High-Confidence Labelling of Pathology Reports using LLM-Based Unanimous Ensembles with Limited Data

Thomas Greatrix[0009−0000−5354−0738], Frank C. Langbein[0000−0002−3379−0323], Roger M. Whitaker[0000−0002−8473−1913], Gualtiero B. Colombo[0000−0001−5883−8517], and Liam D. Turner[0000−0003−4877−5289]

School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, Wales
{GreatrixT,langbeinfc,whitakerrm,colombog,turnerl9}@cardiff.ac.uk

Abstract. Manual labelling of pathology reports is a costly bottleneck for medical data analysis. We propose diverse unanimous ensembles, integrating Large Language Models (LLMs) like GPT-4o with complementary model architectures, for high-confidence automatic labelling of pathology reports, particularly addressing the challenge of labelled training data scarcity. This consensus method yields high precision on an automatically identifiable subset while simultaneously flagging ambiguous cases requiring expert review. Applying this to the public TCGA-Reports dataset, a GPT-4o and DistilBERT ensemble achieved 95.5% accuracy on the 45.5% subset representing a 23.1 percentage point increase over the baseline DistilBERT's overall accuracy on the full dataset. This demonstrates potential for cost-effective data annotation by automatically labelling high-confidence subsets, thereby reserving human effort for ambiguous cases.

Keywords: Automated Labelling · Ensemble Learning · Large Language Models (LLMs) · Natural Language Processing (NLP) · Machine Learning

1 Introduction

Large volumes of unlabelled data exist in many real-world domains, but acquiring labels is often prohibitively expensive or impractical. This is particularly prevalent in healthcare, wherein detailed records and reports, such as pathology reports, are created and stored over decades. Obtaining structured labels from these free-text reports requires manual annotation by medical professionals. This task is complicated by the fact that the data is often incomplete, messy, unstandardized, aimed at humans, and reporting practices may have changed over time. Consequently, manual labelling by domain experts represents a significant bottleneck due to its high cost and time requirements. Unfortunately, this means that large amounts of data are functionally inaccessible for analysis due to the lack of appropriate labels.

The potential value of these large data sources has created a demand for machine learning methods that can achieve high accuracy with low amounts of training data to mitigate the labelling bottleneck. This, in turn, has raised questions on the viability of weakly supervised approaches in low-data regimes to aid in automated classification. While machine learning models can achieve reasonable performance, obtaining high prediction confidence often desired for clinical applications or reliable downstream analysis remains a significant hurdle. Common techniques used in such scenarios include fine-tuned, pre-trained language models like BERT [6,20] and LLM-assisted labelling [29]. Unfortunately, whilst such methods often work well when the data is clean and relatively standardized, their performance often diminishes when faced with the complexities of real-world clinical texts, such as heterogeneity, incompleteness, and evolving standards. Achieving the high levels of accuracy required for reliable subsequent analysis remains a challenge with these methods alone, especially under severe data scarcity.

To address this, we propose using machine learning ensembles–techniques combining multiple models for a final prediction—aiming for the improved accuracy, robustness, and generalization they typically offer over single models. Specifically, we propose heterogeneous unanimous voting ensembles, leveraging the broad knowledge of pre-trained LLMs (GPT-4o) with models adapted specifically to the task (DistilBERT [25], Dense Neural Networks (DNNs), and Support Vector Machines (SVMs) [5]). This has been less explored for automatically classifying reports when trained from a small set of manually labelled data.

Unanimous ensembles operate on a principle of high confidence through consensus: they only produce a label for a given sample if all constituent models agree on the prediction. While this means the ensemble abstains from labelling samples where there is disagreement, it allows for potentially very high accuracy on the subset of samples it *does* label. Such a trade-off is desirable when large amounts of unlabelled data are available, as automatically labelling even a smaller portion of the dataset accurately gives access to a high-confidence labelled dataset. Such datasets tend to be more useful for real-world applications, and depending on the size and accuracy of the automatically labelled subset, these may be used for subsequent machine learning, analysis, and data mining.

To show that our technique has potential for real-world application, we test it on the TCGA-Reports dataset [14], a publicly available dataset consisting of $9,523$ labelled pathology reports from The Cancer Genome Atlas. This dataset is challenging for machine learning approaches for several reasons:

- **Label scarcity:** many labels appear less than ten times in the entire dataset;
- **High cardinality:** the number of distinct labels is large;
- **Data heterogeneity:** the reports are messy, not rigidly following a specific standardisation.

We limit ourselves to only using 500 randomly sampled reports as training data. This represents a realistic constraint reflecting the practical limitations of manual labelling by medical professionals.

Our experiments using unanimous ensembles containing GPT-4o on this task demonstrate their effectiveness. In isolation, a DistilBERT model achieves an average accuracy of 72.4% on the full test set of this dataset, while a unanimous ensemble of GPT-4o and DistilBERT achieves an accuracy of 95.5% on the subset of reports where both models agree (constituting 45.5% of the test data, as detailed in Section 4). Crucially, while DistilBERT evaluated *only* on this specific subset also achieves 95.5% accuracy, its overall performance is much lower. The value of the unanimous ensemble lies in its ability to *identify* this high-certainty subset automatically, without prior knowledge of which samples are 'easy' or 'hard' for a given model.

Knowing where the models in an ensemble agree also means we know where they disagree. Conversely, where the ensemble disagrees and so abstains, highlights samples that are challenging for the constituent models. These abstained samples are candidates for manual review, allowing human experts to focus their efforts effectively. The ability to differentiate between high-confidence automated predictions and uncertain cases requiring review is exceptionally useful, especially in the medical context. This represents an easily understandable metric for prediction reliability, arguably more interpretable than confidence values or prediction probabilities from individual models, which can be difficult to interpret without statistical expertise.

Finally, it is noteworthy that the inclusion of GPT-4o does not just boost the accuracy of DistilBERT–GPT-4o ensembles, but also tends to improve the performance of other ensembles in which it is included. To show this, we ran a series of experiments wherein we tested a collection of different unanimous ensembles on the aforementioned medical labelling task (see Section 4). We hypothesise this is due to the distinct knowledge encoded within GPT-4o's extensive pretraining, complementing the knowledge learned by models fine-tuned or trained solely on the limited task-specific data.

2 Background

2.1 The TCGA-Reports Dataset

The TCGA-Reports dataset comprises 9,523 anonymized pathology reports originally sourced from The Cancer Genome Atlas (TCGA) database. Kefeli et al. [14] extracted these reports, processed them into a more accessible format using optical character recognition (OCR) on the PDF files and custom post-processing, and performed a proof-of-principle cancer-type classification experiment. These reports have a variety of labels attached, namely: age, ethnicity, gender, primary diagnosis, primary site, disease type, patient ID, project ID, project name, and report ID. Many of these labels can be dependent on each other such as the primary diagnosis and disease type. The challenge posed by this dataset is to accurately predict the labels associated with each free-text report. In this paper, we focus on classifying the reports based on the 'primary diagnosis' label with 128 classes. The original work [14] considered proof-of-concept

cancer-type classification with 32 classes over the full dataset, achieving an average test-set AU-ROC (area under receiver-operating characteristic curve) of 0.992 and AU-PRC (area under precision-recall curve) of 0.90.

2.2 Labelling Text Data with Limited Training Data

Many of the approaches associated with labelling text data with limited training data focus on expanding the available training data. Traditionally, a common technique for labelling textual data with highly limited train data is active learning [22]. However, such approaches can be brittle and dependent on model structure [12]. Other techniques typically used in such situations include distant supervised learning [24], which is reliant on having other relevant datasets available. Data augmentation [31] is a further technique, but generating high-quality and diverse augmented or synthetic data is notoriously difficult for text data, particularly in specialized domains like medicine where preserving clinical meaning is crucial. Leveraging clinician expertise to develop string-matching algorithms for AI labeling has been explored [7], but these methods may struggle with the variability and ambiguity of clinical language. It is worth noting that even with small training sets, clinical NLP systems often perform well [13,19,26,28], and achieve diminishing returns from being provided additional data rapidly. As such, data augmentation techniques may not always be necessary for labelling clinical text.

2.3 Language Models for Labelling Data

Before the wider introduction of LLMs, BERT [6] represented the state of the art for a large range of NLP tasks. BERT is a pre-trained transformer-based model designed to excel at a variety of NLP tasks. It was soon discovered that BERT could be fine-tuned with relatively little data to achieve high accuracy on new NLP tasks [20]. However, as seen in this paper, fine-tuning BERT may not always result in sufficiently high accuracy on NLP tasks, especially when training data is limited or domain-specific challenges are significant [9].

Subsequently, LLMs emerged, offering capabilities suited to low-data scenarios where models like BERT may struggle. By leveraging zero-shot and few-shot learning [4], LLMs demonstrate effective and efficient text data labelling, often achieving high agreement with human annotations [10]. This success has driven their use for labelling tasks, including medical report annotation [2,11].

The usage of ensembles that contain LLMs has also been explored. Ensembles of multiple LLMs have shown increased performance on various tasks, including medical NLP focussed tasks [32]. However, the use of heterogeneous ensembles containing LLMs in conjunction with another type of model, such as BERT or SVMs, for medicine-related tasks is limited. Outside of medicine, such diverse ensembles have been used for complex NLP tasks such as detecting AI-generated text [8] and attribute value extraction [32]. More recent papers suggest that repeated search with LLMs can also be used to get better responses over a range of tasks [3]. Our work builds on these ideas by exploring unanimous ensembles of

diverse model types, including an LLM, specifically for high-confidence labelling in a low-data medical context.

3 Method

This section details the methodology employed to achieve high-confidence labelling of pathology reports using limited data. The general approach involves training four distinct classifiers (SVM, DNN, DistilBERT, and predictions from GPT-4o), chosen for their diverse architectures, and combining them using a unanimous ensemble. The ensembling technique employed is unanimous ensembling, a consensus-based approach, wherein a final prediction is returned if and only if all constituent models in the ensemble agree on the predicted label. If disagreement occurs for a sample, the ensemble abstains. Whilst abstention means it is unlikely the entire dataset is labelled, there are key advantages in high-confidence scenarios:

- **High Precision:** By requiring consensus, a very high accuracy is often achieved on the subset of reports for which the ensemble makes predictions.
- **Uncertainty Identification:** As the accuracy of the unanimous ensemble predictions is expected to be very high, samples for which there is no unanimous agreement are effectively flagged as ambiguous or challenging cases, requiring expert review.

We employ unanimous ensembling, rather than strategies like majority voting (see Appendix A.5), because our primary goal is maximizing precision on the labelled subset, accepting abstention on ambiguous cases, to generate labels suitable for high-confidence applications.

In order to test the effectiveness of our ensemble approach, we seek to predict the 'primary diagnosis' tag in the TCGA-Reports dataset [14] with 128 classes. The primary diagnosis is the specific cancer type or condition that the patient in the report is being treated for. This label was chosen for its clinical relevance and as a challenging multi-class classification task suitable for evaluating methods under label scarcity.

To train the models, we randomly shuffled the TCGA-Reports using a set seed, and then selected the first 500 reports as the training set, and the last $1,000$ reports as the held-out test set (see Appendix A.2 for results with smaller training sets). To train a dense neural network (DNN) and a support vector machine (SVM), the reports were then embedded using the pre-trained BioMistral-7B model [18] to generate suitable vector representations associated with the primary diagnosis label. Each embedding vector was of size 4096.

The SVM model used a linear kernel with regularisation $C = 1$ (the default for sklearn [23]). The DNN consists of three layers with ReLu activation of sizes 2056, 1028, and 512 respectively. Dropout with a rate of 0.3 was added after each layer [27], a commonly used value to prevent overfitting. The neural network model used categorical cross-entropy loss and was optimized using ADAM [15], with learning rate 0.001 (the default for TensorFlow [1]). The DNN

was trained with a batch size of 8 over 50 epochs. We fine-tuned DistilBERT on the aforementioned 500 training set reports. DistilBERT has its own tokenizer and embedding layer, processing the text reports directly. In situations where the reports were too long for DistilBERT's context, the reports were truncated to the first 512 characters. DistilBERT was fine-tuned for 30 epochs with a batch size of 16, using the AdamW optimiser [21] with a learning rate of 2×10^{-5} (the default learning rate). The best model at the end was loaded for use in making predictions, based on the model's accuracy on the training data. We conducted additional experiments on all model's performance over a range of train data sizes and over a range of epochs. We found that performance increased with both the number of epochs and the number of train examples, but diminishing returns were rapidly reached (see Appendix A.2). Appendix A.3 analyses the performance of different BERT models, justifying the choice of DistilBERT. Appendix A.4 explores alternative machine learning approaches, justifying the choice of the SVM model.

DistilBERT was trained and ran using the Transformers package, version 4.41.2. The DNN was trained and ran on tensorflow-keras version 2.18.0. All other models were trained and ran on sklearn version 1.2.2.

To obtain the LLM predictions, we used GPT-4o-2024-05-13 (referred to as GPT-4o) in a zero-shot inference mode based on a prompt. To ensure deterministic outputs, a temperature of 0.01 and a top-p of 0.0 are used. Both, top-p and temperature, control how much randomness is in the LLM's answers, with values closer to 0 being less random. The maximum token output was set to $1,500$ tokens to allow for complete responses where one token is approximately 4 characters. The specific prompt used to instruct GPT-4o for the classification task is provided in Appendix A.1.

4 Results

This section focuses on the results of our experiments evaluating individual models and unanimous ensembles. At a high level, the results indicate that unanimous ensembles consistently outperformed individual models in terms of accuracy when evaluated on the subsets where they made predictions. Furthermore, unanimous ensembles containing GPT-4o generally performed better than those which did not.

Table 1 show the performance metrics of all models considered. To enhance readability, we have given each model a one letter tag (D for DistilBERT, G for GPT-4o, N for the DNN, S the SVM). We denote unanimous ensembles by combining the model tags (e.g., GD is the unanimous ensemble between GPT-4o and DistilBERT). Each model was trained five times on different training and test sets (see Section 4.1).

Figure 1 and Table 1 illustrate a clear performance boost from including GPT-4o in a unanimous ensemble, with the top seven performing entries (the highest accuracy on prediction subset) all containing GPT-4o. A notable jump in average accuracy exists between DNS (85.0%, the best-performing ensemble

Table 1. Accuracy comparison of models and ensembles when trained on 500 samples (except for GPT-4o, which was not trained). D is DistilBERT, G is GPT-4o, N is the DNN, S is the SVM (see Section 3). N_p is the portion of the dataset on which predictions have been made. A_p is the average prediction accuracy, and σ_p is the standard deviation over five runs. In the case of the ensembles, the accuracy and standard deviation is the ensemble's accuracy and standard deviation on only the portion of the dataset where predictions were made. The individual model accuracies and standard deviations are on the entire test dataset, and given to 3 significant figures.

Model/ensemble	N_p	A_p	σ_p
DGNS	197	**96.8**%	0.00943
DGS	260	96.7%	0.0101
DGN	223	96.1%	0.0129
DG	455	95.5%	**0.00562**
GNS	213	94.7%	0.0173
GN	249	93.5%	0.0181
GS	288	93.1%	0.0145

Model/ensemble	N_p	A_p	σ_p
DNS	346	85.0%	0.0146
DN	347	83.2%	0.0204
DS	480	82.2%	0.0187
D	**1,000**	72.4%	0.0120
G	**1,000**	60.0%	0.00741
NS	550	56.8%	0.0102
S	**1,000**	42.5%	0.0187
N	**1,000**	36.9%	0.0448

without GPT-4o) and GS (93.1%, the lowest performing ensemble containing GPT-4o), of 8.1 percentage points in subset accuracy. Additionally, we see a large boost in average accuracy between D (the top performing model) and DS (the second worst performing unanimous ensemble) of 9.8 percentage points. Whilst unanimous ensembles with more than two models (i.e. DNS) do perform better in general than unanimous ensembles with only two models (i.e. DS) the difference is often small, and not statistically significant.

The results also highlight the trade-off between accuracy and coverage. For example, the two-model ensemble DG achieves similar accuracy (95.5%) to the four-model ensemble DGNS (96.8%) whilst covering over 45.5% of the dataset ($N_p = 455$), compared to the less than 20% ($N_p = 197$) covered by DGNS. Crucially, the DG ensemble's 95.5% accuracy on this automatically identified subset significantly surpasses the baseline DistilBERT's overall accuracy of 72.4% across the entire test set, demonstrating the practical gain in prediction reliability. As noted previously, the ensemble's key advantage is automatically identifying a high-confidence subset, unlike simply evaluating the baseline on this subset post-hoc.

4.1 Significance of Using an LLM in Unanimous Ensembles

To enable robust statistical comparison to investigate the effect of including GPT-4o in an ensemble, we created five different train/test splits. Each split was generated by randomly selecting 500 train reports and 1,000 test reports from the full TCGA-Reports dataset after a global shuffle (using different seeds for each split). We trained the SVM, DNN, and DistilBERT models anew on each of the five train sets (see Section 3). We then made and recorded predictions from

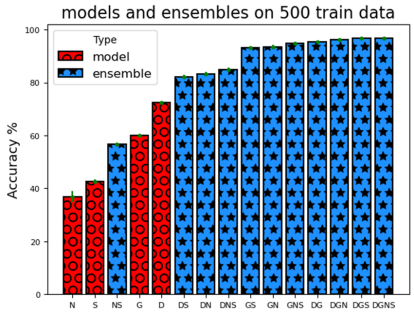

Fig. 1. Accuracy comparison of models and ensembles (trained on 500 samples except for GPT-4o, which was not trained). D is DistilBERT, G is GPT-4o, N is the DNN, S is the SVM (see Section 3). In the case of the ensembles, the accuracy is the ensemble's accuracy on only the portion of the dataset where predictions were made. The individual model accuracies are on the entire test dataset. The y-axis represents subset accuracy for ensembles, overall accuracy for individual models. The green error bars show standard error across five runs for each entry.

each of the four base models (D, G, N, S) on each of the five corresponding test sets (note that GPT-4o was not fine-tuned).

Using these predictions, every possible unanimous ensemble was constructed. We recorded the accuracy achieved (on the subset of agreement for ensembles, and overall for individual models) on each of the five test sets for each model/ensemble resulting in 15 result sets (for four individual models and eleven unanimous ensembles)

Pairwise Mann-Whitney U Test. Due to the small sample size ($n = 5$ runs), standard tests for normality lack reliability. Consequently, we could not assume the accuracy data followed a normal distribution, leading us to employ non-parametric statistical tests. Specifically, we aimed to determine if unanimous ensembles including GPT-4o (of size greater than one) exhibited significantly higher accuracy compared to models/ensembles lacking GPT-4o (of size greater than one). To this end, we performed a series of pairwise, one-sided Mann-Whitney U tests. Each test compared the distribution of accuracy scores (from the 5 runs) for an ensemble containing GPT-4o against the distribution for a model/ensemble without GPT-4o. The null hypothesis (H_0) stated that the distributions were identical, while the alternative hypothesis (H_1) stated that the accuracy distribution for the ensemble containing GPT-4o was stochastically greater than the comparison distribution. Table 2 presents the resulting p-values. Notably, the U-statistic was 0.0 for every test, reflecting complete separation between the accuracy values of the compared groups given the sample size.

Linear Mixed Effects Model. To further investigate the effect of including GPT-4o while accounting for repeated measurements across different ensemble

Table 2. Results (p-values) of our Mann-Whitney U tests. D is DistilBERT, G is GPT-4o, N is the DNN, S is the SVM (see Section 3). Unanimous ensembles are denoted by combining the letter abbreviations, e.g., DG is the unanimous ensemble of DistilBERT and GPT-4o. In each test, we checked to see if the column model/ensembles average is less than the row model/ensemble. All p-values are given to 3 significant figures. The similar values result from the small groups and observing perfect or near perfect separation, with U = 0 for every single test performed. The values of 0.00596 come from situations where there were ties in the dataset.

	N	S	NS	G	D	DS	DN	DNS
GS	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397
GN	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397
GNS	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397
DG	0.00596	0.00596	0.00596	0.00596	0.00596	0.00596	0.00596	0.00596
DGN	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397
DGS	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397
DGNS	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397	0.00397

types, we employed a linear mixed effects (LME) model (using statsmodels version 0.14.4). This accounts for the nested data we are using, however, also makes normality assumptions about the random effects. Accuracy was modelled with the presence/absence of GPT-4o ('Group') as a fixed effect and ensemble identity (N, S, NS, etc.) as a random effect, structured as

$$\text{Accuracy} \sim \text{Group} + \text{Ensemble}.$$

A significant fixed effect was found for Group ($p = 0.00002215 < 0.0001$), supporting that including GPT-4o significantly influences ensemble accuracy.

5 Discussion and Conclusions

Our usage of unanimous ensembles for high-confidence text classification with limited labelled data prioritizes reliability and complements full-dataset classification methods [14] by focusing on maximizing confidence rather than coverage. The unanimous ensembles including GPT-4o automatically identify subsets where very high accuracy (93.1%−96.8%) is achieved, providing high-trust labels essential in certain contexts. For instance, the DistilBERT–GPT-4o combination (DG), performed particularly well, achieving high accuracy (95.5%) whilst still making predictions on a substantial portion (45.5%) of the data.

The enhanced performance of LLM-containing ensembles likely arises from model diversity. LLMs (like GPT-4o) draw on vast pre-training data, while models like DistilBERT are adapted specifically to the small (500-sample) training set. This difference in training suggests distinct knowledge and error profiles. This could explain why the DistilBERT–GPT-4o ensemble (DG, 95.5%) significantly outperforms the DistilBERT–Dense Neural Network ensemble (DN, 83.2%) by about 12 percentage points. The dense neural network, trained on

embeddings from the same limited data, likely shares similar representations and correlated errors with DistilBERT, limiting the ensemble gain compared to the more diverse DistilBERT–GPT-4o combination.

While promising, this study has limitations. Our evaluation uses a single dataset (TCGA-Reports) and task ('primary diagnosis'), and the high confidence achieved via unanimity comes at the cost of reduced data coverage. Furthermore, the best results depend on access to powerful LLMs like GPT-4o.

These findings highlight the potential of using diverse unanimous ensembles, particularly those including large pre-trained models, to generate high-confidence labels for subsets of large datasets, thereby optimizing the use of expert resources for annotating only the ambiguous cases. Future work could explore automated selection of optimal ensemble members and adaptive confidence thresholds.

Acknowledgments. We acknowledge the support of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government and Advanced Research Computing at Cardiff Division. We would also like to acknowledge Piero Gerbino and Yahia Kubrani who both provided input during the creation of the LLM prompts for GPT-4o.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Appendix

A.1 Creating the LLM Prompt

This section details our LLM prompt design, incorporating techniques shown to improve LLM accuracy. We initiated the prompt with role-play ("You are a radiographer") to enhance reasoning and answer quality [17]. Afterwards, we break down the specific task we want the LLM to perform as well as how we want the LLM to perform it in great detail, which is believed to increase the likelihood of the LLM giving a prompt close to what is desired.

We then ask the LLM to provide its reasoning, which leverages principles similar to Chain-of-Thought (CoT) prompting and Zero-shot-CoT, where eliciting step-by-step reasoning or justification before the final answer has improved performance on complex tasks [16,30,33]. We ask the LLM to provide the label predictions within square brackets at the end of the response, which is done to make it easier to locate the LLM prediction afterwards. To ensure use of correct TCGA-Reports labels/formats and prevent custom ones, the full label list was supplied. The prompt concludes by detailing information the model should prioritize, seeking further accuracy improvements. The full prompt follows:

"You are a radiographer. Your task is to identify tumor types from genomic reports. Analyze the provided report and determine the correct tumor type using one of the labels from the list below. A tumor type refers to the classification of

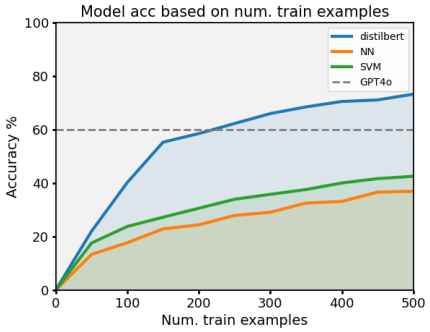

Fig. 2. Comparing model accuracy vs. number of training examples. GPT-4o was *not* fine-tuned nor trained.

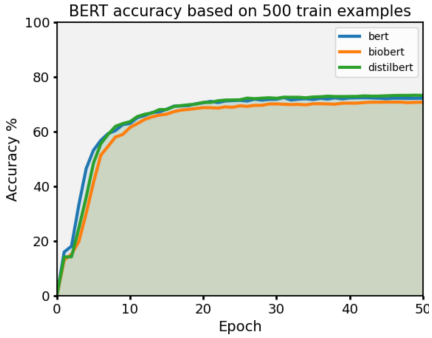

Fig. 3. Comparing model accuracy of different BERT-based models over training epochs.

a tumor based on the characteristics of the cells from which it originates. The presence of mucinous characteristics takes precedence for classification purposes. Provide your reasoning and end your response with the appropriate label in the format [label] with enclosing square brackets. Use only one of the following labels:

`{INSERT LIST OF 128 COMMA-SEPARATED [LABELS]; OMITTED FOR BREVITY}`

Highlight Key Information: Clearly separate and highlight key sections of the report such as Clinical Diagnosis, Specimens Submitted, Diagnosis, and Gross Description. Break Down Diagnosis Details: Break down the diagnosis into simpler points to make it easier to identify the type of tumor. Link Findings to Primary Diagnosis: Explicitly link the findings (e.g., type of carcinoma, size, grade) to the potential disease categories provided. Pay attention to features that specifically capture tumors with histological similarities to synovial tissue."

A.2 Analysing Performance for Different Amounts of Training Data

Since practical applications rarely have fixed training set sizes like 500, we investigated how model performance scales, especially with less data. Figure 2 shows primary diagnosis prediction performance versus training examples. Notably, zero-shot GPT-4o outperformed all models except fine-tuned DistilBERT. Fine-tuning DistilBERT becomes more effective than using GPT-4o beyond approximately 200 training samples.

A.3 Comparing Different Types of BERT Model

There are many variations of BERT that could have been chosen when conducting our experiments. Here, we explore the usage of three different BERT models: the standard BERT model, BioBERT which is fine-tuned on biological information, and DistilBERT which is a smaller BERT model. The results are

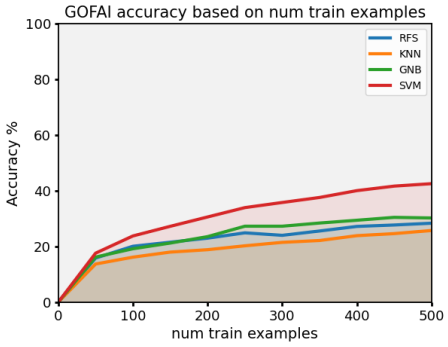

Fig. 4. Comparison of model accuracy of various AI models on primary diagnosis prediction. RFS: Random Forest Search, SVM: Support Vector Machine, GNB: Gaussian Naive Bayes, KNN: K-nearest neighbours.

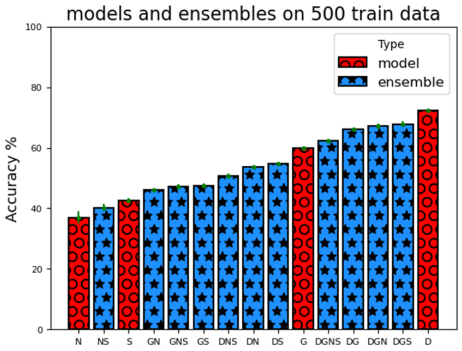

Fig. 5. Comparison of model accuracy for majority voting ensembles trained on 500 samples (GPT-4o was not trained). D: DistilBERT, G: GPT-4o, N: DNN, S: SVM (see Section 3). Green error bars show standard error across five runs.

shown in Figure 3. We note that BERT and DistilBERT have similar performance, whilst BioBERT performs worse in general. As DistilBERT is around 40% smaller than BERT and has similar performance, we find it recommendable for situations with highly limited training data to use DistilBERT over BERT as it is quicker to train and easier to run than BERT.

A.4 Other Machine Learning Models

We tested Random Forest Search (RFS), K-Nearest Neighbours (KNN), Gaussian Naive Bayes (GNB), and support vector machines (SVMs). During our experiments we found that SVMs were the best option for predicting the primary diagnosis from report embeddings, and thus is the one we chose to use when exploring unanimous ensembles. Figure 4 shows that not only was the best approach tested, but it was the best by a substantial margin. It is notable that the KNN model was ran with the number of nearest neighbours, n, equal to 5. SVM used a linear kernel, and that RFS used the default sklearn options.

The data in Figure 4 was collected by averaging the accuracies of 5 instances of each model, each with a different random seed, a different set of training data, and a different set of test data.

A.5 Majority Voting Ensembles

One of the most common types of voting ensemble is the majority voting ensemble. This ensemble allows each model in the ensemble to cast a vote on what it believes the correct answer to be, and the answer with the most votes is returned by the ensemble. Whilst we did experiment with this, the ensembles created did not perform as well as DistilBERT on its own (see Figure 5).

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., *et al.*: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., Sontag, D.: Large language models are few-shot clinical information extractors. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1998–2022 (2022)
3. Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q.V., Ré, C., Mirhoseini, A.: Large language monkeys: Scaling inference compute with repeated sampling. arXiv:2407.21787 (2024). https://doi.org/10.48550/arXiv.2407.21787
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., *et al.*: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901 (2020)
5. Cortes, C.: Support-vector networks. Machine Learning **20**, 273–297 (1995)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018). https://doi.org/10.48550/arXiv.1810.04805
7. Dunnmon, J.A., Ratner, A.J., Saab, K., Khandwala, N., Markert, M., Sagreiya, H., Goldman, R., Lee-Messer, C., Lungren, M.P., Rubin, D.L., et al.: Cross-modal data programming enables rapid medical machine learning. Patterns **1**(2) (2020)
8. El-Sayed, A., Nasr, O.: An ensemble based approach to detecting LLM-generated texts. In: Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association. pp. 164–168 (2023)
9. Gao, S., Alawad, M., Young, M.T., Gounley, J., Schaefferkoetter, N., Yoon, H.J., Wu, X.C., Durbin, E.B., Doherty, J., Stroup, A., Coyle, L., Tourassi, G.: Limitations of transformers on clinical text classification. IEEE Journal of Biomedical Health Informatics **25**(9), 3596–3607 (2021)
10. Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT outperforms crowd-workers for text-annotation tasks. Proceedings of the National Academy of Sciences (PNAS) **120**(26), e2305016120 (2023)
11. Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L.H., Hao, X., Jaber, B., Reddy, S., Kartha, R., et al.: LLMs accelerate annotation for medical information extraction. In: Machine Learning for Health (ML4H). pp. 82–100. PMLR (2023)
12. Hahn, L., Roese-Koerner, L., Cremer, P., Zimmermann, U., Maoz, O., Kummert, A.: On the robustness of active learning. In: Calvanese, D., Iocchi, L. (eds.) Proceedings of the 5th Global Conference on Artificial Intelligence (GCAI). vol. 65, pp. 152–162 (2019)
13. Humbert-Droz, M., Mukherjee, P., Gevaert, O., et al.: Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: case study for the extraction of symptoms from clinical notes. JMIR medical informatics **10**(3), e32903 (2022)
14. Kefeli, J., Tatonetti, N.: TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models. Patterns **5**(3), 100933 (2024)
15. Kingma, D.P., BA, J.: Adam: A method for stochastic optimization. arXiv: 1412.6980 (2014). https://doi.org/10.48550/arXiv.1412.6980
16. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems **35**, 22199–22213 (2022)

17. Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., Dong, X.: Better zero-shot reasoning with role-play prompting. arXiv:2308.07702 (2023). https://doi.org/10.48550/arXiv.2308.07702
18. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: BioMistral: A collection of open-source pretrained large language models for medical domains. arXiv:2402.10373 (2024). https://doi.org/10.48550/arXiv.2402.10373
19. Lamare, J.B., Olatunji, T., Yao, L.: On the diminishing return of labeling clinical reports. arXiv preprint arXiv:2010.14587 (2020)
20. Li, X., Yuan, W., Peng, D., Mei, Q., Wang, Y.: When BERT meets Bilbo: a learning curve analysis of pretrained language model on disease classification. In: IEEE International Conference on Healthcare Informatics (ICHI). pp. 1–2 (2020)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
22. Naseem, U., Khushi, M., Khan, S.K., Shaukat, K., Moni, M.A.: A comparative analysis of active learning for biomedical text mining. Applied System Innovation **4**(1), 23 (2021)
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
24. Roller, R., Stevenson, M.: Making the most of limited training data using distant supervision. In: Proceedings of BioNLP 15. pp. 12–20 (2015)
25. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv:1910.01108 (2019). https://doi.org/10.48550/arXiv.1910.01108
26. Sordo, M., Zeng, Q.: On sample size and classification accuracy: A performance comparison. In: International Symposium on Biological and Medical Data Analysis. pp. 193–201. Springer (2005)
27. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
28. Stember, J.N., Shalu, H.: Deep reinforcement learning with automated label extraction from clinical reports accurately classifies 3d mri brain volumes. Journal of digital imaging **35**(5), 1143–1152 (2022)
29. Wang, Z., Pang, Y., Lin, Y.: Adaptable and reliable text classification using large language models. arXiv:2405.10523 (2024). https://doi.org/10.48550/arXiv.2405.10523
30. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS). pp. 24824–24837 (2022)
31. Yang, D., Parikh, A., Raffel, C.: Learning with limited text data. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. pp. 28–31 (2022)
32. Yang, H., Li, M., Zhou, H., Xiao, Y., Fang, Q., Zhang, R.: One LLM is not enough: Harnessing the power of ensemble learning for medical question answering. medRxiv (2023). https://doi.org/10.1101/2023.12.21.23300380
33. Yugeswardeenoo, D., Zhu, K., O'Brien, S.: Question-analysis prompting improves LLM performance in reasoning tasks. arXiv:2407.03624 (2024). https://doi.org/10.48550/arXiv.2407.03624