Michael R. Berthold
Ad Feelders
Georg Krempl (Eds.)

LNCS 12080

Advances in Intelligent Data Analysis XVIII

18th International Symposium on Intelligent Data Analysis, IDA 2020
Konstanz, Germany, April 27–29, 2020
Proceedings

Springer Open

EXTRAS ONLINE

Lecture Notes in Computer Science 12080

More information about this series at http://www.springer.com/series/7409

Michael R. Berthold · Ad Feelders ·
Georg Krempl (Eds.)

Advances in Intelligent Data Analysis XVIII

18th International Symposium on Intelligent Data Analysis, IDA 2020
Konstanz, Germany, April 27–29, 2020
Proceedings

Springer Open

Editors
Michael R. Berthold
University of Konstanz
Konstanz, Germany

Ad Feelders
Utrecht University
Utrecht, The Netherlands

Georg Krempl
Utrecht University
Utrecht, The Netherlands

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-44583-6 ISBN 978-3-030-44584-3S (eBook)
https://doi.org/10.1007/978-3-030-44584-3

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

We are proud to present the proceedings of the 18th International Symposium on Intelligent Data Analysis (IDA 2020), which was held during April 27–29, 2020, in Konstanz, Germany. The first symposium of this series was organized in 1995 and held biannually until 2009, when the conference switched to being held annually. Following demand expressed by the IDA community in a survey held in 2018, IDA 2020 was the first of the series to take place in spring rather than fall, as was common before.

The switch to April, and a more organized outreach to the community, coincided with an increase in the number of submissions from 65 in 2018, to 114 in 2020. After a rigorous review process, 45 of these 114 submissions were accepted for presentation. Almost all submissions were reviewed by at least three Program Committee (PC) members (only two papers had two reviews) and a substantial number of submissions received more than three reviews. In addition to the PC, the review process also involved program chair advisors – a select set of senior researchers with a multi-year involvement in the IDA symposium series. Whenever a program chair advisor flagged a paper with an informed, thoughtful, positive review due to the paper presenting a particularly interesting and novel idea, the paper was accepted irrespective of the other reviews. Each accepted paper was offered a slot for either oral presentation (15 papers) or poster presentation (30 papers).

We wish to express our gratitude to the authors of all submitted papers for their high-quality contributions; to the PC members and additional reviewers for their efforts in reviewing, discussing, and commenting on all submitted papers; to the program chair advisors for their active involvement; and to the IDA council for their ongoing guidance and support. Many people have helped behind the scenes to make IDA 2020 possible, but this year we are particularly grateful to our publicity chairs who helped spread the word: Daniela Gawehns and Hugo Manuel Proença!

February 2020

Georg Krempl
Ad Feelders
Michael R. Berthold

Organization

Program Chairs

Georg Krempl	Utrecht University, The Netherlands
Ad Feelders	Utrecht University, The Netherlands

Program Chair Advisors

Niall Adams	Imperial College London, UK
Michael R. Berthold	University of Konstanz, Germany
Hendrik Blockeel	Katholieke Universiteit Leuven, Belgium
Elizabeth Bradley	University of Colorado Boulder, USA
Tijl De Bie	Ghent University, Belgium
Wouter Duivesteijn	Eindhoven University of Technology, The Netherlands
Elisa Fromont	Université de Rennes 1, France
Johannes Fürnkranz	Johannes Kepler University Linz, Austria
Jaakko Hollmén	Aalto University, Finland
Frank Höppner	Ostfalia University of Applied Sciences, Germany
Frank Klawonn	Ostfalia University of Applied Sciences, Germany
Arno Knobbe	Leiden University, The Netherlands
Rudolf Kruse	University of Magdeburg, Germany
Nada Lavrač	Jozef Stefan Institute, Slovenia
Matthijs van Leeuwen	Leiden University, The Netherlands
Xiaohui Liu	Brunel University, UK
Panagiotis Papapetrou	Stockholm University, Sweden
Arno Siebes	Utrecht University, The Netherlands
Stephen Swift	Brunel University, UK
Hannu Toivonen	University of Helsinki, Finland
Allan Tucker	Brunel University, UK
Albrecht Zimmermann	Université Caen Normandie, France

Program Committee

Fabrizio Angiulli	DEIS, University of Calabria, Italy
Martin Atzmueller	Tilburg University, The Netherlands
José Luis Balcázar	Universitat Politècnica de Catalunya, Spain
Giacomo Boracchi	Politecnico di Milano, Italy
Christian Borgelt	Universität Salzburg, Austria
Henrik Boström	KTH Royal Institute of Technology, Sweden
Paula Brito	University of Porto, Portugal
Dariusz Brzezinski	Poznań University of Technology, Poland
José Del Campo-Ávila	Universidad de Málaga, Spain

Cassio de Campos	Eindhoven University of Technology, The Netherlands
Andre de Carvalho	University of São Paulo, Brazil
Paulo Cortez	University of Minho, Portugal
Bruno Cremilleux	Université de Caen Normandie, France
Brett Drury	LIAAD-INESC-TEC, Portugal
Saso Dzeroski	Jozef Stefan Institute, Slovenia
Nuno Escudeiro	Instituto Superior de Engenharia do Porto, Portugal
Douglas Fisher	Vanderbilt University, USA
Joao Gama	University of Porto, Portugal
Lawrence Hall	University of South Florida, USA
Barbara Hammer	Bielefeld University, Germany
Martin Holena	Institute of Computer Science, Czech Republic
Tomas Horvath	Eötvös Loránd University, Hungary
Francois Jacquenet	Laboratoire Hubert Curien, France
Baptiste Jeudy	Laboratoire Hubert Curien, France
Ulf Johansson	Jönköping University, Sweden
Alipio M. Jorge	University of Porto, Portugal
Irena Koprinska	The University of Sydney, Australia
Daniel Kottke	University of Kassel, Germany
Petra Kralj Novak	Jozef Stefan Institute, Slovenia
Mark Last	Ben-Gurion University of the Negev, Israel
Niklas Lavesson	Jönköping University, Sweden
Daniel Lawson	University of Bristol, UK
Jefrey Lijffijt	Ghent University, Belgium
Ling Luo	The University of Melbourne, Australia
George Magoulas	Birkbeck University of London, UK
Vlado Menkovski	Eindhoven University of Technology, The Netherlands
Vera Migueis	University of Porto, Portugal
Decebal Constantin Mocanu	Eindhoven University of Technology, The Netherlands
Emilie Morvant	University of Saint-Etienne, LaHC, France
Mohamed Nadif	Paris Descartes University, France
Siegfried Nijssen	Université Catholique de Louvain, Belgium
Andreas Nuernberger	Otto-von-Guericke University of Magdeburg, Germany
Kaustubh Raosaheb Patil	Massachusetts Institute of Technology, USA
Mykola Pechenizkiy	Eindhoven University of Technology, The Netherlands
Jose-Maria Pena	Universidad Politécnica de Madrid, Spain
Ruggero G. Pensa	University of Torino, Italy
Marc Plantevit	LIRIS, Université Claude Bernard Lyon 1, France
Lubos Popelinsky	Masaryk University, Czech Republic
Eric Postma	Tilburg University, The Netherlands
Miguel A. Prada	Universidad de Leon, Spain
Ronaldo Prati	Universidade Federal do ABC, UFABC, Brazil
Peter van der Putten	Leiden University and Pegasystems, The Netherlands
Jesse Read	École Polytechnique, France
Antonio Salmeron	University of Almería, Spain
Vítor Santos Costa	University of Porto, Portugal

Christin Seifert	University of Twente, The Netherlands
Roberta Siciliano	University of Naples Federico II, Italy
Jerzy Stefanowski	Poznań University of Technology, Poland
Frank Takes	Leiden University and University of Amsterdam, The Netherlands
Maguelonne Teisseire	Irstea, UMR Tetis, France
Ljupco Todorovski	University of Ljubljana, Slovenia
Melissa Turcotte	LANL, USA
Cor Veenman	Netherlands Forensic Institute, The Netherlands
Veronica Vinciotti	Brunel University, UK
Filip Zelezny	Czech Technical University, Czech Republic
Leishi Zhang	Middlesex University, UK

Contents

Estimating Uncertainty in Deep Learning for Reporting Confidence: An Application on Cell Type Prediction in Testes Based on Proteomics

Biraja Ghoshal[1(✉)], Cecilia Lindskog[2], and Allan Tucker[1]

[1] Brunel University London, Uxbridge UB8 3PH, UK
biraja.ghoshal@brunel.ac.uk
[2] Department of Immunology, Genetics and Pathology, Rudbeck Laboratory, Uppsala University, 75185 Uppsala, Sweden
https://www.brunel.ac.uk/computer-science

Abstract. Multi-label classification in deep learning is a practical yet challenging task, because class overlaps in the feature space means that each instance is associated with multiple class labels. This requires a prediction of more than one class category for each input instance. To the best of our knowledge, this is the first deep learning study which quantifies uncertainty and model interpretability in multi-label classification; as well as applying it to the problem of recognising proteins expressed in cell types in testes based on immunohistochemically stained images. Multi-label classification is achieved by thresholding the class probabilities, with the optimal thresholds adaptively determined by a grid search scheme based on Matthews correlation coefficients. We adopt MC-Dropweights to approximate Bayesian Inference in multi-label classification to evaluate the usefulness of estimating uncertainty with predictive score to avoid overconfident, incorrect predictions in decision making. Our experimental results show that the MC-Dropweights visibly improve the performance to estimate uncertainty compared to state of the art approaches.

Keywords: Uncertainty estimation · Multi-label classification · Cell type prediction · Human Protein Atlas · Proteomics

1 Introduction

Proteins are the essential building blocks of life, and resolving the spatial distribution of all human proteins at an organ, tissue, cellular, and subcellular level greatly improves our understanding of human biology in health and disease. The testes is one of the most complex organs in the human body [15]. The spermatogenesis process results in the testes containing the most tissue-specific genes than elsewhere in the human body. Based on an integrated 'omics' approach using transcriptomics and antibody-based proteomics, more than 500 proteins with distinct testicular protein expression patterns have previously been identified [10], and transcriptomics data suggests that over 2,000 genes are elevated

M. R. Berthold et al. (Eds.): IDA 2020, LNCS 12080, pp. 223–234, 2020.
https://doi.org/10.1007/978-3-030-44584-3_18

in testes compared to other organs. The function of a large proportion of these proteins are however largely unknown, and all genes involved in the complex process of spermatogenesis are yet to be characterized. Manual annotation provides the standard for scoring immunohistochemical staining pattern in different cell types. However, it is tedious, time-consuming and expensive as well as subject to human error as it is sometimes challenging to separate cell types by the human eye. It would be extremely valuable to develop an automated algorithm that can recognise the various cell types in testes based on antibody-based proteomics images while providing information on which proteins are expressed by that cell type [10]. This is, therefore, a multi-label image classification problem.

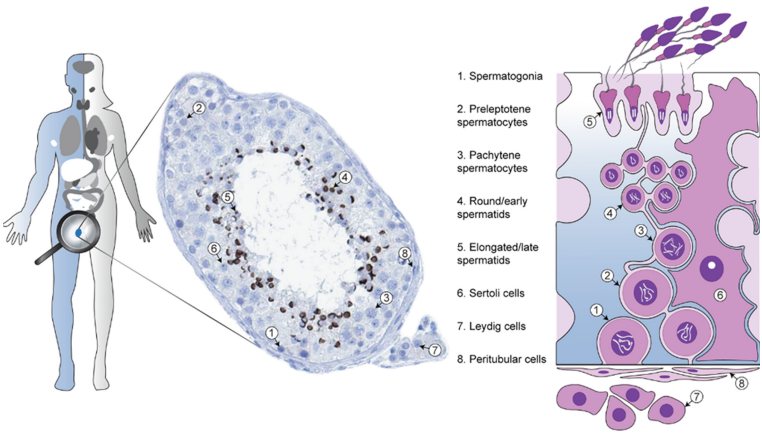


Fig. 1. Schematic overview: cell type-specific expression of testis elevated genes [10]

Exact Bayesian inference with deep neural networks is computationally intractable. There are many methods proposed for quantifying uncertainty or confidence estimates. Recently Gal [5] proved that a dropout neural network, a well-known regularisation technique [13], is equivalent to a specific variational approximation in Bayesian neural networks. Uncertainty estimates can be obtained by training a network with dropout and then taking Monte Carlo (MC) samples of the prediction using dropout during test time. Following Gal [5], Ghoshal et al. [7] also showed similar results for neural networks with Dropweights and Teye [14] with batch normalisation layers in training (Fig. 1).

In this paper, we aim to:

1. Present the first approach in multi-label pattern recognition that can recognise various cell types-specific protein expression patterns in testes based on antibody-based proteomics images and provide information on which cell types express the protein with estimated uncertainty.
2. Show Multi-Label Classification (MLC) is achieved by thresholding the class probabilities, with the Optimal Thresholds adaptively determined by a grid search scheme based on Matthews correlation coefficient.

3. Demonstrate through extensive experimental results that a Deep Learning Model with MC-Dropweights [7] is significantly better than a wide spectrum of MLC algorithms such as Binary Relevance (BR), Classifier Chain (CC), Probabilistic Classifier Chain (PCC) and Condensed Filter Tree (CFT), Cost-sensitive Label Embedding with Multidimensional Scaling (CLEMS) and state-of-the-art MC-Dropout [5] algorithms across various cell types.
4. Develop Saliency Maps in order to increase model interpretability visualizing descriptive regions and highlighting pixels from different areas in the input image. Deep learning models are often accused of being "black boxes", so they need to be precise, interpretable, and uncertainty in predictions must be well understood.

Our objective is not to achieve state-of-the-art performance on these problems, but rather to evaluate the usefulness of estimating uncertainty leveraging MC-Dropweights with predictive score in multi-label classification to avoid over-confident, incorrect predictions for decision making.

2 Multi-label Cell-Type Recognition and Localization with Estimated Uncertainty

2.1 Problem Definition

Given a set of training data D, where $X = \{x_1, x_2 \dots x_N\}$ is the set of N images and the corresponding labels $Y = \{y_1, y_2 \dots y_N\}$ is the cell-type information. The vector $y_i = \{y_{i,1}, y_{i,2} \dots y_{i,M}\}$ is a binary vector, where $y_{i,j} = 1$ indicates that the i^{th} image belongs to the j^{th} cell-type. Note that an image may belong to multiple cell-types, i.e., $1 <= \sum_j y_{i,j} <= M$. Based on $D(X, Y)$, we constructed a Bayesian Deep Learning model giving an output of the predictive probability with estimated uncertainty of a given image x_i belonging to each cell category. That is, the constructed model acts as a function such that $f : X \rightarrow Y$ using weights of neural net parameters ω where $(0 <= \hat{y}_{x,j} <= 1)$ as close as possible to the original function that has generated the outputs Y, output the estimated value $(\hat{y}_{i,1}, \hat{y}_{i,2}, \dots, \hat{y}_{i,M})$ as close to the actual value $(y_{i,1}, y_{i,2}, \dots, y_{i,M})$.

2.2 Solution Approach

We tailored Deep Convolutional Neural Network (DCNN) architectures for cell type detection and localisation by considering a large image capacity, binary-cross entropy loss, sigmoid activation, along with Dropweights in the fully connected layer and Batch Normalization formulation of propagating uncertainty in deep learning to estimate meaningful model uncertainty.

Multi-label Setup: There are multiple approaches to transform the multi-label classification into multiple single-label problems with the associated loss function [8]. In this study, we used immunohistochemically stained testes tissue consisting of 8 cell types corresponding to 512 testis elevated genes.

Therefore, we define a 8-dimensional class label vector $Y = \{y_1, y_2 \dots y_N\}$; $Y \in \{0, 1\}$, given 8 cell types. y_c indicates the presence with respect to according cell type expressing the protein in the image while an all-zero vector [0; 0; 0; 0; 0; 0; 0; 0] represents the "Absence" (no cell type expresses the protein in the scope of any of 8 categories).

Multi-label Classification Cost Function: The cost function for Multi-label Classification has to be different considering the fact that a prediction for a class is not mutually exclusive. So we selected the sigmoid function with the addition of binary cross-entropy.

Data Augmentation: We used Keras' image pre-processing package to apply affine transformations to the images, such as rotation, scaling, shearing, and translation during training and inference. This reduces the epistemic uncertainty during training, captures heteroscedastic aleatoric uncertainty during inference and overall improves the performance of models.

Multi-label Classification Algorithm: In Bayesian classification, the mean of the predictive posterior corresponds to the parameter point estimates, and the width of the posterior reflects the confidence of the predictions. The output of the network is an M-dimensional probability vector, where each dimension indicates how likely each cell type in a given image expresses the protein. The number of cell types that simultaneously express the protein in an image varies. One method to solve this multi-label classification problem is placing thresholds on each dimension. However different dimensions may be associated with different thresholds. If the value of the i^{th} dimension of $\hat{y}$ is greater than a threshold, we can say that the i-th cell-type is expressed in the given tissue. The main problem is defining the threshold for each class label.

A threshold based on Matthews Correlation Coefficient (MCC) is used on the model outcome to determine the predicted class to improve the accuracy of the models.

We adopted a grid search scheme based on Matthews Correlation Coefficients (MCC) to estimate the optimal thresholds for each cell type-specific protein expression [2]. Details of the optimal threshold finding algorithm is shown in Algorithm 1.

The idea is to estimate the threshold for each cell category in an image separately. We convert the predicted probability vector with the estimated threshold into binary and calculate the Matthews correlation coefficient (MCC) between the threshold value and the actual value. The Matthews correlation coefficient for all thresholds are stored in the vector ω, from which we find the index of threshold that causes the largest correlation. The Optimal Threshold for the i^{th} dimension is then determined by the corresponding value. We then leveraged Bias-Corrected Uncertainty quantification method [6] using Deep Convolutional Neural Network (DCNN) architectures with Dropweights [7].

Input: Ground Truth Vector: $\{y_{i,1}, y_{i,2}, \ldots, y_{i,M}\}$;
Estimated Probability Vector: $\{\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,M}\}$;
Upper Bound for threshold = Ω, and Threshold Stride = S
Result: The Optimal Thresholds T = $(ot_1, ot_2, \ldots, ot_M)$
Initialization: The set of threshold T = $(ot_1 = 0, ot_2 = 0, \ldots, ot_M = 0)$;
for $i \leftarrow 1$ ***to*** M **do**
  $j \leftarrow 0$;
  $\omega \leftarrow 0$;
  $\pi \leftarrow 0$;
  for $j < \Omega$ **do**
    Initialize M-dimensional binary vector $\mathbf{v} \leftarrow (v_1 = 0, v_2 = 0, \ldots, v_M = 0)$;
    if $\hat{y}_i > j$ **then**
      $v_i \leftarrow 1$;
    end
    else
      $v_i \leftarrow 0$;
    end
    $\omega \leftarrow \omega.append(MCC(\mathbf{y}[1:i], v))$;
    $\pi = \pi.append(j)$;
    $j = j + S$
  end
  $\hat{m} \leftarrow argmax_m \omega = (\omega_1, \omega_2, \ldots, \omega_m, \ldots)$;
  $ot_i = \pi[\hat{m}]$
end

Algorithm 1. Find Optimal Threshold

Network Architecture: Our models are trained and evaluated using Keras with Tensorflow backend. For the DNN architecture, we used a generic building block containing the following model structure: Conv-Relu-BatchNorm-MaxPool-Conv-Relu-BatchNorm-MaxPool-Dense-Relu-Dropweights and Dense-Relu-Dropweights-Dense-Sigmoid, with 32 convolution kernels, 3×3 kernel size, 2×2 pooling, dense layer with 512 units, 128 units, and 8 feed-forward Dropweights probabilities 0.3. We optimised the model using Adam optimizer with the default learning rate of 0.001. The training process was conducted in 1000 epochs, with mini-batch size 32. We repeated our experiments three times for an algorithm and calculated a mean of the results.

3 Estimating Bias-Corrected Uncertainty Using Jackknife Resampling Method

3.1 Bayesian Deep Learning and Estimating Uncertainty

There are many measures to estimate uncertainty such as softmax variance, expected entropy, mutual information, predictive entropy and averaging predictions over multiple models. In supervised learning, information gain, i.e. mutual

information between the input data and the model parameters is considered as the most relevant measure of the epistemic uncertainty [4,12]. Estimation of entropy from the finite set of data suffers from a severe downward bias when the data is under-sampled. Even small biases can result in significant inaccuracies when estimating entropy [9]. We leveraged Jackknife resampling method to calculate bias-corrected entropy [11].

Given a set of training data D, where $\mathbf{X} = \{x_1, x_2 \dots x_N\}$ is the set of N images and the corresponding labels $\mathbf{Y} = \{y_1, y_2 \dots y_N\}$, a BNN is defined in terms of a prior $p(\omega)$ on the weights, as well as the likelihood $p(D|\omega)$. Consider class probabilities $p(y_{x_i} = c \mid x_i, \omega_t, D)$ with $\omega_t \sim q(\omega \mid D)$ with $\mathcal{W} = (\omega_t)_{t=1}^{T}$, a set of independent and identically distributed (i.i.d.) samples draws from $q(\omega \mid, D)$. The below procedure computes the Monte Carlo (MC) estimate of the posterior predictive distribution, its Entropy and Mutual Information(MI):

$$\sum_{i=1}^{N} \mathbb{I}_{\mathrm{MC}}(y_i; \omega \mid x_i, D) = \mathbb{H}\big(\hat{p}(y_i \mid x_i, D)\big) - \frac{1}{|\mathcal{W}|} \sum_{\omega \in \mathcal{W}} \mathbb{H}\big(p(y_i \mid x_i, \omega, D)\big)\,. \quad (1)$$

where

$$\hat{p}(y_i \mid x_i, D) = \frac{1}{|\mathcal{W}|} \sum_{\omega \in \mathcal{W}} p(y_i \mid x_i, \omega, D)\,. \quad (2)$$

The stochastic predictive entropy is $H[y \mid x, \omega] = \mathbb{H}(\hat{p}) = -\sum_c \hat{p}_c \log(\hat{p}_c)$, where $\hat{p}_c = \frac{1}{T}\sum_t p_{tc}$ is the entire sample maximum likelihood estimator of probabilities.

The first term in the MC estimate of the mutual information is called the plug-in estimator of the entropy. It has long been known that the plug-in estimator underestimates the true entropy and plug-in estimate is biased [11,17].

A classic method for correcting the bias is the Jackknife resampling method [3]. In order to solve the bias problem, we propose a Jackknife estimator to estimate the epistemic uncertainty to improve an entropy-based estimation model. Unlike MC-Dropout, it does not assume constant variance. If $\mathcal{D}(X, Y)$ is the observed random sample, the i^{th} Jackknife sample, x_i, is the subset of the sample that leaves-one-out observation $x_i : x_{(i)} = (x_1, \dots x_{i-1}, x_{i+1} \dots x_n)$. For sample size N, the Jackknife standard error $\hat{\sigma}$ is defined as: $\sqrt{\frac{(N-1)}{N}\sum_{i=1}^{N}(\hat{\sigma}_i - \hat{\sigma}_{(\odot)})^2}$, where $\hat{\sigma}_{(\odot)}$ is the empirical average of the Jackknife replicates: $\frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_{(i)}$. Here, the Jackknife estimator is an unbiased estimator of the variance of the sample mean. The Jackknife correction of a plug-in estimator $\mathbb{H}(\cdot)$ is computed according to the method below [3]:

Given a sample $(p_t)_{t=1}^{T}$ with p_t discrete distribution on $1...C$ classes, T corresponds to the total number of MC-Dropweights forward passes during the test.

1. for each $t = 1...T$
   - calculate the leave-one-out estimator: $\hat{p}_c^{-t} = \frac{1}{T-1}\sum_{j \neq i} p_{jc}$
   - calculate the plug-in entropy estimate: $\hat{H}_{-t} = \mathbb{H}(\hat{p}^{-t})$
2. calculate the bias-corrected entropy $\hat{H}_J = T\hat{H} + \frac{(T-1)}{T}\sum_{t=1}^{T}\hat{H}_{(-i)}$, where $\hat{H}_{(-i)}$ is the observed entropy based on a sub-sample in which the ith individual is removed.

We leveraged the following relation:

$$\mu_{-i} = \frac{1}{T-1}\sum_{j\neq i} x_j = \mu + \frac{\mu - x_i}{T-1}\,.$$

while resolving the i-th data point out of the sample mean $\mu = \frac{1}{T}\sum_i x_i$ and recompute the mean μ_{-i}. This makes it possible to quickly calculate leave-one-out estimators of a discrete probability distribution.

The epistemic uncertainty can be obtained as the difference between the approximate predictive posterior entropy (or total entropy) and the average uncertainty in predictions (i.e: aleatoric entropy):

$$I(\mathbf{y}:\omega) = H_e(\mathbf{y}|\mathbf{x}) = \hat{H}_J(\mathbf{y}|\mathbf{x}) - H_a(\mathbf{y}|\mathbf{x}) = \hat{H}_J(\mathbf{y}|\mathbf{x}) - \mathbb{E}_{q(\omega|\mathbf{D})}[\hat{H}_J(\mathbf{y}|\mathbf{x},\omega)]$$

Therefore, the mutual information $I(\mathbf{y}:\omega)$ i.e. as a measure of bias-corrected epistemic uncertainty, represents the variability in the predictions made by the neural network weight configurations drawn from approximate posteriors. It derives an estimate of the finite sample bias from the leave-one-out estimators of the entropy and reduces bias considerably down to $O(n^{-2})$ [3].

The bias-corrected uncertainty estimation model explains regions of ambiguous data space or difficult to classify, as data distribution with noise in the inputs or model, which was trained with different domain data. Consequently, these inputs should be assigned a higher aleatoric uncertainty. As a result, we can expect high model uncertainty in these regions.

Following Gal [5], we define the stochastic versions of Bayesian uncertainty using MC-Dropweights, where the class probabilities $p(y_{x_i} = c \mid x_i, \omega_t, D)$ with $\omega_t \sim q(\omega \mid D)$ and $\mathcal{W} = (\omega_t)_{t=1}^{T}$ along with a set of independent and identically distributed (i.i.d.) samples drawn from $q(\omega \mid, D)$, can be approximated by the average over the MC-Dropweights forward pass.

We trained the multi-label classification network with all eight classes. We dichotomised the network outputs using optimal threshold with Algorithm 1 for each cell type, with a 1000 MC-Dropweights forward passes at test time. In these detection tasks, $p(y_{x_i} >= 0; OptimalThreshold_i \mid x_i, \omega_t, D)$, where 1 marks the presence of cell type, is sufficient to indicate the most likely decision along with estimated uncertainty.

3.2 Dataset

Our main dataset is taken from The Human Protein Atlas project, that maps the distribution of all human proteins in human tissues and organs [15]. Here, we used high-resolution digital images of immunohistochemically stained testes tissue consisting of 8 cell types: spermatogonia, preleptotene spermatocytes, pachytene spermatocytes, round/early spermatids, elongated/late spermatids, sertoli cells, leydig cells, and peritubular cells, publicly available on the Human Protein Atlas version 18 (v18.proteinatlas.org), as shown in Fig. 2:

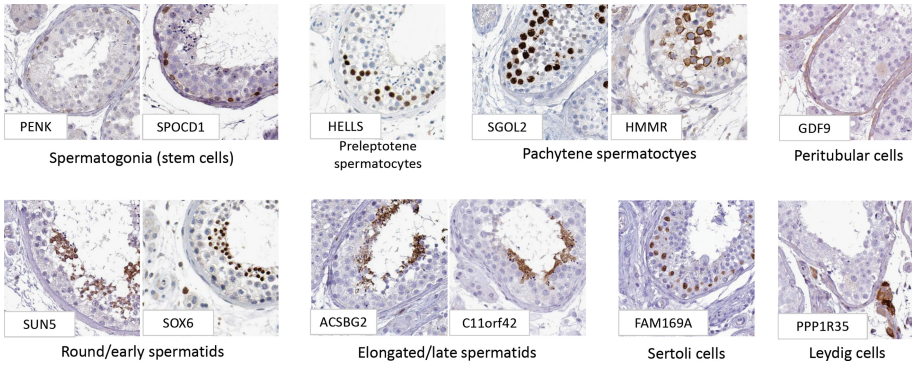


Fig. 2. Examples of proteins expressed only in one cell-type [10]

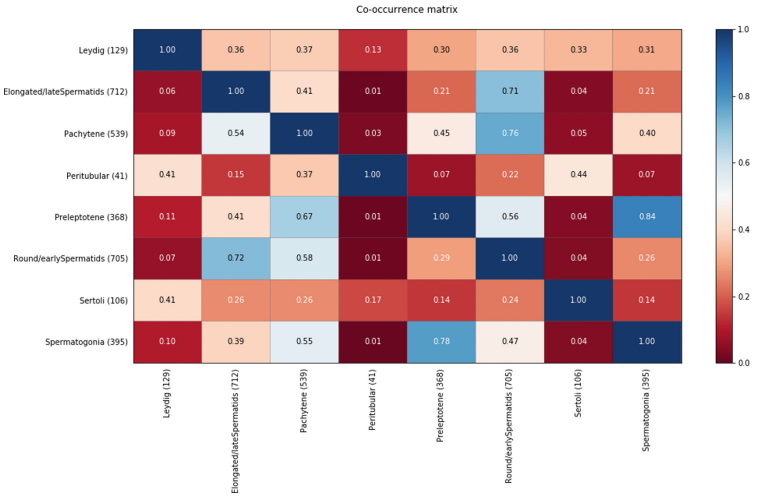


Fig. 3. Annotated heatmap of a correlation matrix between cell types

A relationship was observed between spermatogonia and preleptotene spermatocytes cell types and between round/early spermatids and elongated/late spermatids cell types along with Pachytene spermatocytes cells. Figure 3 illustrates the correlation coefficients between cell types. The observable pattern is that very few cell types are strongly correlated with each other.

3.3 Results and Discussions

We conducted the experiments on Human Protein Atlas datasets to validate the proposed algorithm, MC-Dropweights in Multi-Label Classification.

Multi-label Classification Model Performance: Model evaluation metrics for multi-label classification are different from those used in multi-class (or binary) classification. The performance metrics of multi-label classifiers can be classified as label-based (i.e.: it is assumed that labels are mutually exclusive) and example-based [16]. In this work, example-based measures (Accuracy score, Hamming-loss, F1-Score) and Rank-Loss are used to evaluate the performance of the classifiers.

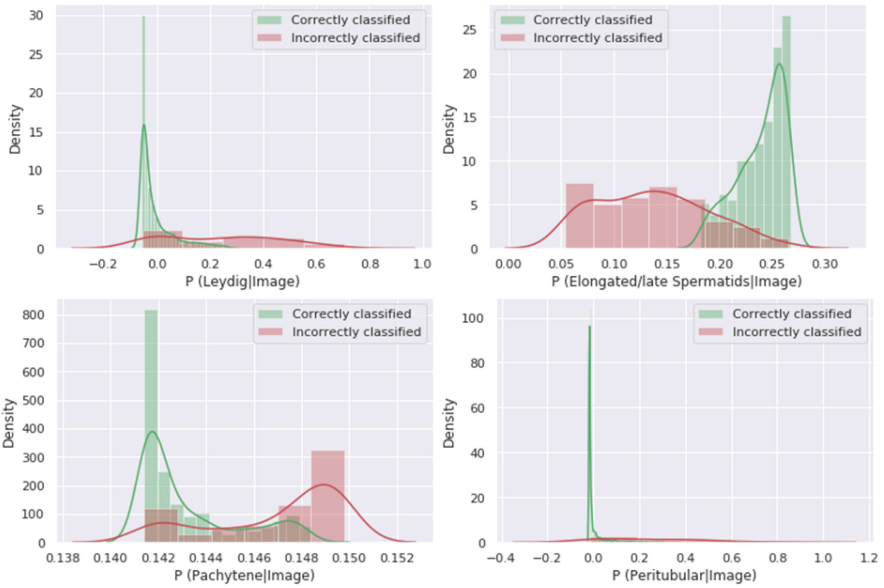
Table 1. Performance metrics

%Metrics	BR	CC	PCC	CFT	CLEMS	MC-Dropout	MC-Dropweights
Hamming loss	0.2445	0.2420	0.2420	0.2375	0.2370	0.207	0.1925
Rank loss	3.6700	3.5740	3.1580	3.2920	3.1120	2.862	2.626
F1 score	0.5038	0.5184	0.5733	0.5373	0.5902	0.6306	0.6627
Avg. accuracy score	0.4236	0.4389	0.4643	0.4573	0.5052	0.6150	0.7067

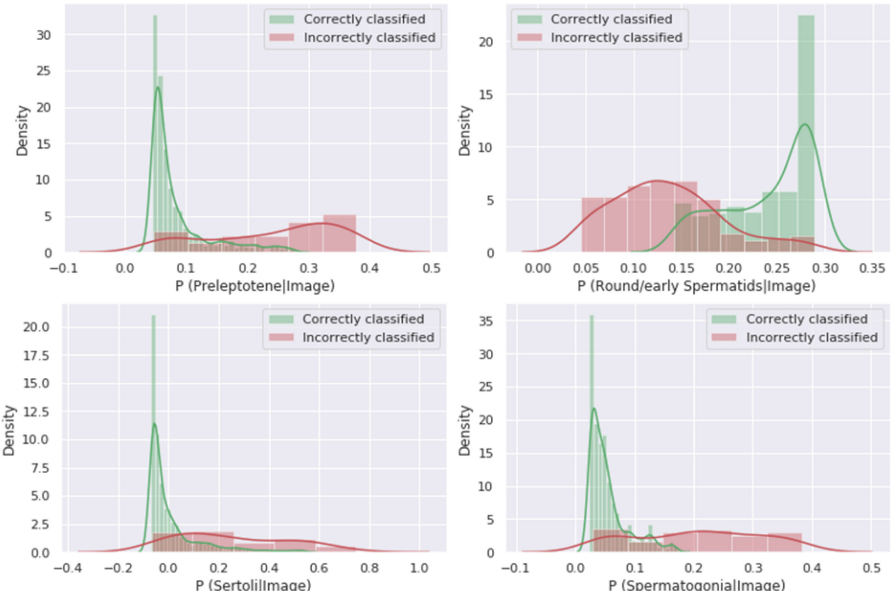
In the first experiment, we compared the MC-Dropweights neural network-based method with five machine learning MLC algorithms introduced in Sect. 1: binary relevance (BR), Classifier Chain (CC), Probabilistic Classifier Chain (PCC) and Condensed Filter Tree (CFT), Cost-Sensitive Label Embedding with Multi-dimensional Scaling (CLEMS) and the MC-Dropout neural network model. Table 1 shows that MC-Dropweights exhibits considerably better performance overall the algorithms, which demonstrates the importance of considering the Dropweights in the neural network.

Cell Type-Specific Predictive Uncertainty: The relationship between uncertainty and predictive accuracy grouped by correct and incorrect predictions is shown in Fig. 4. It is interesting to note that, on average, the highest uncertainty is associated with Elongated/late Spermatids and Round/early Spermatids. This indicates that there is some feature which contributes greater uncertainty to the Spermatids class types than to the other cell types.

Cell Type Localization: Estimated uncertainty with Saliency Mapping is a simple technique to uncover discriminative image regions that strongly influence the network prediction in identifying a specific class label in the image. It highlights the most influential features in the image space that affect the predictions of the model [1] and visualises the contributions of individual pixels to epistemic and aleatoric uncertainties separately. We calculated the class activation maps (CAM) [18] using the activations of the fully connected layer and the weights from the prediction layer as shown in Fig. 5.



(a) Leydig, Elongated/Late Spermatids, Pachytene, Peritubular Cell Type



(b) Preleptotene, Round/Early Spermatids, Sertoli, Spermatogonia Cell-Type

Fig. 4. Distribution of uncertainty values for all protein images, grouped by correct and incorrect predictions. Label assignment was based on optimal thresholding (Algorithm 1). For an incorrect prediction, there is a strong likelihood that the predictive uncertainty is also high in all cases except for Spermatids.

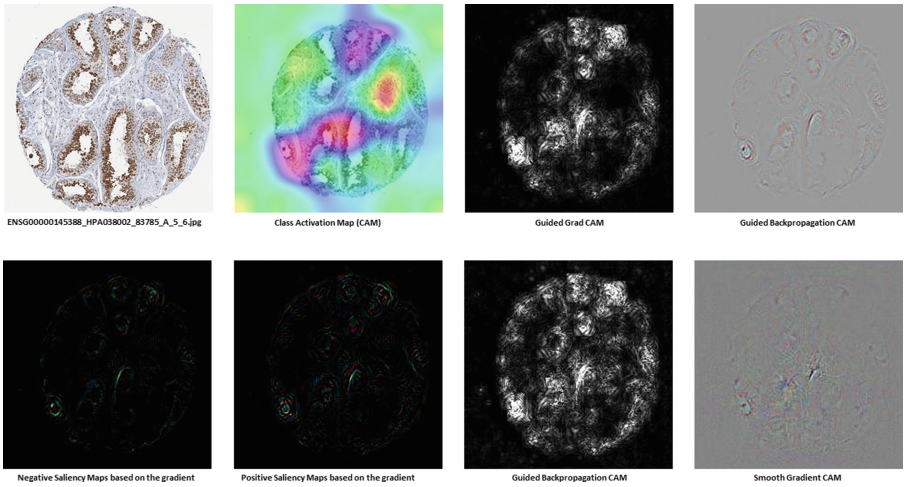


Fig. 5. Saliency maps for some common methods towards model explanation

4 Conclusion and Discussion

In this study, a multi-label classification method was developed using deep learning architecture with Dropweights for the purposes of predicting cell types-specific protein expression with estimated uncertainty, which can increase the ability to interpret, with confidence and make models based on deep learning more applicable in practice. The results show that a Deep Learning Model with MC-Dropweights yields the best performance among all popular classifiers.

Building truly large-scale, fully-automated, high precision, very high dimensional, image analysis system that can recognise various cell type-specific protein expression, specifically for Elongated/Late Spermatids and Round/early Spermatids remains a strenuous task. The properties in the dataset such as label correlations, label cardinality can strongly affect the uncertainty quantification in predictive probability performance of a Bayesian Deep learning algorithm in multi-label settings. There is no systematic study on how and why the performance varies over different data properties; any such study would be of great benefit in progressing multi-label algorithms.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems, pp. 9505–9515 (2018)
2. Chu, W.T., Guo, H.J.: Movie genre classification based on poster images with deep neural networks. In: Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes, pp. 39–45. ACM (2017)
3. DasGupta, A.: Asymptotic Theory of Statistics and Probability. Springer, New York (2008). https://doi.org/10.1007/978-0-387-75971-5
4. Depeweg, S., Hernández-Lobato, J.M., Doshi-Velez, F., Udluft, S.: Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. arXiv preprint arXiv:1710.07283 (2017)

5. Gal, Y.: Uncertainty in deep learning. Ph.D. thesis, University of Cambridge (2016)
6. Ghoshal, B., Tucker, A., Sanghera, B., Wong, W.: Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. In: Computational Intelligence - Special Issue on Foundations of Biomedical (Big) Data Science, vol. 1 (2019)
7. Ghoshal, B., Tucker, A., Sanghera, B., Wong, W.: Estimating uncertainty in deep learning for reporting confidence to clinicians when segmenting nuclei image data. 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), vol. 1, pp. 318–324, June 2019. https://doi.org/10.1109/CBMS.2019.00072
8. Huang, K.H., Lin, H.T.: Cost-sensitive label embedding for multi-label classification. Mach. Learn. **106**(9–10), 1725–1746 (2017)
9. Macke, J., Murray, I., Latham, P.: Estimation bias in maximum entropy models. Entropy **15**(8), 3109–3129 (2013)
10. Pineau, C., et al.: Cell type-specific expression of testis elevated genes based on transcriptomics and antibody-based proteomics. J. Proteome Res. **18**, 4215–4230 (2019)
11. Quenouille, M.H.: Notes on bias in estimation. Biometrika **43**(3/4), 353–360 (1956)
12. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
13. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal Mach. Learn. Res. **15**(1), 1929–1958 (2014)
14. Teye, M., Azizpour, H., Smith, K.: Bayesian uncertainty estimation for batch normalized deep networks. arXiv preprint arXiv:1802.06455 (2018)
15. Uhlén, M., et al.: Tissue-based map of the human proteome. Science **347**(6220), 1260419 (2015)
16. Wu, X.Z., Zhou, Z.H.: A unified view of multi-label performance measures. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 3780–3788. JMLR. org (2017)
17. Yeung, R.W.: A new outlook on Shannon's information measures. IEEE Trans. Inf. Theory **37**(3), 466–474 (1991)
18. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)