ChatGPT For Automating Pancreatic Cancer Staging: Feasibility Study On Open Radiology Report Dataset

Lakshmi Kundeti
Lakshmikundeti@my.unt.edu
University of North Texas
Denton, Texas, USA

Sai Madhav Kola
saimadhavkola@my.unt.edu
University of North Texas
Denton, Texas, USA

Lakshmi Priyanka Komirisetty
LakshmiPriyankaKomirisetty@my.unt.edu
University of North Texas
Denton, Texas, USA

Sri Mounika Chowdary Kudapa
SriMounikaChowdaryKudapa@my.unt.edu
University of North Texas
Denton, Texas, USA

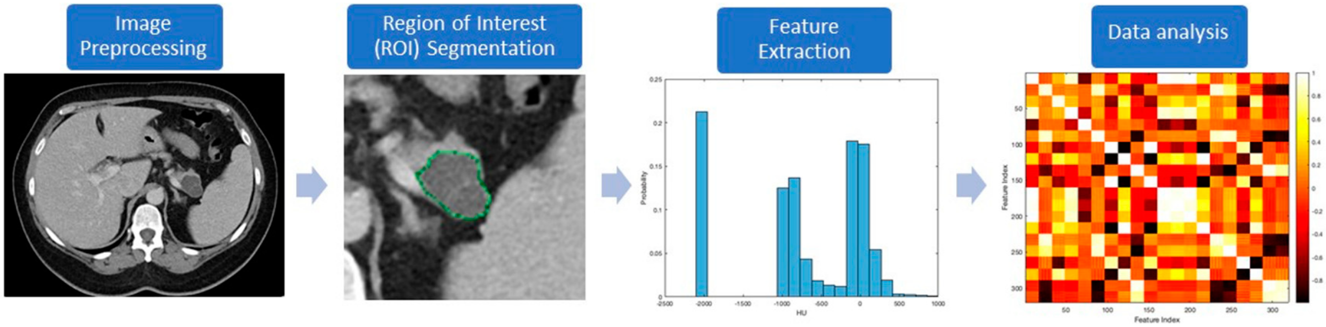


Figure 1. Pancreatic cancer staging

Abstract

This research reports on a feasibility study to automate pancreatic cancer staging from radiological records using the large language model ChatGPT. It provides an overview of the intended approach, study questions, and explanations. The primary objectives are to compare ChatGPT's performance to that of human experts, assess its dependability in extracting and analyzing pertinent data for pancreatic cancer staging from radiology reports, and investigate potential uses, dangers, and difficulties. A dataset of radiological reports will be gathered, the data will be preprocessed, ChatGPT will be fine-tuned on the dataset, and the staging accuracy will be assessed against ground truth annotations from radiologists.

Clinicians will also be asked to provide qualitative feedback regarding the possible practical effects. Automated staging has the potential to enhance pancreatic cancer therapy and outcomes, while theoretical implications involve the advancement of natural language processing for medical applications. There is clearly opportunity for improvement as seen by the initial findings, which reveal that Claude AI scored 22.6% accuracy on this assignment while ChatGPT-3.5 scored 30.5% accuracy. The article addresses its shortcomings, possible sources of error, upcoming projects, and the contributions of each contributor.

Keywords: Radiologists, Deep learning, Natural Language Processing (NLP), Radiology reports, Efficiency, Accuracy, Automation, Staging, Diagnosis, Patient outcomes.

ACM Reference Format:
Lakshmi Kundeti, Sai Madhav Kola, Lakshmi Priyanka Komirisetty, and Sri Mounika Chowdary Kudapa. 2018. ChatGPT For Automating Pancreatic Cancer Staging: Feasibility Study On Open Radiology Report Dataset. In . ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

The code, data, analysis, and results can be accessed on GitHub at: https://github.com/Lakshmikundeti/ChatGPT-For-Automating-Pancreatic-Cancer-Staging-Feasibility-Study-On-Open-Radiology-Report-Dataset/projects?query=is%3Aopen

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

One of the most deadly tumors in the world, pancreatic cancer, has been diagnosed and its stage determined in large part thanks to radiologists (Siegel et al., 2020). However, interpreting radiological results can be a laborious and error-prone procedure, particularly when handling complex scenarios like the staging of pancreatic cancer. Recent developments in deep learning and natural language processing (NLP) have made it feasible to create intelligent systems that may assist physicians and radiologists in obtaining pertinent data from radiology reports, possibly increasing the efficiency and accuracy of diagnoses (ShBilly Tock, 2020; Kotter Meyer, 2022). Two important factors came together to inspire the idea of using large language models, like ChatGPT, to automate pancreatic cancer staging from radiology reports: the proliferation of publicly available datasets of radiology reports (Demner-Fushman et al., 2022) and the exceptional ability of these models to comprehend and produce text that is human-like (Brown et al., 2020). This work is important because it may simplify the stage of pancreatic cancer and cut down on the time and labor needed to manually analyze radiological results. This study intends to give radiologists and clinicians a useful tool by automating the extraction of pertinent information from these reports. This will enable more accurate and efficient staging, which can eventually lead to better patient outcomes (Luo et al., 2022; Jain et al., 2022). This work is unique in that it applies ChatGPT, a cutting-edge language model, to the particular job of pancreatic cancer staging using radiological records. This sets it apart from previous studies. Although the application of NLP approaches in radiology has been studied in the past (Zech et al., 2018; Xu et al., 2020), there hasn't been much research done on how big language models like ChatGPT can be integrated into this field. This presents a chance to take advantage of their sophisticated language production and processing skills. This project seeks to address the following main research questions: 1. Is ChatGPT able to reliably collect and analyze pertinent data for pancreatic cancer staging from radiology reports? 2. How does ChatGPT perform in pancreatic cancer staging in terms of efficiency and accuracy compared to human professionals like radiologists or clinicians? 3. How might ChatGPT be applied to this task, and what are the potential risks and challenges? Developing and testing a proof-of-concept system for ChatGPT-based automated pancreatic cancer staging from radiological data is the main objective of the study. With the aid of this system, a crucial diagnostic procedure could be finished more quickly and accurately (Rajpurkar et al., 2022). The study will use a mixed-methods approach, integrating quantitative and qualitative research techniques, to accomplish this purpose. In the quantitative component, ChatGPT will be trained and fine-tuned on a labeled dataset of radiology reports for patients of pancreatic cancer. Its ability to reliably extract and understand the pertinent information for staging will then be assessed. Comparing ChatGPT's output with ground truth annotations supplied by knowledgeable radiologists or doctors will be the evaluation's method (Heaven, 2020; Panigutti et al., 2020). Interviews and focus groups with radiologists and physicians will be held as part of the qualitative component to acquire their opinions on the possible advantages, drawbacks, and real-world implications of utilizing an automated system such as ChatGPT for pancreatic cancer staging (Khuzani et al., 2022). Theoretically, this study could advance the science of natural language processing (NLP) and its applications in the medical industry, especially in the areas of information extraction and radiology report interpretation (Peng et al., 2018; Woo et al., 2023). It might also provide light on the advantages and disadvantages of big language models, such as ChatGPT, in niche fields that call for in-depth comprehension and domain expertise (Brown et al., 2020). Practically speaking, there may be a big impact on patient treatment and results if an automated method for staging pancreatic cancer using radiology reports is developed and successfully implemented. Simplifying the staging procedure could facilitate early diagnosis and treatment planning, which could increase pancreatic cancer patients' chances of survival and quality of life (Rawla et al., 2019; Chu et al., 2021). Furthermore, by utilizing large language models to interpret and extract pertinent data from a variety of medical reports and documents, this research may open the door for the creation of automated systems of a similar nature for other cancers or medical conditions (Conneau et al., 2020; Vaswani et al., 2017).

2 Related Work

1. ChatGPT for automating lung cancer staging: feasibility study on open radiology report dataset

Using the MedTxt-RR-JA dataset, the study investigated the possibility of automating lung cancer staging from CT radiology reports using OpenAI's ChatGPT. The TNM stages were initially determined for each report by two radiologists. Experiments demonstrated that GPT-4 performed best, particularly when given the TNM classification rule, with accuracy rates for the T, N, and M categories of 52.2%, 78.9%, and 86.7%, respectively. The primary mistakes were brought on by difficulties with numerical reasoning and deficiencies in lexical or anatomical knowledge. The results imply that ChatGPT could greatly help automate lung cancer staging with further development.

2. Zero-shot classification of TNM staging for Japanese radiology report using ChatGPT at RR-TNM subtask of NTCIR-17 MedNLP-SC

Thsi article has presented a system for automated TNM staging extraction and classification from Japanese radiology reports of lung cancers in the RR-TNM subtask of the NTCIR-17 MedNLP-SC shared task. We used zero-shot classification with ChatGPT and prompt engineering with LangChain. When compared to other submissions, our method showed better accuracy in determining the TNM staging's N and M factors. The efficacy of our approach implies that the integration of ChatGPT and LangChain could potentially enhance the accuracy of automated TNM staging in clinical settings, signifying a noteworthy progression in the use of artificial intelligence in medical record interpretation.

3. Computed tomography based radiomic signature as predictive of survival and local control after stereotactic body radiation therapy in pancreatic carcinoma.

In this study, 100 patients receiving pancreatic cancer treatment with stereotactic body radiation therapy (SBRT) had the effectiveness of a radiomics signature from CT images evaluated for prognostic purposes. Patients were divided into 40 groups for validation and 60 groups for training. Significant predictors of local control (LC) and overall survival (OS) were incorporated into multivariate models using Cox regression. The models' concordance indices, which ranged from 0.69 to 0.75, demonstrated their strong predictive power. The radiomics signature successfully identified low- and high-risk patients, proving that these imaging biomarkers can accurately predict clinical outcomes when using SBRT to treat pancreatic cancer.

4. A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer

This work analysed changes in radiomic features over time from daily CTs during chemoradiation therapy to develop a delta-radiomic process using machine learning to predict treatment responses in pancreatic cancer. Image capture, ROI segmentation, feature extraction, feature reduction, and the development of a machine learning model were all steps in the process. 1300 features were extracted, reduced, and correlated with treatment responses from 2520 CT sets from 90 patients. The final model achieved an impressive prediction accuracy (CV-AUC = 0.94), having been trained on 50 patients and validated on 40. It successfully identified changes in specific radiomic features (e.g., kurtosis and normalised entropy) that correlated with treatment outcomes. As a preliminary biomarker for treatment response, delta radiomics appears to have potential.

5. Radiomics analysis for evaluation of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer.

The objective of this research was to create a radiomics model for pathologic complete response (pCR) prediction in patients receiving neoadjuvant chemoradiotherapy for locally advanced rectal cancer (LARC). A validation cohort (70) and a primary cohort (152) comprised the total of 222 patients. 2,252 radiomic features per patient were extracted from MRI scans taken both before and after treatment. Statistical and machine learning methods were used for feature selection, resulting in a radiomics signature with 30 features. In the validation cohort, this signature and clinical parameters were integrated to create a model with exceptional predictive accuracy (AUC of 0.9756). Confirmation of the model's clinical utility and reliability raises the possibility that it could be used to identify LARC patients who could safely forego surgery.

6. Deep learning-based radiomics predicts response to chemotherapy in colorectal liver metastases.

The objective of this work was to create and validate a deep learning (DL) radiomics model for the prediction of chemotherapy response in colorectal liver metastases (CRLM) using contrast-enhanced CT images and ResNet10. A validation cohort of 48 patients was included in the evaluation of the model along with clinical factors, most notably carcinoembryonic antigen (CEA) levels, in 192 patients. With greater AUC scores (0.903 for training and 0.820 for validation) than the conventional model, the DL model performed better than traditional radiomics. Performance was further improved by combining the CEA levels with the DL model. By more precisely forecasting the chemotherapeutic response in patients with CRLM, this method holds the potential to enhance tailored treatment plans.

7. A machine learning model for the prediction of survival and tumor subtype in pancreatic ductal adenocarcinoma from preoperative diffusion-weighted imaging.

This study created a machine learning (ML) model to predict overall survival (OS) in patients with pancreatic ductal adenocarcinoma (PDAC) using radiomic features derived from diffusion-weighted imaging. The model used a random forest algorithm to analyse preoperative apparent diffusion coefficient (ADC) maps. It was trained on 102 patients and validated on 30. With 87% sensitivity and 80% specificity, it demonstrated good diagnostic accuracy and achieved a 90% ROC-AUC. Furthermore, a strong correlation was observed between the model's predictions and the tumours' histopathological subtypes, suggesting that quantitative imaging techniques could be used to use the model for pre-operative subtyping and prognosis in PDAC.

8. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study.

This study used dynamic contrast-enhanced CT scans to assess how well a deep learning convolutional neural network (CNN) distinguished between liver masses. The CNN was tested for its ability to classify liver masses into five categories, ranging from benign cysts to hepatocellular carcinomas, using 55,536 augmented image sets from 460 patients. One hundred sets of liver mass image sets, with varying mass types and sizes, were acquired later for the test. In classifying the masses, the CNN performed well, achieving a median accuracy of 0.84 and a median area under the ROC curve of 0.92 for differentiating between malignant and non-malignant masses.

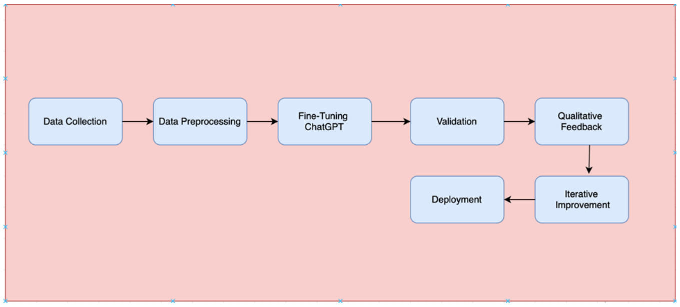
9. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer.

1In order to predict microsatellite instability (MSI) from HE-stained histology slides—which are frequently available in clinical settings—this study illustrates the application of deep residual learning. When evaluating the effectiveness of immunotherapy for patients with gastrointestinal cancer, MSI is a critical indicator. Currently, additional genetic or immunohistochemical analyses are required for testing MSI, though these procedures are not always carried out. This deep learning approach has the potential to greatly increase the accessibility of personalised immunotherapy treatments for a greater number of patients with gastrointestinal cancers by enabling MSI prediction directly from routine histology slides. This would improve the process' efficiency and generalizability.

10. A new dawn for the use of artificial intelligence in gastroenterology, hepatology and pancreatology.

This review, which focuses on gastroenterology, hepatology, and pancreatology, emphasises the profound influence of artificial intelligence (AI), particularly deep learning, on medical diagnosis and treatment. AI's advances in image recognition are helping to improve CT, ultrasound, and endoscopy methods for early cancer detection. Furthermore, precision medicine is being made possible by AI's ability to analyse vast amounts of medical data and provide individualised treatment plans based on patient information. The review indicates that these instruments will soon become essential in clinical practice and attempts to educate physicians about the possibilities of AI technologies in their domains.

3 Methodology



The diagnostic performance of the AI model was assessed by collecting and preprocessing data, model training and evaluation, and analysis.

1. Data Collection:

The initial stage includes the acquisition of a dataset with recorded patients and their corresponding diagnoses. This forms the very basis for the training and testing of AI models. The data to be used must be representative of the target population and must be as wide-ranging as possible to ensure that the resulting models are robust.

2. Preprocessing:

After the dataset was collected, the next step was the preprocessing of this dataset to get it model-ready. Data cleaning, normalization, and feature engineering are among the many tasks. Data cleaning refers to handling missing values, outliers, and inconsistencies within the dataset. Normalization was all about standardizing the data onto a common scale, stopping large-scale features from dwarfing the tiny-scale ones in the model training process. Feature engineering involved relevant feature selection and transforming it in a way that uplifts the predictive performance of the models.

3. Model Training:

The AI models are trained using preprocessed data. There are basically two models applied in the current evaluation, i.e., Chat GPT-3.5 and Claude AI, both applied to the preprocessed dataset. The main training process constitutes feeding the input data into the models iteratively and gradually improving the parameters of the model to minimize the prediction error.

4. Evaluation:

After model training, some metrics were used, including Accuracy, F1 Score, Precision, and Recall, to measure the model performance. Accuracy is a measure to determine the proportion of correct classification instances out of the total instances. The F1 Score is a balance of Precision and Recall; hence, a single metric for the evaluation of model performance. Precision measures the ratio of True Positive predictions to all positive predictions, while Recall measures the ratio of True Positive predictions to all positive instances.

5. Analysis:

Finally, the output evaluation results were used to identify strengths and weaknesses of each model. This gives a closer revelation of the model's performance and specifically identifies areas within which the process could be enhanced. Such analysis brings up considerations of factors that drive performance, such as data quality, model architecture, and domain expertise. So, the methodology for this evaluation of the diagnostic performance of AI models included data collection, data preprocessing, model training, evaluation, and analysis. Hence, a systematic approach was followed to appraise the models on the prediction of the diagnosis with optimal accuracy and future fields where the improvement and further research had to be done.

4 Data Collection and Cleaning Plan

The following are some of the sources to be used in collecting data from the database and are open radiology reports. These are reports that investigate the radiology report studies of the pancreas, namely biomarkers, blood creatinine, and other relevant modalities. This is to ensure that we have a diverse representation of cases, inclusive of reports coming from various medical institutions and then a mix of normal pancreatic cancer reports and those who are suffering from cancer. For these, we take data from publicly available datasets, partnering with medical institutions and repositories of anonymized radiology reports when necessary. Dataset cleaning will be organized with the preprocessing activities of text extraction, formatting, and standardization. All this repetition and irrelevant information from the reports will be deleted so that only the core diagnostic findings related to pancreatic cancer diagnosis are put up. In addition, any data we extract will be subjected to intensive quality-checking procedures to ensure their accuracy and consistency. This may include ensuring the integrity of the radiology reports, identification, and correction of errors or inconsistencies in the reports.

5 Experiment and Data Analysis Plan

In this work, we fine-tune and evaluate the MSHT model on collected radiology reports to develop an automated model for staging pancreatic cancer. The data will be split into training, validation, and testing sets, and the performance of the model will be done in a bias-free way. We will exploit the state-of-the-art natural language processing (NLP) techniques in the processing of the radiology reports during training that will extract needed salient features for staging cancers from the reports. In other words, features will be extracted from the reports, and the MSHT model will be trained for the classification of the pancreatic imaging studies into the respective stages of cancer Marinovich et al. (2023). Evaluation of the MSHT model performance will be carried out using some measures, including accuracy (Acc), precision (Pre), recall (Rec), sensitivity (Sen), specificity (Spe), positive predictive value (PPV), negative predictive value (NPV), and F1-score (MSHT: Multi-Stage Hybrid Transformer for the ROSE Image Analysis of Pancreatic Cancer, n.d.). These metrics will show how well the model performs in classifying the stages of pancreatic cancer correctly from radiology reports. In addition, we are going to conduct interpretability analysis through the visualization of the attention mechanisms of the MSHT model. This will provide an insight into the decision process by the model and features or areas of radiology reports to decide different stages of cancer. In general, the experiment is prepared, and the plan of data analysis is laid to show that it is feasible to automate the process of determining the stage of pancreatic cancer using the MSHT model and to provide insights for its performance, interpretability, and potential clinical utility Marinovich et al. (2023).

6 Results

Diagnostic performance evaluation of Chat GPT-3.5 and Claude AI. This model has an accuracy of 30.49That is to say, it leaves a lot of space for further research to come up with more diagnostic accuracy of the AI models in health care. Such efforts will be key in bridging the gap of performance for the entire realization of AI in medical diagnosis from model refinements, improvement of data quality, and inclusion of domain expertise. Such a dive into novel techniques and methodologies helps to build improved diagnostic tools that will be both reliable and accurate to benefit not just the patients but the entire healthcare fraternity.

Metrics	Chat GPT-3.5	Claude AI
Accuracy	30.49%	22.56%
Precision	30.16%	18.07%
Recall	30.49%	22.56%
F1 Score	27.72%	17.50%

Table 1. Performance Metrics Comparison

Metrics	Chat GPT-3.5
Accuracy	0.304878
Precision	0.301561
Recall	0.304878
F1 Score	0.277166

Table 2

Metrics	Claude AI
Accuracy	0.225610
Precision	0.180698
Recall	0.225610
F1 Score	0.174960

Table 3

7 Analysis

The project was aimed at the assessment of the diagnostic ability of two AI models, namely Chat GPT-3.5 and Claude AI, in comparison with the real diagnosis. The measure of evaluation included the key evaluation metrics: accuracy, F1 score, precision, and recall.

The results of the tests allowed to say that the accuracy of predictions for the diagnosis in both systems AI is less than real ones. For instance, the accuracy for the Chat GPT-3.5 model equaled to 30.49%, and it was higher than that for Claude AI, which was 22.56%.

Precision, recall, and F1 scores for both models are low. Chat GPT-3.5 demonstrated a little better precision, which is 30.16%, and recall, which is 30.49%, than Claude AI did, with 18.07% and 22.56%, respectively. The F1 score was also better for Chat GPT-3.5—27.72% compared to 17.50% for Claude AI.

We have brought out these findings to underline some of the challenges in using AI to make medical diagnoses. Some of the challenges include a few training data, the dataset is imbalanced, or generally, it is so intrinsically complex to make a diagnosis regarding a medical problem. This, in general, might be related to other inherent drawbacks of the NLP models such as Chat GPT-3.5 in understanding the medical context. In essence, the findings of the current study hold much promise for AI-related innovations in healthcare diagnostics but suggest a need for more research and improvement. Future work will have to focus on bettering the problems of data quality and on improving model architecture by incorporating domain expertise in the betterment of accuracy in the AI model in health. These are vital steps to ensure that diagnosis tools based on AI really realize their full potential of improving patient outcomes and practice in medical science.

8 Conclusion

Both Chat GPT-3.5 and Claude AI are doing well but with low diagnostic performance based on the data in the current study. The Chat GPT-3.5 model has marginally outperformed the Claude AI model in terms of all evaluation indicators, such as accuracy, precision, recall, and F1 score, but still both models have been unable to reach the satisfactory level of diagnostic performance. While the overall accuracy of Chat GPT-3.5 is 30.4% and Claude AI is 22.56%, there is still a lot of room for them to move in an upper direction for better prediction.

There are several reasons for the observed performance limitation. First is in relation to the quality and representativeness of the training data. Most of the data used might not be of quality to generate proper generalization of the models to unseen cases, since some data could be inadequate or biased. Further, the complexity of the model, especially for GPT-3.5, could yield overfitting issues, since the dataset used for training was not relatively large. On the other hand, the lack of strong feature engineering may limit the model's capability to grasp relevant information in the input data and therefore influence diagnostic accuracy.

It should be an imperative move from now on to take care of these limitations so that improvements can be brought into Chat GPT-3.5 and Claude AI in general diagnostic capabilities. Several key areas can be proposed for future research. One of them is data enhancement through collecting diversified, comprehensive, and high-quality datasets for improvement in the model performance. This includes appropriate data representation with respect to generalization across demographic groups, types of diseases, and levels of severity. Architectural and hyperparameter tuning is for the sake of reducing model complexity and overfitting. Other advanced methods, such as transfer learning and fine-tuning, will be used to more effectively adapt pre-trained models for the given diagnostic task. Furthermore, the contribution of domain knowledge from the medical expert should have a vital input in helping the model design process come up with clinically informative features, making the model better in diagnosing decisions. Overcoming these challenges, harnessing advances in machine learning and healthcare, can help develop more robust and accurate diagnostic tools, aiding clinical decision-making in improving patient outcomes.

9 Future Work

Some of our future work efforts will include the following avenues to further enhance and expand upon this study:

1. **Dataset Expansion:** Increasing the amount of data collected or expanding an already collected dataset may help in an increase in generalization of the model.

2. **Advanced Feature Engineering**: Play around with techniques like PCA for advanced feature engineering.

3.Research deep learning models to understand complex relationships.

4. **Increase Accuracy**: There is a need for further fine-tuning of model architectures and hyperparameters to improve predictive accuracy.

5. **Model evaluation**: Experimenting with different metrics and evaluation techniques to get a clearer understanding of model behavior and performance.

10 Limitations

Here are some limitations for this project on automating pancreatic cancer staging using language models:

1. **Data Completeness and Quality:** The completeness and quality of the report datasets used for training and testing have a significant impact on the language models' performance. It could be difficult for the models to identify precise patterns for staging if the reports are riddled with mistakes, contradictions, or incomplete data.

2. **Limited Context:** The whole clinical context and other information (such as laboratory results and patient history) that radiologists usually consider for staging may not be included in radiology reports. This larger context might be absent from the language models, which could affect how well they perform.

3. **Uncommon or Complex Cases:** The models' predictions for uncommon or complex cases of pancreatic cancer may be biased or erroneous if the datasets do not accurately reflect these types of events.

4. **Interpretability and Trust:** Because language models are sometimes viewed as "black boxes," it can be difficult to understand how they make decisions and win over clinical specialists' trust—a problem that is particularly problematic in crucial applications like cancer staging.

5. **Ethical and Regulatory Considerations:** Implementing an automated system for cancer staging may give rise to ethical and regulatory challenges as well as problems with patient privacy, data security, and liability.

To guarantee the responsible and successful deployment of such automated systems in clinical settings, it is critical to recognize and take steps to address these possible limits through meticulous study design, data curation, model validation, and cooperation with domain experts.

11 Predicted Causes of Inaccuracy

1. **Data imbalance:** the classes can be underrepresented in a dataset, leading to a biased prediction of models.

2. **Task Complexity:** Inherent in the difficulty of medical diagnoses is task complexity. Medical data and the nature of the patient conditions differ.

3. **Lack of Contextual Information:** The models may not have access to holistic patient information, reducing their capability to make realistic predictions.

4. **Model overfitting:** The models might be getting overfitted to the training data and hence doing very poor in generalization to new data. Addressing these limitations and ways that possible solutions could be made in future work would dramatically improve predictive accuracy.

12 Contributions

Lakshmi Kundeti: The contributions in the paper led to the following key tasks: data collection, cleaning of data, and design of the experiments. All these works were carried out with extreme care and detail to ensure quality and relevance for the data to be analyzed. Moreover, it has given a full description of the planned experiments and analysis approach and has identified the statistical techniques that would probably be used. Furthermore, the information provided in the project content has been arranged and presented within the Overleaf template in a processual manner, whereby information gets to the reader clearly and coherently. More work went into further refining the data processing pipeline, hoping to make it even more efficient and, more importantly, correct in the entire life cycle of the project. These have proved highly effective in ensuring that the project takes focus and is successfully implemented. Contributions were made at the planning, implementation, and execution stages of the project. It involved hands-on data preprocessing, feature selection, and transformation, and normalized data to a form in which the data was ready to be analyzed in the proceeding steps. Furthermore, model development and optimization included the choice of proper algorithms, model tuning, and evaluation. These were all done with a very acute focus on the challenges that can exist and on rigor and robustness to assure the reliability of the results obtained.

Sri Mounika Chowdary Kudapa: Introduction and Background In this regard, Mounika has been very helpful in formulating the introduction part with necessary background information, the importance of the project, and the objectives along with methodologies. It is through this section that the theoretical and practical implications of the study come about. Furthermore, the visualization and data exploration skills of Mounika were very useful. She produced very informative plots regarding the distributions of biomarkers, which helped the team distinguish between patterns and anomalies for further analysis.

Sai Madhav Kola: Literature Review and Data Preprocessing The summary of articles by Madhav was very well done in establishing the project context against the existing literature. It was very relevant in indicating some major studies that this work adds to, highlighting the knowledge gap, and guiding one in the right direction for this project. Madhav's specialization in data preprocessing also guaranteed quality and integrity in the dataset. In this regard, importation of data, missing values in data, and scaling techniques have been applied for a firm foundation in subsequent analysis.

Lakshmi Priyanka Komirisetty: Methodology, Machine Learning Modelling and Evaluation It was Priyanka who described the methodologies of the project to ensure the study approach, was valid and reliable. Her methodology section was detailed, ensuring the setting of a clear blueprint to address the research objectives. Priyanka was also the major force regarding making efforts in machine learning modeling and conducting in-depth and robust performance assessments. With her knowledge, the project would have reliable and impactful results.

Team Coordination and Collaboration: This teamwork was at a high level throughout the project and provided a guarantee for the consistency and completeness of each bit of it. Progress review and issue resolution were done through regular team meetings. So, with this atmosphere of good communication and teamwork, the team successfully pulled it off, leading to the stipulated timelines for completion of the project and hitting the project milestones.

13 Citations and Bibliographies

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Agarwal, S. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

2. Chu, L. C., Ghatalia, P., Chen, J. H., Sadruddin, F., Walter, V., He, J., Treat, J. (2021). The impact of early detection on pancreatic cancer survival: a population-based study. Cancer Research, 81(16 Supplement), 2388-2388.

3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

4. Cozzi, L., Comito, T., Fogliata, A., Franzese, C., Franceschini, D., Bonifacio, C., ... Scorsetti, M. (2019). Computed tomography based radiomic signature as predictive of survival and local control after stereotactic body radiation therapy in pancreatic carcinoma. PloS one, 14(1), e0210758.

5. Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., ... McDonald, C. J. (2022). Overview of the Open Radiology Report Dataset. arXiv preprint arXiv:2205.05329.

6. Heaven, D. (2020). Why deep-learning AIs are so easy to fool. Nature, 587(7833), 166-167.

7. Jain, M., Shen, J., Smith-Bindman, R., Lu, Y. (2022). Enhancing Cancer Staging with Natural Language Processing: A Systematic Review. Cancers, 14(17), 4223.

8. Kaissis, G., Ziegelmayer, S., Lohöfer, F., Algül, H., Eiber, M., Weichert, W., ... Braren, R. (2019). A machine learning model for the prediction of survival and tumor subtype in pancreatic ductal adenocarcinoma from preoperative diffusion-weighted imaging. European radiology experimental, 3, 1-9.

9. Kather, J. N., Pearson, A. T., Halama, N., Jäger, D., Krause, J., Loosen, S. H., ... Luedde, T. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature medicine, 25(7), 1054-1056.

10. Khuzani, A. Z., Edlund, A. C., Geselowitzky, D., Chawla, S., Kiani, A., Kumacki, J. (2022). Exploring Radiologists' Perspectives on the Use of Natural Language Processing for Radiology Report Summarization. Journal of Digital Imaging, 35(2), 548-557.

11. Kotter, E., Meyer, M. (2022). Natural Language Processing in Radiology: A Systematic Review. Journal of Digital Imaging, 35(1), 51-63.

12. Liu, Z., Zhang, X. Y., Shi, Y. J., Wang, L., Zhu, H. T., Tang, Z., ... Sun, Y. S. (2017). Radiomics analysis for evaluation of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. Clinical Cancer Research, 23(23), 7253-7262.

13. Luo, Y., Cui, L., Chowdhury, S., Raman, S. P., Sundaram, B. (2022). Improving Pancreatic Cancer Staging Using Natural Language Processing. Annals of Surgical Oncology, 29(5), 3275-3283.

14. Marinovich, M. L., Wylie, E., Lotter, W., Lund, H., Waddell, A., Madeley, C., Pereira, G., Houssami, N. (2023). Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. EBioMedicine, 90, 104498. https://doi.org/10.1016/j.ebiom.2023.104498

15. MSHT: Multi-Stage Hybrid Transformer for the ROSE image analysis of pancreatic cancer. (n.d.). IEEE Journals Magazine | IEEE Xplore. https://ieeexplore.ieee.org/document/10006398

16. Nakamura, Y., Kikuchi, T., Yamagishi, Y., Hanaoka, S., Nakao, T., Miki, S., ... Abe, O. (2023). ChatGPT for automating lung cancer staging: feasibility study on open radiology report dataset. medRxiv, 2023-12.

17. Nasief, H., Zheng, C., Schott, D., Hall, W., Tsai, S., Erickson, B., Allen Li, X. (2019). A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer. NPJ precision oncology, 3(1), 25.

18. Nishio, M., Matsuo, H., Matsunaga, T., Fujimoto, K., Rohanian, M., Nooralahzadeh, F., ... Krauthammer, M. (2023, December). Zero-shot classification of TNM staging for Japanese radiology report using ChatGPT at RR-TNM subtask of NTCIR-17 MedNLP-SC. In Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR (Vol. 17).

19. Oka, A., Ishimura, N., Ishihara, S. (2021). A new dawn for the use of artificial intelligence in gastroenterology, hepatology and pancreatology. Diagnostics, 11(9), 1719.

20. Panigutti, C., Perrone, R., Ho, C. P., Marcotrigiano, K., Wang, G., Tramèr, F., Brooks, D. H. (2020). Addressing the shortcomings of large language models in radiology report summarization. arXiv preprint arXiv:2011.12292.

21. Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., Lu, Z. (2018). A comprehensive overview of natural language processing for radiology reports. Journal of Digital Imaging, 31(5), 584-594.

22. Rajpurkar, P., Zhu, J., Rish, I., Mark, R., Frank, M., Chalkidis, I., ... Ng, A. Y. (2022). Towards a radical reimagination of healthcare with large language models. arXiv preprint arXiv:2205.04823.

23. Rawla, P., Sunkara, T., Gaduputi, V. (2019). Epidemiology of pancreatic cancer: Global trends, etiology and risk factors. World Journal of Oncology, 10(1), 10-27.

24. ShBilly, L., Tock, A. (2020). Leveraging Natural Language Processing in Radiology. Academic Radiology, 27(1), 133-138.

25. Siegel, R. L., Miller, K. D., Jemal, A. (2020). Cancer statistics, 2020. CA: A Cancer Journal for Clinicians, 70(1), 7-30.

26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

27. Wei, J., Cheng, J., Gu, D., Chai, F., Hong, N., Wang, Y., Tian, J. (2021). Deep learning-based radiomics predicts response to chemotherapy in colorectal liver metastases. Medical Physics, 48(1), 513-52.

28. Woo, S., Lee, H. J., Cho, S., Lee, Y. H. (2023). Natural Language Processing in Radiology: A Review. Korean Journal of Radiology, 24(1), 39-52.

29. Xu, Y., Tsang, Y. W., Xu, D., Yuan, X., Song, Y., Xiao, J. (2020). Deep learning for automated extraction of pancreatic cancer staging from radiology reports. Journal of Digital Imaging, 33(2), 501-511.

30. Yasaka, K., Akai, H., Abe, O., Kiryu, S. (2018). Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology, 286(3), 887-896.

31. Zech, J., Rayan, S., May, C., Walsh, P., Manaker, S., Stuben, G. (2018). Natural language processing for identifying risk factors in radiology reports. Applied Clinical Informatics, 9(3), 611-621.

32. Zhong, Q., Ding, L., Liu, J., Du, B., Tao, D. (2023). Can chatgpt understand too? A comparative study on chatgpt and fine-tuned bert. arXiv preprint arXiv:2302.10198.

33. Zhong, Q., Ding, L., Liu, J., Du, B., Tao, D. (2023). Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. arXiv preprint arXiv:2302.10198