

AI vs. Rule-based Text-Mining for Cancer Diagnosis Extraction: Which Performs Better?


Introduction, Motivation and Goals

- Background:** The Small Animal Surveillance Network (SAVSNET, University of Liverpool) has received almost 3 millions (free-text) electronic pathology records (EPR) from veterinary diagnosis laboratories over the last 14 years.
- Rule-based text-mining (TM) techniques (specific curated dictionaries, regex, etc.) have been applied to extract cancer-related information from those EPR to create the SAVSNET Tumor Registry (TR), a structured and normalized dataset that currently contains about 1.3 million animal tumours reported.
- Previous validation** (2021) showed TM achieved an average accuracy of 95% in extracting diagnosis from EPR, demonstrating its reliability in structured datasets.
- TM is a deterministic approach that requires both biological and code expertise to create rules to deal with context, typos and false positives/negatives.
- Question to explore:** *Would a fine-tuned large language model (LLM) be able to perform cancer information extraction more effectively than traditional rules-based TM while requiring less specialized expertise?*
- Goal:** To compare a *Stochastic* (LLM) with a *Deterministic* (rules-based) approach for the task of extracting cancer-diagnosis related information from free-text EPR.

Material and Methods

- PathologyBERT** (110M parameters), pretrained on pathology-specific vocabulary, was fine-tuned using a representative labeled SAVSNET TR sample dataset.
- Representative dataset (200.000 labeled examples) include all possible diagnostic labels (around 300), contributions of all data providers (diagnostic labs), and both single and multiple diagnosis cases. Additionally, a few thousands Spanish EPRs were also incorporated to evaluate multilingual capabilities.
- GPT-4** (550B parameters) API was used to generate up to three different versions of narratives with uncommon diagnosis (augmentation without duplication) to prevent overlap between training and test sets. Dataset was split 80% for training, 10% for validation, and 10% for testing and optimization used Binary Cross-Entropy (BCE) loss function, a 0.5 threshold, and early stopping to avoid overfitting.

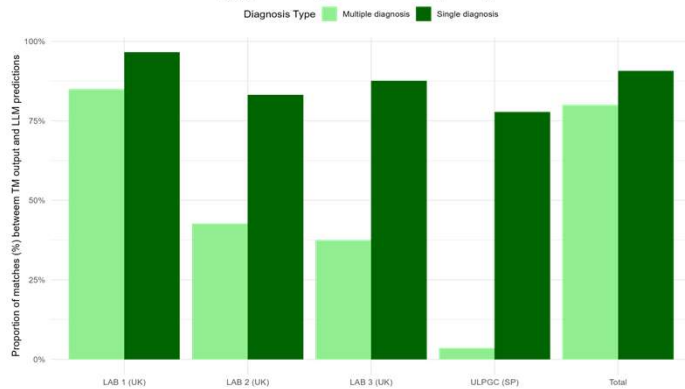
- Performance metrics > 0.9: accurate and comprehensive diagnosis extraction
- Consistent reduction in validation loss: solid generalization without overfitting



Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.005700	0.005460	0.654324	0.781077	0.703487	0.724905
10	0.000900	0.002163	0.871288	0.925731	0.905329	0.913363

Results

Analysis A) Comparison of diagnosis predicted by the LLM with outputs obtained from rules-based TM methodology (test set; n = 19230 reports).



- Previous TM validation showed 96% accuracy for single and 89% for multiple diagnosis.
- Most discrepancies were due to partial matches (“Lipoma” vs. “Lipoma (multiple)”).
- Both methods provided correct diagnosis, with no consistent advantage for either.

Analysis B) Predictions obtained from narratives where TM failed to detect any reported tumour (n = 1213815 reports).

LLM and TM False negatives (FN): In 65% of these narratives (n = 797,508) LLM made no prediction	
Inconclusive diagnosis: uncertain narratives	
Neoplasm (generic)	Cytologies: differentials in preliminary clinical comments Histologies: differentials in hystopathology comments
LLM false positives (FP): Undesired learning patterns	
Lipoma	<i>adipose tissue</i>
Mast cell tumour	<i>perivascular mast cells</i>
Lymphoma	<i>generalized lymphadenopathy</i>
Peripheral odontogenic fibroma	<i>fibrous/gingival hyperplasia</i>
LLM true positives (TP): Typos in diagnosis and keywords (TM FN)	
<i>Sacorma</i> instead of Sarcoma	
<i>Trichopithelioma</i> instead of Trichoepithelioma	
<i>iagnosis</i> instead of Diagnosis	
LLM (TP): Absent of keywords indicating a condition (<i>Diagnosis/Interpretation</i>) (TM FN)	

Conclusions and practical implications

- Improving information extraction from large veterinary datasets is essential for accurately assessing disease impacts on animal populations.
- Rule-based TM excels in structured databases with consistent labeling but often struggles with domain-based texts like electronic pathology records where free-text content is commonly unavoidable. In these scenarios, the generalization capabilities of LLMs could provide more useful results.
- However, LLMs (especially last-generation generative models) come with significant computational and ecological costs. This emphasizes the need for efficient solutions that minimize carbon footprint while keeping AI technologies accessible.
- In this study, we leveraged a BERT-based LLM for the primary task while limiting the use of advanced, large-scale models like GPT-4 to auxiliary tasks.
- TM efficiently extracts high-level data in structured contexts but struggles with typos and complex narratives;
- LLMs excel in understanding nuanced language but risk overfitting to irrelevant patterns, leading to false positives or inaccurate diagnosis.
- Future Directions:** We propose developing a hybrid LLM-TM pipeline to combine the strengths of both methods—balancing efficiency, accuracy, and ecological sustainability—to maximize data extraction quality and improve future research outcomes.