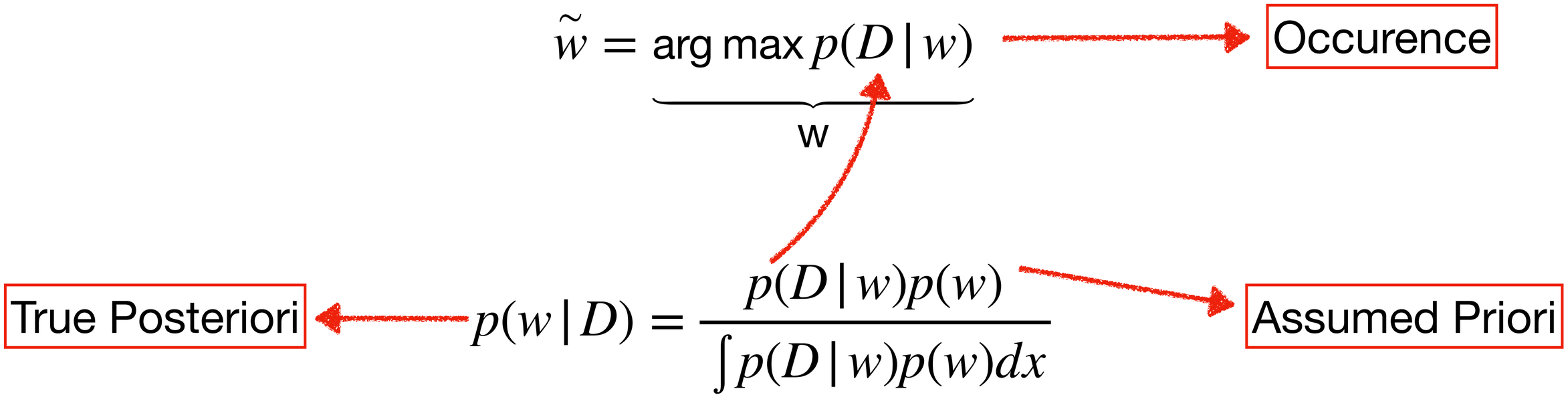
Learn the parameters w such that the probability of occurrence of data given the model parameters is **maximised**.

Maximum likehood estimate

$$\tilde{w} = \underbrace{\arg\max p(D \mid w)}_{\text{w}}$$

Bayesian Neural Network

BNN



Baye's theorem

Bayesian Neural Network

BNN

Solution of $p(w \mid D)$ is *controllable* & *feasible* by

minimising KL divergence

between

variational posterior and true posterior.

$$q(w \mid \delta) = \prod_j N(w_j \mid \mu_j, \sigma^2)$$

q's parameters

w_j's mean and variance

Bayesian Neural Network

BNN

1. Minimise the difference between the distributions $q(w \mid \delta)$ & $p(w \mid D)$.

2. Maximise the probability of occurrence $p(D|w)$.

What is q ?

The probability distribution that the model extracts.

What is p ?

The probability distribution that already exists.

Example
KL Divergence

$$D_{KL}(P||Q) = \sum_{x \in X} P(x)\, ln(\frac{P(x)}{Q(x)})$$

$$= \frac{9}{25} ln(\frac{9/25}{1/3}) + \frac{12}{25} ln(\frac{12/25}{1/3}) + \frac{4}{25} ln(\frac{4/25}{1/3})$$

$$\sim 0.0852996$$

X	0	1	2
P(X)	9/25	12/25	4/25
Q(X)	1/3	1/3	1/3

Source: Table 2.1, Kullback, Solomon (1959), Information Theory and Statistics

(Information loss?) => Loss function ??

Subtract from $log(N)$ to get Shannon Entropy

Bayesian Neural Network

BNN

δ's estimate from q

Kullback-Leibler divergence

$$\tilde{\delta} = \arg\min \mathrm{KL}_{\delta}[q(w \mid \delta) \,||\, p(w \mid D)]$$

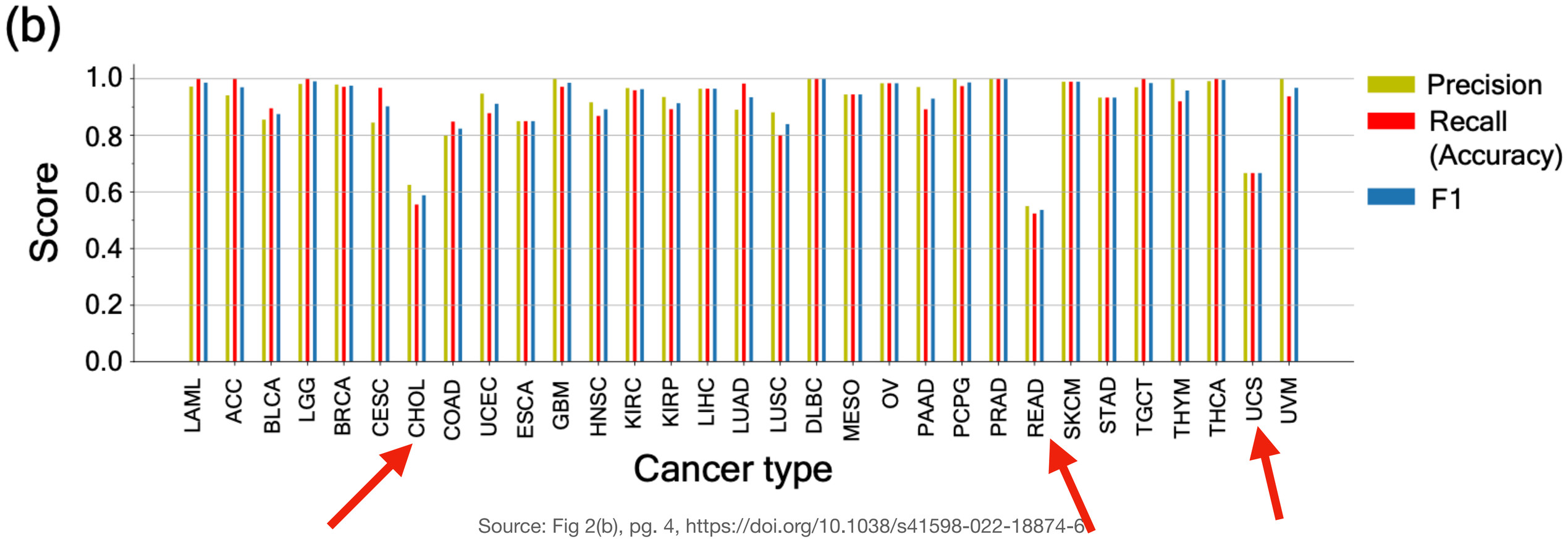
Reference distribution

p is the true posterior

q is the variational posterior

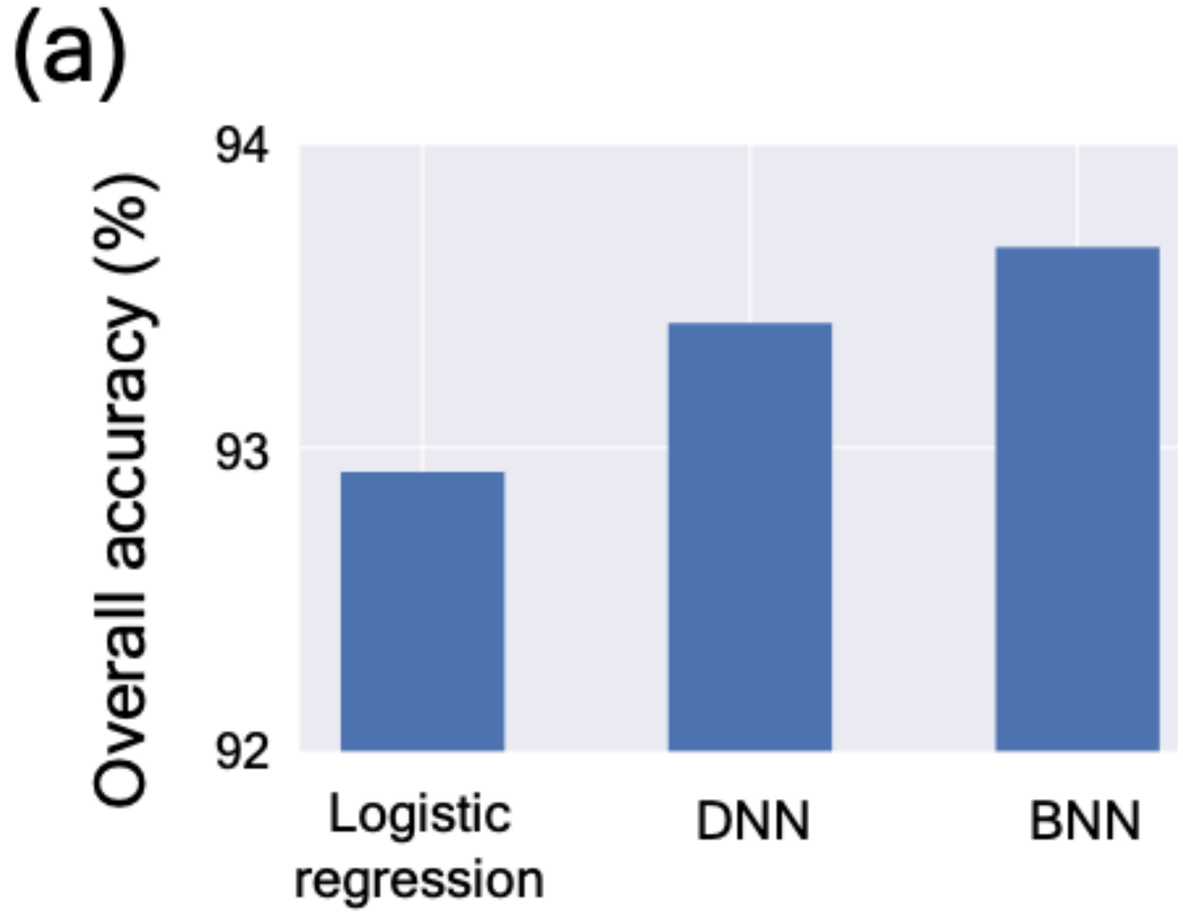
Evaluation Metrics

BNN: prediction of individual cancer types before correction



Source: Fig 2(b), pg. 4, https://doi.org/10.1038/s41598-022-18874-6

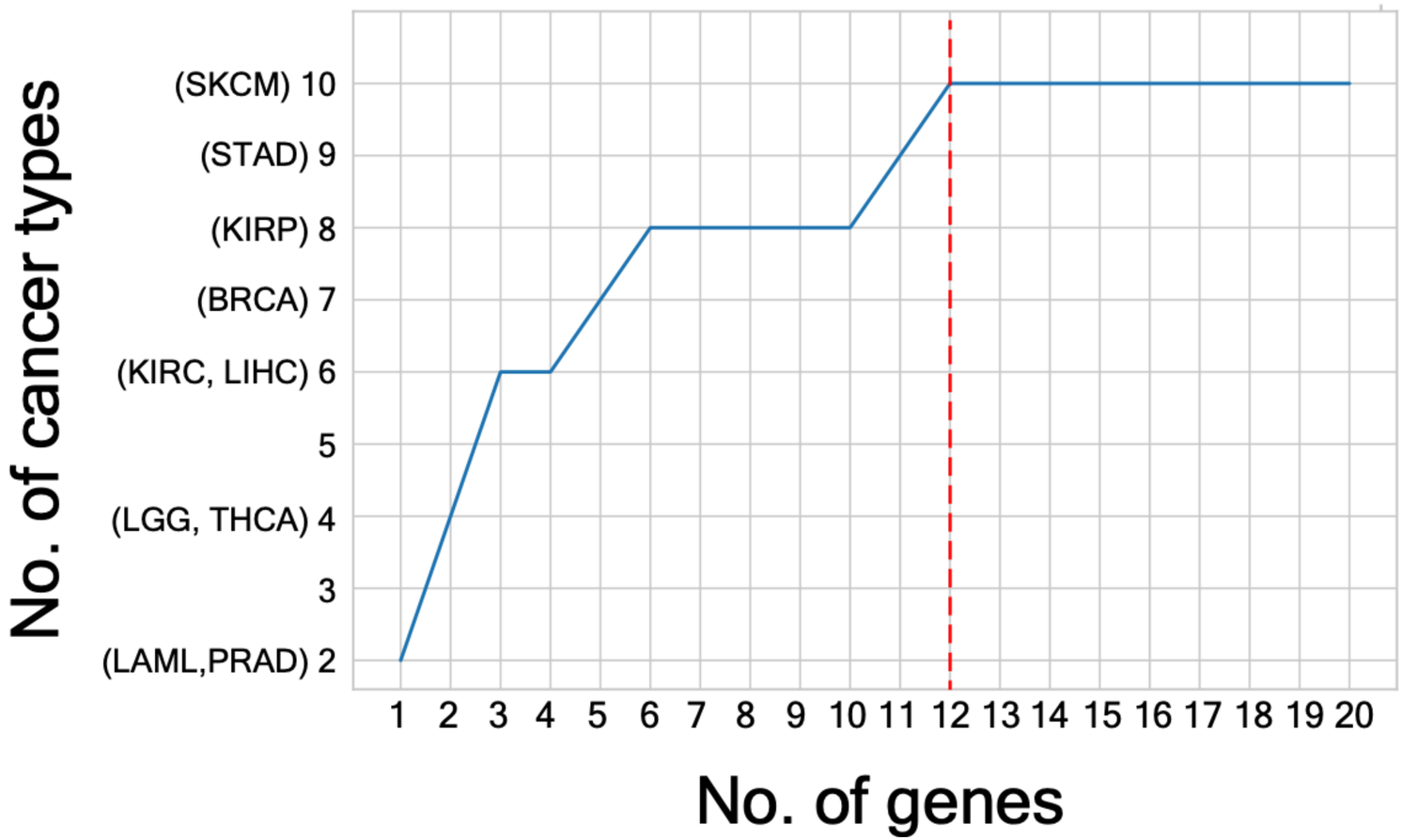
Overall Accuracy comparison



Source: Fig 2(a), pg. 4, https://doi.org/10.1038/s41598-022-18874-6

Precision & Recall > 0.75

Top 20 genes (Highest F1 scores)



Source: Fig S4, Supplementary Material, https://doi.org/10.1038/s41598-022-18874-6

Uncertainty

Estimation

Why ? Getting an idea about the **confidence of predictions** of a model.

Mean value for 500 iterations

$$\text{Epistemic Uncertainity, } \xi_i = \frac{1}{T}\sum_{t=1}^{t=T}(\hat{p}_t^i - \bar{p}^i)^2$$

Index of the class

Softmax prediction for t_{th} monte-carlo iteration

Uncertainty

Correction

1. Fit a linear model between the log-odds of $E[\hat{p}_i]$ and $\sqrt{\xi_i}$.

$$f(x) = ln(\frac{x}{1-x})$$

$$f(E[\widehat{p_i}]) = \alpha + \beta\sqrt{\xi_i} + \epsilon,\ error \sim N(0,\sigma^2)$$

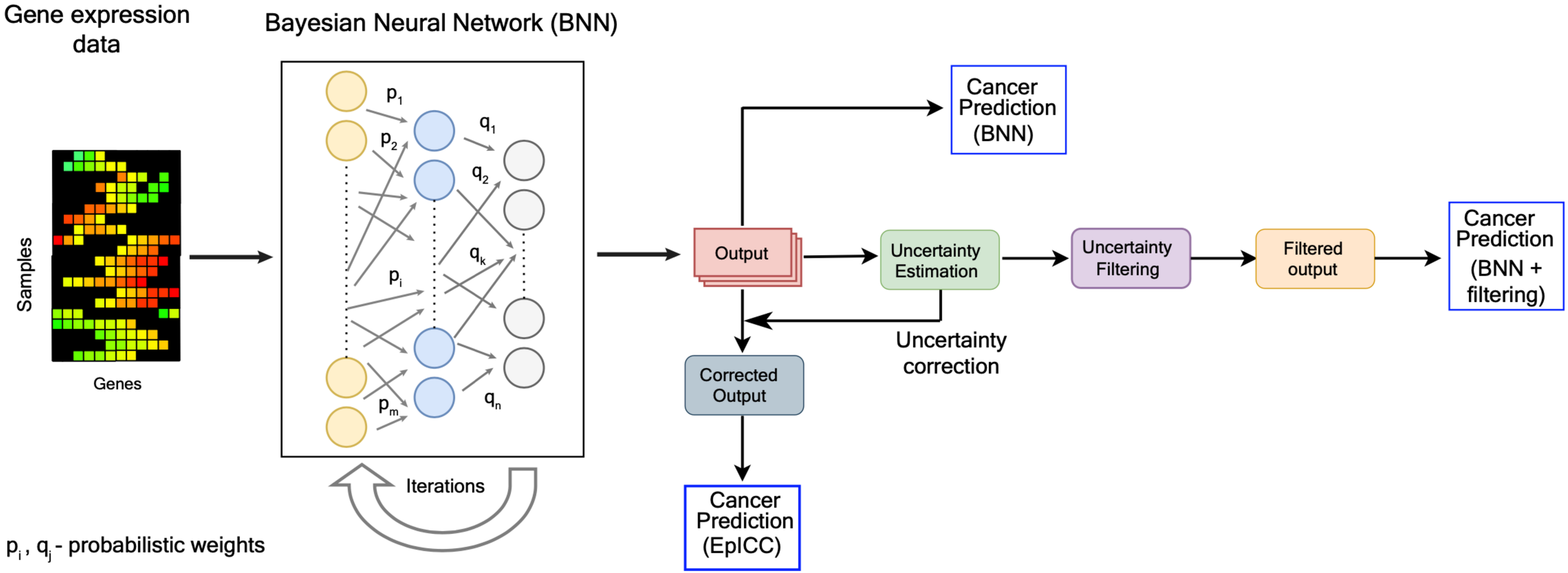
2. Calculate the coefficients α, β of the linear model using OLS.

3. Calculate the corrected prediction probabilities for each cancer class.

$$\widehat{p_{corr,i}} = f^{-1}(E[\widehat{p_i}] - \beta\,\xi_i)$$

3-layered BNN + uncertainty corrected

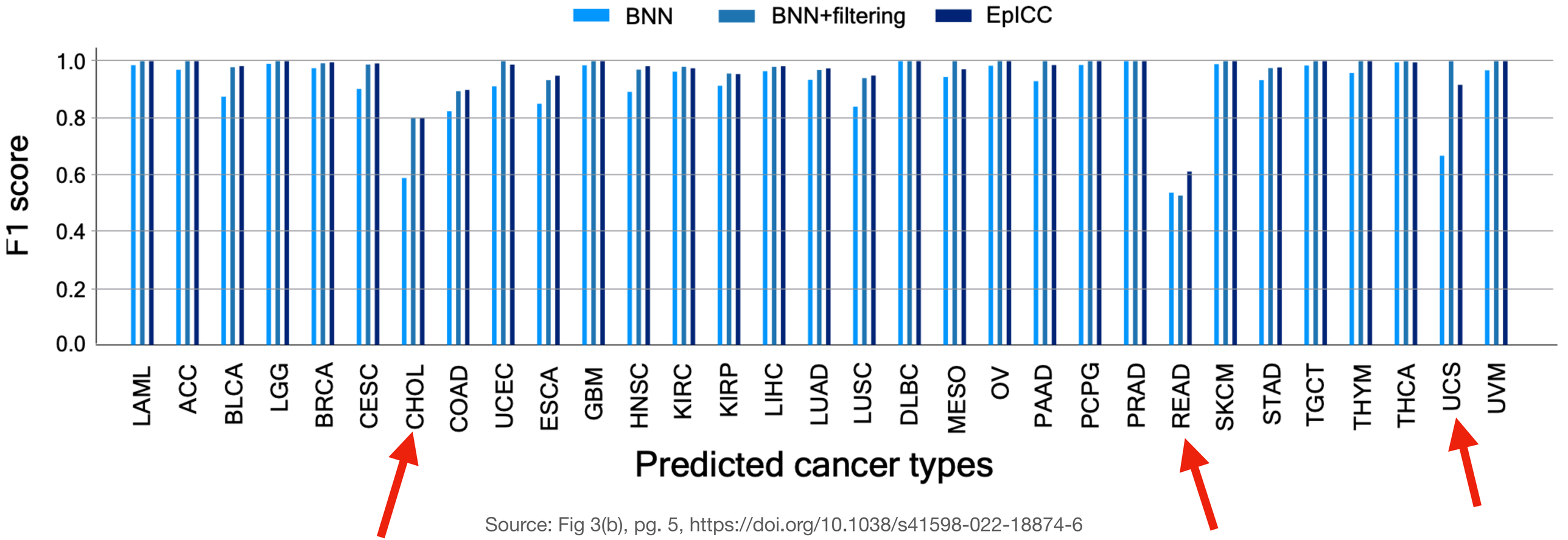
First Layer - 250 units, Hidden layer - 95 units, Output layer - 31 units



Source: Figure 1, pg. 3, https://doi.org/10.1038/s41598-022-18874-6

Evaluation Metrics

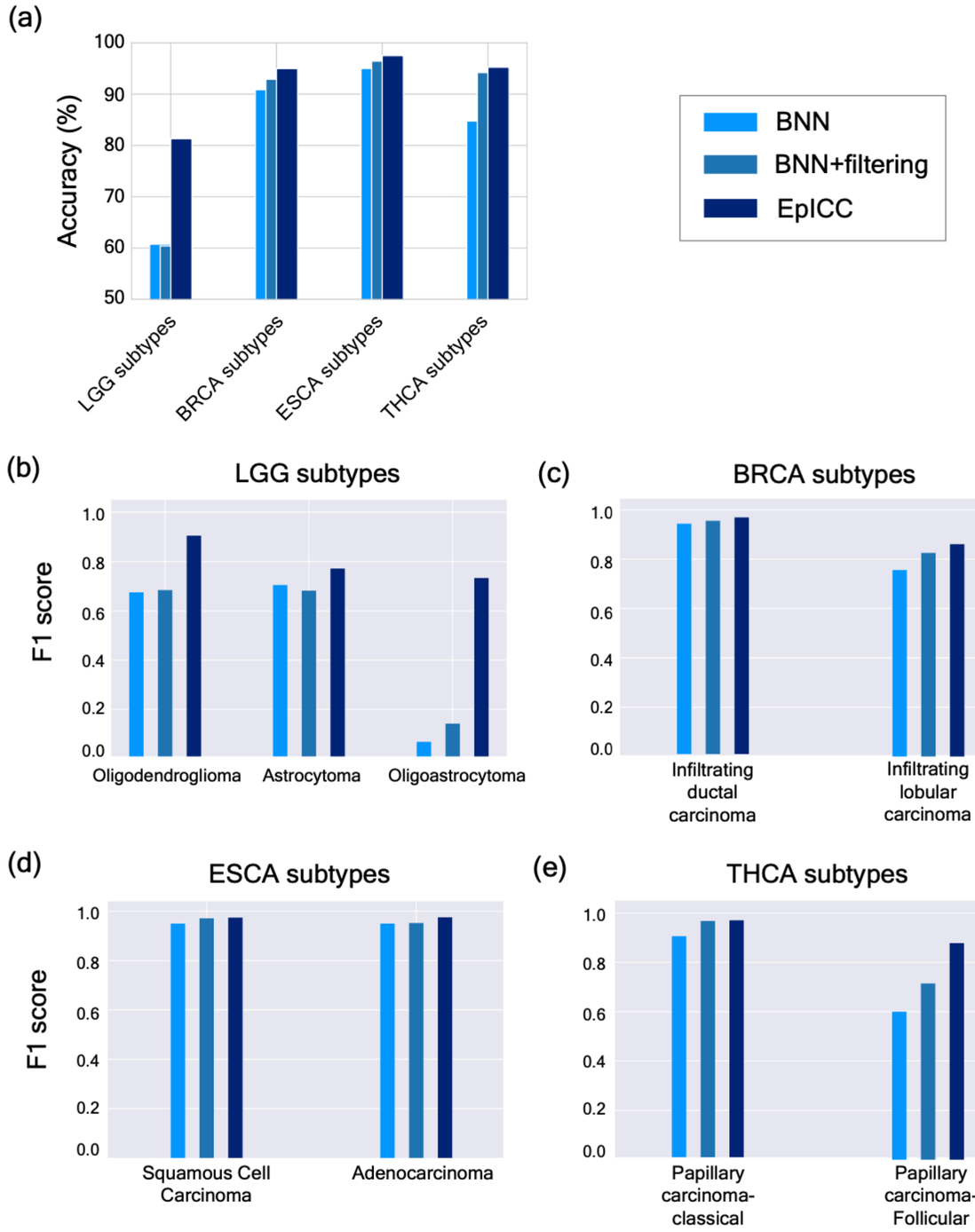
EpiCC: Comparison of F1 scores after uncertainty correction



Source: Fig 3(b), pg. 5, https://doi.org/10.1038/s41598-022-18874-6

Subtype Classification

Accuracy & F1 Score



$$Accuracy = \frac{TP + FN}{TP + FN + FP + TN}$$

$$Precision(P) = \frac{TP}{TP + FP}$$

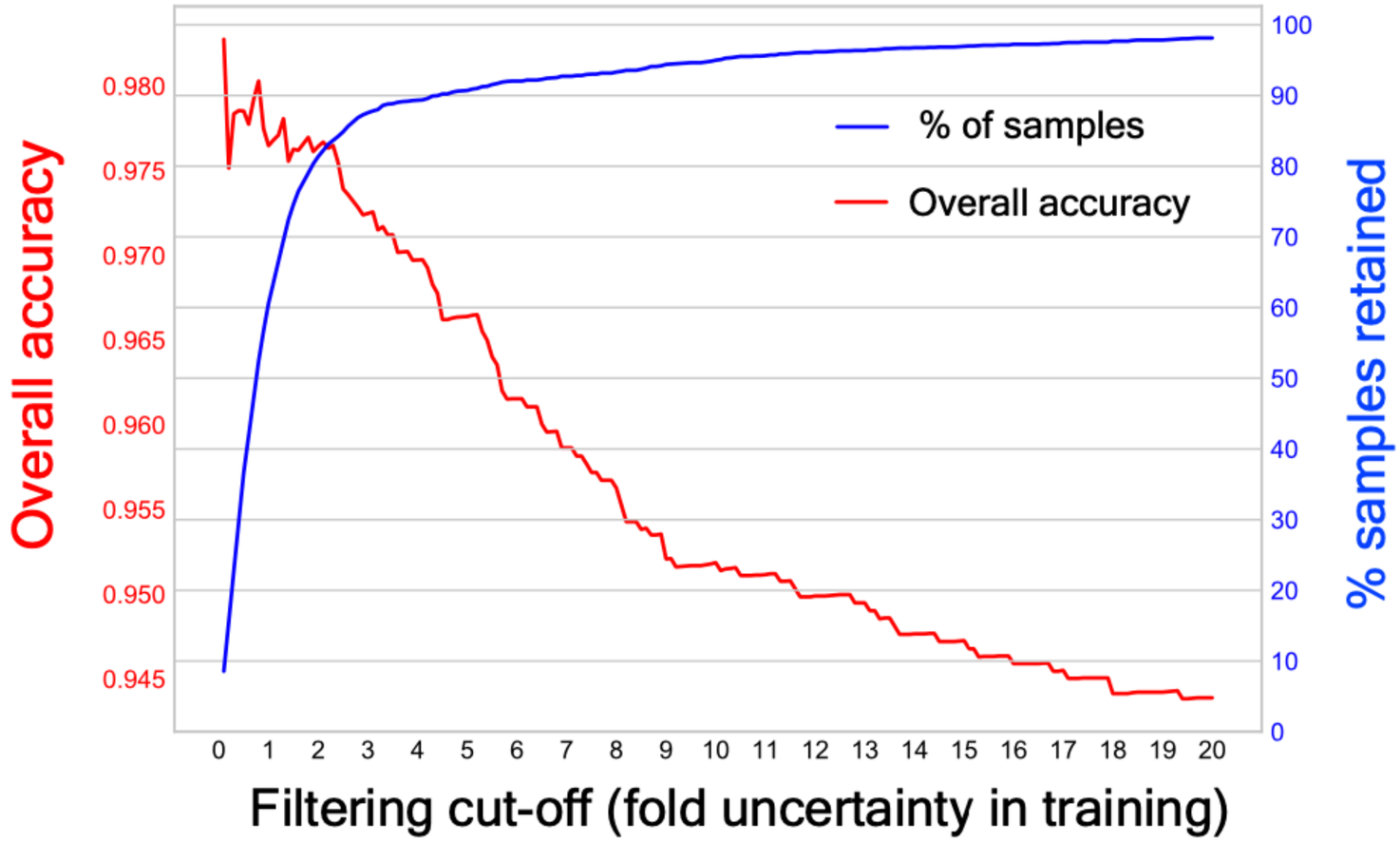
$$Recall(R) = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{2PR}{P + R}$$

Source: Fig 4, pg. 7, https://doi.org/10.1038/s41598-022-18874-6

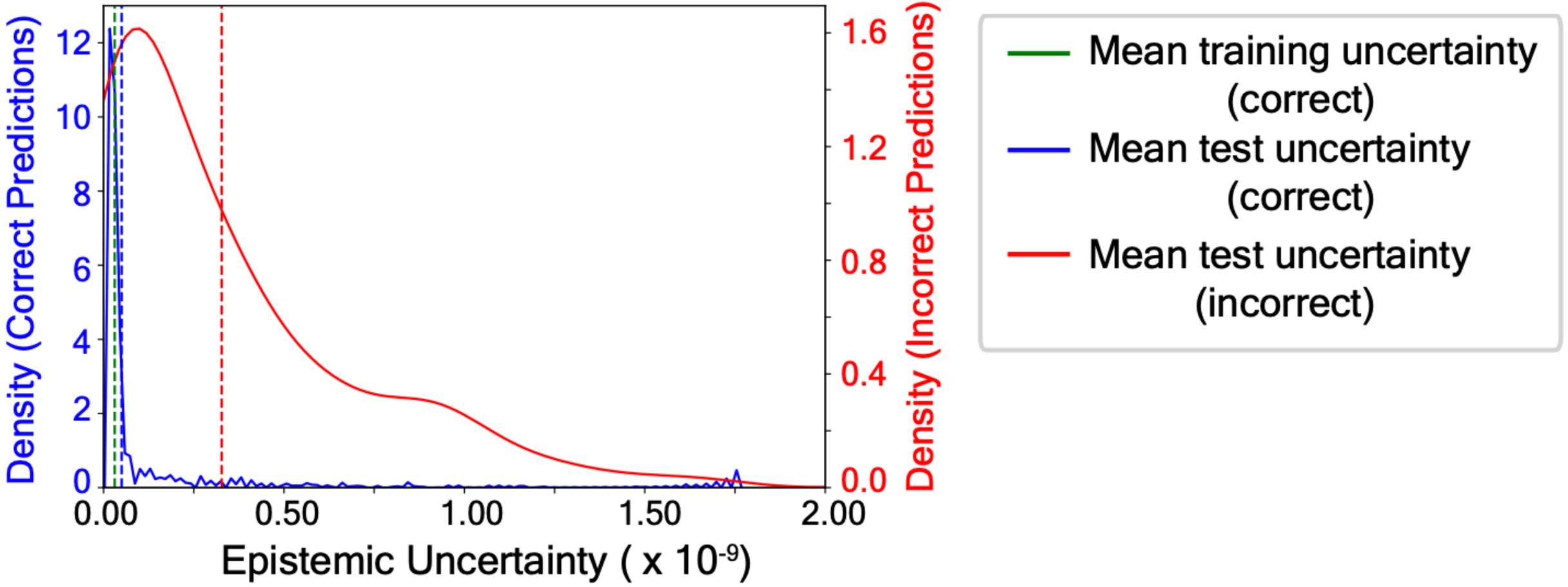
Filtering Cutoff

Accuracy & Number of samples



Source: Fig 3(c), pg. 5, https://doi.org/10.1038/s41598-022-18874-6

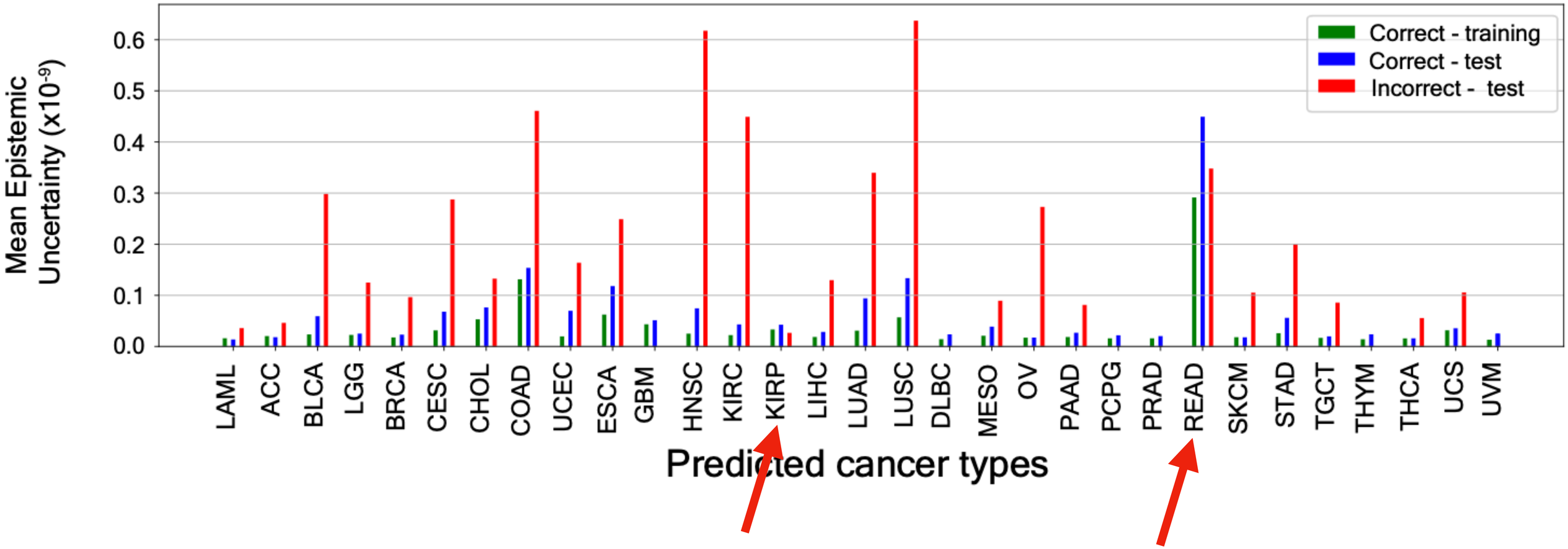
Distribution of Epistemic Uncertainty



Source: Fig. S5, Supplementary Materials, https://doi.org/10.1038/s41598-022-18874-6

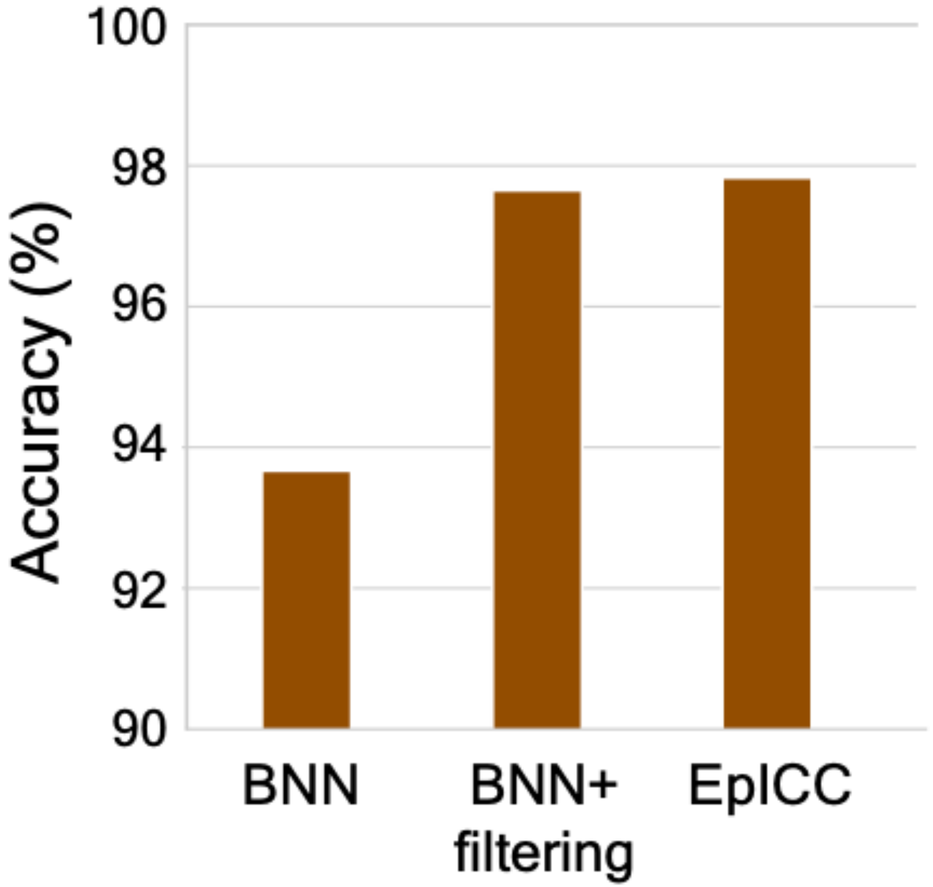
Comparison of Mean Uncertainty

correct and incorrect predictions



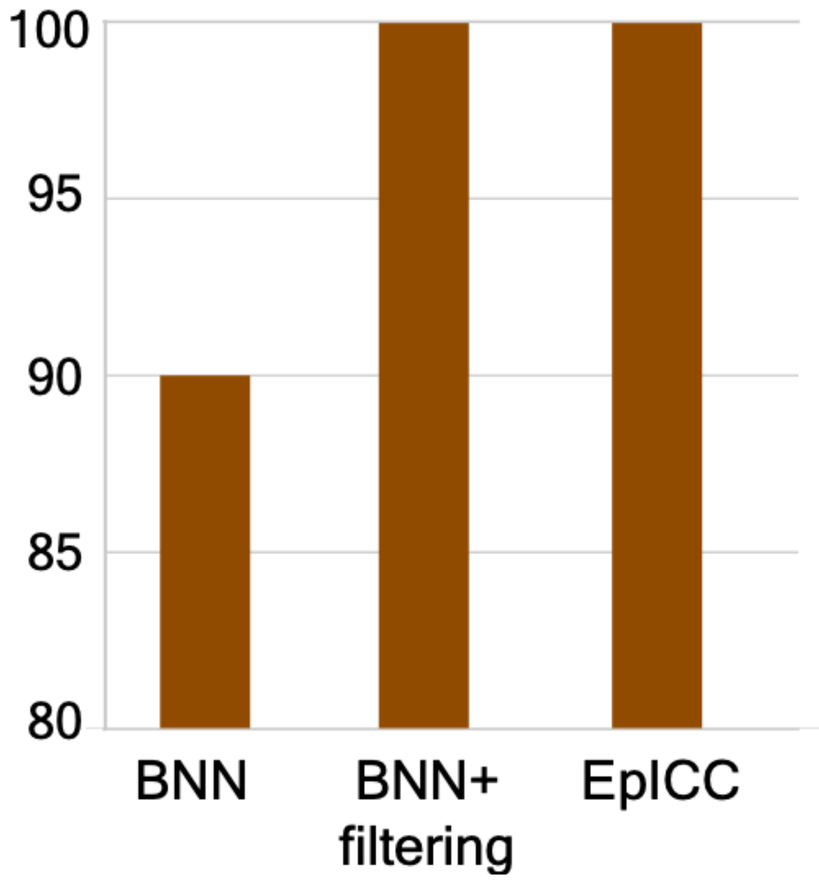
Independent validation: Accuracy

UCSC XENA



Source: Fig 3(d), pg. 5, https://doi.org/10.1038/s41598-022-18874-6

BRCA (external cohort)



Source: Fig 3(e), pg. 5, https://doi.org/10.1038/s41598-022-18874-6

Performance Comparison

Accuracy

Study	Classification accuracy (%)				
	Cancer types	LGG subtypes	BRCA subtypes	ESCA subtypes	THCA subtypes
Lyu and Haque[29]	95.59% (33)	NA	NA	NA	NA
Kim et al.[31]	91.74% (21)	NA	NA	NA	NA
Xiao et al.[25]	96%-99% (3)	NA	NA	NA	NA
Ramirez et al.[49]	94.70% (33)	NA	NA	NA	NA
Sun et al.[48]	97.47% (12)	NA	NA	NA	NA
Pei et al.[50]	NA	63.90 (3)	NA	NA	NA
Couture et al.[51]	NA	NA	94 (2)	NA	NA
EpICC	97.83% (31)	81.31 (3)	94.98 (2)	97.5 (3)	95.24 (2)

Source: Table 1, pg. 8, https://doi.org/10.1038/s41598-022-18874-6

Remarks

Combining transcriptomic data with epigenetic modification patterns in cancers can increase subtype classification accuracy.

This work* demonstrates the value of modelling uncertainty in cancer classification.

„Die Grenzen meiner Sprache sind die Grenzen meiner Welt”

Ludwig Wittgenstein (1922, ’Tractatus logigo-philosphicus’)

References
& Credits

*Joshi, P., Dhar, R. EpICC: A Bayesian neural network model with uncertainty correction for a more accurate classification of cancer. *Sci Rep* **12**, 14628 (2022). https://doi.org/10.1038/s41598-022-18874-6. Repo: https://github.com/pjoshi-hub/Bayesian_classification_model

Theodoridis, K. (2007). Kripke and the Predicament of Epistemic Invariance. *Philosophical Inquiry*, *29*(1/2), 72-83.

Presentation software ‘Keynote’ was used to develop slides. Background images on the first, and last slide belongs to the presenter.