Published in final edited form as:
J Biomed Inform. 2024 January ; 149: 104576. doi:10.1016/j.jbi.2023.104576.

Deep learning uncertainty quantification for clinical text classification☆

Alina Peluso[a,*], **Ioana Danciu**[a], **Hong-Jun Yoon**[a], **Jamaludin Mohd Yusof**[b], **Tanmoy Bhattacharya**[b], **Adam Spannaus**[a], **Noah Schaefferkoetter**[a], **Eric B. Durbin**[c], **Xiao-Cheng Wu**[d], **Antoinette Stroup**[e], **Jennifer Doherty**[f], **Stephen Schwartz**[g], **Charles Wiggins**[h], **Linda Coyle**[i], **Lynne Penberthy**[j], **Georgia D. Tourassi**[a], **Shang Gao**[a]

[a]Oak Ridge National Laboratory, Oak Ridge, TN 37830, United States

[b]Los Alamos National Laboratory, Los Alamos, NM 87545, United States

[c]University of Kentucky, Lexington, KY 40536, United States

[d]Louisiana State University, New Orleans, LA 70112, United States

[e]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08901, United States

[f]University of Utah, Salt Lake City, UT 84132, United States

[g]Fred Hutchinson Cancer Research Center, Seattle, WA 98109, United States

[h]University of New Mexico, Albuquerque, NM 87131, United States

[i]Information Management Services Inc., Calverton, MD 20705, United States

[j]National Cancer Institute, Bethesda, MD 20814, United States

Abstract

Introduction: Machine learning algorithms are expected to work side-by-side with humans in decision-making pipelines. Thus, the ability of classifiers to make reliable decisions is of paramount importance. Deep neural networks (DNNs) represent the state-of-the-art models to

☆Notice: This manuscript has been authored in part by UT-Battelle LLC under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The publisher acknowledges the US government license to provide public access under the DOE Public Access Plan (http://energy.gov/downloads/doe-public-accessplan).

[*]Corresponding author. pelusoa@ornl.gov (A. Peluso).

CRediT authorship contribution statement
Alina Peluso: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Ioana Danciu:** Conceptualization, Investigation, Resources, Writing – review & editing. **Hong-Jun Yoon:** Methodology. **Jamaludin Mohd Yusof:** Validation, Writing – review & editing. **Tanmoy Bhattacharya:** Validation, Writing – review & editing. **Adam Spannaus:** Visualization. **Noah Schaefferkoetter:** Data curation. **Eric B. Durbin:** Data curation. **Xiao-Cheng Wu:** Data curation. **Antoinette Stroup:** Data curation. **Jennifer Doherty:** Data curation. **Stephen Schwartz:** Data curation. **Charles Wiggins:** Data curation. **Linda Coyle:** Data curation. **Lynne Penberthy:** Funding acquisition. **Georgia D. Tourassi:** Funding acquisition. **Shang Gao:** Conceptualization, Investigation, Resources, Writing – review & editing.

Declaration of competing interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary information
Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104576.

address real-world classification. Although the strength of activation in DNNs is often correlated with the network’s confidence, in-depth analyses are needed to establish whether they are well calibrated.

Method: In this paper, we demonstrate the use of DNN-based classification tools to benefit cancer registries by automating information extraction of disease at diagnosis and at surgery from electronic text pathology reports from the US National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) population-based cancer registries. In particular, we introduce multiple methods for selective classification to achieve a target level of accuracy on multiple classification tasks while minimizing the rejection amount—that is, the number of electronic pathology reports for which the model’s predictions are unreliable. We evaluate the proposed methods by comparing our approach with the current in-house deep learning-based abstaining classifier.

Results: Overall, all the proposed selective classification methods effectively allow for achieving the targeted level of accuracy or higher in a trade-off analysis aimed to minimize the rejection rate. On *in-distribution* validation and holdout test data, with all the proposed methods, we achieve on all tasks the required target level of accuracy with a lower rejection rate than the deep abstaining classifier (DAC). Interpreting the results for the *out-of-distribution* test data is more complex; nevertheless, in this case as well, the rejection rate from the best among the proposed methods achieving 97% accuracy or higher is lower than the rejection rate based on the DAC.

Conclusions: We show that although both approaches can flag those samples that should be manually reviewed and labeled by human annotators, the newly proposed methods retain a larger fraction and do so without retraining—thus offering a reduced computational cost compared with the in-house deep learning-based abstaining classifier.

Keywords

Selective classification; Deep learning; Abstaining classifier; Text classification; Uncertainty quantification; Accuracy; DNN; CNN; HiSAN; Pathology reports; NCI SEER

1. Introduction

Cancer is a major threat to human lives: large morbidity rates of about 1.9 million new cancer diagnoses and over 600,000 cancer deaths are recorded in the United States per year. Therefore, surveillance of cancer incidence is essential for monitoring public health. The task, however, requires manual coding and review of clinical documents, and the associated time and monetary costs make it impossible to perform such tasks in real time on a large scale. Electronic pathology reports are an essential data source used by National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) program–sponsored population-based registries to document the diagnosis of cancer. Cancer registrars in the United States are responsible for collecting cancer incidence data—including extracting and classifying information about diagnosis, treatment modalities, and survival data from unstructured notes and reports in a hybrid, AI-assisted manner. The vital records registries with the state health departments are responsible for death data, including cancer deaths.

The natural language processing (NLP) field for cancer applications is still dominated by rule-based systems. Deep neural networks (DNNs) have been very successful at addressing many real-world classification problems [1,2]. As a result, DNN-based classification tools are being deployed in situations where their decisions impact everyday life. The multi-task (MT) convolutional neural network (CNN) and more advanced models have shown great performance in information extraction of disease classification from electronic cancer pathology reports [3–5]. In this context, the ability of AI classifiers to make reliable decisions is critically important. As new architectures and components are introduced, research has focused on improving the accuracy and speed of networks. Yet far less attention has been given to determining when predictions can be trusted and when they cannot. In other words, neural networks are effective at providing output that is correct most of the time but, importantly, are less effective at identifying the extent to which the output can be trusted, including when they amount to little more than educated guesses [6]. When the accuracy of the model is imperfect, a human reviewer is required to verify that the AI classifications are correct, reducing the cost savings meant to be realized through use of AI-based tools [7]. Even if such a model is used only to assist manual classification, we face the problem that inaccurate second opinions may distract – or worse, bias – human registrars and degrade their performance. Therefore, accurate uncertainty quantification (UQ) is vital to developing trust in an AI-based model. In particular, with calibrated UQ, we can trust a machine learning (ML) model's high-confidence decisions while minimizing human labor.

The rest of the paper is organized as follows. In Section 2, the context of our work within recent literature is provided. In Section 3, we discuss the dataset for this study – namely, the NCI SEER cancer pathology reports – and the two different model architectures we test with our rejection methods. Section 4 details the findings of our experiments on the different model architectures and rejection methods compared with the baseline model; experimental findings on out-ofdistribution (OOD) data are also provided. In Section 5, we discuss the methodological contributions and potential generalizability for informatics problems. Lastly, we present conclusions and provide directions for future work in Section 6.

2. Background

2.1. Related UQ work

A well-calibrated DNN model should demonstrate good confidence in its predictions, such that it is accurate and indicates high uncertainty when producing inaccurate predictions, thereby making it reliable and easy to interpret. In spite of recent advances in probabilistic deep learning (DL) to improve model robustness, obtaining accurate quantification of uncertainty estimates from DNNs is still an open research problem [8]. Selective prediction [9] is closely related to confidence estimation, as well as OOD detection [10,11] and uncertainty calibration of DNNs [6,12]—that is, the accurate representation of predictive probabilities with respect to true likelihood, which is a challenging problem because of the unavailability of ground truth uncertainty estimates [8]. A distinction among these topics is that *calibration* focuses on adjusting the overall confidence level of a model, globally increasing or decreasing the model's confidence on all samples, whereas *selective prediction* is based on relative confidence among the samples: the rejection rate in selective prediction results from model uncertainty rather than modelagnostic data uncertainty [13]. Existing research to improve predictive uncertainty in DNNs for multiclass classification tasks can be broadly classified into three categories, which aim to (1) improve the model input by training the model with data augmentation; (2) improve the model parameters with Bayesian and non-Bayesian probabilistic methods; and (3) improve the model accuracy with trainable calibration measures or with post-processing confidence calibration.

Training the model with data augmentation—Methods aimed at improving the model input by training the model with data augmentation produce extra samples during training by augmenting the samples' labels. Although data augmentation methods, most notably Mixup [14] and AugMix [15], produce better-calibrated output, improve model robustness, and can be effective OOD detectors, it is difficult in practice to introduce a wide spectrum of perturbations and corruptions during training that comprehensively represent real-world deployment conditions, especially when the samples' label distribution is highly skewed and data are high-dimensional.

Bayesian and non-Bayesian probabilistic methods—Probabilistic methods aimed at improving the model parameters can be categorized as Bayesian and non-Bayesian.

Non-Bayesian probabilistic methods such as ensemble-based methods estimate confidence based on the statistics of the ensemble model's output. Most notably among this class of models, deep ensembles [16] propose training an ensemble of neural networks from different random initializations; while training, adversarial samples are generated to improve model robustness and provide calibrated confidence [17]. However, ensembles are computationally practical for small models only, as they introduce additional overhead associated with training multiple models and significant memory complexity during testing.

In contrast, probabilistic Bayesian methods assume a prior distribution over the deterministic parameters of the DNN and obtain confidence estimates through the posterior. Predictive uncertainty is estimated as probability distributions over the output label probabilities instead of a single scalar probability. Approximate Bayesian inference methods for DNNs have been proposed, as computing the true posterior is intractable—most notably in this class of methods are variational inference approaches [18], stochastic gradient variants of Markov Chain Monte Carlo [19], Monte Carlo dropout [20] and stochastic weight-averaging Gaussian (SWAG) [21]. Approximate Bayesian inference methods are promising as they are equivalent to using an ensemble for confidence estimation, but they require no actual training and storing of multiple models. Nevertheless, they may fail to provide calibrated uncertainty between separated regions of observations, as they tend to fit an approximation to a local mode and do not capture the complete true posterior [22], potentially causing the model to be overconfident under distributional shift.

Trainable confidence calibration—Trainable calibration methods are proposed to integrate model calibration into classification training. One of the earliest trainable approaches is entropy regularization [23]. The method proposes to use entropy as a regularization term in loss functions for model calibration. One disadvantage of entropy regularization is that the final classification loss depends on a very sensitive weight scalar.

[24] propose to express calibration error as a tractable integral probability measure, that is, the maximum mean calibration error (MMCE) computed in a reproducing kernel Hilbert space. MMCE is an accurate method to minimize calibration error metrics while maximally preserving the number of high-confidence predictions. [25] propose to add the difference between confidence and accuracy (DCA) as an auxiliary loss term to the cross-entropy loss for classification tasks. DCA estimates the expected calibration error by minimizing the difference between the predicted confidence and the neural networks' accuracy. A similar, albeit different class of trainable calibration methods introduces abstention specific cost into the loss function or learns to abstain so that the performance of the model reaches a specific target [26]. Earlier work by [27] focuses on learning with abstention for binary classification. Next, [28] developed a Structured Output Learning with Abstention (SOLA) framework to allow abstaining from predicting, thus increasing the reliability of model predictions. [29] proposed a deep abstaining classifier (DAC) that uses a DNN trained with an extra abstention class for detecting OOD and novel samples. The DAC allows for auto-tuning of a hyperparameter expressing the degree of penalty for abstention (see Section 2.1 of [29]) while also providing a separate abstention class that aids interpretability, as the features supporting abstention can be interrogated. Although all these trainable calibration strategies are effective, they require retraining for each desired confidence level.

Post-processing confidence calibration—Post-processing calibration includes temperature scaling [12] and similar approaches such as Dirichlet calibration [30]. Temperature scaling is a widely used calibration method, which treats model calibration as a post-processing task by scaling the output by a temperature parameter. Although the optimization process of the temperature parameter is inexpensive, this method globally increases or decreases the model's confidence on all samples, resulting in unchanged ranking of all samples' confidence; thus, the calibration on the independent and identically distributed validation dataset does not guarantee calibration under distributional shift.

In the same class of methods, *a–posteriori* confidence estimation, also known as *selective classification*, is done by pairing a standard classifier with a confidence estimator. The simplest approach considers the highest score from the probability distribution of the final output layer of DNN models as a proxy for predictive confidence. [31] provides empirical evidence that for DNN classifiers, in-distribution predictions do tend to have higher winning scores than OOD samples, thus empirically justifying the use of softmax thresholding as a useful baseline. Nevertheless, though the strength of activations in DNNs is often correlated with the network's confidence, in-depth analyses are needed to establish whether the network's outputs are well calibrated and can thus be employed to measure the uncertainty of their prediction. Even so, this post-processing optimization process is inexpensive in terms of time and cost of training, and it is also flexible in defining the uncertainty, or accepting or rejecting the model predictions at a certain threshold level. Moreover, effective confidence estimators can also be employed in active learning strategies in the context of uncertainty sampling [32]. Finally, this approach easily allows coupling uncertainty and explainability to align prediction probability with the actual accuracy in test data.

2.2. Contribution of this work

In the context of our study, a specific accuracy level – higher, in practice, than what can be achieved by the trained model for predictions on all data – must be met by the model employed to perform information extraction from electronic text pathology reports from the population-based cancer registries participating in the US NCI SEER program. Therefore, it is crucial to identify the reports in which the model is less confident (i.e., the rejection rate) so that its accuracy on the remaining retained set reaches the accuracy specified. This ideally requires confidence to be calibrated to the accuracy and the confidence score to be maximally discriminative—separating the correct predictions from the incorrect ones. With infinite training data, we expect the neural net to extract all relevant features from the input and for the prediction scores to converge to the Bayes risk, thus satisfying this optimality criterion. However, it is an empirical question whether better scores can be designed in particular applications. Therefore, we pair our classifier with four confidence estimators for UQ to meet the accuracy requirements for our models. We test the proposed methods (1) by evaluating the discriminating power as measured by the rejection rate while maintaining a minimum targeted accuracy on the retained predictions within our four target classification tasks, as well as (2) by comparing the rejection rate with the current in-house DL-based abstaining classifier that is designed to use a trainable DL calibration technology to extract features specifically associated with confidence [29]. We show that though both methods can flag the samples that require manual review, the newly proposed methods retain a larger fraction and can do so without retraining. Thus, the method proposed herein results in a reduced computational cost compared with that of the in-house DL-based abstaining classifier.

3. Materials and methods

3.1. Cancer pathology report data

The data for our information extraction task comprise electronic cancer pathology reports collected by the NCI SEER program from seven different cancer registries: California (CA), New Mexico (NM), Kentucky (KY), Louisiana (LA), New Jersey (NJ), Seattle (SA), and Utah (UT). Each new primary tumor diagnosis is assigned a unique tumor ID and pathology reports consisting of highly technical, partially structured text characterizing the tumor at the time of the diagnosis (i.e., biopsy pathology reports) or at the time of surgery (i.e., surgical pathology reports). Notably, the structure of the text varies somewhat among pathology laboratories and from registry to registry: standard structured analysis systems often do not generalize very well. The ground truth labels for these reports are manually annotated by certified tumor registrars at the tumor level in the Cancer/Tumor/Case (CTC) database, which stores all diagnostic, staging, and treatment information for reportable cancers in the SEER Data Management System (SEER*DMS). In particular, the pathologists write the free-form text and may have options for drop-down entry. The reports are then sent to registrars at a SEER registry, who use standardized reporting guidelines instituted by the SEER program[1] to code the written reports. The standardized International Classification of Diseases for Oncology (ICD-O-3) terminology is used for the four tasks of interest in

[1] https://seer.cancer.gov/tools/codingmanuals/.

this study: primary cancer site (Site, 70 CTC classes), primary anatomic subsite (Subsite, 324 CTC classes), laterality (Laterality, 7 CTC classes), and histological type (Histology, 620 CTC classes).[2] An example of CTC codes for primary anatomic subsite is shown in Supplementary Figure 1. Even though most pathology reports are associated with only a single tumor, the converse is not true (i.e., a single tumor may have more than one pathology report); therefore, the ground truth being available only at the tumor level means that any particular report may not have the information to decide on the correct classification, leading to an irreducible Bayes' error in the predictions of the trained model.

3.2. Deep learning models for text classification

This paper compares DNNs with differing combinations of architectures and activation functions. See Fig. 1 for a visual representation of the different DNN architectures that we consider. The first model considered is a TextCNN [1,33], which is one of the most successful and widely used CNN models for text classification. It consists of three parts: word embedding, 1D convolution, and a fully connected decision layer. *Word embedding* is a learned representation of terms that maps a set of words onto vectors of numerical representations with the same semantic meaning and similar observation. The 1D convolution layer has a series of 1D convolution filters that have latent representations to articulate the features in the word vectors of documents. The identified features are passed to the fully connected layer to make inferences. MTCNN [4] is an extension of TextCNN that applies the multi-task learning (MTL) mechanism [34] to the decision layer. A classifier learns multiple tasks simultaneously and finds an optimal latent representation to solve a series of related tasks. The MTL helps find more generalized solutions than single-task models, thus yielding higher task performance. The second architecture is a hierarchical self-attention network (MTHiSAN), which is the current state-of-the-art model for classifying electronic cancer pathology reports [3,35,36]. This DL architecture is composed of two hierarchies, each containing several self-attention layers. The word-level hierarchy takes in word embeddings and generates a line embedding representation for each line in the pathology report. Next, the line-level hierarchy utilizes line embeddings to generate a document embedding representation that can then be used for classification. A multi-task decision layer is then used to simultaneously classify on all relevant tasks for each input document. For this study, both the MTCNN and the MTHiSAN architectures are developed in a multilabel classification (MLC) setting aimed to simultaneously output a prediction across all existing classes.

The pathology report, the written description of a histological examination, is the most accurate method for diagnosing cancer. Examples of pathology reports can be found on the NCI SEER website.[3] SEER cancer registries collect demographic, tumor (including pathology reports), treatment, and outcome data for all cancer cases diagnosed within their catchment area. Tumor registrars abstract information from pathology and other medical reports to organize, summarize, and categorize information about each tumor. The manually coded categorical information is used for our training labels. The pathology report text

[2] The class counts reported here were obtained with consideration of the universe of CTC labels across all SEER registries; a lower count class is reported in our experiments, as we are considering samples from only a few registries out of all SEER registries.

[3] https://training.seer.cancer.gov/abstracting/procedures/pathological/histologic/operative/example/.

is free-form and generally messy. Therefore, we implemented the following cleaning procedures before training the model. Before being used in the neural networks, each electronic pathology report was subjected to a preprocessing step to minimize formatting inconsistencies across the dataset. This step consists of converting all text to lowercase, stripping all documents of any hex escapes and unicode characters, and replacing numerical values greater than 100 with either 'large_int_token' or 'large_float_token' depending on the format of the value being replaced. The resulting sequence of words in the processed documents is then tokenized to a sequence of integer indices, so that each word in the vocabulary is mapped to a unique integer value. So that reports can be used as input to each network, we reverse the word order of each and truncate to a uniform length of 3000 tokens. Mathematically, for documents shorter than 3000 words (99% of the documents are less than 3000 tokens), reversing word order has no effect on model performance: both MTCNN and MTHiSAN are immune to reversing word order if both train/val/test documents are reversed the same way. For documents longer than 3000 words, the last 3000 tokens are used because the final diagnosis is detailed at the end of the report and tends to have the most useful information. Next, the documents are padded with 0s, which map to the <pad> token, which MTCNN and MTHiSAN are designed to ignore. Lastly, the integer tokens are passed either through the convolutional and maxpool layers in the MTCNN or through the hierarchical word and line embedding layers in MTHiSAN to extract the features passed to either the softmax or sigmoid activation function. When a softmax activation function is used, the probabilities by class will always sum to one because they are modeled as a joint distribution. However, when a sigmoid activation function is used, the outcomes are modeled independently, and, therefore, the resulting probabilities are not constrained to sum to one. The assumption of independence of the outcome may play a key role in the decision to abstain.

The models are trained on aggregated data from five registries (LA + KY + UT + NJ + SA) and evaluated on the most recent (all reports after 2017) holdout data from these same five registries as well as samples from two OOD registries (CA and NM). The support of these sets is detailed in Table 1.[4] This study design is relevant in the context of our application, as new data from current registries are available every year, and collaboration with new registries is ongoing. The study was executed according to the Institutional Review Board (IRB) protocol DOE000619, approved by Central DOE Institutional Review Board on April 6, 2021 (initial approval on September 23, 2016).

3.3. Proposed UQ methods

We use a DAC [29] as the multitask abstaining classifier for classifying electronic text pathology reports. The DAC is a regular DNN but with an additional unlabeled abstention class and a modified version of the standard cross-entropy loss function required to allow abstention during the training. The DAC is trained such that if the model is not highly confident on a particular sample, it will predict the abstention class, indicating that the model should abstain on that sample. In addition, since the abstention is determined during

[4] Although we document the count of reports from each single registry, we do not explicitly account for this split in the model; therefore, we refer to the data as the LAKYUTNJSA set.

training, confusing examples tend to be down-weighted during training, leading to a better model. Further details regarding the loss function and the development of the DAC can be found in [37].

Nevertheless, the relative performance of a DAC to other UQ strategies is an empirical question that can depend on the problem instance. Therefore, this work investigates multiple selective classification strategies as a comparison to the DAC performance. Examination of more flexible strategies that require no model retraining upon changing the desired accuracy may reduce the computational costs of training the model and increase flexibility for downstream users. The training of DL models is typically the most computationally expensive part of the process, making retraining impractical for other practitioners with limited resources. Moreover, if new modeling strategies are flexible enough to adapt to OOD data, barriers of use in new datasets will be reduced.

Let us consider the output from our DL model, yielding the predicted probabilities $p(y_{ij}) = p(y_{11}), \ldots, p(y_{nk})$ for the ith pathology report (with $i = 1, \ldots, n$) of belonging to any one of the k classification labels (with $j = 1, \ldots, k$) for each classification task considered in the study. The strategies proposed below are all calculated at the pathology report level; therefore, to simplify the notation, we will drop the i index.

Fixed confidence score.—This is probably the most simplistic way to apply a thresholding criteria. No transformation is applied to the predicted probabilities $p(y_1), \ldots, p(y_k)$ and among these we select the highest one. We abstain on all reports with a highest predicted probability less than or equal to a threshold value in the interval [0, 1]. For perfect training, this probability should guarantee the accuracy on the retained set of reports; however, in practice, one may need to calibrate it using a validation set.

Delta difference confidence score.—This method is motivated by the expectation that when the model is confident in predicting, differences between the highest predicted probability and all the remaining probabilities should be large. In practice, such differences are well captured by the two highest predicted probabilities. Therefore, we compute a confidence score as the difference between the two top predicted probabilities. We do not retain samples with a confidence score less than or equal to a threshold value in the interval [0, 1], with the threshold being tuned to reach the target accuracy on the retained set.

Entropy ratio confidence score.—This method considers variations among all the predicted probabilities. First, we calculate the Shannon entropy $-\Sigma_{j=1}^{k} \; p(y_j)\log_2(p(y_j))$; this measure can be used as an adjustment factor because it informs on the confidence of the model's predictions (i.e., an entropy score close to 0 indicates that the model is confident in its prediction). Therefore, we compute the confidence score as the ratio between the highest predicted probability and the Shannon entropy. We abstain on samples with confidence scores less than or equal to a threshold value in the interval $\left[0, \frac{\max[p(y_1), \ldots, p(y_k)]}{-\sum_{j=1}^{k} p(y_j)\log_2(p(y_j))}\right]$, which again must be tuned.

Bayes beta confidence score.—This strategy is quite different from the methods presented above, as it relies on parametric assumptions necessary to describe the distribution of the correctly and incorrectly classified labels. Specifically, the confidence score is estimated as the conditional probability of a correct classification, which is based on the Bayes theorem for a binary variable, where being correct and incorrect are mutually exclusive outcomes:

$$p(\text{correct} \mid y_j) = \frac{p(y_j \mid \text{correct}) \cdot p(\text{correct})}{p(y_j \mid \text{correct}) \cdot p(\text{correct}) + p(y_j \mid \text{incorrect}) \cdot p(\text{incorrect})} .$$

The marginal probabilities $p(\text{correct})$ and $p(\text{incorrect})$ are called *priors* and are estimated as the corresponding relative frequencies: the total number of correct or incorrect decisions divided by the total number of cases. The conditional probabilities $p(y_j \mid \text{correct})$ and $p(y_j \mid \text{incorrect})$ are estimated from the data assuming $p(y_j \mid \text{correct}) \sim \text{Beta}(\alpha_{\text{correct}}, \gamma_{\text{correct}})$ and $p(y_j \mid \text{incorrect}) \sim \text{Beta}(\alpha_{\text{incorrect}}, \gamma_{\text{incorrect}})$, where the parameters α and γ are obtained via maximum likelihood estimation (MLE). We will abstain on samples with confidence scores less than or equal to a threshold value in the interval [0, 1]. Again, though this method is expected to provide guarantees on the accuracy depending on the threshold value, in real examples, one may need to calibrate the score using validation data.

The accuracy and the rejection rate are correlated, and we want to minimize rejections while achieving the desired accuracy. Therefore, we measure the performance of our models by comparing the rejection rate while maintaining the required accuracy level on the retained predictions within our four classification tasks. The overall accuracy is the chosen evaluation metric, as we are mainly interested in the count of correctly predicted labels (i.e., true positive). Moreover, we compute the accuracy along with a 95% confidence interval from an exact test based on the binomial distribution (see Clopper–Pearson intervals [38]).

In this particular study, we know the ground truth for each pathology report in the training, validation, and test sets. Generally, however, tuning the threshold values for our proposed *a posteriori* rejection methods on a target dataset (e.g., a new cancer registry or new incoming reports for existing cancer registries) may be infeasible in real-world settings due to a lack of ground truth labels. Instead, these values may need to be tuned on existing labeled training data and then applied to the target dataset. Therefore, for our study, we utilize the validation set to tune the threshold values for the *a posteriori* rejection methods and then use these tuned values to quantify the rejection rate on the test set where we expect to achieve the target accuracy (or higher) with the lowest possible rejection rate.

4. Results

For this study, we set a target accuracy level of 97%, a value high enough to consider the prediction scores trustworthy – comparable to the level at which registrars can manually annotate (~97%) – while mitigating human annotators' workload for state cancer registries. We present the results for a combination of architectures (MTCNN and MTHiSAN) and activation functions (i.e., softmax and sigmoid).

The baseline performance of our models for the four tasks considered is presented in Fig. 2. For all tasks – primary cancer site, primary anatomic subsite, laterality, and histological type – and for all combinations of architectures and activation functions, the baseline accuracy without rejection is lower than the target 97% on the validation data and on the test data for in-distribution, more recent holdout data (UTNJKYLASA) and for OOD data (CA and NM). For this reason, we leverage the selective classifications methods to increase the accuracy to the targeted level or higher in a trade-off analysis aimed to minimize the rejection rate.

4.1. Experimental study 1

This first study tested whether the proposed thresholding methods enable an accuracy level in the test set equal to or higher than the accuracy level in the validation set; it also exploits the performance of the different methods within the proposed architecture and activation function strategies. A specific tuning approach is required for the accuracy level on the validation data, whereas the current implementation of the DAC is self-tuning. For this reason, the DAC is excluded from this comparison.

For each of the proposed methods, Fig. 3 shows the tuning to the 97% accuracy level performed on the validation data to identify the threshold values to be used on the test set. The rejection rate associated with the 97% accuracy level is displayed next to the accuracy level. The results for the test set are presented in Fig. 4. As expected, with all the proposed methods and for all combinations of architecture and activation functions, we achieve a higher level of accuracy than 97% for all tasks.

The best performing method by task is selected as the one achieving 97% accuracy or higher and the lowest rejection rate. In both Figs. 3 and 4, the best performing method by task is denoted with an asterisk (*), whereas the worst performing method is denoted with an (x). The MTHiSAN architecture typically retains higher accuracy than the MTCNN architecture. The sigmoid is the selected architecture for laterality only, whereas for all other tasks the softmax architecture is selected based on lower rejection rate. Among the proposed approaches, the delta difference generally leads to lower rejection rate.

It should be noted that though a retention rate on the test set of 31.1% for histological type, or 40.3% for primary anatomic subsite, might seem low, the pathology coding API is currently in production in 13 SEER registries processing more than 3 million reports annually, which translates to a huge time savings of person-hours compared with that required by manual coding.

4.2. Experimental study 2

This second study aims to compare the performance of the proposed methods with the current in-house DAC. Therefore, we tune all methods on the validation set for the same self-tuning accuracy selected by the DAC, as shown at the top of Fig. 5. Next, the same thresholds found on the validation set are applied to the test set, and the results are presented at bottom of Fig. 5. Recalling that from the results of our first experimental study, the MTHiSAN model with the softmax activation function typically demonstrates higher accuracy at a lower rejection rate, in this second experimental study we present results on this selected architecture only. The full set of results, including the MTCNN model and with both activation functions, is reported in the Supplementary file.

Again, as expected, with all the proposed methods, we achieve an accuracy level higher than 97% on the holdout test set. On both the validation and the test sets, the proposed methods guarantee the 97% accuracy target level, but the best of them also lead to lower rejection rates than those from the DAC method.

We further extend the comparative analysis to two registries not included in the training or the validation set, namely CA (Fig. 6, top) and NM (Fig. 6, bottom). This is a strong test of the generalization of the proposed methods, as the distribution of the labels in these OOD registries is not necessarily represented by the registries included in the training set. Nevertheless, in both OOD cases, the rejection rate from the best among the proposed methods is lower than the abstention based on the DAC.

Overall, although the differences in retention rates compared to the current in-house DAC may seem small, such differences translate to tens of thousands of reports being automatically classified on all sets (training/validation/test) because of the large number of reports included in the study. Also, because the registries receive several million reports annually, significant savings in person-hours can be achieved when the proposed methods are deployed.

Furthermore, in this second experimental study, we also investigate how the retained predictions vary among classes in terms of the count of predicted classes retained at the specified accuracy and corresponding rejection rate. Our results show that our model predicts on multiple classes: at the very high 97% accuracy level required, the retained predictions vary across multiple classes. Moreover, all the proposed *a posteriori* methods better preserve the class distribution compared to the DAC by retaining samples that vary among a higher number of predicted classes out of all ground truth CTC classes at a larger or equal rate of classes (number of retained predicted classes vs. ground truth CTC classes) compared to the DAC—both for in-distribution data (Fig. 5) and for OOD test data (Fig. 6). Additional details on the number of retained samples by categories of classes (rare classes, common classes, and most prevalent classes) are provided in Section 2 of the Supplementary file.

As a case example, we can consider the results for the test set for anatomic subsite presented in Fig. 5 (bottom): with the best performing out of the proposed methods (i.e., entropy), 132 out of 312 classes are represented when optimizing the rejection rate while maintaining an arbitrarily fixed prediction accuracy of 0.97 or higher. Precisely, 138,949 ({100–62.63}*371,820/100) samples can be correctly predicted at an accuracy level of 0.98 (95% CI: [0.979;0.981]); these samples are retained among 132 classes of the original 312 CTC classes. On the other hand, with the DAC, a lower number of samples – 134,152 ({100–63.92}*371,820/100) – can be correctly predicted at the same accuracy level of 0.98 (95% CI: [0.979;0.981]) but representing only 108 classes out of 312 CTC classes.

A larger number of retained classes achieved by using the proposed methods instead of the DAC is encouraging, as the model learned to classify examples while treating the classes with equal importance; nevertheless, classes do not have the same frequencies, so it is difficult to obtain the same accuracy in each class, and the DAC seems more oriented to increasing the number of correct predictions from the majority class than from the minority ones. Conversely, the *a posteriori* methods better preserve the class distribution retained at the specified accuracy and corresponding rejection rate.

5. Discussion

In this study we proposed methods for UQ of DL models for text classification from electronic pathology reports.

We measure the performance of our models by optimizing the rejection rate while maintaining an arbitrarily fixed prediction accuracy level. The overall accuracy is the chosen evaluation metric, as we are mainly interested in the count of correctly predicted labels, or true positives. Identifying the samples where the model is less confident (i.e., the rejection rate) allows us to achieve the required level of accuracy on the remaining retained set of samples and thus obtain a real-world–deployable model in a high-risk setting, in which incorrect decisions have the potential to adversely affect patient and population health. All of the methods presented are flexible and can be tuned to fit the user’s need.

We evaluated the proposed methods (1) by comparing the rejection rate while maintaining a minimum targeted accuracy on the retained predictions, as well as (2) by comparing the rejection rate with the current in-house DAC model for text classification.

The following conclusions represent recommendations to achieve the required target (or higher) level of accuracy and the lowest rejection rate: (1) the use of the MTHiSAN architecture leads to better performance than MTCNN; (2) the softmax activation function is preferred to the sigmoid activation function for the majority of tasks; and (3) the delta and entropy ratio methods are most effective, indicating that differences among the predicted probabilities of all class labels are important.

Overall, all the proposed methods effectively allow for achieving the targeted level accuracy or higher in a trade-off analysis aimed to minimize the rejection rate. On validation and holdout test data, with all the proposed methods, we achieve on all tasks the required target level of accuracy with a lower rejection rate than the DAC. Interpreting the results for the OOD test data is more complex; nevertheless, in this case as well, the rejection rate from the best among the proposed methods achieving 97% accuracy or higher is lower than the rejection rate based on the DAC.

Also, compared to the DAC and other trainable confidence calibration approaches, the proposed post-processing optimization methods are inexpensive in terms of training time and cost while remaining flexible in defining the uncertainty—that is, accepting or rejecting the model predictions at a certain threshold level.

Additionally, all the proposed *a posteriori* methods better preserve the class distribution compared to the DAC, as the class rate (no. of predicted classes vs. no. of ground truth CTC classes) retained at the specified accuracy and corresponding rejection rate is higher when the reports are selected with any of the proposed *a posteriori* methods compared to the DAC.

This research product impacts real-world biomedical applications, directly benefiting the NCI by reducing the time to solution when training new models, as the UQ methods can be calculated *a posteriori*. The new methods also reduced the rejections across all prediction tasks, resulting in higher rates of auto-coding and expedited processing.

In addition, this work highlights DOE's capability to provide robust solutions for real-time health data analytics, a vital component for response to national biosecurity threats.

Furthermore, this work possesses general applicability in biomedical informatics, as the proposed methods can be applied across disciplinary domains, DL architectures (e.g., MTCNN, MTHiSaN, transformers), and case studies in which it is not possible to re-train on the original data.

While our proposed approach offers a practical way to incorporate the rejection option for pre-trained classifiers, we acknowledge that there might be an implicit cost as the good global performance might come at the cost of unwanted behaviors across subgroups. In fact, it has been shown that marginal calibration does not account for differences between sub-populations or individuals [39], which is relevant for both fairness and personalized decision-making.

6. Conclusions

Accurate UQ is vital to developing trust in AI-based models while minimizing human labor. In this work, we developed computationally efficient and reliable methods for assessing the predictions from DL models classifying pathology reports in population-level data. In particular, selective classification methods for *a posteriori* UQ were developed and compared to the current standard, DAC.

In this study, we evaluated each selective classification method independently, but future work will explore combined approaches such as a majority vote rule, where robustness of the selection is achieved by retaining all reports where all or the majority of methods agree on the prediction. Moreover, although overall accuracy is the metric of interest for this study, other metrics such as precision, recall, and F-measure values might be more appropriate for similar studies.

In addition, although the sigmoid-based architecture in this analysis was not directly linked to the best performances on all tasks, such an implementation will allow us to extend the current study to more specific multi-label tasks for which jointly modeling multiple outcomes is required, such as biomarker extraction and clinical trials matching, where there may be zero, one, or multiple correct answers for a given report. In such studies, we will generalize the inclusion criteria of selecting only a single top prediction instead of selecting a set of predictions that make it past the threshold.

Another extension of the current work will focus on improving the robustness of the proposed methods to many OOD test sets by tuning the accuracy in the validation set to achieve just about the required target level of accuracy in the test set, thereby minimizing the rejection rate. For in-training methods such as the DAC, the tuning of the accuracy is performed within the model; this approach produces quite liberal estimates that result in a high rejection rate, estimates that we showed were outperformed by the proposed methods.

Furthermore, since the loss function for the DAC balances the cross-entropy on the true classes with the cost of abstention, it might be that thresholding on the cross-entropy of the base model recovers the DAC performance. Effectively, this would imply that the activations on the true classes of the DAC match the base classifier. If that is indeed the case, then this work suggests a method to test new forms of the loss function that might retain more reports.

Lastly, future work will focus on in-depth analysis to understand the pattern of differences in predictions at the report and class levels among the newly proposed methods, as well as in a comparative analysis with the DAC, offering potential perspectives for accounting for class imbalance and, therefore, to improve further on the model's performance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the Joint Design of Advanced Computing Solutions for Cancer program established by Department of Energy and the National Cancer Institute of the National Institutes of Health. This work was performed under the auspices of DOE by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725. KCR data were collected with funding from NCI's SEER Program (HHSN261201800013I), the CDC's National Program of Cancer Registries (NPCR) (U58DP 00003907), and the Commonwealth of Kentucky. LTR data were collected using funding from NCI's SEER program (HHSN261201800007I), the CDC's NPCR (NU58DP 006332-02-00), and the State of Louisiana. NJSCR data were collected using funding from NCI's SEER program (HHSN261201300021I), the CDC's NPCR (NU58DP 006279-02-00), as well as the State of New Jersey and the Rutgers Cancer Institute of New Jersey. NMTR's participation in this project was supported by Contract HHSN261201800014I, Task Order HHSN26100001 from the NCI's SEER program. The Cancer Surveillance System is supported by the NCI's SEER program (HHSN261291800004I) with additional funds provided by the Fred Hutchinson Cancer Center and the State of Washington. UCR is funded by the NCI's SEER program (HHSN261201800016I) and the CDC's NPCR (NU58DP0063200) with additional support from the University of Utah and Huntsman Cancer Foundation.

References

[1]. Qiu JX, Yoon H-J, Fearn PA, Tourassi GD, Deep learning for automated extraction of primary sites from cancer pathology reports, IEEE J. Biomed. Health Inform 22 (1) (2017) 244–251. [PubMed: 28475069]

[2]. Hughes M, Li I, Kotoulas S, Suzumura T, Medical text classification using convolutional neural networks, in: Informatics for Health: Connected Citizen-Led Wellness and Population Health, IOS Press, 2017, pp. 246–250.

[3]. Gao S, Qiu JX, Alawad M, Hinkle JD, Schaefferkoetter N, Yoon H-J, Christian B, Fearn PA, Penberthy L, Wu X-C, et al. , Classifying cancer pathology reports with hierarchical self-attention networks, Artif. Intell. Med 101 (2019) 101726. [PubMed: 31813492]

[4]. Alawad M, Gao S, Qiu JX, Yoon HJ, Blair Christian J, Penberthy L, Mumphrey B, Wu X-C, Coyle L, Tourassi G, Automatic extraction of cancer registry reportable information from free-text

pathology reports using multitask convolutional neural networks, J. Am. Med. Inform. Assoc 27 (1) (2020) 89–98. [PubMed: 31710668]

[5]. Yoon H-J, Peluso A, Durbin EB, Wu X-C, Stroup A, Doherty J, Schwartz S, Wiggins C, Coyle L, Penberthy L, Automatic information extraction from childhood cancer pathology reports, JAMIA open 5 (2) (2022) ooac049. [PubMed: 35721398]

[6]. Jiang H, Kim B, Guan M, Gupta M, To trust or not to trust a classifier, Adv. Neural Inf. Process. Syst 31 (2018).

[7]. Kompa B, Snoek J, Beam AL, Second opinion needed: communicating uncertainty in medical machine learning, NPJ Digit. Med 4 (1) (2021) 1–6. [PubMed: 33398041]

[8]. Krishnan R, Tickoo O, Improving model calibration with accuracy versus uncertainty optimization, Adv. Neural Inf. Process. Syst 33 (2020) 18237–18248.

[9]. Geifman Y, El-Yaniv R, Selective classification for deep neural networks, Adv. Neural Inf. Process. Syst 30 (2017).

[10]. Schölkopf B, Williamson RC, Smola A, Shawe-Taylor J, Platt J, Support vector method for novelty detection, Adv. Neural Inf. Process. Syst 12 (1999).

[11]. Liang S, Li Y, Srikant R, Enhancing the reliability of out-of-distribution image detection in neural networks, 2017, arXiv preprint arXiv:1706.02690.

[12]. Guo C, Pleiss G, Sun Y, Weinberger KQ, On calibration of modern neural networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330.

[13]. Xin J, Tang R, Yu Y, Lin J, The art of abstention: Selective prediction and error regularization for natural language processing, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 1040–1051.

[14]. Thulasidasan S, Chennupati G, Bilmes JA, Bhattacharya T, Michalak S, On mixup training: Improved calibration and predictive uncertainty for deep neural networks, Adv. Neural Inf. Process. Syst 32 (2019).

[15]. Hendrycks D, Mu N, Cubuk ED, Zoph B, Gilmer J, Lakshminarayanan B, Augmix: A simple method to improve robustness and uncertainty under data shift, in: International Conference on Learning Representations, Vol. 1, 2020, p. 6.

[16]. Lakshminarayanan B, Pritzel A, Blundell C, Simple and scalable predictive uncertainty estimation using deep ensembles, Adv. Neural Inf. Process. Syst 30 (2017).

[17]. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B, Snoek J, Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, Adv. Neural Inf. Process. Syst 32 (2019).

[18]. Graves A, Practical variational inference for neural networks, Adv. Neural Inf. Process. Syst 24 (2011).

[19]. Welling M, Teh YW, Bayesian learning via stochastic gradient langevin dynamics, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), Citeseer, 2011, pp. 681–688.

[20]. Gal Y, Ghahramani Z, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.

[21]. Maddox WJ, Izmailov P, Garipov T, Vetrov DP, Wilson AG, A simple baseline for bayesian uncertainty in deep learning, Adv. Neural Inf. Process. Syst 32 (2019).

[22]. Heek J, Well-calibrated bayesian neural networks, University of Cambridge, 2018.

[23]. Pereyra G, Tucker G, Chorowski J, Kaiser Ł, Hinton G, Regularizing neural networks by penalizing confident output distributions, 2017, arXiv preprint arXiv:1701.06548.

[24]. Kumar A, Sarawagi S, Jain U, Trainable calibration measures for neural networks from kernel mean embeddings, in: International Conference on Machine Learning, PMLR, 2018, pp. 2805–2814.

[25]. Liang G, Zhang Y, Wang X, Jacobs N, Improved trainable calibration method for neural networks on medical imaging classification, 2020, arXiv preprint arXiv:2009.04057.

[26]. Zhang X, Chan FT, Mahadevan S, Explainable machine learning in image classification models: An uncertainty quantification perspective, Knowl.-Based Syst 243 (2022) 108418.

[27]. Cortes C, DeSalvo G, Mohri M, Boosting with abstention, Adv. Neural Inf. Process. Syst 29 (2016).

[28]. Garcia A, Clavel C, Essid S, d'Alché Buc F, Structured output learning with abstention: Application to accurate opinion prediction, in: International Conference on Machine Learning, PMLR, 2018, pp. 1695–1703.

[29]. Thulasidasan S, Bhattacharya T, Bilmes J, Chennupati G, Mohd-Yusof J, Knows when it doesn't know: Deep abstaining classifiers, 2018.

[30]. Kull M, Perello Nieto M, Kängsepp M, Silva Filho T, Song H, Flach P, Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration, Adv. Neural Inf. Process. Syst 32 (2019).

[31]. Hendrycks D, Gimpel K, A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2016, arXiv preprint arXiv:1610. 02136.

[32]. De Angeli K, Gao S, Alawad M, Yoon H-J, Schaefferkoetter N, Wu X-C, Durbin EB, Doherty J, Stroup A, Coyle L, et al. , Deep active learning for classifying cancer pathology reports, BMC Bioinform. 22 (1) (2021) 1–25.

[33]. Kim Y, Convolutional neural networks for sentence classification, 2014, arXiv preprint arXiv:1408.5882.

[34]. Yoon H-J, Ramanathan A, Tourassi G, Multi-task deep neural networks for automated extraction of primary site and laterality information from cancer pathology reports, in: INNS Conference on Big Data, Springer, 2016, pp. 195–204.

[35]. Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon H-J, Wu X-C, Durbin EB, Doherty J, Stroup A, et al. , Limitations of transformers on clinical text classification, IEEE J. Biomed. Health Inform (2021).

[36]. Gao S, Alawad M, Schaefferkoetter N, Penberthy L, Wu X-C, Durbin EB, Coyle L, Ramanathan A, Tourassi G, Using case-level context to classify cancer pathology reports, PLoS One 15 (5) (2020) e0232840.

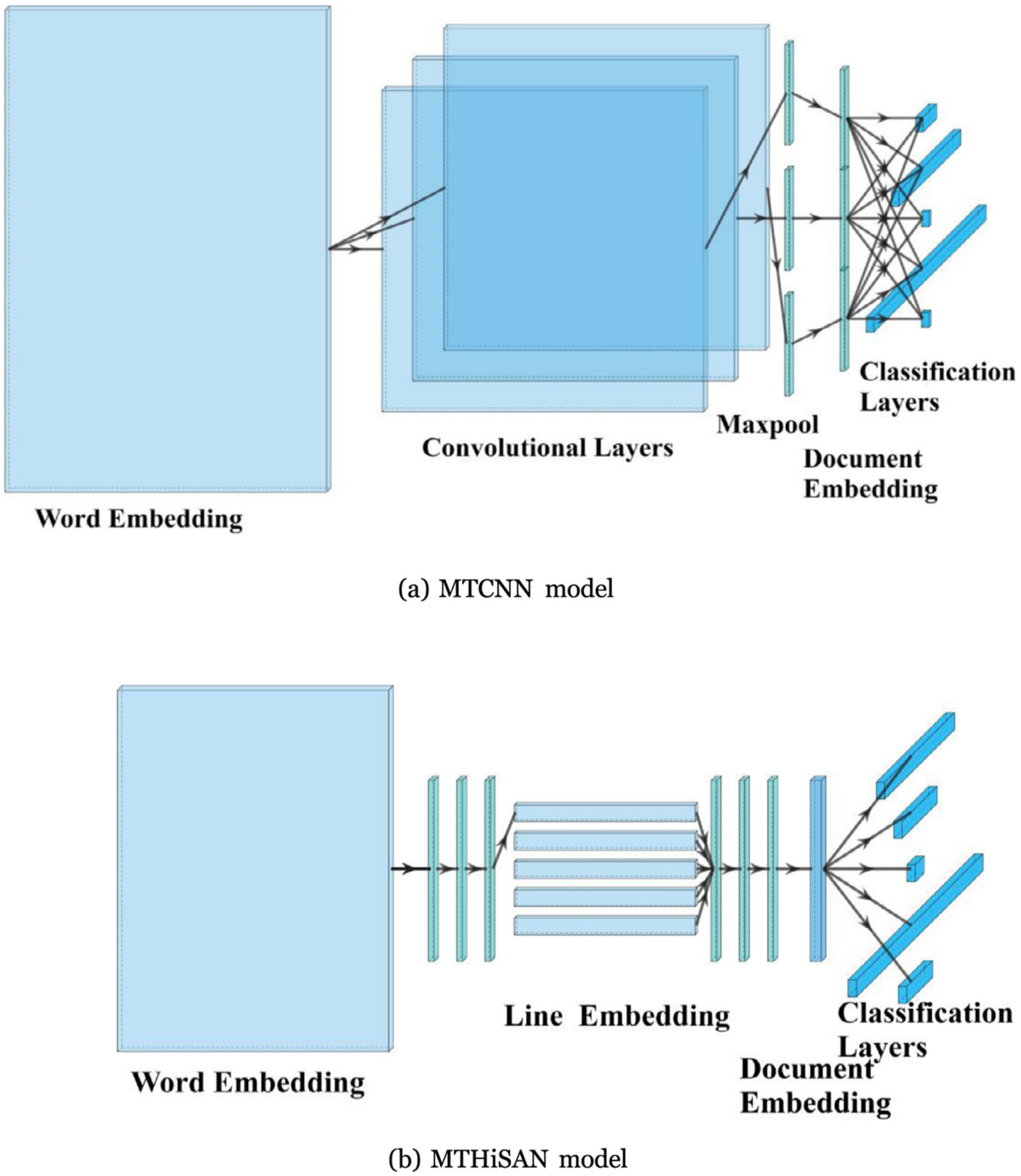
[37]. Dhaubhadel S, Mohd-Yusof J, Ganguly K, Chennupati G, Thulasidasan S, Hengartner N, Mumphrey BJ, Durban EB, Doherty JA, Lemieux M, et al. , Why I'm not answering: Understanding determinants of classification of an abstaining classifier for cancer pathology reports, 2020, arXiv preprint arXiv:2009.05094.

[38]. Clopper CJ, Pearson ES, The use of confidence or fiducial limits illustrated in the case of the binomial, Biometrika (1934) 404–413.

[39]. Jones E, Sagawa S, Koh PW, Kumar A, Liang P, Selective classification can magnify disparities across groups, 2020, arXiv preprint arXiv:2010.14134.

Statement of significance

Summary	Description
Problem	Deep neural networks (DNNs) show great performance on extraction and classification of information about tumor characterization from unstructured notes in electronic pathology reports, but uncertainty quantification is vital to develop a measure of the reliability of the model’s predictions.
What is already known	Performance gain from modified DNNs such as the current in-house deep learning–based abstaining classifier (DAC) comes from the majority classes, with this method being limited by its high rejection rate in minority classes.
What this paper adds	We show how multiple a posteriori methods for selective classification can achieve the desired level of accuracy or higher and often the lowest rejection rate compared with the DAC; moreover, they also require no model retraining based upon the desired accuracy, resulting in higher rates of auto-coding and expedited processing.

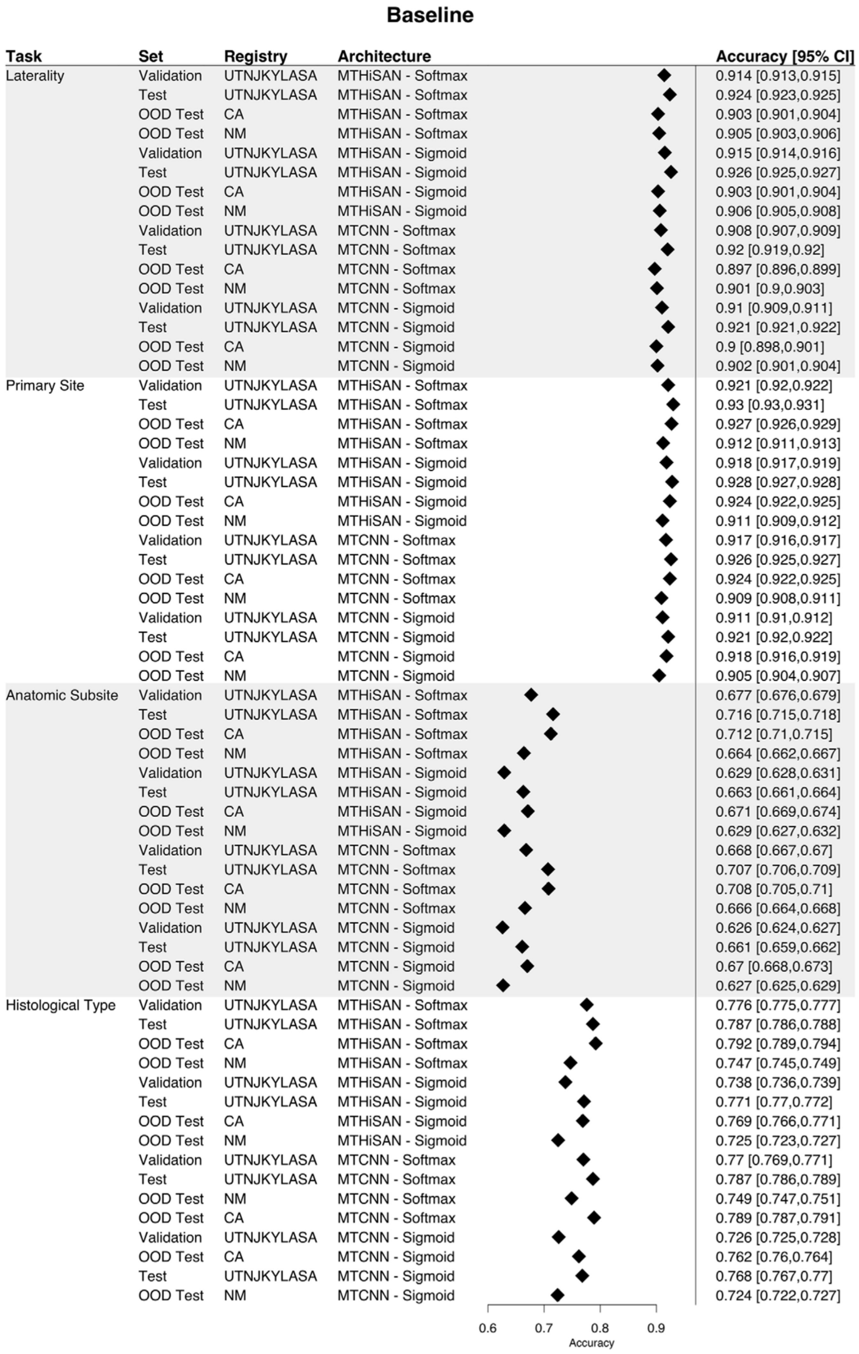


(a) MTCNN model

(b) MTHiSAN model

Fig. 1.

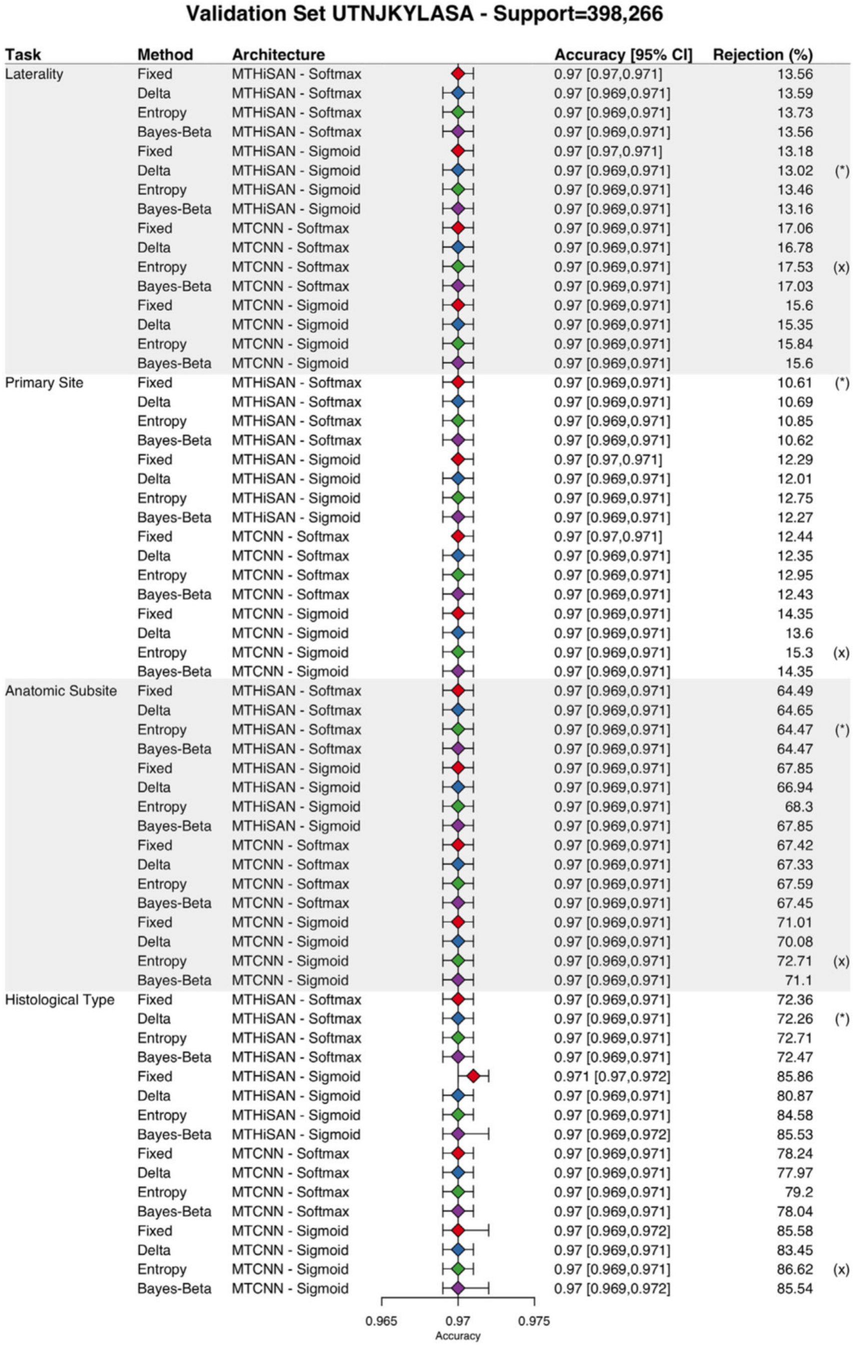
Subfigure (a) shows the architecture of the MTCNN for multi-task classification. The model has three parallel filters with a different window size for each filter. The output from these filters is fed into a maxpooling layer and is then concatenated before a final softmax or sigmoid function is applied for each classification task. In subfigure (b), the MTHiSAN architecture is presented, showing how the different layers of word embeddings create a word hierarchy which is connected to a self-attention and target attention, respectively. The output of these attention mechanisms is directly connected to similar hierarchical attention mechanisms, creating a hierarchy over the lines in a pathology report. These features create the document embedding, which are the extracted features used in the final classification layer.



Task	Set	Registry	Architecture	Accuracy [95% CI]
Laterality	Validation	UTNJKYLASA	MTHiSAN - Softmax	0.914 [0.913,0.915]
	Test	UTNJKYLASA	MTHiSAN - Softmax	0.924 [0.923,0.925]
	OOD Test	CA	MTHiSAN - Softmax	0.903 [0.901,0.904]
	OOD Test	NM	MTHiSAN - Softmax	0.905 [0.903,0.906]
	Validation	UTNJKYLASA	MTHiSAN - Sigmoid	0.915 [0.914,0.916]
	Test	UTNJKYLASA	MTHiSAN - Sigmoid	0.926 [0.925,0.927]
	OOD Test	CA	MTHiSAN - Sigmoid	0.903 [0.901,0.904]
	OOD Test	NM	MTHiSAN - Sigmoid	0.906 [0.905,0.908]
	Validation	UTNJKYLASA	MTCNN - Softmax	0.908 [0.907,0.909]
	Test	UTNJKYLASA	MTCNN - Softmax	0.92 [0.919,0.92]
	OOD Test	CA	MTCNN - Softmax	0.897 [0.896,0.899]
	OOD Test	NM	MTCNN - Softmax	0.901 [0.9,0.903]
	Validation	UTNJKYLASA	MTCNN - Sigmoid	0.91 [0.909,0.911]
	Test	UTNJKYLASA	MTCNN - Sigmoid	0.921 [0.921,0.922]
	OOD Test	CA	MTCNN - Sigmoid	0.9 [0.898,0.901]
	OOD Test	NM	MTCNN - Sigmoid	0.902 [0.901,0.904]
Primary Site	Validation	UTNJKYLASA	MTHiSAN - Softmax	0.921 [0.92,0.922]
	Test	UTNJKYLASA	MTHiSAN - Softmax	0.93 [0.93,0.931]
	OOD Test	CA	MTHiSAN - Softmax	0.927 [0.926,0.929]
	OOD Test	NM	MTHiSAN - Softmax	0.912 [0.911,0.913]
	Validation	UTNJKYLASA	MTHiSAN - Sigmoid	0.918 [0.917,0.919]
	Test	UTNJKYLASA	MTHiSAN - Sigmoid	0.928 [0.927,0.928]
	OOD Test	CA	MTHiSAN - Sigmoid	0.924 [0.922,0.925]
	OOD Test	NM	MTHiSAN - Sigmoid	0.911 [0.909,0.912]
	Validation	UTNJKYLASA	MTCNN - Softmax	0.917 [0.916,0.917]
	Test	UTNJKYLASA	MTCNN - Softmax	0.926 [0.925,0.927]
	OOD Test	CA	MTCNN - Softmax	0.924 [0.922,0.925]
	OOD Test	NM	MTCNN - Softmax	0.909 [0.908,0.911]
	Validation	UTNJKYLASA	MTCNN - Sigmoid	0.911 [0.91,0.912]
	Test	UTNJKYLASA	MTCNN - Sigmoid	0.921 [0.92,0.922]
	OOD Test	CA	MTCNN - Sigmoid	0.918 [0.916,0.919]
	OOD Test	NM	MTCNN - Sigmoid	0.905 [0.904,0.907]
Anatomic Subsite	Validation	UTNJKYLASA	MTHiSAN - Softmax	0.677 [0.676,0.679]
	Test	UTNJKYLASA	MTHiSAN - Softmax	0.716 [0.715,0.718]
	OOD Test	CA	MTHiSAN - Softmax	0.712 [0.71,0.715]
	OOD Test	NM	MTHiSAN - Softmax	0.664 [0.662,0.667]
	Validation	UTNJKYLASA	MTHiSAN - Sigmoid	0.629 [0.628,0.631]
	Test	UTNJKYLASA	MTHiSAN - Sigmoid	0.663 [0.661,0.664]
	OOD Test	CA	MTHiSAN - Sigmoid	0.671 [0.669,0.674]
	OOD Test	NM	MTHiSAN - Sigmoid	0.629 [0.627,0.632]
	Validation	UTNJKYLASA	MTCNN - Softmax	0.668 [0.667,0.67]
	Test	UTNJKYLASA	MTCNN - Softmax	0.707 [0.706,0.709]
	OOD Test	CA	MTCNN - Softmax	0.708 [0.705,0.71]
	OOD Test	NM	MTCNN - Softmax	0.666 [0.664,0.668]
	Validation	UTNJKYLASA	MTCNN - Sigmoid	0.626 [0.624,0.627]
	Test	UTNJKYLASA	MTCNN - Sigmoid	0.661 [0.659,0.662]
	OOD Test	CA	MTCNN - Sigmoid	0.67 [0.668,0.673]
	OOD Test	NM	MTCNN - Sigmoid	0.627 [0.625,0.629]
Histological Type	Validation	UTNJKYLASA	MTHiSAN - Softmax	0.776 [0.775,0.777]
	Test	UTNJKYLASA	MTHiSAN - Softmax	0.787 [0.786,0.788]
	OOD Test	CA	MTHiSAN - Softmax	0.792 [0.789,0.794]
	OOD Test	NM	MTHiSAN - Softmax	0.747 [0.745,0.749]
	Validation	UTNJKYLASA	MTHiSAN - Sigmoid	0.738 [0.736,0.739]
	Test	UTNJKYLASA	MTHiSAN - Sigmoid	0.771 [0.77,0.772]
	OOD Test	CA	MTHiSAN - Sigmoid	0.769 [0.766,0.771]
	OOD Test	NM	MTHiSAN - Sigmoid	0.725 [0.723,0.727]
	Validation	UTNJKYLASA	MTCNN - Softmax	0.77 [0.769,0.771]
	Test	UTNJKYLASA	MTCNN - Softmax	0.787 [0.786,0.789]
	OOD Test	NM	MTCNN - Softmax	0.749 [0.747,0.751]
	OOD Test	CA	MTCNN - Softmax	0.789 [0.787,0.791]
	Validation	UTNJKYLASA	MTCNN - Sigmoid	0.726 [0.725,0.728]
	OOD Test	CA	MTCNN - Sigmoid	0.762 [0.76,0.764]
	Test	UTNJKYLASA	MTCNN - Sigmoid	0.768 [0.767,0.77]
	OOD Test	NM	MTCNN - Sigmoid	0.724 [0.722,0.727]

Fig. 2.

Baseline accuracy for our models on validation and test data for both in-distribution, more recent holdout data (UTNJKYLASA) as well as OOD data (CA and NM). In all instances, the accuracy is lower than the required 97% level.

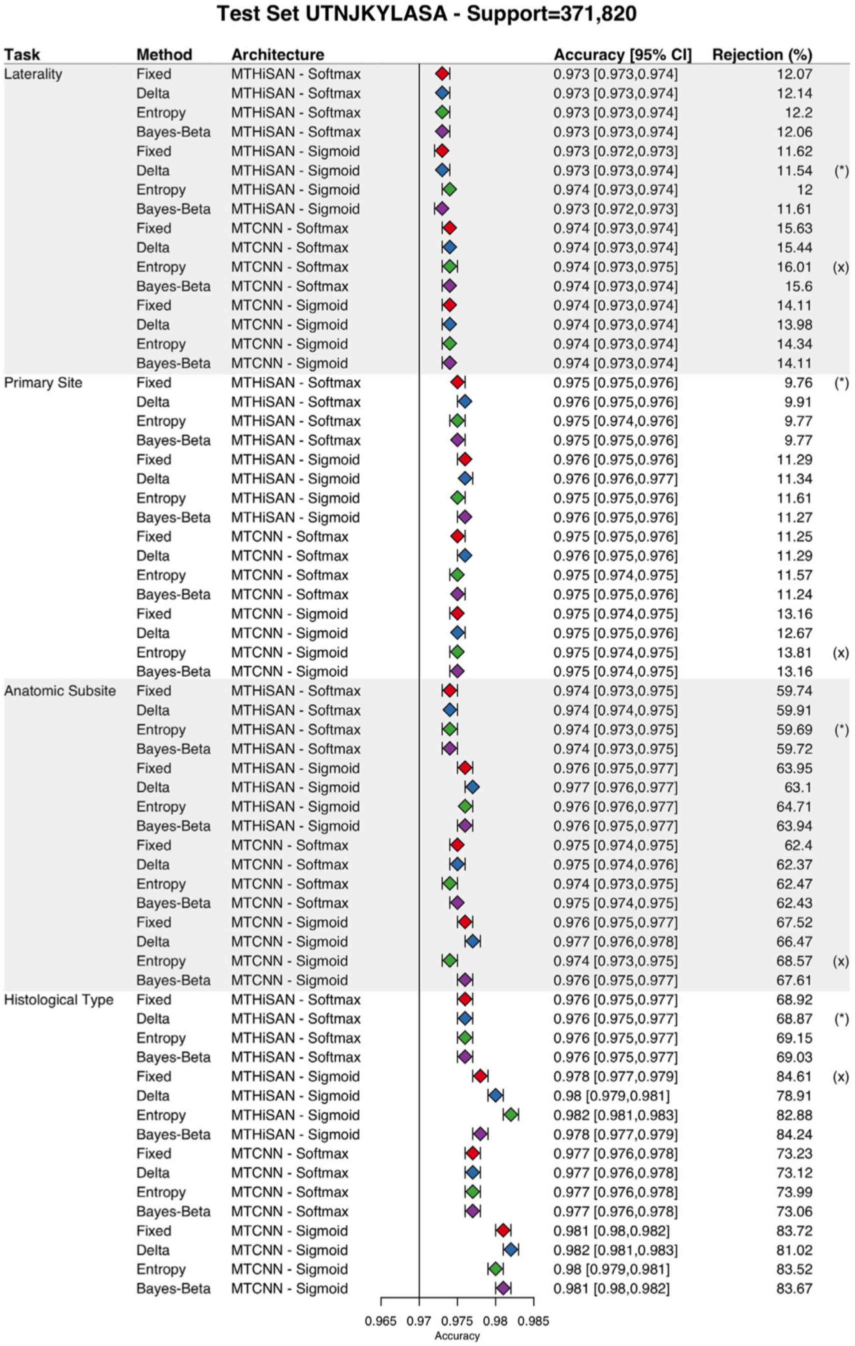


Validation Set UTNJKYLASA - Support=398,266

Task	Method	Architecture	Accuracy [95% CI]	Rejection (%)	
Laterality	Fixed	MTHiSAN - Softmax	0.97 [0.97,0.971]	13.56	
	Delta	MTHiSAN - Softmax	0.97 [0.969,0.971]	13.59	
	Entropy	MTHiSAN - Softmax	0.97 [0.969,0.971]	13.73	
	Bayes-Beta	MTHiSAN - Softmax	0.97 [0.969,0.971]	13.56	
	Fixed	MTHiSAN - Sigmoid	0.97 [0.97,0.971]	13.18	
	Delta	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	13.02	(*)
	Entropy	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	13.46	
	Bayes-Beta	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	13.16	
	Fixed	MTCNN - Softmax	0.97 [0.969,0.971]	17.06	
	Delta	MTCNN - Softmax	0.97 [0.969,0.971]	16.78	
	Entropy	MTCNN - Softmax	0.97 [0.969,0.971]	17.53	(x)
	Bayes-Beta	MTCNN - Softmax	0.97 [0.969,0.971]	17.03	
	Fixed	MTCNN - Sigmoid	0.97 [0.969,0.971]	15.6	
	Delta	MTCNN - Sigmoid	0.97 [0.969,0.971]	15.35	
	Entropy	MTCNN - Sigmoid	0.97 [0.969,0.971]	15.84	
	Bayes-Beta	MTCNN - Sigmoid	0.97 [0.969,0.971]	15.6	
Primary Site	Fixed	MTHiSAN - Softmax	0.97 [0.969,0.971]	10.61	(*)
	Delta	MTHiSAN - Softmax	0.97 [0.969,0.971]	10.69	
	Entropy	MTHiSAN - Softmax	0.97 [0.969,0.971]	10.85	
	Bayes-Beta	MTHiSAN - Softmax	0.97 [0.969,0.971]	10.62	
	Fixed	MTHiSAN - Sigmoid	0.97 [0.97,0.971]	12.29	
	Delta	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	12.01	
	Entropy	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	12.75	
	Bayes-Beta	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	12.27	
	Fixed	MTCNN - Softmax	0.97 [0.97,0.971]	12.44	
	Delta	MTCNN - Softmax	0.97 [0.969,0.971]	12.35	
	Entropy	MTCNN - Softmax	0.97 [0.969,0.971]	12.95	
	Bayes-Beta	MTCNN - Softmax	0.97 [0.969,0.971]	12.43	
	Fixed	MTCNN - Sigmoid	0.97 [0.969,0.971]	14.35	
	Delta	MTCNN - Sigmoid	0.97 [0.969,0.971]	13.6	
	Entropy	MTCNN - Sigmoid	0.97 [0.969,0.971]	15.3	(x)
	Bayes-Beta	MTCNN - Sigmoid	0.97 [0.969,0.971]	14.35	
Anatomic Subsite	Fixed	MTHiSAN - Softmax	0.97 [0.969,0.971]	64.49	
	Delta	MTHiSAN - Softmax	0.97 [0.969,0.971]	64.65	
	Entropy	MTHiSAN - Softmax	0.97 [0.969,0.971]	64.47	(*)
	Bayes-Beta	MTHiSAN - Softmax	0.97 [0.969,0.971]	64.47	
	Fixed	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	67.85	
	Delta	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	66.94	
	Entropy	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	68.3	
	Bayes-Beta	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	67.85	
	Fixed	MTCNN - Softmax	0.97 [0.969,0.971]	67.42	
	Delta	MTCNN - Softmax	0.97 [0.969,0.971]	67.33	
	Entropy	MTCNN - Softmax	0.97 [0.969,0.971]	67.59	
	Bayes-Beta	MTCNN - Softmax	0.97 [0.969,0.971]	67.45	
	Fixed	MTCNN - Sigmoid	0.97 [0.969,0.971]	71.01	
	Delta	MTCNN - Sigmoid	0.97 [0.969,0.971]	70.08	
	Entropy	MTCNN - Sigmoid	0.97 [0.969,0.971]	72.71	(x)
	Bayes-Beta	MTCNN - Sigmoid	0.97 [0.969,0.971]	71.1	
Histological Type	Fixed	MTHiSAN - Softmax	0.97 [0.969,0.971]	72.36	
	Delta	MTHiSAN - Softmax	0.97 [0.969,0.971]	72.26	(*)
	Entropy	MTHiSAN - Softmax	0.97 [0.969,0.971]	72.71	
	Bayes-Beta	MTHiSAN - Softmax	0.97 [0.969,0.971]	72.47	
	Fixed	MTHiSAN - Sigmoid	0.971 [0.97,0.972]	85.86	
	Delta	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	80.87	
	Entropy	MTHiSAN - Sigmoid	0.97 [0.969,0.971]	84.58	
	Bayes-Beta	MTHiSAN - Sigmoid	0.97 [0.969,0.972]	85.53	
	Fixed	MTCNN - Softmax	0.97 [0.969,0.971]	78.24	
	Delta	MTCNN - Softmax	0.97 [0.969,0.971]	77.97	
	Entropy	MTCNN - Softmax	0.97 [0.969,0.971]	79.2	
	Bayes-Beta	MTCNN - Softmax	0.97 [0.969,0.971]	78.04	
	Fixed	MTCNN - Sigmoid	0.97 [0.969,0.972]	85.58	
	Delta	MTCNN - Sigmoid	0.97 [0.969,0.971]	83.45	
	Entropy	MTCNN - Sigmoid	0.97 [0.969,0.971]	86.62	(x)
	Bayes-Beta	MTCNN - Sigmoid	0.97 [0.969,0.972]	85.54	

Fig. 3.

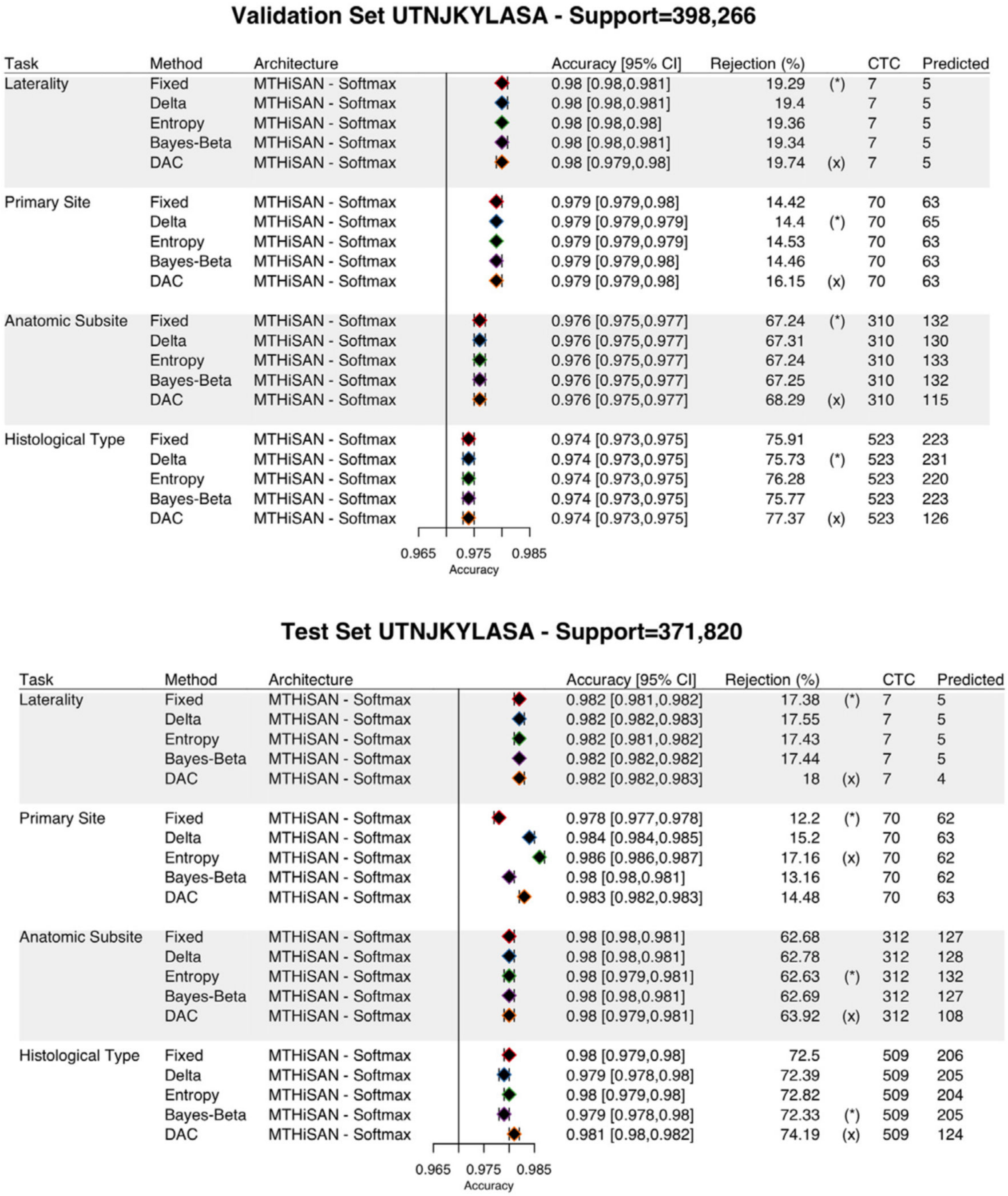
Experimental study 1: validation data. The accuracy level of about 97% is achieved with the displayed rejection rate. (*) and (x) represent the lowest and highest rejection rate by task.



Task	Method	Architecture	Accuracy [95% CI]	Rejection (%)	
Laterality	Fixed	MTHiSAN - Softmax	0.973 [0.973,0.974]	12.07	
	Delta	MTHiSAN - Softmax	0.973 [0.973,0.974]	12.14	
	Entropy	MTHiSAN - Softmax	0.973 [0.973,0.974]	12.2	
	Bayes-Beta	MTHiSAN - Softmax	0.973 [0.973,0.974]	12.06	
	Fixed	MTHiSAN - Sigmoid	0.973 [0.972,0.973]	11.62	
	Delta	MTHiSAN - Sigmoid	0.973 [0.973,0.974]	11.54	(*)
	Entropy	MTHiSAN - Sigmoid	0.974 [0.973,0.974]	12	
	Bayes-Beta	MTHiSAN - Sigmoid	0.973 [0.972,0.973]	11.61	
	Fixed	MTCNN - Softmax	0.974 [0.973,0.974]	15.63	
	Delta	MTCNN - Softmax	0.974 [0.973,0.974]	15.44	
	Entropy	MTCNN - Softmax	0.974 [0.973,0.975]	16.01	(x)
	Bayes-Beta	MTCNN - Softmax	0.974 [0.973,0.974]	15.6	
	Fixed	MTCNN - Sigmoid	0.974 [0.973,0.974]	14.11	
	Delta	MTCNN - Sigmoid	0.974 [0.973,0.974]	13.98	
	Entropy	MTCNN - Sigmoid	0.974 [0.973,0.974]	14.34	
	Bayes-Beta	MTCNN - Sigmoid	0.974 [0.973,0.974]	14.11	
Primary Site	Fixed	MTHiSAN - Softmax	0.975 [0.975,0.976]	9.76	(*)
	Delta	MTHiSAN - Softmax	0.976 [0.975,0.976]	9.91	
	Entropy	MTHiSAN - Softmax	0.975 [0.974,0.976]	9.77	
	Bayes-Beta	MTHiSAN - Softmax	0.975 [0.975,0.976]	9.77	
	Fixed	MTHiSAN - Sigmoid	0.976 [0.975,0.976]	11.29	
	Delta	MTHiSAN - Sigmoid	0.976 [0.976,0.977]	11.34	
	Entropy	MTHiSAN - Sigmoid	0.975 [0.975,0.976]	11.61	
	Bayes-Beta	MTHiSAN - Sigmoid	0.976 [0.975,0.976]	11.27	
	Fixed	MTCNN - Softmax	0.975 [0.975,0.976]	11.25	
	Delta	MTCNN - Softmax	0.976 [0.975,0.976]	11.29	
	Entropy	MTCNN - Softmax	0.975 [0.974,0.975]	11.57	
	Bayes-Beta	MTCNN - Softmax	0.975 [0.975,0.976]	11.24	
	Fixed	MTCNN - Sigmoid	0.975 [0.974,0.975]	13.16	
	Delta	MTCNN - Sigmoid	0.975 [0.975,0.976]	12.67	
	Entropy	MTCNN - Sigmoid	0.975 [0.974,0.975]	13.81	(x)
	Bayes-Beta	MTCNN - Sigmoid	0.975 [0.974,0.975]	13.16	
Anatomic Subsite	Fixed	MTHiSAN - Softmax	0.974 [0.973,0.975]	59.74	
	Delta	MTHiSAN - Softmax	0.974 [0.974,0.975]	59.91	
	Entropy	MTHiSAN - Softmax	0.974 [0.973,0.975]	59.69	(*)
	Bayes-Beta	MTHiSAN - Softmax	0.974 [0.973,0.975]	59.72	
	Fixed	MTHiSAN - Sigmoid	0.976 [0.975,0.977]	63.95	
	Delta	MTHiSAN - Sigmoid	0.977 [0.976,0.977]	63.1	
	Entropy	MTHiSAN - Sigmoid	0.976 [0.976,0.977]	64.71	
	Bayes-Beta	MTHiSAN - Sigmoid	0.976 [0.975,0.977]	63.94	
	Fixed	MTCNN - Softmax	0.975 [0.974,0.975]	62.4	
	Delta	MTCNN - Softmax	0.975 [0.974,0.976]	62.37	
	Entropy	MTCNN - Softmax	0.974 [0.973,0.975]	62.47	
	Bayes-Beta	MTCNN - Softmax	0.975 [0.974,0.975]	62.43	
	Fixed	MTCNN - Sigmoid	0.976 [0.975,0.977]	67.52	
	Delta	MTCNN - Sigmoid	0.977 [0.976,0.978]	66.47	
	Entropy	MTCNN - Sigmoid	0.974 [0.973,0.975]	68.57	(x)
	Bayes-Beta	MTCNN - Sigmoid	0.976 [0.975,0.977]	67.61	
Histological Type	Fixed	MTHiSAN - Softmax	0.976 [0.975,0.977]	68.92	
	Delta	MTHiSAN - Softmax	0.976 [0.975,0.977]	68.87	(*)
	Entropy	MTHiSAN - Softmax	0.976 [0.975,0.977]	69.15	
	Bayes-Beta	MTHiSAN - Softmax	0.976 [0.975,0.977]	69.03	
	Fixed	MTHiSAN - Sigmoid	0.978 [0.977,0.979]	84.61	(x)
	Delta	MTHiSAN - Sigmoid	0.98 [0.979,0.981]	78.91	
	Entropy	MTHiSAN - Sigmoid	0.982 [0.981,0.983]	82.88	
	Bayes-Beta	MTHiSAN - Sigmoid	0.978 [0.977,0.979]	84.24	
	Fixed	MTCNN - Softmax	0.977 [0.976,0.978]	73.23	
	Delta	MTCNN - Softmax	0.977 [0.976,0.978]	73.12	
	Entropy	MTCNN - Softmax	0.977 [0.976,0.978]	73.99	
	Bayes-Beta	MTCNN - Softmax	0.977 [0.976,0.978]	73.06	
	Fixed	MTCNN - Sigmoid	0.981 [0.98,0.982]	83.72	
	Delta	MTCNN - Sigmoid	0.982 [0.981,0.983]	81.02	
	Entropy	MTCNN - Sigmoid	0.98 [0.979,0.981]	83.52	
	Bayes-Beta	MTCNN - Sigmoid	0.981 [0.98,0.982]	83.67	

Fig. 4.

Experimental study 1: more recent, hold-out test set. The tuning on the validation set resulted in higher accuracy on the test set than the target level of 97%, corresponding to the displayed rejection rate. (*) and (x) represent the lowest and highest rejection rate by task.



Validation Set UTNJKYLASA - Support=398,266

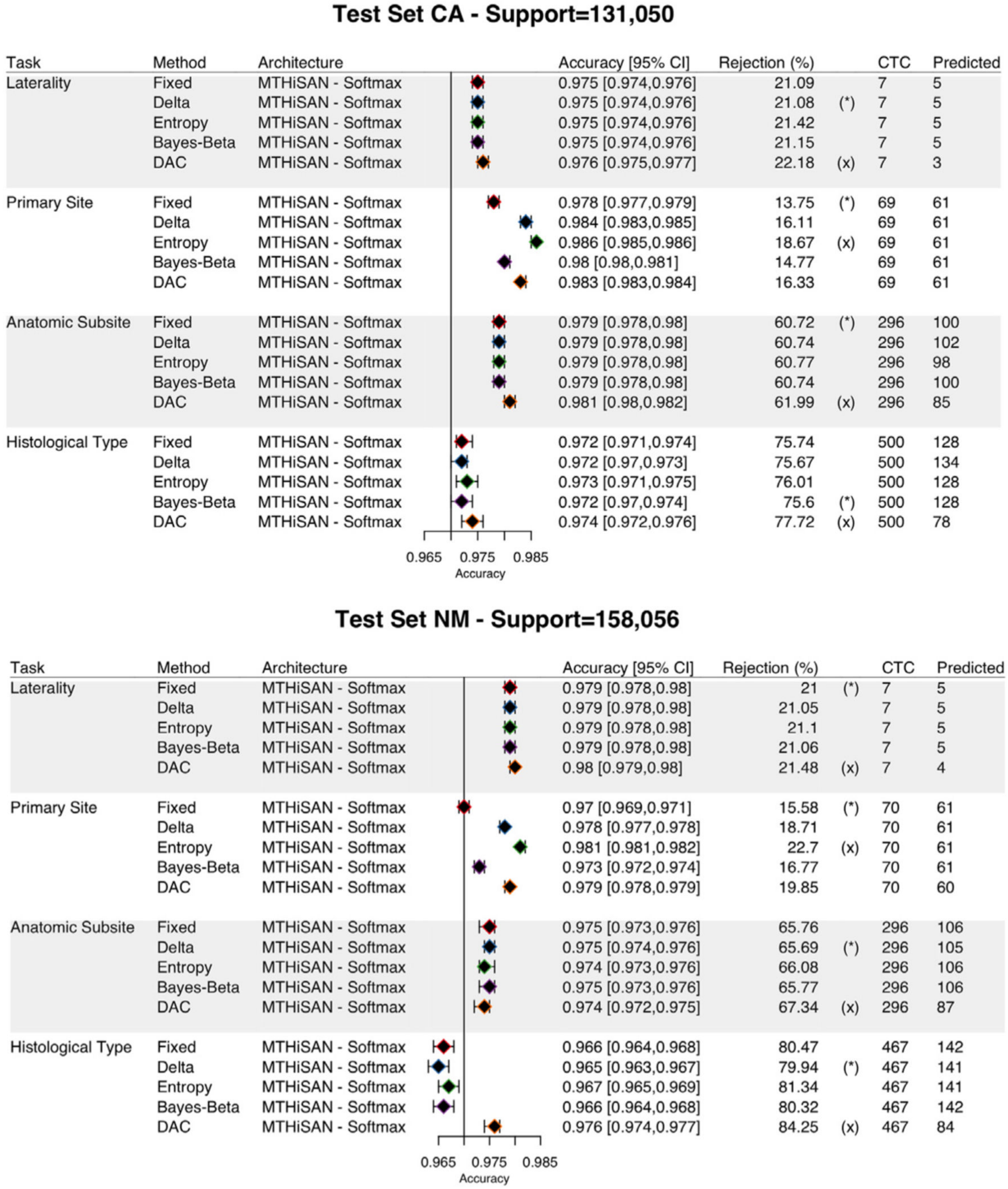
Task	Method	Architecture	Accuracy [95% CI]	Rejection (%)		CTC	Predicted
Laterality	Fixed	MTHiSAN - Softmax	0.98 [0.98,0.981]	19.29	(*)	7	5
	Delta	MTHiSAN - Softmax	0.98 [0.98,0.981]	19.4		7	5
	Entropy	MTHiSAN - Softmax	0.98 [0.98,0.98]	19.36		7	5
	Bayes-Beta	MTHiSAN - Softmax	0.98 [0.98,0.981]	19.34		7	5
	DAC	MTHiSAN - Softmax	0.98 [0.979,0.98]	19.74	(x)	7	5
Primary Site	Fixed	MTHiSAN - Softmax	0.979 [0.979,0.98]	14.42		70	63
	Delta	MTHiSAN - Softmax	0.979 [0.979,0.979]	14.4	(*)	70	65
	Entropy	MTHiSAN - Softmax	0.979 [0.979,0.979]	14.53		70	63
	Bayes-Beta	MTHiSAN - Softmax	0.979 [0.979,0.98]	14.46		70	63
	DAC	MTHiSAN - Softmax	0.979 [0.979,0.98]	16.15	(x)	70	63
Anatomic Subsite	Fixed	MTHiSAN - Softmax	0.976 [0.975,0.977]	67.24	(*)	310	132
	Delta	MTHiSAN - Softmax	0.976 [0.975,0.977]	67.31		310	130
	Entropy	MTHiSAN - Softmax	0.976 [0.975,0.977]	67.24		310	133
	Bayes-Beta	MTHiSAN - Softmax	0.976 [0.975,0.977]	67.25		310	132
	DAC	MTHiSAN - Softmax	0.976 [0.975,0.977]	68.29	(x)	310	115
Histological Type	Fixed	MTHiSAN - Softmax	0.974 [0.973,0.975]	75.91		523	223
	Delta	MTHiSAN - Softmax	0.974 [0.973,0.975]	75.73	(*)	523	231
	Entropy	MTHiSAN - Softmax	0.974 [0.973,0.975]	76.28		523	220
	Bayes-Beta	MTHiSAN - Softmax	0.974 [0.973,0.975]	75.77		523	223
	DAC	MTHiSAN - Softmax	0.974 [0.973,0.975]	77.37	(x)	523	126

Test Set UTNJKYLASA - Support=371,820

Task	Method	Architecture	Accuracy [95% CI]	Rejection (%)		CTC	Predicted
Laterality	Fixed	MTHiSAN - Softmax	0.982 [0.981,0.982]	17.38	(*)	7	5
	Delta	MTHiSAN - Softmax	0.982 [0.982,0.983]	17.55		7	5
	Entropy	MTHiSAN - Softmax	0.982 [0.981,0.982]	17.43		7	5
	Bayes-Beta	MTHiSAN - Softmax	0.982 [0.982,0.982]	17.44		7	5
	DAC	MTHiSAN - Softmax	0.982 [0.982,0.983]	18	(x)	7	4
Primary Site	Fixed	MTHiSAN - Softmax	0.978 [0.977,0.978]	12.2	(*)	70	62
	Delta	MTHiSAN - Softmax	0.984 [0.984,0.985]	15.2		70	63
	Entropy	MTHiSAN - Softmax	0.986 [0.986,0.987]	17.16	(x)	70	62
	Bayes-Beta	MTHiSAN - Softmax	0.98 [0.98,0.981]	13.16		70	62
	DAC	MTHiSAN - Softmax	0.983 [0.982,0.983]	14.48		70	63
Anatomic Subsite	Fixed	MTHiSAN - Softmax	0.98 [0.98,0.981]	62.68		312	127
	Delta	MTHiSAN - Softmax	0.98 [0.98,0.981]	62.78		312	128
	Entropy	MTHiSAN - Softmax	0.98 [0.979,0.981]	62.63	(*)	312	132
	Bayes-Beta	MTHiSAN - Softmax	0.98 [0.98,0.981]	62.69		312	127
	DAC	MTHiSAN - Softmax	0.98 [0.979,0.981]	63.92	(x)	312	108
Histological Type	Fixed	MTHiSAN - Softmax	0.98 [0.979,0.98]	72.5		509	206
	Delta	MTHiSAN - Softmax	0.979 [0.978,0.98]	72.39		509	205
	Entropy	MTHiSAN - Softmax	0.98 [0.979,0.98]	72.82		509	204
	Bayes-Beta	MTHiSAN - Softmax	0.979 [0.978,0.98]	72.33	(*)	509	205
	DAC	MTHiSAN - Softmax	0.981 [0.98,0.982]	74.19	(x)	509	124

Fig. 5.

Experimental study 2: validation set (top) and more recent, hold-out test set (bottom). The tuning on the validation set for the same self-tuning accuracy selected by the DAC resulted in a higher accuracy than the target level of 97%. (*) and (x) represent the lowest and highest rejection rate by task. Also, all the proposed a posteriori methods retain a larger or equal rate of classes (i.e., number of retained predicted classes vs. ground truth CTC classes) compared to the DAC.



Test Set CA - Support=131,050

Task	Method	Architecture	Accuracy [95% CI]	Rejection (%)		CTC	Predicted
Laterality	Fixed	MTHiSAN - Softmax	0.975 [0.974,0.976]	21.09		7	5
	Delta	MTHiSAN - Softmax	0.975 [0.974,0.976]	21.08	(*)	7	5
	Entropy	MTHiSAN - Softmax	0.975 [0.974,0.976]	21.42		7	5
	Bayes-Beta	MTHiSAN - Softmax	0.975 [0.974,0.976]	21.15		7	5
	DAC	MTHiSAN - Softmax	0.976 [0.975,0.977]	22.18	(x)	7	3
Primary Site	Fixed	MTHiSAN - Softmax	0.978 [0.977,0.979]	13.75	(*)	69	61
	Delta	MTHiSAN - Softmax	0.984 [0.983,0.985]	16.11		69	61
	Entropy	MTHiSAN - Softmax	0.986 [0.985,0.986]	18.67	(x)	69	61
	Bayes-Beta	MTHiSAN - Softmax	0.98 [0.98,0.981]	14.77		69	61
	DAC	MTHiSAN - Softmax	0.983 [0.983,0.984]	16.33		69	61
Anatomic Subsite	Fixed	MTHiSAN - Softmax	0.979 [0.978,0.98]	60.72	(*)	296	100
	Delta	MTHiSAN - Softmax	0.979 [0.978,0.98]	60.74		296	102
	Entropy	MTHiSAN - Softmax	0.979 [0.978,0.98]	60.77		296	98
	Bayes-Beta	MTHiSAN - Softmax	0.979 [0.978,0.98]	60.74		296	100
	DAC	MTHiSAN - Softmax	0.981 [0.98,0.982]	61.99	(x)	296	85
Histological Type	Fixed	MTHiSAN - Softmax	0.972 [0.971,0.974]	75.74		500	128
	Delta	MTHiSAN - Softmax	0.972 [0.97,0.973]	75.67		500	134
	Entropy	MTHiSAN - Softmax	0.973 [0.971,0.975]	76.01		500	128
	Bayes-Beta	MTHiSAN - Softmax	0.972 [0.97,0.974]	75.6	(*)	500	128
	DAC	MTHiSAN - Softmax	0.974 [0.972,0.976]	77.72	(x)	500	78

Test Set NM - Support=158,056

Task	Method	Architecture	Accuracy [95% CI]	Rejection (%)		CTC	Predicted
Laterality	Fixed	MTHiSAN - Softmax	0.979 [0.978,0.98]	21	(*)	7	5
	Delta	MTHiSAN - Softmax	0.979 [0.978,0.98]	21.05		7	5
	Entropy	MTHiSAN - Softmax	0.979 [0.978,0.98]	21.1		7	5
	Bayes-Beta	MTHiSAN - Softmax	0.979 [0.978,0.98]	21.06		7	5
	DAC	MTHiSAN - Softmax	0.98 [0.979,0.98]	21.48	(x)	7	4
Primary Site	Fixed	MTHiSAN - Softmax	0.97 [0.969,0.971]	15.58	(*)	70	61
	Delta	MTHiSAN - Softmax	0.978 [0.977,0.978]	18.71		70	61
	Entropy	MTHiSAN - Softmax	0.981 [0.981,0.982]	22.7	(x)	70	61
	Bayes-Beta	MTHiSAN - Softmax	0.973 [0.972,0.974]	16.77		70	61
	DAC	MTHiSAN - Softmax	0.979 [0.978,0.979]	19.85		70	60
Anatomic Subsite	Fixed	MTHiSAN - Softmax	0.975 [0.973,0.976]	65.76		296	106
	Delta	MTHiSAN - Softmax	0.975 [0.974,0.976]	65.69	(*)	296	105
	Entropy	MTHiSAN - Softmax	0.974 [0.973,0.976]	66.08		296	106
	Bayes-Beta	MTHiSAN - Softmax	0.975 [0.973,0.976]	65.77		296	106
	DAC	MTHiSAN - Softmax	0.974 [0.972,0.975]	67.34	(x)	296	87
Histological Type	Fixed	MTHiSAN - Softmax	0.966 [0.964,0.968]	80.47		467	142
	Delta	MTHiSAN - Softmax	0.965 [0.963,0.967]	79.94	(*)	467	141
	Entropy	MTHiSAN - Softmax	0.967 [0.965,0.969]	81.34		467	141
	Bayes-Beta	MTHiSAN - Softmax	0.966 [0.964,0.968]	80.32		467	142
	DAC	MTHiSAN - Softmax	0.976 [0.974,0.977]	84.25	(x)	467	84

Fig. 6.

Experimental study 2: OOD test set – CA (top) and NM (bottom). The tuning on the validation set led to a higher accuracy on the test set than the target level of 97%. (*) and (x) represent the lowest and highest rejection rate by task. Also, all the proposed *a posteriori* methods retain a larger or equal rate of classes (i.e., number of retained predicted classes vs. ground truth CTC classes) compared to the DAC.

Table 1

Number of electronic pathology reports considered in the study split by training, validation, and test data. Specifically, the models are trained using data from five registries (LA+KY+UT+NJ+SA) and evaluated on most recent (all reports after 2017) holdout data from these same five registries as well as two OOD registries (CA and NM).

Registry	Training samples	Validation samples	Test samples
LAKYUTNJSA	1,864,099	398,266	371,820
LA	363,367	77,623	76,305
KY	361,981	76,920	80,588
UT	140,026	30,298	28,436
NJ	369,441	78,874	74,742
SA	629,284	134,551	111,749
CA	n/a	n/a	131,050
NM	n/a	n/a	158,056