BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Validation of large language models for detecting pathologic complete response in breast cancer using population-based pathology reports

Ken Cheligeer[1,2], Guosong Wu[1,3], Alison Laws[4,5], May Lynn Quan[3,4,5], Andrea Li[1], Anne-Marie Brisson[6], Jason Xie[1] and Yuan Xu[1,3,4,5*]

Abstract

Aims The primary goal of this study is to evaluate the capabilities of Large Language Models (LLMs) in understanding and processing complex medical documentation. We chose to focus on the identification of pathologic complete response (pCR) in narrative pathology reports. This approach aims to contribute to the advancement of comprehensive reporting, health research, and public health surveillance, thereby enhancing patient care and breast cancer management strategies.

Methods The study utilized two analytical pipelines, developed with open-source LLMs within the healthcare system's computing environment. First, we extracted embeddings from pathology reports using 15 different transformer-based models and then employed logistic regression on these embeddings to classify the presence or absence of pCR. Secondly, we fine-tuned the Generative Pre-trained Transformer-2 (GPT-2) model by attaching a simple feed-forward neural network (FFNN) layer to improve the detection performance of pCR from pathology reports.

Results In a cohort of 351 female breast cancer patients who underwent neoadjuvant chemotherapy (NAC) and subsequent surgery between 2010 and 2017 in Calgary, the optimized method displayed a sensitivity of 95.3% (95%CI: 84.0–100.0%), a positive predictive value of 90.9% (95%CI: 76.5–100.0%), and an F1 score of 93.0% (95%CI: 83.7–100.0%). The results, achieved through diverse LLM integration, surpassed traditional machine learning models, underscoring the potential of LLMs in clinical pathology information extraction.

Conclusions The study successfully demonstrates the efficacy of LLMs in interpreting and processing digital pathology data, particularly for determining pCR in breast cancer patients post-NAC. The superior performance of LLM-based pipelines over traditional models highlights their significant potential in extracting and analyzing key clinical data from narrative reports. While promising, these findings highlight the need for future external validation to confirm the reliability and broader applicability of these methods.

Keywords Large Language Models (LLMs), Breast cancer, Machine learning in healthcare, Natural language processing, Clinical pathology information extraction, Pathologic Complete Response (pCR)

*Correspondence:
Yuan Xu
yuxu@ucalgary.ca
Full list of author information is available at the end of the article

Introduction

Pathologic complete response (pCR) is characterized by the absence of residual invasive malignant cells, whether with or without in situ disease [1]. It is a known prognostic factor for long-term outcomes of breast cancer patients and is used to guide the course of adjuvant systemic therapy [2–4]. The presence of pCR is not explicitly documented in pathology reports for patients treated with neoadjuvant chemotherapy (NAC); however, utilizing such information is crucial for further reporting and research purposes. Detecting pCR from Electronic Medical Records (EMR) primarily serves the purposes of comprehensive reporting, advancing health research, and supporting public health surveillance. The challenge lies in the fact that pCR data is not always stored in a structured format within EMRs. Consequently, manually extracting pCR from narrative reports is both time-consuming and costly, particularly due to its implicit nature, which creates a significant burden when handling large, population-based datasets.

Using advanced algorithms and data mining techniques can facilitate the pCR detection from unstructured medical text [5]. This method is essential for classifying patients into clinical categories automatically, thereby enabling the detection of the presence of pCR in breast cancer patients to become feasible [6].

Traditional rule-based methods like keyword matching and regular expression-based techniques are straightforward and interpretable within medical contexts [7]. However, their development and maintenance present numerous challenges, such as handling abbreviations, negations, conditional or uncertain statements, and conditions based on their presence or historical context [8–11]. Additionally, the variability in grammatical and syntactic structures of medical texts, along with keyword variations across different cases and institutions, restricts the generalizability of these methods [11]. Furthermore, developing and testing these systems necessitates a deep understanding of medical and clinical knowledge and close collaboration with experts [12]. While expert involvement can enhance system performance, it also results in significant maintenance costs [13].

In response to these challenges, we explored the integration of Large Language Models (LLMs), which are trained with billions of parameters and capable of handling a diverse set of natural language tasks [14]. Representing a promising development in pathology informatics, LLMs have the potential to enhance the comprehension and analysis of pathology reports significantly [15]. Models like bidirectional encoder transformers (BERT) [16] and generative pretrained transformers (GPT) [17] are trained to predict the likelihood of word sequences based on contextual understanding, enabling the generation of coherent and contextually relevant outputs.

To harness the full potential of these LLMs in pathology reports analysis, we incorporate LLM embeddings within our data pipelines. These embeddings are vector representations of text, which capture deep linguistic and semantic relationships. By transforming complex medical jargon into dense numerical vectors, embeddings enable our models to process and analyze text with heightened accuracy and efficiency. The use of embeddings not only improves the contextual comprehension of the text but also enriches the feature set available for training our predictive models, enhancing the overall predictive performance.

Therefore, leveraging both the predictive capabilities and the embeddings generated by LLMs, this study aims to develop and validate robust data pipelines for the identification of pCR. Such advancements could facilitate data collection and analysis automation in research settings, potentially leading to more personalized and timely therapeutic strategies.

Methods

Cohort and data

This retrospective cohort study included all female non-metastatic invasive breast cancer patients who underwent NAC and subsequent curative-intent surgery during their admission at all four tertiary acute care hospitals in Calgary, Alberta, Canada, between 1 January 2010 and 31 December 2017. The study excluded patients with a diagnosis of multiple breast primary tumors. The study followed the Standards for Reporting of Diagnostic Accuracy Study (STARD) [18] and was approved by Health Research Ethics Board of Alberta–Cancer Committee. A waiver of consent was granted.

The patient cohort was identified from the Alberta Cancer Registry (ACR) database which captures all new cancer diagnoses in Alberta. Text data were retrieved from Sunrise Clinical Manager (SCM), which is an EMR system universally applied in all four acute care hospitals in Calgary. Final surgical pathology reports, containing comprehensive biomarker and histopathology evaluations in raw free-text format, were utilized for the development of pCR phenotyping algorithms. Patient text pathology reports without a lymph node evaluation report were excluded because true pCR achievement cannot be verified without lymph node evaluation. All databases then were linked by the patient's personal health care number (PHN) and unique lifetime identifiers (ULIs). Patient records without a valid PHN or ULI were excluded.

Definition and ascertainment of pCR

In our dataset, "pCR" is not explicitly mentioned in all reports. Therefore, to ensure accuracy in identifying

pCR, each pathology report was meticulously reviewed by a breast radiology fellow who served as the gold standard. pCR was defined as the complete absence of residual invasive malignant cells, regardless of the presence of in situ disease [6].

All uncertainties raised from chart review were discussed and resolved to ensure that they satisfied the case definition. The data extraction agreement between the physician and a senior pathologist was tested to confirm the presence of pCR for the first 10 charts, and the result was excellent (kappa = 1).

Pipeline development

To assess the capabilities of LLM embeddings in interpreting natural text within pathology reports, our research encompassed two innovative pipelines. Firstly, we ventured into adapting LLMs within a local computing environment under a healthcare institution firewall to capture the unique linguistic characteristics of pathology texts. Second, our approach expanded to a tailored system, where we embedded LLMs as a crucial layer in a neural network architecture. This network was meticulously fine-tuned to enhance its proficiency in understanding complex medical terminology, demonstrating the diverse potential of LLMs in medical text analysis.

Both pipelines consist of preprocessing, feature extraction, classification, and evaluation process (shown in Fig. 1) and have minor differences in features extraction and classification stage when it comes to using default LLMs and fine-tuned LLMs.

Pipeline A—identify pCR with custom machine learning pipelines using LLMs embedding models on local environments with original patient reports

We evaluate and assess the effectiveness of various pre-trained LLM embedding models, without further fine-tuning at this stage, to determine which ones possess better inherent knowledge of healthcare, particularly in comprehending pathology reports in our local healthcare environment. To achieve this, we utilized locally deployable models from BERT [16], BART [19], T5 [20] and GPT [21] families, which have shown considerable promise in EMR text classification tasks [22, 23].

Preprocessing These locally deployable LLM embedding models have a limitation on the input length of text. To address this limitation, we implemented a strategy that segments longer texts into manageable chunks.

Our chunking method processes the input text by tokenizing it into discrete units, assessing the total token count, and segmenting the text to fit the model's maximum token capacity while maintaining contextual continuity through designed overlaps. Each segment includes special boundary tokens that ensure seamless integration for processing by the specific LLM. This method strikes a balance between computational efficiency and information integrity, thus improving model performance across large text datasets without sacrificing quality.

The overlapping strategy preserves the inherent contextual information that might be lost at the boundaries of individual segments. Segments that do not reach the maximum token capacity are extended with zeros to ensure a consistent input length.

Feature extraction The chunked pathology reports are processed by various transformer-based language models, which generate embeddings for each token. These token embeddings are subsequently analyzed and averaged to produce a consolidated vector for each segment, effectively capturing its core semantic content. Then, we use mean pooling to aggregate these segment vectors into a holistic representation of the entire document. Mean

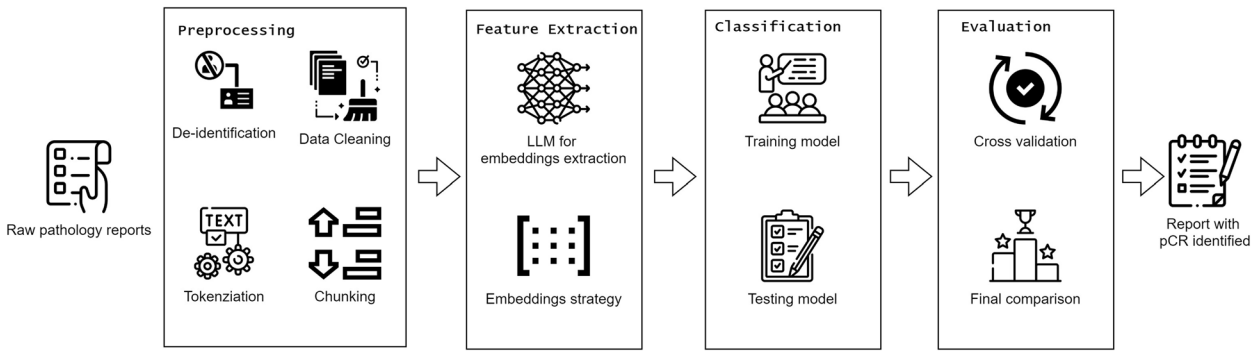

Fig. 1 Analytical pathways for processing text data in pathology reports. *GPT: Generative Pre-trained Transformer; LLM: Large Language Model; ML: Machine Learning; pCR: Pathologic Complete Response. Designed using images from [Flaticon.com]

pooling involves averaging the vectors of all segments, which simplifies the data while retaining important features and reducing noise. The output of this mean pooling is a comprehensive embedding vector that encapsulates the full essence of the text. This vector then served as a foundation for various downstream applications, ensuring that the insights derived were both accurate and reflective of the document's entirety.

Classification The extracted embedding vectors were fed to down streaming binary classifier. To identify whether a pCR presented in the given pathology report can be transferred to a binary classification task. To ensure a comprehensive assessment, the primary dataset was divided into training and testing subsets with an 80–20 split, stratified by class label (pCR vs. non-pCR). This strategic split was crucial to ensure the models achieved both high accuracy and robust generalizability to unseen data. After processing the data for classification, we employed fivefold cross-validation on the training set to ensure a robust evaluation of our model's performance. Additionally, we use Bayesian optimization [24] to choose optimal hyperparameters for a logistic regression classifier. This method builds a probabilistic model of the objective function, guiding the selection of hyperparameters like regularization strength, learning rate, and tolerance. The data flow for model selection and evaluation is illustrated in Fig. 2.

In addition, we noted that the number of patients with the presence of pCR was relatively lower than those without it. This is a common scenario in machine learning projects dealing with medical data, where a balanced dataset is crucial for accurate analysis. To rectify this, we used the Synthetic Minority Over-sampling Technique (SMOTE) [25], which allows the minority class to be over-sampled by synthesizing new examples in the feature space.

Data pipeline B—locally hosted LLMs with advanced fine-tuning to unlock its potential on understanding textual data

In addition to validating pre-training LLMs' embeddings, we endeavored to explore the potential enhancements in model performance through optimization, particularly focusing on fine-tuning techniques. Fine-tuning is a critical method in machine learning where a pre-trained model, such as the GPT-2 used in this study, is further trained (or "fine-tuned") on a specific, smaller dataset relevant to a particular task—in our case, pCR detection. To this end, the GPT-2 model served as the foundational language model. The main idea behind this was to harness the inherent capabilities of GPT-2 and augment it for our specific pCR detection task.

Model architecture The neural architecture attached to the GPT-2 consisted of two main components. After

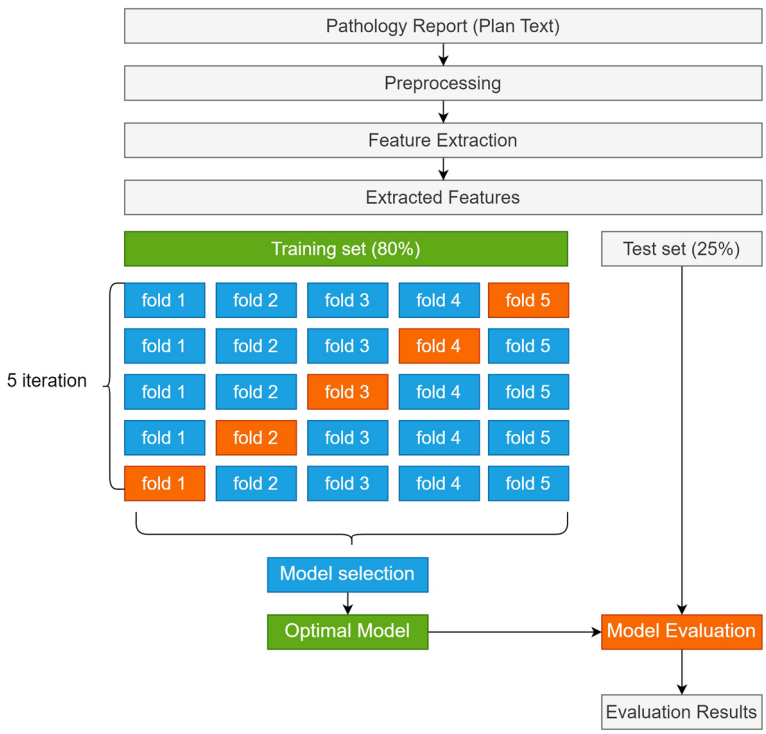

Fig. 2 Data processing and model evaluation workflow

processing the input sequence through the GPT-2 model, its output underwent a mean pooling operation to condense the sequence representation. These pooled embedding vectors were subsequently fed into a fully connected neural network layer. The weights of this layer were initialized using the Xavier normal initializer [26] to ensure optimal backpropagation. The output of this layer was then passed through a sigmoid activation function, producing a probability score representative of the pCR classification.

Fine-tuning procedure For fine-tuning purposes, we adopted a technique called Low-Rank Adaptation of Large Language Models (LoRA) [27]. LoRA curtails the count of adjustable parameters for specialized tasks by incorporating trainable rank decomposition matrices into the Transformer's every layer. This significant reduction in adjustable parameters and computational resource requirements allows for agile task transitioning during deployment without added latency. This greatly reduces the number of trainable parameters and computational resource requirements for LLMs adapted to specific tasks, enabling efficient task-switching during deployment without introducing inference latency.

Moreover, to fine-tune our model's learning rate, we utilized the Adam optimizer [28] in tandem with a learning rate scheduler. This scheduler systematically modified the learning rate, enhancing the training regimen throughout the learning phase.

Model evaluation and statistical analysis

Descriptive analysis of text data was summarized, with median and interquartile range (IQR). Each pipeline was evaluated against the chart reviewed pCR data using various metrics: Positive Predictive Value (PPV), sensitivity, specificity, Negative Predictive Value (NPV), F1-score, and accuracy. To further ensure the robustness and reliability of our results, we employed stratified fivefold cross-validation, where the data was divided into 5 distinct subsets, and the model was trained and tested 5 times, each time with a different subset reserved as the test set and the remaining 4 subsets used for training. This helps in ensuring that our evaluation metrics are not overly optimistic and are indicative of the model's performance on unseen data.

In addition to cross-validation, we enhanced the reliability of our results by employing a bootstrapping resampling method to determine the 95% confidence interval for each of these metrics. Bootstrapping involved resampling with replacement from the original data and recalculating the metrics for each resampled dataset. This procedure was repeated 10,000 times, and the confidence intervals were calculated from the empirical distribution of these metrics. The computational analyses were carried out on an isolated health authority-approved system equipped with an NVIDIA Tesla V100 16GB graphics processing unit (GPU). All statistical analyses were conducted using Python 3.10, NumPy [29], SciPy [30], and PyTorch [31].

Results

Data characteristics

The pathology text data was processed following the flow of Fig. 1. There were 425 patient records linked to EMR data and 74 were excluded due to either the presence of multiple tumors or missing lymph node evaluation. The final cohort included 351 female breast cancer patients. Of them, 102 (29%) patients achieved pCR after NAC as ascertained by manual chart review. The flow of patient cohorts, detailed patient demographics, and clinical characteristics have been previously reported [6].

The median report length was 1,316 words (IQR: 925, 1,631). Patients who achieved pCR tended to have shorter reports compared to those who did not achieve pCR. The median unique word count per report was 583 (IQR: 425, 754), with patients who achieved pCR exhibiting a lower median count of 436 (IQR: 335, 518) compared to patients who did not achieve pCR (M = 684, IQR: 517, 788). After the removal of unnecessary characters, punctuation, and special symbols, the median token count was 2,293 (IQR: 1,540, 2,786) per report.

Performance of data pipeline A

We tested 15 LLMs in pipeline A (Table 1). The sensitivity ranged from 76.2% to 100.0%, while the PPV ranged from 64.0% to 87.0%. The overall performance of the F1 score ranged from 69.6% to 91.3% and the GPT-2 Large model performed the best (highest F1 score and narrow 95% CI). Specifically, BERT-based models exhibited a sensitivity ranging from 90.5% to 100.0%, PPV ranging from 73.1% to 87.0%, and F1 scores ranging from 80.8% to 90.9%. Encoder-to-decoder models such as BART and T5 demonstrated a sensitivity range of 76.2% to 100.0%, PPV ranging from 64.0% to 84.0%, and F1 scores from 69.6% to 91.3%. GPT-based decoder models achieved a sensitivity range of 95.2% to 100.0%, PPV ranging from 80.0% to 84.0%, and F1 scores of 87.0% to 91.3%.

Performance of data pipeline B

The fine-tuning of GPT-2 with LORA significantly improved the performance of the LLMs achieving a high

Table 1 LLMs performance statistics with 95% confidence interval

LLMs	Sensitivity	PPV	Specificity	NPV	Accuracy	AUC ROC	F1 score
Encoder-only models							
BERT base model (uncased) [16]	100.0 (100.0—100.0)	75.0 (56.0—90.5)	86.0 (75.5—94.7)	100.0 (100.0—100.0)	90.1 (81.7—95.8)	93.0 (87.8—97.1)	85.7 (73.2—95.0)
BERT base model (cased) [16]	95.2 (84.2—100.0)	76.9 (59.1—92.0)	88.0 (78.9—96.2)	97.8 (92.5—100.0)	90.1 (83.1—95.8)	91.6 (84.1—97.2)	85.1 (70.6—94.7)
DistilBERT base model (uncased) [32]	95.2 (83.3—100.0)	74.1 (57.1—90.6)	86.0 (75.5—95.7)	97.7 (92.3—100.0)	88.7 (81.7—95.8)	90.6 (83.3—96.2)	83.3 (69.8—93.3)
BioClinicalBERT [33]	90.5 (76.5—100.0)	79.2 (60.9—95.2)	90.0 (80.0—98.0)	95.7 (89.3—100.0)	90.1 (83.1—97.2)	90.2 (81.3—97.1)	84.4 (71.1—94.5)
Tiny BERT [34]	95.2 (84.6—100.0)	76.9 (58.3—92.0)	88.0 (77.5—96.2)	97.8 (92.7—100.0)	90.1 (83.1—97.2)	91.6 (85.1—97.2)	85.1 (72.4—94.7)
BERT multilingual base model (cased) [16]	95.2 (84.2—100.0)	87.0 (70.6—100.0)	94.0 (86.0—100.0)	97.9 (93.0—100.0)	94.4 (88.7—98.6)	94.6 (87.9—99.1)	90.9 (81.1—98.0)
GatorTronS [35]	100.0 (100.0—100.0)	75.0 (56.5—90.9)	86.0 (75.0—94.4)	100.0 (100.0—100.0)	90.1 (83.1—95.8)	93.0 (88.0—97.1)	85.7 (73.2—94.7)
Encoder- decoder models							
BART (base-sized model) [19]	100.0 (100.0—100.0)	84.0 (66.7—96.2)	92.0 (83.6—98.1)	100.0 (100.0—100.0)	94.4 (88.7—98.6)	96.0 (91.5—99.1)	91.3 (81.1—98.0)
BART (large-sized model) [19]	95.2 (84.6—100.0)	80.0 (62.5—95.5)	90.0 (81.8—98.0)	97.8 (92.6—100.0)	91.5 (84.5—97.2)	92.6 (85.7—98.0)	87.0 (74.3—96.3)
BART-large-mnli [19]	90.5 (76.2—100.0)	76.0 (57.1—91.3)	88.0 (77.5—96.2)	95.7 (88.9—100.0)	88.7 (81.7—95.8)	89.2 (80.9—96.3)	82.6 (69.2—92.7)
FLAN-T5 small [36]	76.2 (56.2—94.4)	64.0 (44.4—82.6)	82.0 (70.5—92.0)	89.1 (79.1—97.6)	80.3 (70.4—88.7)	79.1 (68.0—89.2)	69.6 (52.6—82.6)
T5-Large [20]	90.5 (76.2—100.0)	70.4 (51.4—86.2)	84.0 (72.9—93.8)	95.5 (88.1—100.0)	85.9 (77.5—93.0)	87.2 (77.5—94.7)	79.2 (64.9—90.6)
T5-Small [20]	90.5 (75.0—100.0)	70.4 (50.0—86.4)	84.0 (73.3—93.8)	95.5 (88.1—100.0)	85.9 (77.5—93.0)	87.2 (78.4—95.0)	79.2 (64.0—90.9)
Decoder-only models							
GPT-2 Large [21]	**100.0 (100.0—100.0)**	84.0 (66.7—96.2)	92.0 (83.6—98.1)	**100.0 (100.0—100.0)**	94.4 (88.7—98.6)	**96.0 (91.5—99.1)**	91.3 (81.1—98.0)
GPT-2 [21]	85.7 (68.8—100.0)	78.3 (58.3—94.7)	90.0 (80.8—97.9)	93.8 (86.0—100.0)	88.7 (81.7—95.8)	87.9 (79.4—96.0)	81.8 (68.4—92.7)
Baseline models							
Decision Tree based method [6]	*90.5 (69.6–98.9)*	*76.0 (59.6–87.2)*	*87.8 (75.2–95.4)*	*93.8 (86.0—100.0)*	*88.6 (78.7–94.9)*	*87.9 (79.4—96.0)*	*81.8 (68.4—92.7)*
Fine-tuned models (Pipeline B)							
GPT-2 fine-tuned	*95.3 (84.0—100.0)*	*90.9 (76.5–100.0)*	*96.0(90.0–100.0)*	*98.0 (93.3- 100.0)*	***95.8 (90.1–100.0)***	*95.6(89.4–100.0)*	***93.0 (83.7–100.0)***

LLMs Large Language Models, *NPV* Negative Predictive Value, *PPV* Positive Predictive Value

sensitivity (95.3%, 95CI%: 84.0–100.0%) and PPV (90.9, 95CI%: 76.5 -100%). In addition, the NPV and accuracy reached 90.9% (95% CI: 76.5–100%) and 95.6% (95% CI: 89.4–100%), respectively. The F1 score outperformed all other models, peaking at 93.0% (95% CI: 83.7–100%). The decreasing training and validation loss, along with the improving model performance metrics, signify effective learning and enhanced classification abilities (Fig. 3).

Discussion

Our study designed a novel application of LLMs in digital health for the determination of pCR among breast cancer patients who underwent NAC and subsequent curative-intent surgery. We developed and validated two pipelines for processing pathology text data. Our findings demonstrated that LLMs outperformed traditional ML models in the task of pCR detection, confirming their potential utility in extracting critical information from textual pathology data.

Compared to keyword-based methods, our methods present distinct advantages, particularly in simplifying the development process and reducing the dependency on extensive rule-based programming. Firstly, these models significantly lower the threshold of medical knowledge required from developers. Unlike traditional methods, which necessitate deep domain expertise to accurately model the nuances of medical language, LLMs learn from vast datasets, capturing these complexities inherently. This capability allows developers to focus more on application integration and less on the underlying complexities of medical terminology and language patterns.

A prevalent concern regarding the use of AI revolves around its reliability [37]. In our study, the optimized fine-tuned model achieved high sensitivity and NPV. This implies that the model can effectively identify all possible presence of pCR cases or exclude the absence of pCR. Comparatively, this model surpassed the tree classifiers developed from the same dataset, showcasing higher performance metrics [6] (Table 1). The fine-tuned model, designed specifically for interpreting pathology reports, achieved near-perfect accuracy in the test dataset, suggesting it could be effectively implemented for pCR cohort retrieval in future studies.

From a data privacy and security perspective, our analysis evaluated various LLMs across different transformer families to determine their effectiveness and feasibility for interpreting medical text in a secure, locally governed healthcare environment. Our proposed method is designed to integrate seamlessly into existing healthcare infrastructures, effectively addressing potential data security concerns.

Based on the strengths described above, the pipelines developed in this study can be utilized for large cohort studies or clinical trials to evaluate interventions and treatment outcomes, significantly reducing the time required compared to manual chart reviews. They can also be integrated into local clinical information collection systems to accelerate diagnosis and enable personalized treatment strategies. Additionally, these pipelines can support population-based surveillance and facilitate cohort studies in specific clinical contexts.

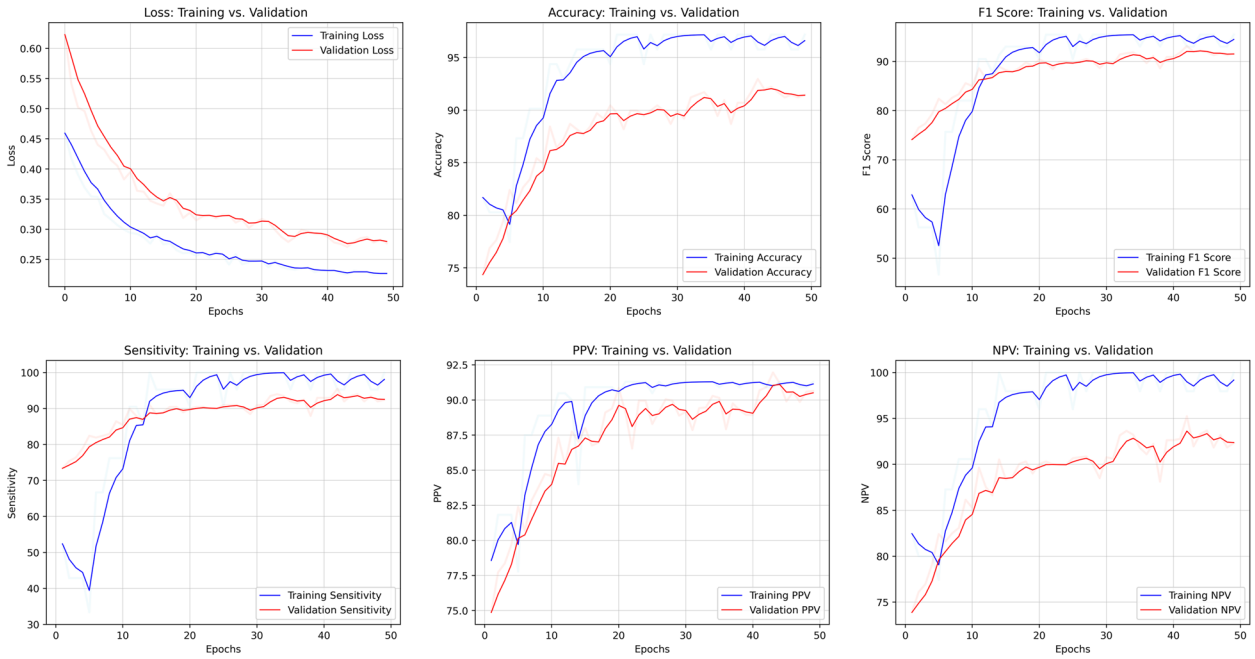

Fig. 3 Training and validation loss along with performance metrics in fine-tuning logistic regression classifier

We acknowledge several limitations in our study. First, our evaluation was limited to 15 commonly used LLMs in the field of medicine, and the performance of other pre-trained LLMs remains unexplored. Specifically, Dialog-based AI such as ChatGPT was not included in our evaluation due to privacy and security concerns. While it is anticipated that as an advanced language model, ChatGPT may offer better performance than GPT-2, it is hosted by OpenAI which hindered its integration into our research framework and local system [37].

Secondly, while our models have been effectively validated within this study, integrating them into broader clinical information systems might require further validation. This need arises because our models were developed using hospital data from a single region. To ensure they work accurately and efficiently in different clinical settings, additional compatibility assessments are necessary.

Third, in our study, we adopted a chunking text processing method for pathology report preprocessing. This approach was necessary due to the data intake constraints of the transformer neural network architecture of the different LLMs. Chunking helps in managing large volumes of text data more efficiently. However, it's important to note that this method might segment the text in a way that disrupts its natural continuity. As a result, the models might face challenges in fully grasping wider context and nuances within the text, which is a crucial aspect to consider in multidisciplinary applications where contextual understanding is key [38].

Lastly, as our validation was performed using internal cross-validation, it is important to acknowledge certain limitations in our approach. While internal validation is a valuable method for assessing model robustness, it may not fully capture the potential variability inherent in datasets from different centers or pathologists.

Future research should focus on exploring the robustness of the model against a more diverse and independent dataset, potentially with the inclusion of center or pathologist-level information, to better ensure generalizability. Moreover, studies that have access to such detailed data could implement center-based or pathologist-based validation strategies to provide a more rigorous assessment of model performance in varied clinical settings.

Conclusion

Our study demonstrates the efficacy of LLMs in digital pathology for precise pCR determination in breast cancer patients post-NAC treatment. The superior performance of the developed pipelines over traditional ML models. These findings highlight LLMs' potential in extracting key clinical data from narrative reports, although external validation is needed in the future.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02677-y.

Supplementary Material 1.

Authors' contributions

K.C. and G.W. contributed equally to this work and are considered co-first authors. Y.X., K.C., G.W., M.Q., A. Laws and A.B contributed to the conceptualization and study design. A.Li, A.B., and J.X. contributed data collection. K.C. and G.W. contributed to the development of the analysis plan. K.C. implemented the pipelines and analyzed the results. K.C. and G.W. drafted the first version of the manuscript. All authors contributed result interpretation, the revision of the manuscript, and the final approval of the manuscript.

Funding

This research was funded by the CIHR (Grant Reference Number: PJT191963).

Availability of data and materials

The data underlying this article were obtained from the local health authority and cannot be shared externally due to privacy and confidentiality restrictions. The complete code and fine-tuned models used in this study are available from the corresponding author upon reasonable academic request.

Declarations

Ethics approval and consent to participate

This study received approval from the Health Research Ethics Board of Alberta – Cancer Committee, with a waiver of informed consent granted.

Consent for publication

Not applicable as this manuscript does not contain any individual person's data in any form that could be used to identify them.

Competing interests

The authors declare no competing interests.

Author details

[1]The Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, Canada. [2]Provincial Research Data Services, Alberta Health Services, Calgary, Canada. [3]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Canada. [4]Department of Surgery, Cumming School of Medicine, University of Calgary, Calgary, Canada. [5]Department of Oncology, Cumming School of Medicine, University of Calgary, Calgary, Canada. [6]Department of Radiology, Cumming School of Medicine, University of Calgary, Calgary, Canada.

Received: 1 March 2024 Accepted: 9 September 2024
Published online: 03 October 2024

References

1. Cortazar P, Geyer CE. Pathological complete response in neoadjuvant treatment of breast cancer. Ann Surg Oncol. 2015;22:1441–6.
2. Mamounas EP. Impact of neoadjuvant chemotherapy on locoregional surgical treatment of breast cancer. Ann Surg Oncol. 2015;22:1425–33.
3. Cortazar P, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet. 2014;384:164–72.
4. Korn E, Sachs M, McShane L. Statistical controversies in clinical research: assessing pathologic complete response as a trial-level surrogate endpoint for early-stage breast cancer. Ann Oncol. 2016;27:10–5.
5. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assn. 2013;20:E206–11. https://doi.org/10.1136/amiajnl-2013-002428.
6. Wu G, Cheligeer C, Brisson AM, Quan ML, Cheung WY, Brenner D, et al. A new method of identifying pathologic complete response

after neoadjuvant chemotherapy for breast cancer patients using a population-based electronic medical record system. Ann Surg Oncol. 2023;30(4):2095–103. https://doi.org/10.1245/s10434-022-12955-6.
7. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2:160. https://doi.org/10.1007/s42979-021-00592-x.
8. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. J Am Med Inform Assoc. 2017;24:607–13. https://doi.org/10.1093/jamia/ocw144.
9. Sheikhalishahi S, et al. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform. 2019;7:e12239. https://doi.org/10.2196/12239.
10. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. J Am Med Inform Assoc. 2017;24(5):986–91.
11. Perera S, Sheth A, Thirunarayan K, Nair S, Shah N. Challenges in understanding clinical notes: Why NLP engines fall short and where background knowledge can help. In Proceedings of the 2013 international workshop on Data management & analytics for healthcare; 2013. p. 21–6.
12. van Baalen S, Boon M, Verhoef P. From clinical decision support to clinical reasoning support systems. J Eval Clin Pract. 2021;27:520–8. https://doi.org/10.1111/jep.13541.
13. Wei WQ, et al. Improving reporting standards for phenotyping algorithm in biomedical research: 5 fundamental dimensions. J Am Med Inform Assn. 2024;31:1036–41. https://doi.org/10.1093/jamia/ocae005.
14. Thirunavukarasu AJ, et al. Large language models in medicine. Nat Med. 2023;29:1930–40.
15. Hart SN, et al. Organizational preparedness for the use of large language models in pathology informatics. J Pathol Inform. 2023;14:100338.
16. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Naacl Hlt 2019), vol. 1. 2019. p. 4171–86.
17. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018.
18. Bossuyt PM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Ann Clin Biochem. 2003;40:357–63. https://doi.org/10.1258/000456303766476986.
19. Lewis M, et al. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. 2019.
20. Raffel C, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21:1–67.
21. Radford A, et al. Language models are unsupervised multitask learners. OpenAI blog. 2019;1:9.
22. Cheligeer C, et al. BERT-based neural network for inpatient fall detection from electronic medical records: retrospective cohort study. JMIR Med Inform. 2024;12:e48995. https://doi.org/10.2196/48995.
23. Lu HX, Ehwerhemuepha L, Rakovski C. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. Bmc Med Res Methodol. 2022;22:181. https://doi.org/10.1186/s12874-022-01665-y.
24. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst. 2012;25. https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.
25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. https://doi.org/10.1613/jair.953.
26. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings; 2010. p. 249–56.
27. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685. 2021.
28. Kingma DP. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.
29. Harris CR, et al. Array programming with NumPy. Nature. 2020;585:357–62. https://doi.org/10.1038/s41586-020-2649-2.
30. Virtanen P, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python (vol 33, pg 219, 2020). Nat Methods. 2020;17:352–352. https://doi.org/10.1038/s41592-020-0772-5.
31. Paszke A, et al. PyTorch: an imperative style, high-performance deep learning library. Adv Neur In. 2019;32.
32. Sanh V. DistilBERT, A distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. 2019.
33. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. 2019.
34. Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, et al. Tinybert: distilling BERT for natural language understanding. arXiv preprint arXiv:1909.10351. 2019.
35. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194.
36. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. J Mach Learn Res. 2024;25(70):1–53.
37. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388:1233–9.
38. Ramkumar P, et al. Chunking as the result of an efficiency computation trade-off. Nat Commun. 2016;7:12176.

Publisher's Note