Original Paper

Development and Application of Deep Learning Models for Case Finding From Pathology Reports in a Hospital-Based Cancer Registry: Algorithm Development and Validation

Chanavee Toh[1*], MD; Chanon Kongkamol[2*], MD, PhD; Sarayut Lucien Geater[3*], MD, PhD; Thammasin Ingviya[1*], MD, PhD

[1]Department of Clinical Research and Medical Data Sciences, Faculty of Medicine, Prince of Songkla University, Songkhla, Thailand
[2]Department of Family Medicine and Preventive Medicine, Faculty of Medicine, Prince of Songkla University, Songkhla, Thailand
[3]Department of Internal Medicine, Faculty of Medicine, Prince of Songkla University, Songkhla, Thailand
*all authors contributed equally

Corresponding Author:

Thammasin Ingviya, MD, PhD
Department of Clinical Research and Medical Data Sciences
Faculty of Medicine, Prince of Songkla University
15 Kanchanavanich Road, Medical Research and Technology Transfer Building, Hat Yai District
Songkhla 90110
Thailand
Phone: 66 918477271
Email: thammasin.i@psu.ac.th

Abstract

Background: Few studies have investigated automated case finding in hospital-based cancer registries despite their crucial role in supporting population-based surveillance. In these settings, data completeness is often compromised by referral patterns, yet the influence of weakly supervised learning strategies on model performance remains underexplored.

Objective: This study aimed to evaluate the effects of different supervision approaches (inaccurate, inexact, and report-level labeling) on deep learning model performance for cancer case finding in a hospital-based registry setting and assess model generalizability across patient-level and report-level classification tasks.

Methods: Hospital-based cancer registry and pathology report data from Songklanagarind Hospital (2005-2023) were used for model development and evaluation. Three labeling approaches were implemented: inaccurate (all reports from patients with cancer labeled positive), inexact (concatenated reports from patients with cancer labeled positive), and report-level (only registrar-annotated tumor-related reports labeled positive). Model performance was evaluated on max-pooled, concatenated, and individual report classification tasks using a simple word embedding model, multilayer perceptron, recurrent neural network, convolutional neural network (CNN), and hierarchical self-attention network architectures.

Results: Models trained using weakly supervised labeling achieved the highest performance across patient-level evaluation tasks, and CNN models consistently showed superior performance. While report-level labeling yielded superior accuracy for individual reports, it suffered substantial degradation in patient-level evaluations. This suggests that models trained on curated data overfitted to primary tumor patterns, failing to recognize secondary evidence (eg, metastatic or surveillance findings). Conversely, weakly supervised approaches enabled the detection of patients with cancer via diverse report types. Error analysis revealed that false negatives were primarily driven by insufficient institutional data—specifically, referred patients lacking primary diagnostic reports—rather than algorithmic failure.

Conclusions: In this single tertiary referral hospital setting, weakly supervised labeling outperformed report-level human annotation for patient-level case finding, whereas report-level labeling performed best for individual report classification. The CNN architecture achieved the highest or near-highest area under the receiver operating characteristic curve in most configurations.

JMIR Med Inform 2026;14:e92814; doi: 10.2196/92814

Keywords: deep learning; cancer case finding; hospital-based cancer registry; weakly supervised learning; pathology reports; natural language processing

Introduction

A cancer registry is a system developed to collect and manage the clinical and pathological information of patients with cancer within a defined population. To efficiently and correctly identify patients with cancer within the population, a sufficient number of qualified registrars working under rigorous case finding processes are necessary to ensure high-quality screening and data extraction [1]. Among the various sources of documents suitable for case finding, pathology reports are the most critical, as 64% to 97% of cancer cases are morphologically verified via histology or cytology [2]. Accordingly, the Surveillance, Epidemiology, and End Results Program emphasizes the review of all pathology reports as a vital component of the case finding process [3].

However, with the rising incidence of cancer, the expansion of early detection programs, and increasing routine health examinations, pathology laboratories face a continuous upward trend in workload [4,5]. As the global cancer burden increases, cancer registries encounter challenges such as high patient volumes, limited infrastructure, and staffing shortages. These factors can compromise data quality and hinder use, especially in resource-constrained countries where hospital-based cancer registries serve as the essential data source for national health policy planning [6].

Several studies have explored AI to support cancer registry processes, ranging from rule-based systems to deep neural networks and large language models (LLMs). Early efforts such as that by Hanauer et al [7] used cancer-specific terms and negative phrase detection to screen pathology reports, achieving 100% sensitivity and 85% specificity. Similarly, Spini et al [8] combined free-text keywords with Systematized Nomenclature of Medicine codes to identify non–small cell lung cancer cases in a region-wide registry. More recently, deep learning models have gained attention due to their ability to learn from both historical and new data without the need for constant rule updates or manual adjustments while also achieving higher performance on complex, unstructured text. For example, Preston et al [9] implemented deep learning models, including convolutional neural networks (CNNs), gated recurrent units, and hierarchical attention networks, to identify patients with cancer and extract oncology attributes from multiple sources (pathology reports, radiology reports, and operative notes) across 28 cancer care centers in the United States, achieving an F_1-score of 97.3 for the case finding task. Gondara et al [10] also demonstrated that fine-tuned transformer-based language models (GatorTron and BlueBERT) could classify tumor reportability from unstructured pathology reports in a population-based registry, maintaining a 98% true reportable rate in prospective deployment. More recently, Krawczuk et al [11] compared hierarchical self-attention network (HiSAN) with a general-purpose LLM (Llama 3-8B) for metastasis detection from pathology reports at scale, with HiSAN outperforming the LLM with a recall of 0.894 compared to 0.824. Despite these advancements, hospital-based settings remain understudied and present distinct challenges compared to their population-based counterparts. A critical limitation is data incompleteness driven by referral patterns; institutional databases often contain only secondary findings (eg, metastatic biopsies and surveillance specimens) rather than primary diagnostic reports. This incompleteness affects supervised learning models requiring well-labeled training data. Additionally, ground truth labeling is problematic because cancer registries organize data by patients or tumors rather than by individual reports, making it difficult to accurately match multiple pathology reports from a single patient to their corresponding cancer attributes.

Prior studies have used various strategies to mitigate this label-input mismatch. For instance, Gao et al [12] used tumor-level annotations as ground truth labels for all reports with the same tumor ID even though these labels did not always align with the content of individual pathology reports. Preston et al [9] addressed this issue by concatenating all pathology reports in chronological order and using the combined document as input for the machine learning model. Another approach is to include only patients with no more than a single pathology report in the training dataset. However, this significantly reduces the sample size and potentially introduces algorithmic bias [12,13]. The most accurate and also the most expensive approach is to manually annotate pathology reports at the report level [14]. Although the manual annotation approach yields the most valid training data, it is the most resource-intensive of all methods. It often results in limited numbers of patients and reports, which can be a major constraint in developing reliable machine learning models.

On the basis of the analysis, no study has systematically evaluated weakly supervised learning as a solution for these hospital settings where limited access to comprehensive pathology report data may increase noise and errors in the dataset, further compromising model performance. This study aimed to compare the performance of various supervision strategies and analyze the specific failure modes associated with each approach.

Methods

Data Sources

Data were retrieved from the hospital information system database at Songklanagarind Hospital, the largest tertiary hospital in southern Thailand, and comprised all pathology reports recorded between January 1, 2005, and December 31, 2023. Each record included the patient's hospital number, specimen number, date of tissue retrieval, and findings written by the pathologist. All reports were written in English; example reports are shown in Multimedia Appendix 1. These reports were linked to the hospital-based cancer registry

using unique hospital numbers to retrieve associated cancer attributes; the data linkage structure is illustrated in Multimedia Appendix 1.

In addition to structured morphology and topography codes, cancer registrars routinely record a free-text variable labeled "Pathology detail." This field encompasses the content of specific pathology reports that registrars deemed relevant for diagnosis (ie, reports confirming cancer morphology, topography, grade, or stage). Maintained primarily to facilitate data abstraction and quality audits, this "Pathology detail" variable served as the matching criterion for the registrar-annotated approach in our report-level labeling.

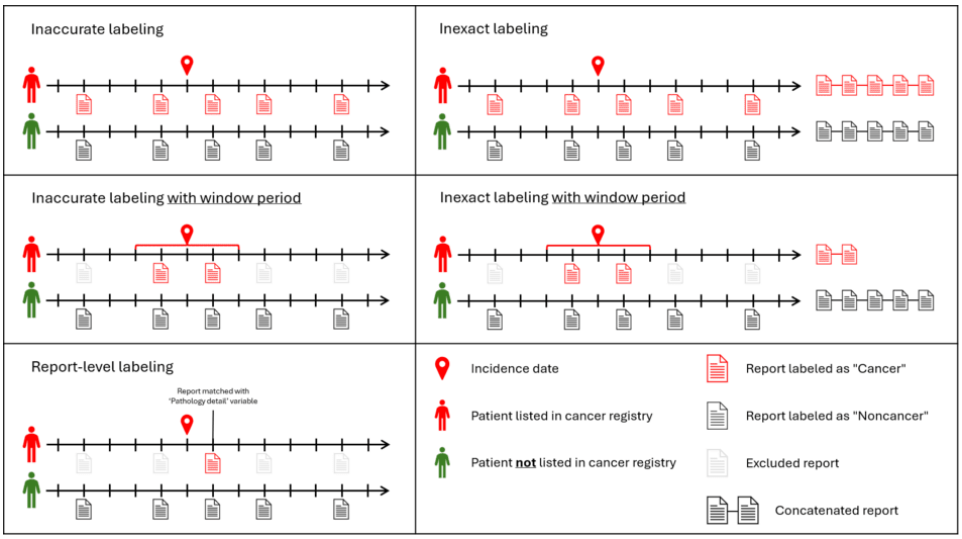
Data Preprocessing and Partitioning

Pathology reports underwent natural language preprocessing, including lower casing, punctuation normalization, and tokenization via the WordPiece algorithm using the *keras_nlp* Python package (version 0.25.0). Numerical values were masked using special tokens ("decimal" or "large_integer">100) to reduce the vocabulary space following the strategy described by Gao et al [12], and sequences were right padded to the maximum token length in each batch. The WordPiece vocabulary was derived from the training and validation sets with a vocabulary size of 10,000 tokens. To ensure temporal independence and assess model robustness to natural variation in clinical documentation over time, including shifts in reporting style and wording patterns arising from pathologist turnover, patients were assigned to splits based on their most recent report date: those with reports from 2020 onward formed the test set used for model validation, while pre-2020 patients were randomly split into training (80%) and validation (20%) sets used for model development.

Labeling Strategies for Development

We implemented 3 labeling approaches (Figure 1). "Inaccurate labeling" propagated the patient's cancer status to all of that patient's individual reports, with variants restricting data to 7-day or 30-day windows around the diagnosis date to mitigate noise. "Inexact labeling" combined all patient reports chronologically (separated by "SEP" tokens) into a single input labeled by patient status; this also included 7-day and 30-day window variants. "Report-level labeling" used the registrar-curated "Pathology detail" field to isolate specific tumor-confirming reports as ground truth. All other reports from patients with cancer were excluded under this approach. This approach prioritized label accuracy but reduced dataset size by excluding nondiagnostic reports from patients with cancer.

Figure 1. Data labeling strategies used to assign cancer classification labels to pathology reports for model training.



Performance Evaluation

Performance was assessed on the test dataset using 3 methods corresponding to distinct real-world use cases and ground truth definitions. First, max-pooled evaluation classified patients based on the maximum score derived from processing reports individually, simulating a real-time screening alert system where any single report flagging cancer was sufficient to identify the patient; ground truth was defined at the patient level using cancer registry status. Second, concatenated evaluation processed a patient's full report history as a single input, simulating a scenario where the complete longitudinal record was available and testing whether models could synthesize information across multiple reports to improve patient-level identification; ground truth was similarly defined at the patient level. Third, individual report evaluation measured the model's ability to identify specific tumor-confirming reports against registrar-annotated "Pathology detail" entries as ground truth. No temporal window restriction was applied to the test dataset. Although the report-level labeling approach offers the least label noise arising from discordance between report content and ground truth, it does not reflect real-world case finding workflows in tertiary referral settings, where primary diagnostic tissue is frequently unavailable locally due to referral patterns, leaving only secondary or surveillance reports as the available evidence.

Model Development

Five neural network architectures were developed (Multimedia Appendix 1). Simple word-embedding model (SWEM) [15] and multilayer perceptron (MLP) served as baselines, representing the simplest class of neural text classifiers that process text nonsequentially, treating the document as an unordered set of word embeddings. CNN [16] and recurrent neural network (RNN) [17] were used to evaluate architectures that exploit local sequential structure. Finally, HiSAN [12] was selected to evaluate an architecture capable of modeling long-range dependencies across the entire document. The first 4 models did not require a fixed token length, so no further padding or truncation was applied. The HiSAN model, in contrast, required a fixed token length to be specified in its architecture. For the HiSAN model, all sequences were right padded to a length of 8192 tokens to ensure coverage of the token lengths of the dataset. All models used randomly initialized embedding layers. Models were implemented using TensorFlow (version 2.18.0; Google Brain Team) with Python (version 3.12.0; Python Software Foundation). Model training was performed on a single NVIDIA RTX 6000 Ada Generation graphics processing unit.

Hyperparameter optimization was performed using the Tree-Structured Parzen Estimator algorithm [18], with the search ranges detailed in Table 1. Balanced class weights were applied to account for class imbalance using the following formula: $W_j=\frac{N}{K\times n_j}$, where W_j is the weight for class j, N is the number of samples in the dataset, K is the number of classes, and n_j is the number of samples in class j. The best-performing model configuration was selected based on the F_1-score achieved on the validation dataset.

Table 1. Hyperparameter search ranges for 5 model architectures. The search was optimized using the Tree-Structured Parzen Estimator algorithm[a].

Model and hyperparameter	Value range
All models	
Batch size	128
Embedding dimension	128-1024
Dropout rate	0.0-0.9
Loss function	Cross-entropy and focal cross-entropy [19]
Learning rate	0.01-0.0001
SWEM[b]	
Pooling layer	Max pooling, average pooling, and concatenated
Layer width	32-1024
Activation function	Linear, tanh, ReLU[c], and swish
MLP[d]	
Pooling layer	Max pooling, average pooling, and concatenated
Layer depth	1-4
Layer width	32-1024
Activation function	Linear, tanh, ReLU, and swish
RNN[e]	
RNN architecture	LSTM[f] and GRU[g]
Layer width	32-512
CNN[h]	
Kernel size	3, 4, 5, 6, and 7
Activation function	Linear, tanh, ReLU, and swish
Number of filters	24-768
Pooling layer	Max pooling and average pooling
HiSAN[i]	
Tokens per segment	8, 16, 32, 64, 128, 256, and 512
Number of attention heads	1-4
Attention size	128-1024

[a]Focal cross-entropy loss parameters were fixed at default values: alpha=0.25 and gamma=2.0.
[b]SWEM: simple word-embedding model.
[c]ReLU: rectified linear unit.
[d]MLP: multilayer perceptron.
[e]RNN: recurrent neural network.
[f]LSTM: long short-term memory.
[g]GRU: gated recurrent unit.
[h]CNN: convolutional neural network.
[i]HiSAN: hierarchical self-attention network.

Statistical Analysis

The selected models were evaluated on the held-out temporally split test dataset. Performance metrics included accuracy, precision, recall, binary F_1-score (harmonic mean of precision and recall of the positive class), and area under the receiver operating characteristic curve (AUC). Classification thresholds for accuracy, precision, recall, and F_1-score were determined by selecting the cutoff that maximized the Youden index (sensitivity + specificity − 1) on the validation dataset. The 95% CIs for all metrics were calculated using bootstrap resampling with 1000 bootstrap iterations. The resampling unit for bootstrapping was at the patient level for max-pooled and concatenated evaluations and at the individual report level for the individual report evaluation.

To investigate failure modes, we conducted an error analysis by randomly sampling 100 cases from the discordant cases in the test dataset. These cases underwent detailed manual review to characterize the causes of disagreement between model predictions and registry labels.

Ethical Considerations

This study was approved by the Institutional Review Board of the Faculty of Medicine, Prince of Songkla University (REC.67-533-51-4) under an expedited review procedure. Informed consent was waived as the study involved retrospective secondary analysis of deidentified registry and pathology data with no direct patient contact. All data were anonymized prior to analysis, and no personally identifiable information was retained. No compensation was provided. No patient-identifiable information was included in this manuscript.

Results

Dataset Characteristics

Between 2005 and 2023, Songklanagarind Hospital generated 675,933 pathology reports, which were included in this dataset (Figure 2). Reports were derived from 270,446 unique patients, of whom 72,446 (26.8%) were listed in the cancer registry, contributing 38.9% (262,774/675,933) of the total volume of reports. Table 2 shows the characteristics of the study dataset. The overall cohort was predominantly female (192,076/270,446, 71.0%), with a mean age of 47 (SD 19) years. Demographic characteristics were well balanced across the development and test sets. Among patients with cancer, the most common tumor sites were female genital organs (15,280/72,446, 21.1%), digestive organs (12,633/72,446, 17.4%), and the respiratory system (10,012/72,446, 13.8%). Unknown stage comprised the largest group (21,959/72,446, 30.3%), followed by stage 4 (16,490/72,446, 22.8%), while the test set contained a notably higher proportion of stage 1 cases (2388/13,223, 18.1% vs 6560/59,223, 11.1%) and a lower proportion of stage 4 cases (2444/13,223, 18.5% vs 14,046/59,223, 23.7%). Regarding report token length, individual reports had a median of 109 tokens (IQR 96-121, maximum 1562) in the development set and 110 tokens (IQR 98-127, maximum 1538) in the test set. When reports were concatenated, as in the inexact labeling approach and concatenated report evaluation, the median token length was 166 tokens (IQR 107-321, maximum 6099) in the development set and 287 tokens (IQR 133-606, maximum 6175) in the test set.

Figure 2. Flow diagram of data partitioning. Patients with reports from 2020 onward were assigned to the test set; pre-2020 patients were randomly split 80:20 into training and validation sets. n represents the number of records. *Percentage of records labeled as cancer. Records are at the patient level for the inexact labeling approach, and at the individual report level for the inaccurate and report level labeling approaches.

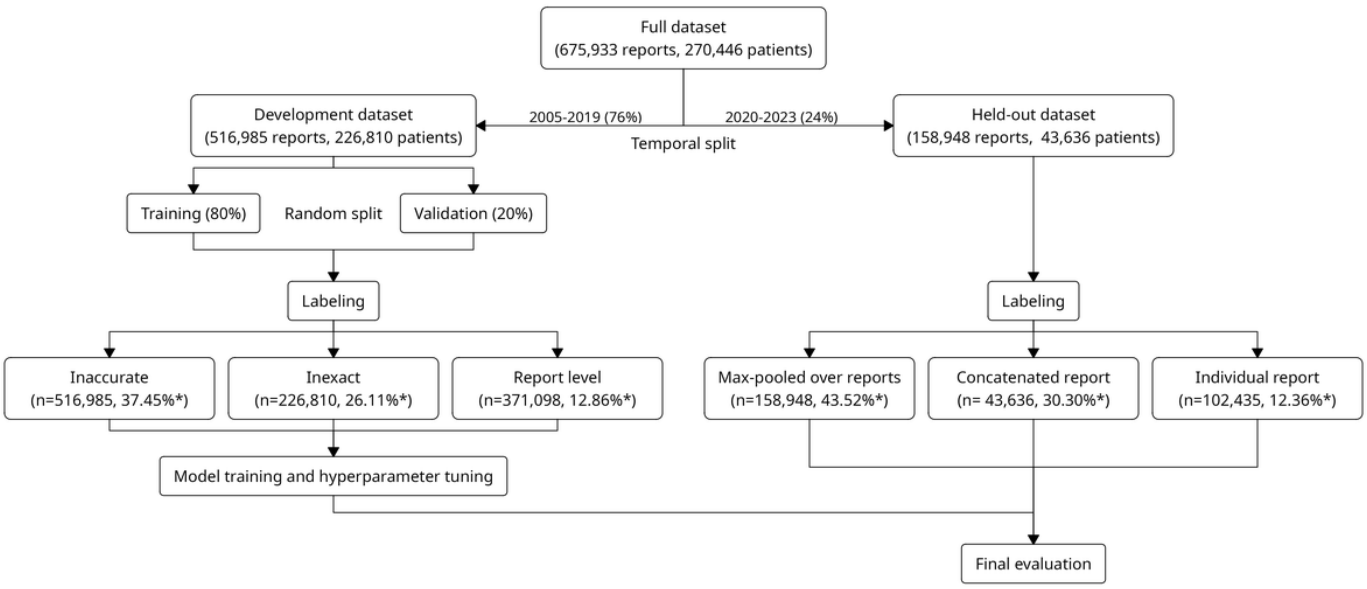
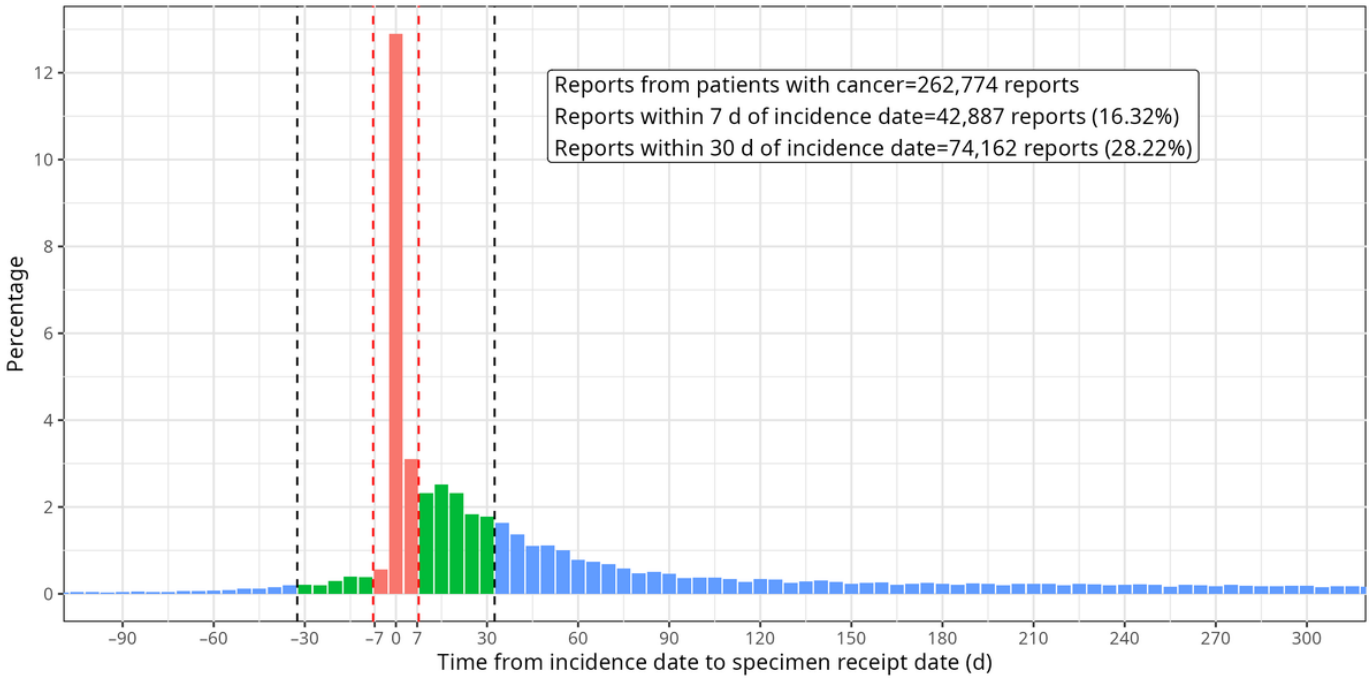


Table 2. Patient and tumor characteristics of the study population. Tumor sites were classified according to the *International Classification of Diseases for Oncology, Third Edition* (N=270,446).

Patient characteristics	Overall	Development (n=226,810)	Test (n=43,636)
Sex, n (%)			
Female	192,076 (71.0)	160,745 (70.9)	31,331 (71.8)
Male	78,354 (29.0)	66,052 (29.1)	12,302 (28.2)
Unknown	16 (0.0)	13 (0.0)	3 (0.0)
Age (y), mean (SD)	47 (19)	47 (19)	48 (17)
Tumor characteristics, n/N (%)			
Tumor site			
Lip, oral cavity, and pharynx (C00-C14)	6064/72,446 (8.4)	5321/59,223 (9.0)	743/13,223 (5.6)
Digestive organs (C15-C26)	12,633/72,446 (17.4)	10,373/59,223 (17.5)	2260/13,223 (17.1)
Respiratory system and intrathoracic organs (C30-C39)	10,012/72,446 (13.8)	8735/59,223 (14.7)	1277/13,223 (9.7)
Bones, joints, and articular cartilage (C40-C41)	488/72,446 (0.7)	399/59,223 (0.7)	89/13.223 (0.7)
Hematopoietic and reticuloendothelial systems (C42)	3525/72,446 (4.9)	2956/59,223 (5.0)	569/13,223 (4.3)
Skin (C44)	2178/72,446 (3.0)	1843/59,223 (3.1)	335/13,223 (2.5)
Peripheral nerves and autonomic nervous system (C47)	59/72,446 (0.1)	46/59,223 (0.1)	13/13,223 (0.1)
Retroperitoneum and peritoneum (C48)	226/72,446 (0.3)	199/59,223 (0.3)	27/13,223 (0.2)
Connective, subcutaneous, and soft tissues (C49)	1181/72,446 (1.6)	985/59,223 (1.7)	196/13,223 (1.5)
Breast (C50)	7925/72,446 (10.9)	6040/59,223 (10.2)	1885/13,223 (14.3)
Female genital organs (C51-C58)	15,280/72,446 (21.1)	11,706/59,223 (19.8)	3574/13,223 (27.0)
Male genital organs (C60-C63)	2144/72,446 (3.0)	1764/59,223 (3.0)	380/13,223 (2.9)
Urinary tract (C64-C68)	2150/72,446 (3.0)	1782/59,223 (3.0)	368/13,223 (2.8)
Eye, brain, and central nervous system (C69-C72)	2531/72,446 (3.5)	2037/59,223 (3.4)	494/13,223 (3.7)
Thyroid and endocrine glands (C73-C75)	3019/72,446 (4.2)	2333/59,223 (3.9)	686/13,223 (5.2)
Other and ill-defined sites (C76)	123/72,446 (0.2)	112/59,223 (0.2)	11/13,223 (0.1)
Lymph nodes (C77)	2126/72,446 (2.9)	1892/59,223 (3.2)	234/13,223 (1.8)
Unknown primary site (C80)	782/72,446 (1.1)	700/59,223 (1.2)	82/13,223 (0.6)
Tumor stage			
0	2148/72,446 (3.0)	1692/59,223 (2.9)	456/13,223 (3.4)
1	8948/72,446 (12.4)	6560/59,223 (11.1)	2388/13,223 (18.1)
2	10,660/72,446 (14.7)	8428/59,223 (14.2)	2232/13,223 (16.9)
3	12,241/72,446 (16.9)	9907/59,223 (16.7)	2334/13,223 (17.7)
4	16,490/72,446 (22.8)	14,046/59,223 (23.7)	2444/13,223 (18.5)
Unknown	21,959/72,446 (30.3)	18,590/59,223 (31.4)	3369/13,223 (25.5)

In the development (training and validation) set, 47,711 reports were identified via the "Pathology detail" variable and served as the positive class in the training data for the report-level labeling approach. The test dataset was structured to support 3 evaluation tasks: the max-pooled application (158,948 individual reports), the concatenated application (43,636 patients), and the individual report prediction task (102,435 reports); detailed sample counts by class label, dataset split, and labeling approach including class weights and token length statistics are provided in Multimedia Appendix 1. Regarding temporal distribution, only 16% (42,887/262,774) and 28% (n=74,162) of pathology reports from patients with cancer occurred within 7 and 30 days of the incidence date, respectively (Figure 3).

Figure 3. Temporal distribution of pathology reports from patients with cancer relative to their registry-recorded cancer incidence date. Positive values indicate reports generated after the incidence date; negative values indicate reports prior to diagnosis. Vertical dashed lines indicate the 7- and 30-day temporal windows applied in labeling strategy variants.



Model Performance

Figure 4 presents the AUC scores on the held-out test dataset. Models trained with inexact labeling (no window period) performed consistently well across all patient-level evaluation tasks. For concatenated report classification, the CNN and RNN achieved AUCs of 98.16 (95% CI 98.00-98.30) and 98.17 (95% CI 98.01-98.31), respectively. For max-pooled classification, the CNN attained an AUC of 97.86 (95% CI 97.70-98.02), while HiSAN trained with inaccurate labeling and no window period recorded the highest AUC of 97.94 (95% CI 97.79-98.10). In the individual report classification task, the CNN demonstrated robust performance, with an AUC of 99.80 (95% CI 99.76-99.83).

While report-level labeling yielded the highest accuracy for individual report classification (CNN AUC=99.89), it suffered substantial degradation when applied to patient-level tasks. Specifically, AUCs dropped to 92.86 for max-pooled classification and 92.11 for concatenated classification compared with the best-performing models for these tasks, which achieved AUCs of 97.94 and 98.17, respectively.

Inaccurate labeling showed a trade-off across window period constraints. When a window period was applied, the models demonstrated high performance on the individual report classification task, with the CNN achieving an AUC of 99.78 using a 30-day window. In contrast, without a window period, the HiSAN yielded AUCs of 97.94 and 97.95 for max-pooled and concatenated classification, respectively.

Figure 5 shows the receiver operating characteristic curves of each model evaluated on the held-out test dataset using max-pooled classification to further explore the performance discrepancies between models. Models trained using report-level labels, as well as those trained under weak supervision with window periods, revealed a marked notch in the upper portion of the curves, indicating a sharp decline in specificity when high sensitivity was desired. This limits these models' practical utility as a screening tool as achieving the sensitivity thresholds necessary would result in unacceptably low specificity, thereby generating excessive false positives. In contrast, the model trained using weak supervision without a window period did not show this notched pattern, allowing selection of a threshold that achieves high sensitivity while maintaining acceptable false positive rates. Additional performance metrics (precision, recall, accuracy, and F_1-score) are shown in Multimedia Appendix 1.

Figure 4. Area under the receiver operating characteristic curve (AUC) with 95% bootstrapped CIs for 5 model architectures trained under 7 labeling strategies evaluated on the held-out test set across 3 evaluation tasks. Columns represent the 7 labeling approaches; rows represent the 3 evaluation tasks. The x-axis represents model architectures. CNN: convolutional neural network; HiSAN: hierarchical self-attention network; MLP: multilayer perceptron; RNN: recurrent neural network; SWEM: simple word-embedding model.

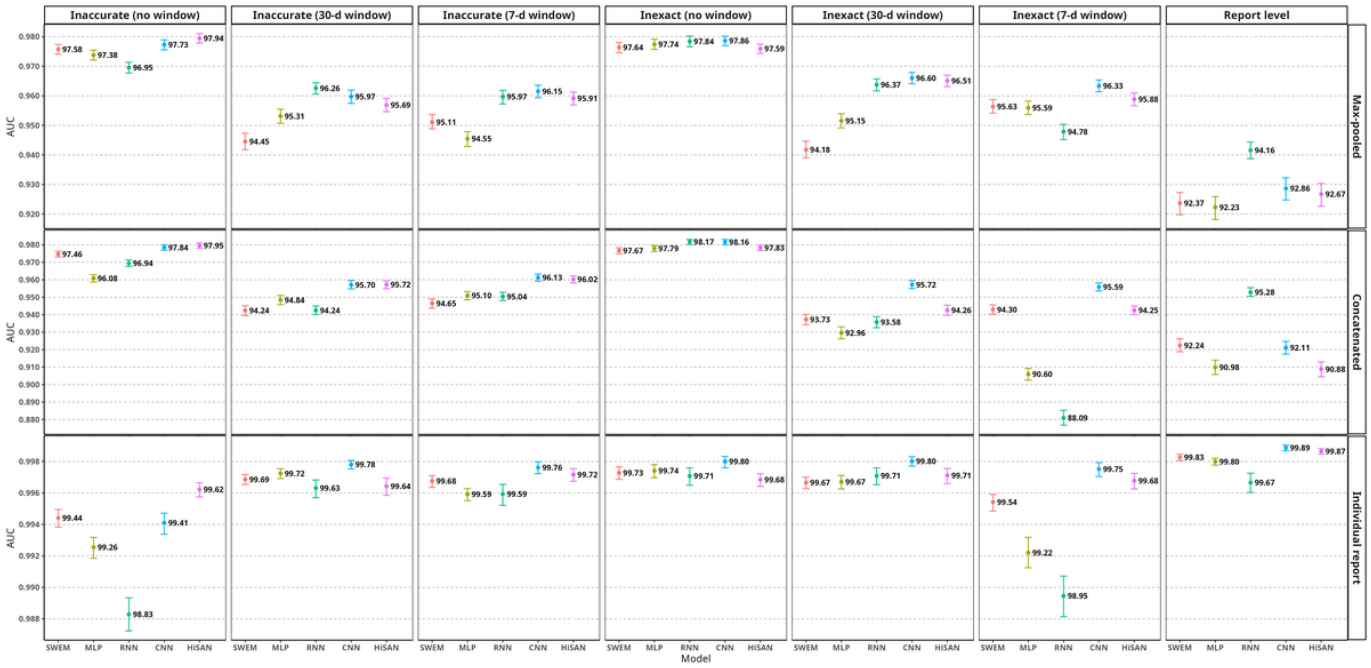
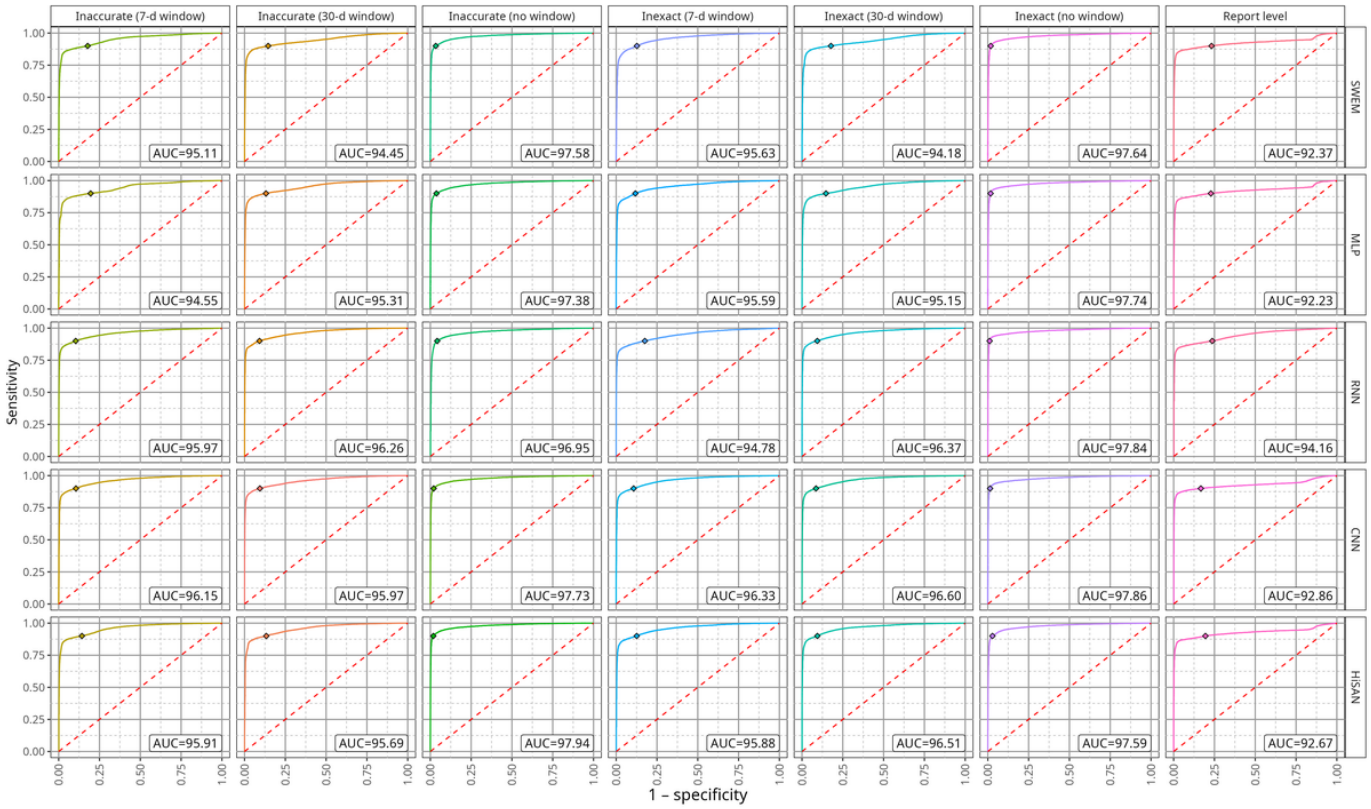


Figure 5. Receiver operating characteristic (ROC) curves for 5 model architectures trained under 7 labeling strategies evaluated on the max-pooled classification task. Columns represent the 7 labeling approaches; rows represent the 5 model architectures. Area under the receiver operating characteristic curve (AUC) values are displayed within each panel. CNN: convolutional neural network; HiSAN: hierarchical self-attention network; MLP: multilayer perceptron; RNN: recurrent neural network; SWEM: simple word-embedding model.



Effect of Window Periods

The application of window periods produced contrasting effects depending on the evaluation level. For patient-level tasks (max-pooled and concatenated classification), window periods consistently reduced performance. For example, applying 30- and 7-day windows to the CNN (inaccurate labeling) decreased the concatenated AUC from 97.84 to 95.70 and 96.13, respectively.

However, the pattern was reversed for individual report classification. Models trained using inaccurate labeling (no window) performed relatively poorly (AUC=99.41), whereas applying 30- and 7-day windows substantially improved performance to 99.78 and 99.76, respectively. Notably, the inexact labeling approach showed minimal sensitivity to window periods for individual report classification, maintaining AUCs at or above 99.75 across all configurations for CNN models.

Comparison Across Model Architectures

The CNN architecture consistently outperformed simpler models (SWEM and MLP) in 19 of 21 configurations. CNN demonstrated particular robustness across labeling approaches and evaluation tasks, achieving the highest or near-highest AUC in most configurations.

HiSAN showed unexpected performance patterns. Despite its hierarchical design intended to leverage information across multiple reports, it did not demonstrate superior performance in concatenated report evaluation, where such capability would be advantageous. Instead, HiSAN only outperformed CNN and RNN when trained using inaccurate labeling approaches—achieving an AUC of 97.95 (95% CI 97.80-98.10) for concatenated classification and 97.94 (95% CI 97.79-98.10) for max-pooled classification compared to CNN's 97.84 and 97.73, respectively. For inexact and report-level labeling approaches, HiSAN did not surpass the performance of CNN or RNN models.

Error Analysis

A detailed review of 100 discordant cases from the best-performing CNN model revealed that most classification disagreements stemmed from data limitations and linguistic ambiguity rather than algorithmic failure. The primary source of error was insufficient institutional data (45%), where referred patients lacked primary diagnostic reports in the local system, leaving only benign surveillance documents. False positives (41%) were largely driven by benign reports containing cancer-associated terminology or equivocal clinical impressions. Notably, the model identified registry omissions in 5% of cases, correctly flagging malignancies (eg, polycythemia vera) missed by human registrars. Only 9% of cases represented true false negatives, specifically involving benign central nervous system tumors and myelodysplastic syndromes.

Discussion

Real-World Data and Labeling Strategies

While Thailand's national cancer incidence is 154 per 100,000 population [20], pathology reports from patients with cancer comprised nearly 39% of our dataset. This disproportionate volume reflects clinical reality: cancer care involves a longitudinal trajectory of histological examinations, ranging from initial diagnosis to staging, surveillance, and investigation of recurrence or metastasis. This pattern not only underscores the central role of pathology reports in hospital-based registries but also highlights the challenge that supervised learning models face when mapping multiple reports to a single patient-level diagnosis, a challenge similarly noted in recent studies [21].

A key finding of this study is the counterintuitive performance of labeling strategies. Models trained on report-level data (the cleanest labels, hand-picked by registrars) suffered a substantial performance decline when applied to patient-level tasks (AUC dropping from 99.89 to approximately 92.11-92.86). This occurred because registrars typically annotate only the definitive primary diagnostic report. Consequently, models trained on this narrow distribution overfitted to perfect diagnostic language and failed to recognize the secondary evidence of cancer—such as metastatic workups or surveillance biopsies—which are often the only records available for referred patients.

In contrast, inaccurate and inexact labeling approaches demonstrated robust transferability (AUC of 97.73-98.16). By treating all reports from a patient with cancer as "positive" (or effectively "associated with cancer"), these models learned to identify a diverse array of cancer signatures beyond the primary tumor site. This capability is critical for hospital-based registries, where referred patients frequently lack a primary diagnostic report in the local database and must be identified via secondary findings. This robustness is consistent with broader evidence that deep learning models can remain effective under systematic label noise when appropriate mitigation strategies are applied [22]. Hsu et al [21] demonstrated that case-level inference—evaluating each pathology report in the context of other reports from the same case—substantially improved performance on data abstraction tasks, suggesting that patient-level context is valuable at inference time. In our case finding setting, however, max-pooled and concatenated evaluations yielded comparable performance, indicating that case finding relies less on cross-document information synthesis than data abstraction does. Rather, the key challenge is exposure to the full diversity of cancer-associated report types during training, providing a conceptual rationale for inaccurate and inexact labeling as a preparation strategy for the varied report landscape encountered in real-world case finding. Notably, the label noise introduced by inaccurate and inexact labeling differs structurally from the noise typically studied in the literature, which has focused predominantly on computer vision tasks, with comparatively few studies addressing natural language processing applications [22]. Prior work

has generally attributed label noise to data entry errors, machine-extracted labels, or disagreement among experts. In our study, however, noise arises primarily from referral patterns that produce systematic content mismatch: reports labeled positive under inaccurate and inexact labeling may reflect surveillance, metastatic, or secondary findings rather than primary diagnostic evidence. This represents a more complex, institutionally specific pattern of noise than the random or case difficulty–based errors typically examined in prior studies.

Model Architecture Comparison

CNN models consistently outperformed nonsequential architectures (SWEM and MLP), confirming that local pattern recognition is essential for extracting clinical concepts [15]. CNN showed particular robustness across labeling approaches and applications, achieving the highest or near-highest AUC in most configurations. The inclusion of CNN and HiSAN in our model evaluation is further motivated by their deployment in the National Cancer Institute and Department of Energy production data abstraction pipeline reported by Hsu et al [21], where both architectures are used across 19 Surveillance, Epidemiology, and End Results registries. However, while both architectures perform well in their data abstraction context, our results indicate that CNN consistently outperformed HiSAN for case finding, suggesting a task-specific distinction: data abstraction may benefit from HiSAN's hierarchical attention to synthesize information across document segments, whereas case finding is primarily a pattern detection task dependent on local recognition of cancer-associated terminology. HiSAN showed a marginal advantage only under the noisiest labeling conditions, suggesting that its attention mechanisms may provide some benefit in filtering irrelevant segments, but long-range dependency modeling confers limited additional gain for case finding.

The Window Period Trade-Off

The analysis of temporal window periods reveals a fundamental trade-off between reducing label noise and preserving information completeness. It is observable that while narrowing the temporal window improved performance for individual report classification by eliminating nondiagnostic "noise" and preventing spurious associations, it conversely degraded patient-level case finding. This decline occurs because strict temporal filtering strips away valuable longitudinal context—such as metastatic workups or recurrence checks occurring years after diagnosis—which are often the only signals available for referred patients. Consequently, while precise temporal filtering optimizes models for specific diagnostic verification, it hinders holistic screening models that rely on the full, albeit messy, clinical history to capture the complete disease trajectory.

Error Analysis and Model Limitations

Manual review revealed that the models relied heavily on cancer-associated terminology rather than deeper diagnostic reasoning. False positives frequently arose from benign findings described with complex cancer-like vocabulary or equivocal language. However, the identification of 5 registry-missed cases demonstrates the model's potential as a safety net for human registrars.

The predominance of false negatives due to insufficient institutional data highlights the intrinsic limitation that no algorithm can identify a patient with cancer if the relevant diagnostic evidence is in an external hospital's database.

Finally, the difficulty in classifying "gray zone" entities, such as central nervous system tumors described as "benign" despite reportability and myelodysplastic syndromes, suggests that future work must incorporate more nuanced, registry-specific inclusion criteria. Reports with mixed language ("negative for malignancy" in some sections alongside positive findings) confused classification logic, suggesting refinement opportunities in handling registry-specific criteria and multicomponent reports. Deployment strategies such as OR logic, in which a case is classified as reportable if at least one model predicts it as reportable [10], or confidence-based thresholding to defer low-confidence cases for human review [11] represent promising directions for minimizing false negatives while managing registrar workload in practice.

Limitations

First, this study was conducted at a single tertiary hospital in southern Thailand, limiting generalizability to settings with different case mixes, referral patterns, documentation practices, or reportability criteria. Smaller hospital-based registries with fewer referred-in cases may experience less labeling noise as primary diagnostic reports are more likely to be available locally. Multicenter validation is needed to assess broader applicability.

Second, potential label noise in the test set stems from 2 sources: registry omissions undetected at the time of data collection and the reliance on pathology reports alone for annotation, whereas real-world registrars additionally consult physician notes, radiology reports, and other clinical documentation. Labels were not retroactively corrected, reflecting the naturalistic data collection process. While this noise may cause performance metrics to underestimate true model capability, its consistent effect across all models preserves the validity of internal comparisons. Results should be interpreted with caution when generalized beyond this study or compared against externally validated benchmarks.

Third, this study evaluated convolutional, recurrent, and hierarchical self-attention architectures but did not include pretrained transformer-based models such as bidirectional encoder representations from transformers or its domain-specific variants (eg, ClinicalBERT or PathologyBERT). The primary scope of this study was to systematically compare labeling strategies across established task-specific architectures rather than benchmark pretrained language models; the substantially higher parameter count and computational demands of transformer-based models were secondary considerations in resource-constrained registry settings. Recent studies have demonstrated the effectiveness of transformer-based models for cancer registry tasks [23], including tumor reportability classification and tumor group

assignment from pathology reports [10,24,25]. Whether these models maintain robustness under weakly supervised or noisy label conditions as evaluated in this study remains an open question and warrants further investigation.

Future Work

Future studies should validate these models across multiple institutions with varying referral patterns and documentation practices to assess generalizability. Prospective evaluation with registrar-annotated gold-standard labels rather than retrospective registry data subject to omissions would also provide a more reliable benchmark for model performance.

Conclusions

In this single tertiary referral hospital setting, weakly supervised labeling outperformed report-level human annotation for patient-level case finding, whereas report-level labeling performed best for individual report classification. The CNN architecture achieved the highest or near-highest AUC in most configurations.

Acknowledgments

The authors would like to thank the cancer registrars at Songklanagarind Hospital for their dedicated work in maintaining the hospital-based cancer registry. Their routine data collection made this study possible. This work was partially supported by the PSU-Graduate Studies Scholarship.

Generative AI Use

During the preparation of this work, the authors used AI (Claude Haiku 4.5; Anthropic) to improve writing clarity and grammar. After using this tool, the authors reviewed and edited the outputs as needed and take full responsibility for the content of the published article.

Funding

This work was partially supported by the PSU-Graduate Studies Scholarship.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available due to patient privacy and ethical restrictions but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: CT, CK, SLG, TI
Data curation: CT
Formal analysis: CT
Investigation: CT
Methodology: CT
Project administration: TI
Resources: TI
Software: CT
Supervision: TI
Visualization: CT
Writing—original draft: CT, TI
Writing—review and editing: CT, TI

Conflicts of Interest

None declared.

Multimedia Appendix 1

Neural network architectures, data linkage, pathology report examples, additional dataset characteristics, and model performance.
[DOCX File (Microsoft Word File), 382 KB-Multimedia Appendix 1]

References

1. Solomon DJ, Henry RC, Hogan JG, Van Amburg GH, Taylor J. Evaluation and implementation of public health registries. Public Health Rep. 1991;106(2):142-150. [Medline: 1902306]
2. Bashar MA, Thakur JS, Budukh A. Evaluation of data quality of four new Population Based Cancer Registries (PBCRs) in Chandigarh and Punjab, North India- a quality control study. Asian Pac J Cancer Prev. May 1, 2021;22(5):1421-1433. [doi: 10.31557/APJCP.2021.22.5.1421] [Medline: 34048170]
3. Pathology reports. Surveillance, Epidemiology, and End Results Program at the National Cancer Institute. URL: https://training.seer.cancer.gov/casefinding/sources/pathology.html [Accessed 2025-05-08]
4. Walsh E, Orsi NM. The current troubled state of the global pathology workforce: a concise review. Diagn Pathol. Dec 21, 2024;19(1):163. [doi: 10.1186/s13000-024-01590-2] [Medline: 39709433]

5. Pathology laboratories market (2025 - 2030). Grand View Research; 2024. URL: https://www.grandviewresearch.com/industry-analysis/pathology-laboratories-market-report [Accessed 2026-11-10]
6. Jedy-Agba EE, Curado MP, Oga E, et al. The role of hospital-based cancer registries in low and middle income countries-the Nigerian case study. Cancer Epidemiol. Oct 2012;36(5):430-435. [doi: 10.1016/j.canep.2012.05.010] [Medline: 22704971]
7. Hanauer DA, Miela G, Chinnaiyan AM, Chang AE, Blayney DW. The registry case finding engine: an automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. J Am Coll Surg. Nov 2007;205(5):690-697. [doi: 10.1016/j.jamcollsurg.2007.05.014] [Medline: 17964445]
8. Spini A, Rosellini P, Bellan C, et al. Development and validation of a case-finding algorithm for the identification of non-small cell lung cancers in a region-wide Italian pathology registry. PLoS One. 2022;17(6):e0269232. [doi: 10.1371/journal.pone.0269232] [Medline: 35675338]
9. Preston S, Wei M, Rao R, et al. Towards structuring real-world data at scale: deep learning for extracting key oncology information from clinical text with patient-level supervision. arXiv. Preprint posted online on Mar 20, 2022. [doi: 10.48550/arXiv.2203.10442]
10. Gondara L, Simkin J, Arbour G, Devji S, Ng R. Classifying tumor reportability status from unstructured electronic pathology reports using language models in a population-based cancer registry setting. JCO Clin Cancer Inform. Nov 2024;8:e2400110. [doi: 10.1200/CCI.24.00110] [Medline: 39561305]
11. Krawczuk P, Fox ZR, Petkov V, et al. Large-scale deep learning for metastasis detection in pathology reports. JAMIA Open. 2025;8(4):ooaf070. [doi: 10.1093/jamiaopen/ooaf070] [Medline: 40655537]
12. Gao S, Qiu JX, Alawad M, et al. Classifying cancer pathology reports with hierarchical self-attention networks. Artif Intell Med. Nov 2019;101:101726. [doi: 10.1016/j.artmed.2019.101726] [Medline: 31813492]
13. De Angeli K, Gao S, Alawad M, et al. Deep active learning for classifying cancer pathology reports. BMC Bioinformatics. Mar 9, 2021;22(1):113. [doi: 10.1186/s12859-021-04047-1] [Medline: 33750288]
14. Causa Andrieu P, Golia Pernicka JS, Yaeger R, et al. Natural language processing of computed tomography reports to label metastatic phenotypes with prognostic significance in patients with colorectal cancer. JCO Clin Cancer Inform. Sep 2022;6(6):e2200014. [doi: 10.1200/CCI.22.00014] [Medline: 36103642]
15. Shen D, Wang G, Wang W, et al. Baseline needs more love: on simple word-embedding-based models and associated pooling mechanisms. In: Gurevych I, Miyao Y, editors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2018:440-450. [doi: 10.18653/v1/P18-1041]
16. Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv. Preprint posted online on Oct 13, 2015. [doi: 10.48550/arXiv.1510.03820]
17. Wicaksono AF, Adriani M. Single-output recurrent neural networks for sentence binary classification. In: 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE; 2016:293-296. [doi: 10.1109/ICACSIS.2016.7872723]
18. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: NIPS'11: Proceedings of the 25th International Conference on Neural Information Processing Systems. Curran Associates Inc; 2011:2546-2554. [doi: 10.5555/2986459.2986743]
19. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. arXiv. Preprint posted online on Aug 7, 2017. [doi: 10.48550/arXiv.1708.02002]
20. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. May 2021;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
21. Hsu E, Hanson H, Coyle L, Stevens J, Tourassi G, Penberthy L. Machine learning and deep learning tools for the automated capture of cancer surveillance data. J Natl Cancer Inst Monogr. Aug 1, 2024;2024(65):145-151. [doi: 10.1093/jncimonographs/lgae018] [Medline: 39102883]
22. Wei Y, Deng Y, Sun C, Lin M, Jiang H, Peng Y. Deep learning with noisy labels in medical prediction problems: a scoping review. J Am Med Inform Assoc. Jun 20, 2024;31(7):1596-1607. [doi: 10.1093/jamia/ocae108] [Medline: 38814164]
23. Hands I, Kavuluru R. A survey of NLP methods for oncology in the past decade with a focus on cancer registry applications. Artif Intell Rev. 2025;58(10):314. [doi: 10.1007/s10462-025-11316-5] [Medline: 40688631]
24. Gondara L, Simkin J, Devji S, Arbour G, Ng R. ELM: a hybrid ensemble of language models for automated tumor group classification in population-based cancer registries. arXiv. Preprint posted online on Mar 24, 2025. [doi: 10.48550/ARXIV.2503.21800]
25. Santos T, Tariq A, Das S, et al. PathologyBERT - pre-trained vs. a new transformer language model for pathology domain. AMIA Annu Symp Proc. 2023;2022:962-971. [Medline: 37128387]

Abbreviations

AUC: area under the receiver operating characteristic curve
CNN: convolutional neural network
HiSAN: hierarchical self-attention network
LLM: large language model
MLP: multilayer perceptron
RNN: recurrent neural network
SWEM: simple word-embedding model

Edited by Andrew Coristine; peer-reviewed by Heber Anandan, Oleh Pitsun, Yi Lu; submitted 06.Feb.2026; initial comments to author 12.Jun.2026; final revised version received 15.Oct.2026; accepted 20.Oct.2026; published 20.Nov.2026

Please cite as:
Toh C, Kongkamol C, Geater SL, Ingviya T
Development and Application of Deep Learning Models for Case Finding From Pathology Reports in a Hospital-Based Cancer Registry: Algorithm Development and Validation
JMIR Med Inform 2026;14:e92814
URL: https://medinform.jmir.org/2026/1/e92814
doi: 10.2196/92814

© Chanavee Toh, Chanon Kongkamol, Sarayut Lucien Geater, Thammasin Ingviya. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 20.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.