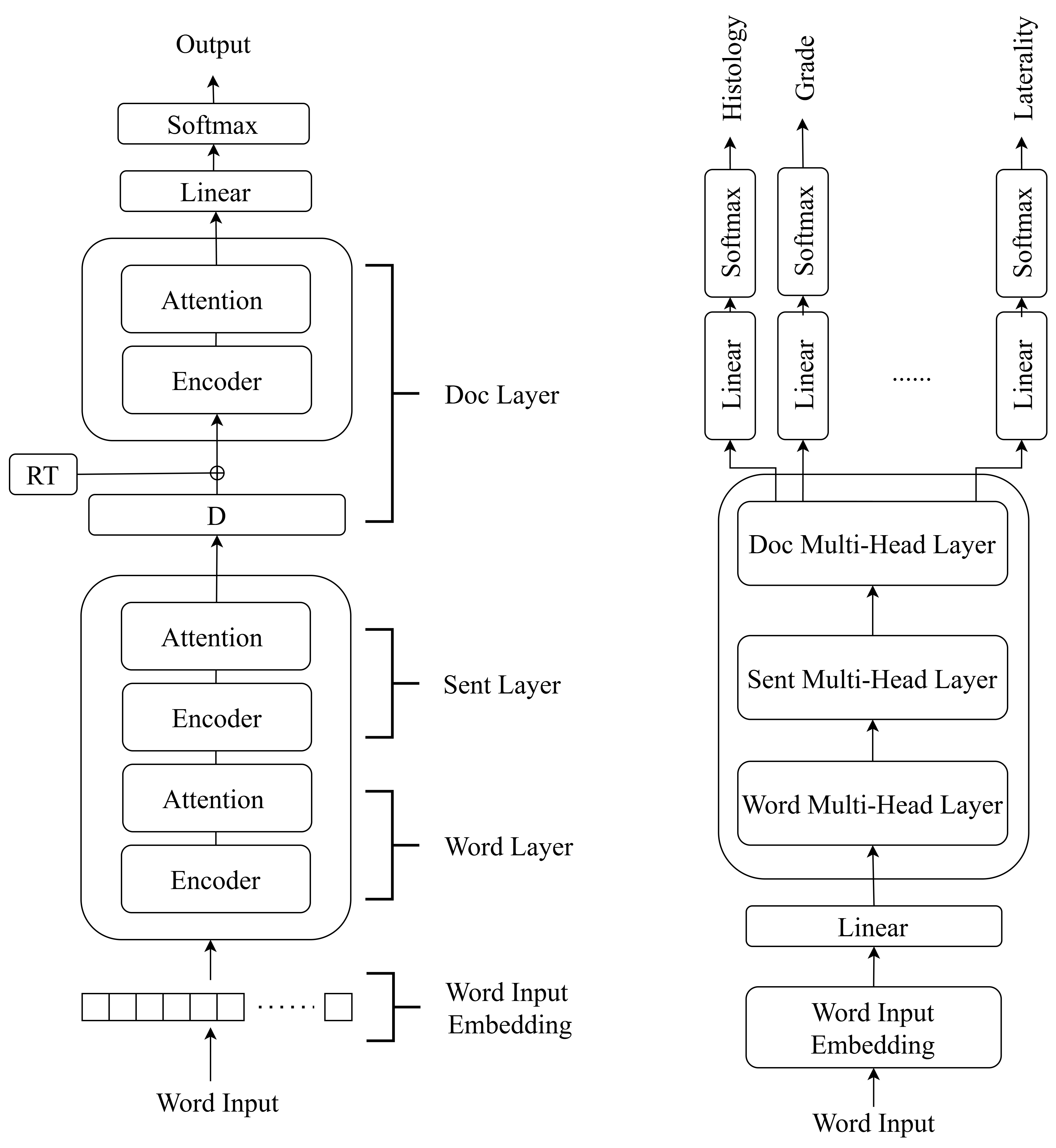
## A. Report Types Included in This Study

The report types considered in this study including: angiogram reports, cardiac ultrasound (US) reports, colonoscopy reports, computed tomography reports, echocardiogram reports, endoscopy reports, intervention reports, endoscopic retrograde cholangiopancreatography examination reports, lower limb arterial US reports, lower limb venous US reports, magnetic resonance imaging reports, mammography reports, nuclear medicine reports, pathology reports, plain file reports, special reports and transesophageal echocardiogram reports.

## B. Dataset Curation Procedure and Statistical Information of the Curated Data

TABLE B.1 summarizes the coding items considered in this study. As reported by Kao, et al. [1], among a random sample of 700 cases from 25 hospitals using long-form registry reporting, the accuracy rate of overlapping coding items between 2014 and 2016 was 95.29%, reflecting strong agreement among expert abstractors. To compile the training, validation, and test set for the presented abstraction task, we applied a data filtering procedure based on the recommendations of expert cancer registrars. This procedure aimed to prevent data quality issues that could affect subsequent model performance assessments. In the first step, we collected patients’ registry records and the corresponding reports from the cancer registry database of KMUH. Only patients with a single registry record were included to exclude cases involving multiple primary cancers where a patient is diagnosed with more than one type of cancer. In the second step, we applied the following excluding criteria:



(a) (b)

Fig. 1. (a) Architecture of the proposed HTAN. (b) Architecture of the proposed MHHTAN.

* Non-in-hospital-diagnosed or non-in-hospital-treated patients: Records that included reports from external hospitals were excluded due to concerns that incomplete cancer reports might lead to missing data.

TABLE B.I

Summary of Coding Items Considered in This Study.

|  |  |
| --- | --- |
| **Item** | **Description** |
| **The Edition and Chapter of AJCC Cancer Staging (AJCC)** | Version and chapter of the AJCC cancer staging system. |
| **Behavior (BH)** | The BH or the 5th digit of the M-code code in the pathological diagnosis. |
| **Chemotherapy at**  **This Facility (CHEMO)** | The status of chemotherapy in the patient's first treatment course. |
| **Diagnostic Confirmation (DC)** | The most precise basis of diagnosis. |
| **Grade/Differentiation Pathological (PATH\_G)** | Grading/differentiation of the solid tumor after surgery at the PS. |
| **Histology (HS)** | Structure of the primary tumor cells under the microscope. |
| **Hormone/Steroid Therapy at This Facility (HT)** | The details of hormone/steroid therapy administered during the first treatment course at the reporting hospital. |
| **Immunotherapy at This Facility (IT)** | The details of immunotherapy administered during the first treatment course at the reporting hospital. |
| **Laterality (LL)** | The side of the body or paired organs where the cancer originates. |
| **Lymph vessels or Vascular Invasion (LVI)** | Indicates whether lymphatic or vascular invasion was reported at the PS. |
| **Minimally Invasive Surgery (MIS)** | Records whether minimally invasive or robotic surgery was performed. |
| **Surgical Margins of the PS (SM)** | The final status of the surgical margins after tumor resection at the PS. |
| **Surgical Procedure of PS**  **at This Facility (OPTYPE)** | The type of surgery performed on the PS by the reporting hospital. |
| **Pathologic T (PATH\_T)** | Size or extent of the primary tumor. |
| **Pathologic N (PATH\_N)** | Regional lymph node metastasis and its extent. |
| **Pathologic M (PATH\_M)** | The presence of distant metastasis. |
| **Pathologic Stage (Prefix/Suffix) Descriptor (PATH\_SD)** | Descriptive symbols for the prefix and suffix of AJCC pathological staging. |
| **Pathologic Stage Group (PATH\_SG)** | The anatomical extent of disease based on pathological T, N, and M. |
| **Perineural Invasion (PI)** | Records the presence of perineural invasion at the PS. |
| **Primary Site (PS)** | The PS of the cancer. |
| **Scope of Regional Lymph Node Surgery at This Facility (SRLN)** | The extent of surgery at the PS, including regional lymph node dissection, biopsy, or aspiration. |
| **Surgical Diagnostic and Staging Procedure**  **at This Facility (SD)** | Surgical procedures performed by the reporting hospital for diagnosis or staging. |
| **Targeted therapy at This Facility (TARGET)** | The details of targeted therapy administered during the first treatment course at the reporting hospital. |

* Patients diagnosed before January 1, 2018: This criterion was based on the cancer registry handbook published by the Taiwan cancer registry [2], which followed the guidelines set by the American joint committee on cancer (AJCC). As of January 1, 2018, the 8th edition of the cancer staging system was adopted in Taiwan. Patients diagnosed before this date were staged under earlier versions and were therefore excluded from this study.
* Patients with BH codes of 0 or 6: Following expert recommendations, we excluded patients with a BH code of 0 (indicating benign tumors) or 6 (indicating metastatic cancer with an unknown PS).

The third step involved ensuring the completeness of the cancer registration data by establishing inclusion criteria: 1) Patients must have at least one pathology report and one examination report; 2) The patient’s first report must be earlier than the initial diagnosis date, and the date of the last report must be later than the date of the first surgery. Finally, the resulting patients were classified into six cancer categories based on their PS information.

References

[1] C.-W. Kao *et al.*, "Accuracy of long-form data in the Taiwan cancer registry," *Journal of the Formosan Medical Association,* vol. 120, no. 11, pp. 2037-2041, 2021.

[2] 衛生福利部國民健康署. (2022). 台灣癌症登記長表摘錄手冊.