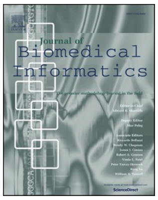
Journal of Biomedical Informatics 125 (2022) 103957

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types[☆]

Kevin De Angeli[a,b,*], Shang Gao[a], Ioana Danciu[a,c], Eric B. Durbin[d], Xiao-Cheng Wu[e], Antoinette Stroup[f], Jennifer Doherty[g], Stephen Schwartz[h], Charles Wiggins[i], Mark Damesyn[j], Linda Coyle[k], Lynne Penberthy[l], Georgia D. Tourassi[a], Hong-Jun Yoon[a]

[a] *Oak Ridge National Laboratory, 1 Bethel Valley Rd, Oak Ridge, TN 37830, USA*
[b] *The Bredesen Center, The University of Tennessee, 821 Volunteer Blvd. Knoxville, TN 37996, USA*
[c] *Department of Biomedical Informatics, Vanderbilt University, 2525 West End Avenue, Nashville, TN 37203, USA*
[d] *College of Medicine, University of Kentucky, Lexington, KY 40536, USA*
[e] *Louisiana State University Health Sciences Center, School of Public Health, New Orleans, LA 70112, USA*
[f] *Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, 08901, USA*
[g] *Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84132, USA*
[h] *Fred Hutchinson Cancer Research Center, Epidemiology Program, Seattle, WA 98109, USA*
[i] *University of New Mexico, Albuquerque, NM 87131, USA*
[j] *California Department of Public Health, Sacramento, CA 59814, USA*
[k] *Information Management Services Inc., Calverton, MD 20705, USA*
[l] *National Cancer Institute, Bethesda, MD 20814, USA*

ARTICLE INFO

Keywords:
Deep Learning
CNN
Class Imbalance
Text Classification
Ensemble
NLP

ABSTRACT

In the last decade, the widespread adoption of electronic health record documentation has created huge opportunities for information mining. Natural language processing (NLP) techniques using machine and deep learning are becoming increasingly widespread for information extraction tasks from unstructured clinical notes. Disparities in performance when deploying machine learning models in the real world have recently received considerable attention. In the clinical NLP domain, the robustness of convolutional neural networks (CNNs) for classifying cancer pathology reports under natural distribution shifts remains understudied. In this research, we aim to quantify and improve the performance of the CNN for text classification on out-of-distribution (OOD) datasets resulting from the natural evolution of clinical text in pathology reports. We identified class imbalance due to different prevalence of cancer types as one of the sources of performance drop and analyzed the impact of previous methods for addressing class imbalance when deploying models in real-world domains. Our results show that our novel class-specialized ensemble technique outperforms other methods for the classification of rare cancer types in terms of macro F1 scores. We also found that traditional ensemble methods perform better in top classes, leading to higher micro F1 scores. Based on our findings, we formulate a series of recommendations for other ML practitioners on how to build robust models with extremely imbalanced datasets in biomedical NLP applications.

1. Introduction

One of the tasks of the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI) is to provide statistics and analyze cancer trends in the US. Every year, cancer registries receive thousands of electronically transmitted pathology reports

[☆] This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

[*] Corresponding author at: Oak Ridge National Laboratory, 1 Bethel Valley Rd, Oak Ridge, TN 37830, USA.
E-mail address: kevindeangeli94@gmail.com (K. De Angeli).

https://doi.org/10.1016/j.jbi.2021.103957
Received 30 June 2021; Received in revised form 4 November 2021; Accepted 17 November 2021
Available online 22 November 2021

from pathology laboratories. These pathology reports consist of unstructured clinical text. Specialized human annotators are required to extract valuable information. This process is costly and time consuming. Therefore, developing reliable models to classify cancer pathology reports automatically remains one of the priorities of SEER.

Recently, numerous machine learning researchers have shown that models which are trained in labs often exhibit a significant performance drop when they are deployed in the real world. Some researchers have identified certain aspects of the modeling process and model training as the source of performance disparity [1]. Others have perceived the issue as a robustness problem and focused on building models that learn generalizable features so that they can maintain their performance under natural distribution shifts. Performance drop at deployment time is a serious issue that could affect every machine learning practitioner and the reliability of AI systems [1].

Significant research progress has been made in developing deep learning models for information extraction from cancer pathology reports [2–5]. Difficulties in sharing data between healthcare systems has made analyses of the performance of models on out-of-distribution (OOD) datasets very challenging. As registries around the country are legally required to collect cancer pathology reports for all residents of their state, natural statistical variations arise in the datasets. These variations may occur due to different data acquisition pipelines as the pathology reports can come from different laboratories, disparate disease prevalence patterns, or the evolution of language and reporting protocols over time [6].

Class imbalance occurs when the proportion of samples belonging to one or more classes in a dataset varies drastically. Class imbalance is of extreme importance for the clinical NLP domain because many clinical conditions such as certain cancer histologies are very rare. For example, extracting the histology code from pathology reports is a task that involves text classification with over 600 classes. Some types of cancers are extremely common (e.g. adenocarcinoma, ductal carcinoma), while others occur less frequently (e.g. squamous cell carcinoma). These phenomena lead to datasets with extreme levels of class imbalance, which impacts the performance of the classifiers. Models trained with such datasets will exhibit bias towards the most prevalent majority classes because of their higher prior probabilities, while often ignoring the minority classes. Although the effects of class imbalance in machine learning have been well documented [7–9], previous researchers have concurred that class imbalance in the context of deep learning is understudied [10]. Additionally, most of the existing work on class imbalance has been done in the computer vision field [11].

The class imbalance problem can become crucial when deploying models to classify pathology reports on unseen registries from states and geographical regions outside the training data (OOD datasets). The distribution between the minority groups may differ widely across registries, and classifiers' bias towards the top classes leads to deteriorated deployment performance. Few researchers in the clinical NLP domain have focused on characterizing and improving the performance of deep learning models under natural distribution shifts. Our aim is to narrow this literature gap with a specific focus on minority classes (rare cancer types). Our contributions are as follows:

- We show that natural distribution shifts have a considerable impact in performance when classifying cancer pathology reports.
- We identify class imbalance and class distribution as one of the sources of performance drop at deployment time, and analyze some of the existing methods to solve the problems associated with these issues.
- To strengthen the consistency of our results, we compare the methods under two different classification tasks commonly found in cancer pathology reports: histology and subsite.
- We demonstrate that an ensemble of 12 CNNs can improve the generalization power at deployment time. However, we show that most of the performance gain comes from the majority classes.
- We propose a novel implementation of ensemble learning where each model specializes in a different group of classes. Our class-specialized ensemble outperforms other class imbalance techniques in terms of macro F1 scores when testing on unseen registries while maintaining competitive micro F1 scores.

Our research is at the intersection of robustness and class imbalance for clinical NLP. Our novel ensemble model, which improves performance in rare cancers, is generalizable and can also benefit other practitioners working on problems related to bias and fairness in machine learning.

2. Previous work

2.1. Class Imbalance

In this section, we present existing work in the class imbalance literature that is relevant to our research. Generally, class imbalance techniques in the context of machine learning are grouped into data-level techniques and algorithm-level methods. We note that our specific problem involves extreme levels of imbalance (described in detail in our methods 3.7), which are uncommonly observed in previous works. Nevertheless, some of their methods, results, and findings are still applicable.

Data-level techniques focus on manipulating the distribution of the training dataset in order to reduce the imbalance present in the original data. The two most basic paradigms in this group are: (1) random oversampling (ROS), where samples from minority classes are duplicated, and (2) random undersampling (RUS), where samples from majority classes are discarded.

Masko et al. [12] presented a comprehensive study of the effects ROS using CNNs for image classification. They used the MNIST, ImageNet, CIFAR-10, and CIFAR-100 datasets. They performed experiments with relatively small levels of imbalance and showed that ROS improved the baseline scores.

The effects of RUS have also been studied extensively. For example, Kubat et al. [13] presented an algorithm that selectively removed samples from the majority classes. The downside of their study is that they only focus on 2-class datasets. Hulse et al. [14] developed an extensive analysis of seven sampling techniques using 35 benchmark datasets and 11 classifiers. They found that the performance of the sampling techniques is dependent upon the machine learning model and showed that, in some circumstances, RUS can outperform other classical techniques.

Dynamic sampling is a technique which combines both RUS and ROS. Pouyanfar et al. [15] developed this sampling strategy based on the way humans often operate: repeating a certain task until the error is reduced. Thus, the researchers created an algorithm which adjusts the distribution of classes in the training dataset based on a performance metric (e.g. F1 score). As a result, majority classes are expected to be undersampled while minority classes will be oversampled. Chawla et al. [16] proposed a novel technique called SMOTE which also uses a combination of RUS and ROS. It had previously been noticed that simply oversampling minority documents with replacement does not improve minority performance significantly [17]. For this reason, the authors developed a special case of oversampling which selects synthetically created samples from the minority classes.

Although the simplicity and efficiency of methods that combine ROS and RUS may seem appealing, in applications with extreme levels of class imbalance such as ours (where the top classes appears 17 k times and the bottom class appear only once), oversampling repeatedly from the same documents within minority classes will only force the model to memorize features that may not even be useful for the respective classes. In addition, event though SMOTE is a standard class imbalance tool for traditional machine learning [18–20], it still has important limitations when it comes to deep learning models. Some of the challenges come from the implementation of the algorithm itself. For example, in problems such as ours when the input to a model is a matrix composed of word vectors, sampling the K nearest neighbors is not adequate. Moreover, previous researchers have demonstrated that in situations where high-dimensional data are common, SMOTE does not improve model performance [21].

Algorithm-level methods for class imbalance focus on modifying the learning process without altering the distribution of the dataset [10]. The most popular paradigm in this group is cost-sensitive learning, where models are penalized for the classification of certain (minority) classes. The cost associated with the misclassification of each class is assigned using a cost matrix, where an entry $C_{i,j}$ in this matrix represents the cost of predicting class i for the true class j. In the context of text classification, Padurariu et al. showed that cost-sensitive methods can outperform data-sampling methods [22]. Previous researchers have noted that the biggest challenge of cost-sensitive methods is building an effective cost matrix [10]. Depending on the specific problem, experts could use previous knowledge to define costs. However, in complex problems with a lot of classes and extreme level of class imbalance, coming up with an optimal cost matrix is a serious challenge.

Some authors have built novel approaches for class imbalance that borrow ideas from both data and algorithm-level methods. This is the case of Lee et al. [23] who showed that a particular implementation of transfer learning, also known as two-phase learning, can outperform other classical class-imbalance techniques. Their application involves the classification of plankton images using CNNs. During the first phase, they trained a model with a subset of the data using some threshold N. In this subset of data, samples are rejected so that the frequency of each class present in the dataset does not exceed $N = 5000$ (found experimentally). The authors' reasoning is that the model trained with the thresholding data is less biased, and it can learn features that are relevant for the minority classes, but it loses population information. Therefore, to recover the lost information, they fine-tune the model with the entire dataset. The authors compared two-phase learning with other models trained with noise addition, data augmentation, and a combination of both. Our class-specialized ensemble method presented in this paper was partially inspired by their two-phase learning implementation.

For a detailed analysis of previous results in the imbalance literature, we refer to [10], a survey paper where the authors review 15 deep learning methods for class imbalance. Their extensive review discusses all three types of techniques: data-level methods, algorithmic-level methods, and hybrid-methods. For an overview of imbalance methods focusing specifically in text classification, we recommend [22]. Additionally, [8] provides a comparative study of different data-level methods.

2.2. Robustness

Numerous authors have recently identified and evaluated the disparities in a model's performance during deployment [1,24,6,25–30]. From the pool of existing research, notable work includes [1], where the authors identified underspecification as a key factor diminishing the reliability of machine learning systems. In their paper, they performed a series of stress tests and showed how different modeling aspects, even as simple as a random seed, can lead to almost unpredictable performance when deploying a model in the real world. Although the authors provided substantial examples of the performance discrepancy when testing in OOD datasets, a distinct solution was not provided.

In computer vision, prior work has often focused on the ImageNet dataset using CNNs. For example, in [24], the authors analyzed the reliability of robustness techniques which were developed using datasets with synthetic distribution shifts. They showed that most of the existing techniques are not effective under natural distribution shifts, and they found that most improvement comes from data size and diversity. Conversely, Hendrycks et al. [27] argued that using synthetic data can improve the performance of a model on OOD data. In addition, they built three robustness benchmarks for image classification and introduced a new data augmentation technique. Djolonga et al. [28] also used the ImageNet dataset, but their analysis focused on the effects that data/model scale and transfer learning have in OOD performance. Their conclusion is that given the limitations associated with data and model scaling, transfer learning is the most promising approach in the short term. In this study, we analyzed the effects of transfer learning through our two-phase learning implementation.

In the clinical field, Stacke et al. [6] presented a technique to quantify how robust a model is to domain shifts and how to identify new data for which the model would struggle to generalize. They achieve this by measuring the differences in feature representation by an arbitrary model. Their specific application is tumor classification from images. Although the authors provide a useful metric to quantify the robustness of a model, they do not focus on the aspects of the learning process which enhance the models' performance.

In the context of NLP, Wu et al. [29] approached the OOD robustness problem by modifying existing models to produce multiple disentangled representations. They argue that it is important for a model to separate between general, target-specific, and source-specific features. Intuitively, their approach is an ensemble of models combined together into a single architecture. The down-side of their study is that they used datasets with very few classes, making it hard to predict the efficiency of their methods in more challenging problems.

2.3. Ensemble Methods

Ensemble learning [31] is a machine learning technique that solves a given task with multiple models. The purpose of applying multiple models is to obtain collective decisions from them, thus reducing the likelihood of incorrect selections. Aggregating decisions from multiple models adds generalizability to the outcomes, improving overall task performance and avoiding overfitting the training dataset. This is a desirable feature for the classification of under-represented class labels.

Since ensembles of classifiers combine decisions from multiple models, the individual models should exhibit some level of diversity. Bootstrap aggregation [32], also known as bagging, is a popular technique that infuses variability via the bootstrapping of the training samples. However, a recent study [33] demonstrated that the intrinsic variability from the randomized initial values of trainable parameters in artificial neural network-based models adds enough variability.

Ensemble learning does not necessarily require the models to be trained in the same feature space. Combining models of multiple local experts trained by different portions of the feature space is an alternative ensembling technique. In this approach, inferring the final decision can be done with an additive classifier by concatenating the outputs of local experts (stacked generalization) [34]. Another way to infer the final decision is by the use of a gating network to determine a generalized linear rule, a method known as mixture-of-experts (MoE) [35]. Our class specialized ensemble technique borrows some ideas from the MoE method.

In the ensemble learning literature, one particular work that is relevant to our research is [36]. Here, the authors train an ensemble of models where there is a generalist (trained in the entire dataset) and multiple specialists (trained on a confusable set of classes in the dataset). Using the MNIST dataset, the authors shows that the specialist ensemble outperforms their baseline ensemble by $\sim$3%. Their research shares some conceptual similarities with MoE, and therefore it is applicable to our research. However, their implementation of "ensemble of specialists" is completely different: we do not separate the models between generalists and specialists, and we focus specifically on class imbalance and rare classes.

3. Methods

3.1. CNN Architecture

The baseline for our experiments is a standard TextCNN used extensively in previous work involving cancer pathology reports classification [5,37–39] and clinical text classification in general [40–44]. In addition to being an universally used architecture, previous work showed that, for the task of pathology report classification, the TextCNN has competitive performance with other machine learning models [37], including BERT-based approaches [45]. We used this base TextCNN for every model in this study with some training variations described in greater detail in the following subsections. The network consists of an embedding layer followed by three parallel convolution layers with filter sizes of 3,4, and 5 consecutive words and 300 filters each, a global max pooling layer, and a dense layer. The network has ~91 million trainable parameters, where ~90 million of them belong to the embedding layer.

3.2. CNN with Class Weights

When training DL models, one can simply implement cost-sensitive learning by using custom class weights. These weights dictate how the model will be punished by the misclassification of certain classes. Thus, assigning higher weights to minority classes would force the model to pay special attention to these classes. There is a lot flexibility on how to assign weights to each class. After experimenting with an inverse frequency function, we found that giving minority classes too much weight brings the micro score down to non-permissible levels. That is because the proportion of the most rare cancer types in the dataset is extremely low, which leads to excessively high weights for the rarest classes. As a result, the model focuses on learning features for these rare classes and ignores the majority classes, which highly impacts the micro F1 score.

For our class weight implementation, we used a variation of the inverse class frequency where minority classes are assigned non-excessive, larger weights. we set the class weights WC_c following Eqs. 1 and 2.

$$weight_c = \log(\frac{|Y|}{|y_c|}) \tag{1}$$

$$\begin{cases} WC_c = weight_c & weight_c > 1 \\ WC_c = 1 & otherwise \end{cases} \tag{2}$$

In the equations above, $|Y|$ is the total number of samples in the dataset, and $|y_c|$ is the number of samples belonging to class c. Using this rule gives a weight of 1 to the majority classes and a class weight close to 14 for the most rare cases. This approach gives higher importance to the rarest cancer types without ignoring the majority classes.

3.3. Two-phase Learning

We implemented a version of two-phase learning originally introduced by Lee et al. [23]. In their paper, the authors first train the model with a class-normalized dataset which has a thresholded class distribution. Due to extreme imbalance and the large frequency of the top classes, we implemented a variation of this method in which the top 50 classes are completely left out during the first phase of training. The model is then fine-tuned with the entire dataset during the second phase. We tried a standard version of two-phase learning without class weights, and we also tried another version in which class weights are introduced (as described in 3.2) during both learning phases.

3.4. Undersampling

In the simplest form, undersampling methods discard a portion of the majority class to balance the dataset. In problems with moderate levels of class imbalance, one can simply discard majority class samples until reaching equal number of samples with the minority classes. For our specific problem, discarding the top classes based on the frequency of the rarest cancer types is not possible because these rare classes appear at extremely low proportions (see Section 3.7 and 7). Alternatively, we discarded a number of documents from the top classes using a threshold based on certain percentiles (50^{th},90^{th}, and 95^{th}) of class frequency. The specific implementation procedure is described as follows:

- Find number of documents belonging to the class in the respective percentile (50^{th},90^{th}, and 95^{th}). We call this value the undersampling threshold α.
- Discard documents from the dataset so that there are at most α documents in each class. No documents are discarded for classes with fewer than α samples.
- Train model with this smaller, more balanced dataset.

3.5. Class-Specialized Ensemble

Our novel method was inspired by MoE and two-phase learning with class weights. We wanted to create an ensemble of models were each TextCNN would specialize in different group of classes. Thus, we first ordered the classes by frequency based on the training and validation datasets. Then we created groups of 50 (histology) and 28 (subsite) classes based on their frequency order. The reasoning behind forming frequency-based groups is that the imbalance between the individual groups will be reduced, as opposed to creating groups of classes selected randomly. We decided to use group sizes of 50 and 28 because that will keep the ensemble relatively small (12 models). However, one could easily experiment with having larger ensembles which specialize in smaller groups of classes.

During the first learning phase, we let individual members of the ensemble learn features that are key for their assigned class group. Then, each member was fine-tuned with the entire dataset. For example, during the first training phase of the histology task, we trained one TextCNN with the top 50 classes (classes 0–49) and another TextCNN with the second group of 50 classes (classes 50–99), and so on. During the second learning phase (fine-tuning), we trained each of the models with the entire dataset. Fig. 1 shows a general overview of the steps we took to train the class-specialized ensemble.

To aggregate the individual predictions of the ensemble and generate the final prediction, we use a simple multilayer perceptron (MLP) model. This MLP model is trained with input vectors that are created by concatenating the softmax vectors from each of the 12 models in the ensemble and the respective y label associated with the documents. Thus, for the case of histology where there are 645 classes, the input to the MLP is a vector of size 7740 (number of classes multiplied by the number of models). The architecture consists of two dense layers with 4000 and 3000 neurons, respectively. We also included a dropout layer between each of the dense layers (dropout rate $= 0.5$). The number of layers, neurons, and hyperparameters were found experimentally. The MLP network has ~47 million trainable parameters.

3.6. Ensemble

Since our proposed model involves an ensemble of models which naturally presents an advantage against individual models, we implemented two traditional ensemble learning techniques. We used an ensemble of 12 models to be consistent with our class-specialized method and perform a fair comparison.

The first ensemble technique implemented is majority voting. Here, the final prediction is the class that is predicted the most often across the 12-model ensemble (ties are resolved by randomly selecting one of the classes with the most votes). The other technique is softmax averaging. This method consists of taking the average of the softmax vectors across the ensemble. For example, for the ensemble trained in the histology

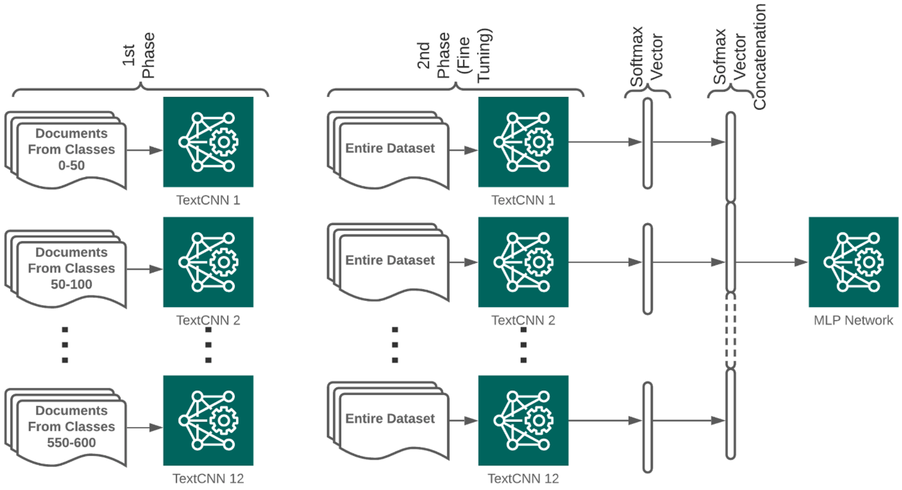


Fig. 1. Training pipeline of the proposed model. The MLP network takes as input a vector of concatenated softmax vectors and their respective Y label.

task, we simply take the average of 12 vectors (ensemble size) of size 645 (number of classes) and then predict the class with the highest softmax value in this average vector.

We note that our selection of ensemble methods can easily be applied by other machine learning practitioners, it is highly parallelizable, and it is computationally cheap compared to other ensemble methods.

3.7. Dataset

The dataset consists of cancer pathology reports from the Louisiana Tumor Registry (LTR), Kentucky Cancer Registry (KCR), Utah Cancer Registry (UCR), New Jersey State Cancer Registry (NJSCR), Seattle Cancer Registry (SCR), New Mexico Cancer Registry (NMCR), and California Cancer Registry (CCR). The total number of pathology reports from these seven registries is 2,059,758 documents. Table 1 shows the size of each of the individual datasets associated with the seven registries. We use numerical values instead of the actual registry names to preserve anonymity. Even though there are other tasks associated with our dataset (site, laterality, and behavior), in this study we focus on the histology and subsite tasks because these are the top priority for NCI; our labels are based on the ICD-O-3 system from the World Health Organization Classification of Tumors [46]. Additionally, histology and subsite have the largest numbers of classes and highest level of class imbalance, making them good targets for our robustness study.

There are 645 histology classes, and the dataset presents extreme cases of class imbalance. For example, 22.0% of the reports belong to the top class (adenocarcinoma in situ/NOS) and 19.0% belong to the second most popular class (duct carcinoma). The top 10 classes constitute 62.8% of the dataset. The least prevalent 635 classes constitute only 37.2% of the data. Some of the cancer types (31 classes) are exceptionally rare and only appear once in the entire dataset. Although removing these classes could make sense from a modeling perspective, these cancer types may still be encountered at deployment time, and they are still part of the classification problem. Therefore, all the classes were considered during training.

Identifying the subsite of a cancer pathology reports is a task with 327 classes. The level of imbalance found in this task is still high but slightly lower than what we observed in the histology task. Here, only 8.9% of the reports belong to the top class (compared to 22.0%), and the top 10 classes constitute 49.5% of the documents (compared to 62.8%). Just as in the histology task, there are cancer subsites in the dataset which are extremely rare. For example, 16 cancer subsites appear less than ten times in the dataset.

Researchers in the class imbalance field often use metrics to quantify the levels of imbalance in the dataset. For example, one common metric is $\rho = \frac{max_i(|C_i|)}{min_i(|C_i|)}$. Where $max_i(|C_i|)$ and $min_i(|C_i|)$ represents the number of samples in the top class and the bottom class, respectively. Computing this value for histology leads to $\rho = 452,363$. We note that this value is substantially larger than what one usually finds in previous work, which further demonstrates the extreme levels of imbalance in our problem.

3.8. Experimental Setup

For our experiments, we took a leave-one-out approach. In each run, we first define which of the seven registries will be the OOD dataset. This dataset is left-out of the training process. The other remaining six registries are then combined and shuffled. The combined dataset represents what a machine learning practitioner may be given to train a model in a lab setting and the left-out registry represents what one may find when deploying the model in the real world. After training the model with the combined dataset, we recorded performance metrics for both the test set from the combined dataset and the left-out (OOD) registry. In order to consider every possible combination case, we repeat this process seven times for each of the two tasks. Thus, every registry is used as the OOD dataset once. This experimental setup leads to a total of 14 individual results (7 possible dataset combinations and two tasks).

We used standard training practices to prevent serious overfitting issues. We performed a 80/10/10 train-validation-test split on the combined dataset. At the end of each epoch, we monitored the validation loss. We let the models train until the validation loss stopped decreasing for five consecutive epochs. Once training stops, we recovered the best set of weights based on the validation loss. To further prevent overfitting, our CNN model uses 50% dropout on the dense layer (Section 3.1).

The parameters and software used in this study are similar to previous work involving pathology report classification [5,37,38]. We used Keras 2.3 with the Adam optimizer, a batch size of 128, and a learning rate of 1e-4.

Table 1
Number of pathology reports in each individual dataset.

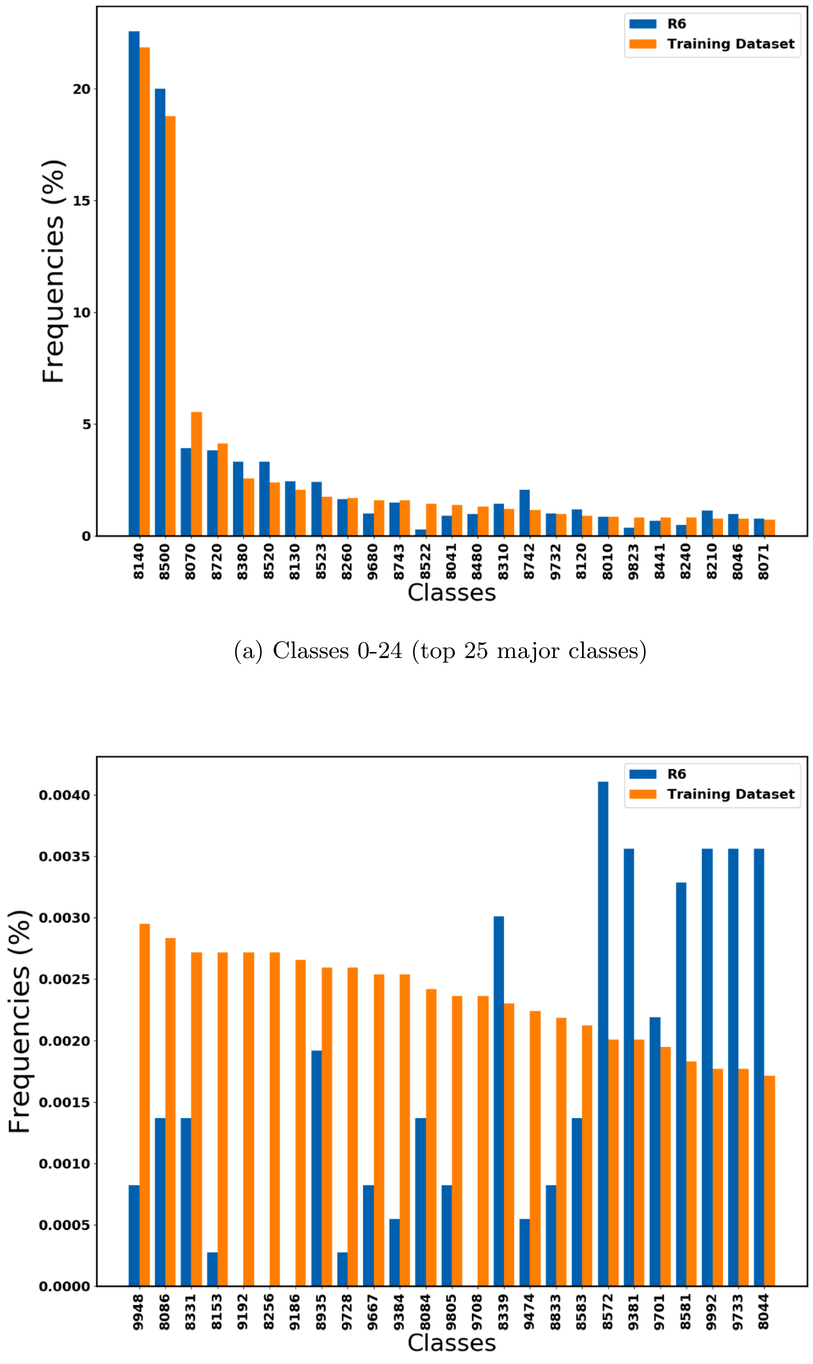
Registry	R1	R2	R3	R4	R5	R6	R7
Dataset Size	85,789	577,094	137,135	92,481	441,732	360,375	365,152

We set the document length size to 1500 words, meaning that longer documents are truncated and shorter documents are zero-padded. The word embeddings consist of vectors of size 300 which were randomly initialized; previous studied showed that random embeddings are as effective as other pre-trained word embeddings when applied to our dataset [47].

All experiments were run on individual NVIDIA V100 GPUs. The ensemble models were trained in parallel, with their output combined to form the final predictions.

3.9. Evaluation Metrics

We evaluated the performance of the models by computing the micro F1 score (Eqs. (3)–(5)) in the test dataset and in the OOD dataset. We note that for our problem, micro F1 score is equivalent to accuracy. When working with highly imbalanced datasets, using micro F1 scores can be misleading. That is because the majority classes will drive a large portion of this score, and information about the model performance on the rare classes is lost. In order to better understand the performance of the model in minority classes, we also calculated macro F1 scores (Eq. 6). The macro F1 score is a common evaluation metric used in problems



(a) Classes 0-24 (top 25 major classes)

(b) Classes 350-374 (a subset of the minority classes)

Fig. 2. Differences in class distribution between the training data and registry 6 (see Section 3.7) for the histology task. The specific class names associated with the encoded labels can be found in the SEER website [46].

with class imbalance because it averages the model's accuracy on individual classes independently of their frequency in the datasets. In other words, this metric gives equal importance to every class.

$$\text{Precision} = \frac{TruePositive}{TruePositives + FalsePositives} \tag{3}$$

$$\text{Recall} = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{4}$$

$$\text{MicroF1} = 2 * \frac{(Precision) \times (Recall)}{Precision + Recall} \tag{5}$$

$$\text{MacroF1} = \frac{1}{|C|}\sum_{C_i}^{C} F1(C_i) \tag{6}$$

In Eq. 6, $F1(C_i)$ is the accuracy score for class i, and $|C|$ represents the total number of classes in the dataset.

3.10. *Performance on Rare Cancer Types*

We performed an additional analysis of the model performances in minority classes. This in-depth study focused on the histology task because it has a larger number of classes and more extreme levels of class imbalance. For this analysis, we sorted the classes by frequency in the dataset and then created groups of 50 classes so that the first group contains the top 50 classes and the last group consists of the 50 least common classes. Then, we used the models that were trained in the entire datasets to predict on each of the specific groups. The motivation for this analysis is to gain more insight into the models' performance on minority classes beyond a single macro F1 score.

4. Results

4.1. *Class distribution in OOD datasets*

Given the large number of classes in the two tasks considered, we hypothesize that the class distribution in the training and OOD datasets will differ considerably for the rare cancer types. In order to test this hypothesis, we plotted the distribution of classes as percentages of their respective datasets in Fig. 2, using registry 6 as the OOD dataset. We note that for the top classes (Fig. 2a), the distribution is similar in both datasets. However, there are substantial differences in the distribution of the least common cancer types (Fig. 2b).

The implications of these distributional differences are one of the focus of this study. Deep learning models are known to be biased towards the top classes since they drive the loss function. Hence, the distribution of classes will affect the features it learns and which classes receive higher priority. The distribution of minority classes differs greatly across registries leading to a large underperformance of the model in OOD datasets in terms of macro scores.

4.2. *F1 Scores*

Our experimental setup yields seven individual sets of results (one for each registry left-out) for each of our two tasks. Here, we present the average scores across the seven registries for histology (Table 2) and subsite (Table 3). The pattern found in this table is representative of the outcomes found in the individual registry results, but the scale may differ slightly (see Appendix A for individual registry results).

We found that training a baseline CNN leads to decent micro scores but low macro scores. This was an indication of the model's bias against the top classes: it learned features that were important to classify common cancer types and reduce the loss, but it ignored patterns that were relevant for the more rare cancer types.

Adding class weights to the model helped the minority classes but it also hurt the performance on the majority classes (in the case of the histology task). We note that this technique is prone to produce a trade-off between micro and macro scores, since giving special attention to minority classes will reduce the performance on top classes.

Table 2
Histology Results. Overall micro and macro scores for the test and the out-of-distribution data (unseen registry). Scores were calculated by taking the average of the individual results for each of the seven registries.

Model	Test Micro	Test Macro	OOD Micro	OOD Macro
CNN	0.8007	0.4089	0.7749	0.3552
CNN w/ Class Weights	0.7885	0.4104	0.7704	0.3677
Two-Phase	0.8071	0.4815	0.7723	0.3624
Two-Phase w/ Class Weights	0.7942	**0.5169**	0.7631	0.3771
Ensemble (Maj.Vot.)	0.8096	0.4373	0.7866	0.3781
Ensemble (Softmax. Avg.)	**0.8119**	0.4458	**0.7876**	0.3841
Class-Specialized Ensemble	0.8085	0.4809	0.7778	**0.4003**

Table 3
Subsite Results. Overall micro and macro scores for the test and the out-of-distribution data (unseen registry). Scores were calculated by taking the average of the individual results for each of the seven registries.

Model	Test Micro	Test Macro	OOD Micro	OOD Macro
CNN	0.7133	0.4005	0.6717	0.3269
CNN w/ Class Weights	0.7077	0.4232	0.6701	0.3371
Two-Phase	0.7230	0.4476	0.6671	0.3232
Two-Phase w/ Class Weights	0.7090	**0.4873**	0.6543	0.3290
Ensemble (Maj.Vot.)	0.7319	0.4320	**0.6902**	0.3462
Ensemble (Softmax. Avg.)	**0.7371**	0.4397	0.6896	0.3492
Class-Specialized Ensemble	0.7253	0.4785	0.6746	**0.3658**

Using two-phase learning without weights effectively improved the test macro scores in both tasks by a significant amount (∼7% and ∼4%). It also improved the micro test scores (∼1% for both tasks). This method also introduced a small decrease in OOD micro. In terms of OOD macro scores, the improvement was small (∼1%) for histology, and there was no improvement for subsite.

Two-phase learning with class weights pushed the macro scores further but resulted in mediocre micro scores that were below the baseline CNN. We also note that the increase in macro scores between the test and OOD datasets was highly disproportional when comparing it with the baseline CNN: for histology, the test macro increased by ∼10% and the OOD macro increased by ∼2%. For subsite, the test macro increased by ∼9% and the OOD macro increased by only 0.2%. In both cases, the test macros were the highest scores across the board. The disproportional increase in macro scores between test and OOD is a potential sign of overfitting in some of the minority classes. Two-phase learning is potentially more susceptible to overfitting the minority classes since these documents are introduced during both training phases.

Ensemble methods outperformed other models in terms of micro test and micro OOD. They also provided high macro scores when compared to single models. Between the two standard ensemble methods that we implemented (majority-voting and softmax-average), the scores were similar for subsite. For histology, taking the average of the softmax vectors across the ensemble showed slightly better performance.

Our class-specialized ensemble method produced the highest OOD macro score across the board, outperforming the baseline CNN results by ∼4% for both tasks. It also produced higher test macro scores (between

∼3% and ∼4%) than the standard ensemble methods. In terms of micro scores, the class-specialized model outperformed the baseline CNN but performed slightly worse than the other ensemble models.

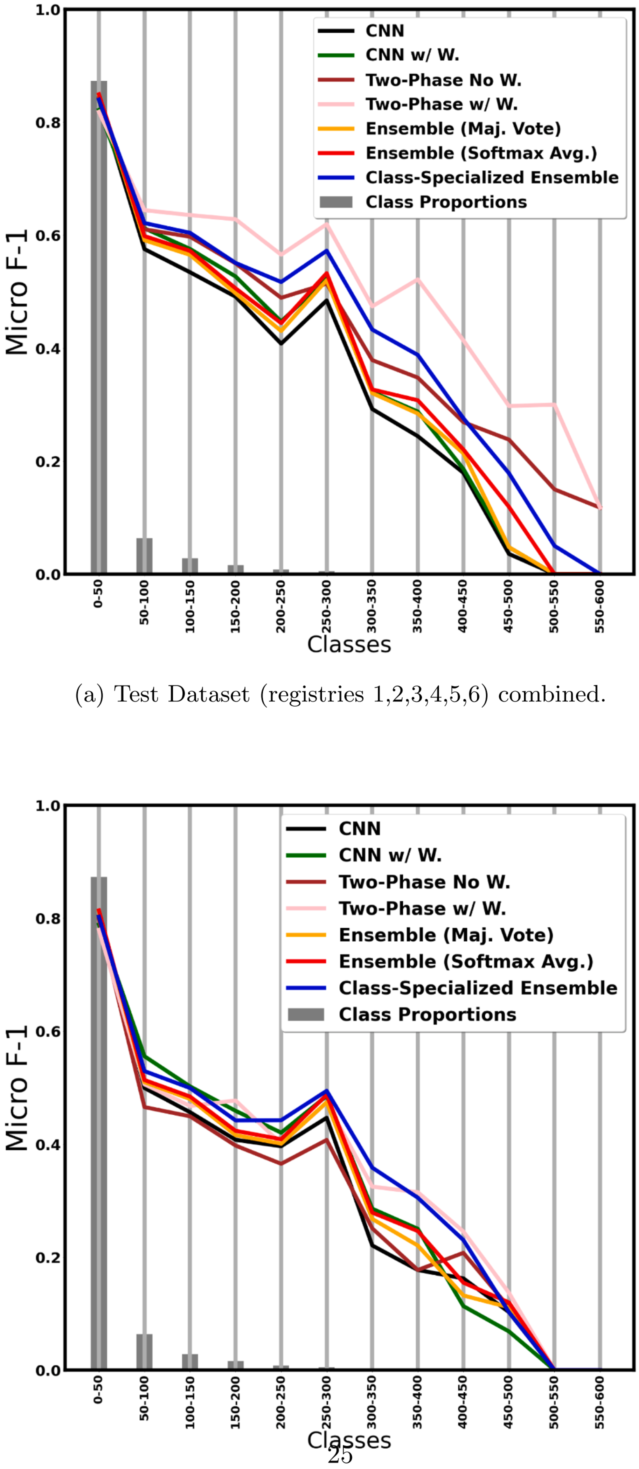
We also compared the performance gap between test and OOD scores for all models (Table 4). Higher values in this table indicates lack of robustness. We observed that two-phase learning led to the highest performance gap. The implication of this result is that methods like two-phase learning can be highly misleading when trained in a lab setting without access to an OOD dataset because lab results may not correlate with real world performance. The table also shows the baseline CNN with class weights resulted in the smallest performance gap.

During our experiments, we found that undersampling is not an effective technique for the classification of cancer pathology reports. Discarding documents from the majority classes diminishes the model micro F1 scores to non-permissible levels with no significant improvement in macro F1 scores. The results table with different undersampling thresholds are included in Appendix B (Section 8, Table 6).

We note that in task such as ours, with extreme class imbalance and a large number of classes, it is common to observe relatively low macro F1 scores. Classifiers are not able to correctly identify features that are relevant for classes that are extremely rare. This is exacerbated by the lengthiness of cancer pathology reports – on average each report is approximately 700 words in length, so it is difficult to distinguish which words are relevant to a particular class when there are very few samples. In our previous work, we demonstrated that low macro scores persist across different deep learning architectures ([37]).

4.3. Classification Performance in Minority Classes

Fig. 3 shows the micro F1 scores obtained when predicting in different class groups ordered by frequency. In the case of the test dataset, we observed that two-phase learning with class weights clearly outperforms other methods for all the groups, except for the first one (the top 50 classes). However, for the OOD dataset, the differences in performance becomes smaller and class-specialized ensemble outperforms other methods for some of the groups. We note that the top 50 classes drive most of the micro F1 score, and we observe that models which excel in this group often show lower performance in the rest of the groups.



(a) Test Dataset (registries 1,2,3,4,5,6) combined.

(b) OOD Dataset (Registry 7)

Fig. 3. Class group performance for the histology task using registry 7 as the OOD dataset. Classes are ordered by frequency which is shown by the gray bars.

Table 4
Absolute differences in micro and macro scores between the corresponding test scores and the OOD score. Bold values represent the largest differences (lack of robustness) while underlined values represent the smallest differences.

Model	Histology		Subsite	
	Test-OOD Mic	Test-OOD Mac	Test-OOD Mic	Test-OOD Mac
CNN	2.57	5.36	4.16	7.35
CNN w/ Class Weights	1.81	4.28	3.76	8.61
Two-Phase	**3.48**	11.92	**5.59**	12.44
Two-Phase w/ Class Weights	3.12	**13.98**	5.48	**15.83**
Ensemble (Maj.Vot.)	2.30	5.91	4.16	8.58
Ensemble (Softmax. Avg.)	2.43	6.17	4.75	9.06
Class-Specialized Ensemble	3.06	8.10	5.08	11.26

Fig. 2a (also see 3.7 for exact percentage values) shows that the top two classes were especially common in the dataset. These two classes have a lot of influence during the training phase because they drive most of the loss function. Models which focus on learning features that are important for top classes will show degraded performance on the rest of the classes while obtaining high F1 micro scores. Therefore, we also analyzed the performance of the models when these two top classes are left out at testing time. The motivation of this experiment was to understand how much bias is involved in the learning process. We were interested in observing which model captures the most characteristics relevant for the non-top classes. The results for this experiment complement the macro F1 score further, and provide a broader insight about models performance on minority classes and their differences in performance with respect to test and OOD datasets.

Table 5 shows the micro and macro F1 scores obtained when predicting in a subset of the datasets which excludes the top two classes. We found that under this experimental setting, the class-specialized

Table 5
Histology Results. Accuracy results when testing in all but the two most frequent classes. The top two classes represent 40.95% of the dataset.

Model	Test Micro	Test Macro	OOD Micro	OOD Macro
CNN	0.6985	0.4090	0.6663	0.3559
CNN w/ Class Weights	0.7118	0.4144	**0.6899**	0.3720
Two-Phase	0.7077	0.4816	0.6590	0.3623
Two-Phase w/ Class Weights	0.6885	**0.5203**	0.6445	0.3797
Ensemble (Maj.Vot.)	0.7077	0.4365	0.6770	0.3779
Ensemble (Softmax. Avg.)	0.7119	0.4453	0.6792	0.3842
Class-Specialized Ensemble	**0.7251**	0.4890	0.6831	**0.3994**

ensemble outperforms other methods in terms of test micro scores and OOD macro scores. Moreover, consistent with previous results, two-phase learning obtains the highest test macro.

5. Discussion

This is the first study which quantifies the performance of the TextCNN for cancer pathology report classification on OOD datasets. In the histology task, we observed drops in performance of up to 3.48% and 13.98% for micro and macro scores, respectively. For the subsite task, the observed values were 5.59% (micro) and 15.83% (macro). These scores demonstrated that the baseline CNN model is not robust under natural distribution shifts when classifying cancer reports.

We note that our methods have different effects in the test and OOD datasets. While we tried to increase the performance of the model on the OOD dataset, we often found that what works well in the OOD dataset also works well (or better) in the test dataset. Thus, there was a consistent gap between the test and OOD dataset. The implication of this result is that improving performance in a closed environment can be highly misleading. That is because improvements in the test dataset do not necessarily correlate with improved performance at deployment time on new data. Other authors have found that larger and diverse datasets can help with robustness [24]. Our models were trained with a large amount of data from different registries across the country, yet we still identified serious drops in performance when predicting on unseen registries.

Through a data profiling analysis, we detected large class distribution differences of the non-common cancer types. Some of the minority classes which appeared in low proportions in the training dataset can often appear much more frequently on the unseen registries. We argue that these class distribution differences are one of the sources of performance drop when deploying the model, especially in terms of macro score. Thus, we explored techniques developed in prior studies to deal with imbalanced datasets and improve the performance on the rare cancer types. While some common methods such as ROS and RUS are not adequate for our specific problem, we showed that other techniques such as two-phase learning and class weights are efficient.

Our experiments showed that the CNN with class weight effectively improves the macro score and provides the lowest difference when comparing the test and OOD scores. The downside of this method was that the micro scores were relatively lower. Using two-phase learning pushed the macro scores further without degrading the micro score performance. In fact, this method provided the highest test macro scores among all experiments. As one would expect when combining multiple models, ensemble methods had the highest micro scores across the board and improved the OOD macro. Finally, our novel class-specialized ensemble method, inspired by the mixture-of-experts model, obtained the highest OOD macro score while maintaining competitive test macro and micro scores.

For two-phase learning, the increase in macro scores was highly disproportional when comparing the test and OOD scores; that is, the increase in OOD macro is relatively low. We hypothesize that the performance of this technique is highly dependent on the class distribution of minority class documents. Two-phase learning is able to capture more features that are relevant for the classification of non-majority classes. Because of the large differences in class distribution with respect to the OOD dataset, the OOD macro score did not experience a meaningful improvement. Another possibility is that two-phase learning is overfitting on the minority documents, which leads to low generalization power.

Our methods comparison provided insight that can help other machine learning practitioners deal with extreme class imbalance and issues caused by natural distribution shifts. Based on our results, we provide the following four recommendations: 1) if micro score or accuracy is the main concern independently of the model's performance in minority classes, then using traditional ensembles is appropriate, 2) if classification of minority classes is a priority, we recommend using our class-specialized ensemble implementation, 3) if one is primarily concerned with minimizing differences in performance between the test and OOD datasets, then simply using the CNN with class weights may be sufficient, and 4) if one cannot afford the computational cost of ensemble methods, then two-phase learning is a simple and effective option. However, the high test macro scores can be misleading in terms of the robustness and generalization of the model.

We acknowledge that our study lacks a formal analysis of the statistical significance of our results. The main reason for this design choice is that we used large datasets (∼2 million documents) which made the experiments computationally expensive (a single model trained on six registries and tested on the seventh unseen registry takes ∼8hs). We considered statistical tests such as the McNemar's test, which are common in circumstances where training multiple models is expensive. However, this test computes a p-value based on the differences in class predictions between two models, and we observed that for the tasks considered in this research (with large number of classes) the p-value is too close to 0. Nevertheless, the main results of this paper (Table 2, 3) are the average of seven individual registries which exhibit the same pattern (micro and macro scores for each individual registry are included in Appendix A). Therefore, we feel confident of the validity of our results.

6. Conclusion

The issue of performance drop at deployment time is complex, and the source of the problem is likely due to multiple factors. In this study, we showed that natural distribution shifts degrade the performance of a TextCNN for the task of classifying cancer pathology reports. We particularly focused on improving the performance of the model on rare cancer types, increasing the macro scores of the model. We presented a novel version of ensemble learning in which each model learns features that are relevant for a specific group of classes. Our class-specialized ensemble model outperformed other techniques implemented in this paper in terms of OOD macro scores while obtaining competitive test macro and micro scores. Our results helped formulate a series of suggestions for other machine learning practitioners working with highly imbalanced datasets and robustness issues.

Our methods are computationally intensive due to ensembles of models operating on millions of pathology reports from seven different states. Although computation at this scale is feasible for supercomputing centers, the average medical center may not have the same volume of data and therefore require the same computational capabilities. Additionally, the majority of resources are needed for training and not for applying the models. And if the trained models are more robust to begin with, less frequent training and deployments are needed. This is of particular importance for models integrated in clinical workflows where minimization of system downtimes and clinical disruptions is a target.

The results presented in this paper form the basis for future research in model robustness on out-of-distribution clinical text. We showed that the class distribution of rare cancers varies widely across different registries, which translates into diminished performance on minority classes and low macro scores. We hypothesize that distinct vocabulary patterns that are unique to individual registries can also contribute to the disparity in performance. In addition, we showed that ensemble methods outperform single models even when testing on OOD datasets. A natural question that follows is whether a model distilled from the ensemble can maintain the performance advantage over baseline models. This would enable us to obtain similar robustness levels while enjoying the low-resource advantages of a single model.

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.

Acknowledgment

This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the US Department of Energy (DOE) Office of Science and the National Nuclear Security Administration.

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by DOE and the NCI 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.

The collection of cancer incidence data used in this study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; Centers for Disease Control and Prevention's (CDC) National Program of Cancer Registries, under cooperative agreement 5NU58DP006344; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201800032I awarded to the University of California, San Francisco, contract HHSN261201800015I awarded to the University of Southern California, and contract HHSN261201800009I awarded to the Public Health Institute. The ideas and opinions expressed herein are those of the author(s) and do not necessarily reflect the opinions of the State of California, Department of Public Health, the National Cancer Institute, and the Centers for Disease Control and Prevention or their Contractors and Subcontractors.

KCR data were collected with funding from NCI Surveillance, Epidemiology and End Results (SEER) Program (HHSN261201800013I), the CDC National Program of Cancer Registries (NPCR) (U58DP00003907) and the Commonwealth of Kentucky.

LTR data were collected using funding from NCI and the Surveillance, Epidemiology and End Results (SEER) Program (HHSN261201800007I), the CDC's National Program of Cancer Registries (NPCR) (NU58DP006332-02–00) as well as the State of Louisiana.

NJSCR data were collected using funding from NCI and the Surveillance, Epidemiology and End Results (SEER) Program (HHSN261201300021I), the CDC's National Program of Cancer Registries (NPCR) (NU58DP006279-02–00) as well as the State of New Jersey and the Rutgers Cancer Institute of New Jersey.

New Mexico Tumor Registry's participation in this project was supported by Contract HHSN261201800014I, Task Order HHSN26100001 from the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) Program.

The Cancer Surveillance System is supported by the National Cancer Institute's SEER Program (Contract Award HHSN261291800004I) and with additional funds provided by the Fred Hutchinson Cancer Research Center.

The Utah Cancer Registry is funded by the National Cancer Institute's SEER Program, Contract No. HHSN261201800016I,and the US Centers for Disease Control and Prevention's National Program of Cancer Registries, Cooperative Agreement No. NU58DP0063200, with additional support from the University of Utah and Huntsman Cancer Foundation.

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the DOE Office of Science under Contract No. DE-AC05-00OR22725.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103957.

References

[1] A. D`Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M.D. Hoffman, F. Hormozdiari, N. Houlsby, S. Hou, G. Jerfel, A. Karthikesalingam, M. Lucic, Y. Ma, C. McLean, D. Mincu, A. Mitani, A. Montanari, Z. Nado, V. Natarajan, C. Nielson, T.F. Osborne, R. Raman, K. Ramasamy, R. Sayres, J. Schrouff, M. Seneviratne, S. Sequeira, H. Suresh, V. Veitch, M. Vladymyrov, X. Wang, K. Webster, S. Yadlowsky, T. Yun, X. Zhai, D. Sculley, Underspecification presents challenges for credibility in modern machine learning (2020). arXiv:2011.03395.

[2] S. Gao, J.X. Qiu, M. Alawad, J.D. Hinkle, N. Schaefferkoetter, H.-J. Yoon, B. Christian, P.A. Fearn, L. Penberthy, X.-C. Wu, L. Coyle, G. Tourassi, A. Ramanathan, Classifying cancer pathology reports with hierarchical self-attention networks, Artif. Intell. Med. 101 (2019) 101726, https://doi.org/10.1016/j.artmed.2019.101726, https://www.sciencedirect.com/science/article/pii/S0933365719303562.

[3] S. Gao, M.T. Young, J.X. Qiu, H.-J. Yoon, J.B. Christian, P.A. Fearn, G.D. Tourassi, A. Ramanthan, Hierarchical attention networks for information extraction from cancer pathology reports, J. Am. Med. Inform. Assoc. 25 (3) (2017) 321–330, arXiv:https://academic.oup.com/jamia/article-pdf/25/3/321/34150614/ocx131.pdf, doi:10.1093/jamia/ocx131. doi: 10.1093/jamia/ocx131.

[4] S. Gao, A. Ramanathan, G. Tourassi, Hierarchical convolutional attention networks for text classification, in: Proceedings of The Third Workshop on Representation Learning for NLP, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 11–23. doi:10.18653/v1/W18-3002. URL https://www.aclweb.org/anthology/W18-3002.

[5] K. De Angeli, S. Gao, M. Alawad, H.-J. Yoon, N. Schaefferkoetter, X.-C. Wu, E. B. Durbin, J. Doherty, A. Stroup, L. Coyle, L. Penberthy, G. Tourassi, Deep active learning for classifying cancer pathology reports, BMC Bioinformatics 22 (1) (2021) 113, https://doi.org/10.1186/s12859-021-04047-1.

[6] K. Stacke, G. Eilertsen, J. Unger, C. Lundström, Measuring domain shift for deep learning in histopathology, IEEE Journal of Biomedical and Health Informatics 25 (2) (2021) 325–336, https://doi.org/10.1109/JBHI.2020.3032060.

[7] X. Guo, Y. Yin, C. Dong, G. Yang, G. Zhou, On the class imbalance problem, in: 2008 Fourth International Conference on Natural Computation, Vol. 4, 2008, pp. 192–201. doi:10.1109/ICNC.2008.871.

[8] E. Rendón, R. Alejo, C. Castorena, F.J. Isidro-Ortega, E.E. Granda-Gutiérrez, Data sampling methods to deal with the big data multi-class imbalance problem, Applied Sciences 10 (4). doi:10.3390/app10041276. https://www.mdpi.com/2076-3417/10/4/1276.

[9] A. Ali, S.M. Shamsuddin, A. Ralescu, Classification with class imbalance problem: A review 7 (2015) 176–204.

[10] J. Johnson, T. Khoshgoftaar, Survey on deep learning with class imbalance, Journal of Big Data 6 (2019) 27, https://doi.org/10.1186/s40537-019-0192-5.

[11] C. Bellinger, R. Corizzo, N. Japkowicz, Remix: Calibrated resampling for class imbalance in deep learning, CoRR abs/2012.02312. arXiv:2012.02312. URL https://arxiv.org/abs/2012.02312.

[12] D. Masko, P. Hensman, The impact of imbalanced training data for convolutional neural networks, 2015.

[13] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: One-sided selection, in: Proceedings of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann, 1997, pp. 179–186.

[14] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano, Experimental perspectives on learning from imbalanced data, in: Proceedings of the 24th International Conference on Machine Learning, ICML '07, Association for Computing Machinery, New York, NY, USA, 2007, p. 935–942. doi:10.1145/1273496.1273614. doi: 10.1145/1273496.1273614.

[15] S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A.S. Kaseb, K. Gauen, R. Dailey, S. Aghajanzadeh, Y.-H. Lu, S.-C. Chen, M.-L. Shyu, Dynamic sampling in convolutional neural networks for imbalanced data classification, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018, pp. 112–117, https://doi.org/10.1109/MIPR.2018.00027.

[16] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357, https://doi.org/10.1613/jair.953.

[17] C. Ling, C.X. Ling, C. Li, Data mining for direct marketing: Problems and solutions, in: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), AAAI Press, 1998, pp. 73–79.

[18] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-smote: A new over-sampling method in imbalanced data sets learning, in: D.-S. Huang, X.-P. Zhang, G.-B. Huang (Eds.), Advances in Intelligent Computing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 878–887.

[19] A. Fernández, S. García, F. Herrera, N.V. Chawla, Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary, J. Artif. Int. Res. 61 (1) (2018) 863–905.

[20] L. Torgo, R.P. Ribeiro, B. Pfahringer, P. Branco, Smote for regression, in: L. Correia, L.P. Reis, J. Cascalho (Eds.), Progress in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 378–389.

[21] R. Blagus, L. Lusa, Smote for high-dimensional class-imbalanced data, BMC bioinformatics 14 (2013) 106, https://doi.org/10.1186/1471-2105-14-106.

[22] C. Padurariu, M.E. Breaban, Dealing with data imbalance in text classification, Procedia Computer Science 159 (2019) 736–745, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019. doi:https://doi.org/10.1016/j.procs.2019.09.229. URL https://www.sciencedirect.com/science/article/pii/S1877050919314152.

[23] H. Lee, M. Park, J. Kim, Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning, in: 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 3713–3717, https://doi.org/10.1109/ICIP.2016.7533053.

[24] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, L. Schmidt, Measuring robustness to natural distribution shifts in image classification (2020). arXiv: 2007.00644.

[25] J. Miller, K. Krauth, B. Recht, L. Schmidt, The effect of natural distribution shift on question answering models (2020). arXiv:2004.14444.

[26] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, M. Hardt, Test-time training with self-supervision for generalization under distribution shifts, in: H.D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 9229–9248. URL http://proceedings.mlr.press/v119/sun20b.html.

[27] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, J. Gilmer, The many faces of robustness: A critical analysis of out-of-distribution generalization (2020). arXiv:2006.16241.

[28] J. Djolonga, J. Yung, M. Tschannen, R. Romijnders, L. Beyer, A. Kolesnikov, J. Puigcerver, M. Minderer, A. D'Amour, D. Moldovan, S. Gelly, N. Houlsby, X. Zhai, M. Lucic, On robustness and transferability of convolutional neural networks (2021). arXiv:2007.08558.

[29] J. Wu, X. Li, X. Ao, Y. Meng, F. Wu, J. Li, Improving robustness and generality of nlp models using disentangled representations (2020). arXiv:2009.09587.

[30] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks, Nature Machine Intelligence 2 (11) (2020) 665–673, https://doi.org/10.1038/s42256-020-00257-z.

[31] R. Polikar, Ensemble learning, in: Ensemble machine learning, Springer, 2012, pp. 1–34.

[32] L. Breiman, Bagging predictors, Machine learning 24 (2) (1996) 123–140.

[33] T. Miyato, S.-I. Maeda, M. Koyama, S. Ishii, Virtual adversarial training: a regularization method for supervised and semi-supervised learning, IEEE transactions on pattern analysis and machine intelligence 41 (8) (2018) 1979–1993.

[34] Z.-H. Zhou, Ensemble learning, Encyclopedia of biometrics 1 (2009) 270–273.

[35] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, arXiv preprint arXiv:1701.06538.

[36] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network (2015). arXiv:1503.02531.

[37] M. Alawad, S. Gao, J.X. Qiu, H.J. Yoon, J. Blair Christian, L. Penberthy, B. Mumphrey, X.-C. Wu, L. Coyle, G. Tourassi, Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks, Journal of the American Medical Informatics Association 27 (1) (2019) 89–98. arXiv:https://academic.oup.com/jamia/article-pdf/27/1/89/34152435/ocz153.pdf, doi:10.1093/jamia/ocz153. doi: 10.1093/jamia/ocz153.

[38] M. Alawad, S. Gao, J. Qiu, N. Schaefferkoetter, J.D. Hinkle, H.-J. Yoon, J. B. Christian, X.-C. Wu, E.B. Durbin, J.C. Jeong, I. Hands, D. Rust, G. Tourassi, Deep transfer learning across cancer registries for information extraction from pathology reports, in: 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), 2019, pp. 1–4, https://doi.org/10.1109/BHI.2019.8834586.

[39] G.K. Savova, I. Danciu, F. Alamudun, T. Miller, C. Lin, D.S. Bitterman, G. Tourassi, J.L. Warner, Use of natural language processing to extract clinical cancer phenotypes from electronic medical records, Cancer Research 79 (21) (2019) 5463–5470. arXiv:https://cancerres.aacrjournals.org/content/79/21/5463.full.pdf, doi:10.1158/0008-5472.CAN-19-0579. https://cancerres.aacrjournals.org/content/79/21/5463.

[40] L. Yao, C. Mao, Y. Luo, Clinical text classification with rule-based features and knowledge-guided convolutional neural networks (2018). arXiv:1807.07425.

[41] M. Hughes, I. Li, S. Kotoulas, T. Suzumura, Medical text classification using convolutional neural networks, Studies in Health Technology and Informatics 235. doi:10.3233/978-1-61499-753-5-246.

[42] A. Rios, R. Kavuluru, Convolutional neural networks for biomedical text classification: Application in indexing biomedical articles, in: Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics, BCB '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 258–267. doi:10.1145/2808719.2808746. doi: 10.1145/2808719.2808746.

[43] B. He, Y. Guan, R. Dai, Classifying medical relations in clinical text via convolutional neural networks, Artificial Intelligence in Medicine 93 (2019) 43–49, extracting and Processing of Rich Semantics from Medical Texts. doi:10.1016/j.artmed.2018.05.001. https://www.sciencedirect.com/science/article/pii/S0933365717305523.

[44] H.S. Yahia, A. Abdulazeez, Medical text classification based on convolutional neural network: A review, International Journal of Science and Business 5 (3) (2021) 27–41, https://EconPapers.repec.org/RePEc:aif:journl:v:5:y:2021:i:3:p:27-41.

[45] S. Gao, M. Alawad, M.T. Young, J. Gounley, N. Schaefferkoetter, H.-J. Yoon, X.-C. Wu, E.B. Durbin, J. Doherty, A. Stroup, L. Coyle, G.D. Tourassi, Limitations of transformers on clinical text classification, IEEE Journal of Biomedical and Health Informatics (2021) 1, https://doi.org/10.1109/JBHI.2021.3062322.

[46] SEER, Icd-0-3 seer site/histology validation llist (2020). https://seer.cancer.gov/icd-o-3/.

[47] J.X. Qiu, H.-J. Yoon, P.A. Fearn, G.D. Tourassi, Deep learning for automated extraction of primary sites from cancer pathology reports, IEEE Journal of Biomedical and Health Informatics 22 (1) (2018) 244–251, https://doi.org/10.1109/JBHI.2017.2700722.