viTBI-BERT: A Vietnamese Language Model for Prediction of Traumatic Brain Injury

Duc-Khiem Doan[1], Thanh Hai Tran[1], Trung-Kien Tran[2], Thi-Lan Le[1], Hai Vu[1], Huu-Khanh Nguyen[3], Van Mao Can[4], Thanh Bac Nguyen[3]

[1] School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam
[2] Military Insitute of Science and Technology, Hanoi, Vietnam
[3] Military Hospital 103, Hanoi, Vietnam
[4] Vietnam Military Medical University, Hanoi, Vietnam

Correspondence: Thanh-Hai Tran, email: hai.tranthithanh1@hust.edu.vn
Communication: received 12 Sep 2024, revised 28 Oct 2024, accepted 04 Mar 2025
DOI: 10.32913/mic-ict-research.v2025.n1.1303

Abstract: Artificial Intelligence (AI) is increasingly utilized for disease prediction by analyzing large and complex datasets. Traditionally, AI-based traumatic brain injury (TBI) prediction relies on multimodal data, including structured information like test results and unstructured data such as CT and/or MRI scans. However, physician conclusions in text form at admission and discharge represent an underutilized source of valuable insights into a patient's condition. This paper proposes a novel approach to classifying TBI severity using these physicians' written conclusions. We introduce viTBI-BERT, a model built on the ViHealthBERT backbone, incorporating pre-processing techniques such as acronym normalization, special character removal, word segmentation, and textual data augmentation. Following pre-processing, the data is passed through the pre-trained ViHealthBERT model, augmented with a fully connected layer and a softmax layer for classification. Evaluated on three sets of medical notes from a self-collected dataset of 503 Vietnamese patients, viTBI-BERT achieved a sensitivity of 71% in classifying four levels of injury severity. These findings demonstrate the potential of language-based analysis, which, when integrated with structured clinical and subclinical data, could further improve TBI classification outcomes.

Keywords: *Traumatic Brain Injury, Large Language Model, Transformer, Classification.*

I. INTRODUCTION

Traumatic brain injury (TBI) is a critical medical condition resulting from external forces impacting the head, leading to varying degrees of brain dysfunction. TBI can be caused by various incidents such as traffic accidents, falls, and sports injuries, with consequences ranging from mild concussions to severe brain damage. TBI significantly influences the patient's quality of life and, in some cases, leads to long-term disability or death.

Current diagnostic and treatment procedures for TBI involve a combination of clinical examinations, imaging techniques such as CT scans or MRIs, and continuous monitoring of the patient's neurological status. However, the overload situation and the limited experience of healthcare professionals, particularly at remote medical facilities, may cause errors in prognosis and treatment, exacerbate patient outcomes, and increase healthcare costs.

In recent years, there has been a growing interest in applying AI models to enhance TBI diagnosis. Most existing AI approaches have focused on analyzing structured clinical information [1]. Some recent works focused on the medical analysis of CT (Computerized Tomography) or MRI (Magnetic Resonance Imaging) brain images [2]. While these models have shown promise in detecting abnormalities in brain images, they often do not provide comprehensive diagnostic conclusions, particularly regarding the severity of the injury.

Accurate conclusions about the severity of TBI are crucial, as they directly influence treatment decisions and prognoses. For example, a conclusion like "Máu tụ dưới màng cứng mạn tính bán cầu phải" ("Chronic subdural hematoma, right hemisphere" in English) can indicate a severe level of TBI. Meanwhile, a conclusion like "Máu tụ dưới màng cứng trán đỉnh phải nghiện rượu" ("Subdural hematoma, right frontal-parietal, alcoholic" in English) may correspond to a moderate TBI case. These conclusions typically appear at the end of the medical report. To our knowledge, there is a gap in the research concerning the analysis of diagnostic conclusions for determining the severity of TBI. Existing studies have not explored the integration of textual information from medical reports into AI models for generating comprehensive diagnostic conclusions. On one hand, no dataset containing the conclusions of TBI patients

is available for research purposes. On the other hand, if such a dataset exists, the linguistic characteristics of the reports must be taken into consideration.

This paper addresses the existing gap by investigating text analysis models and their application in the field of traumatic brain injury diagnosis. We propose a novel approach that categorizes TBI into four severity levels: mild, moderate, severe, and critical, based on the analysis of Vietnamese textual conclusions in medical reports. This study represents the first effort to employ text analysis for TBI classification, aiming to enhance diagnostic accuracy and improve patient outcomes. The proposed framework, called **viTBI-BERT**, consists of two main blocks: a preprocessing block that includes acronym normalization, special character removal, word segmentation, and data augmentation; and a classification block composed of a pretrained ViHealthBERT model [3], followed by an additional fully connected layer and a softmax layer. The method is validated on three sets of medical notes extracted from our self-collected dataset, **TBI_MH**, showing promising results.

In summary, the contributions of this paper are threefold.

- First, we introduce an original idea for TBI outcome prediction using Vietnamese medical notes, which has not been investigated in the literature.
- Second, we propose a framework, **viTBI-BERT**, which leverages ViHealthBERT to analyze TBI Vietnamese medical notes, incorporating specific preprocessing steps.
- Finally, we validate our proposed framework on three sets of TBI medical notes extracted from our self-collected dataset.

The remainder of this paper is organized as follows. In section II, we present some existing works for the automatic prognosis of traumatic brain injury using machine learning. In section III, we present the general framework, and a detailed description of each component is presented. Section IV presents the dataset and experimental results. We conclude and give some ideas for future works in Section V.

II. RELATED WORKS

1. Machine learning for TBI outcome prediction

Recent studies have explored using ML to predict TBI outcomes. One study analyzed data from 1653 TBI patients using algorithms like random forest (RF) and decision tree (DT)[1], identifying key predictors such as the motor component of the Glasgow Coma Scale (GCS), pupillary condition, and age. RF was found to be the best for short-term mortality prediction, while a generalized linear model (GLM) performed best for long-term survival prediction. The study also highlights the potential of deep learning (DL) for handling high-dimensional data but recommends using simple models to avoid overfitting.

In 2020, Matsuo et al., similarly, [4] evaluated the effectiveness of nine machine learning algorithms in predicting outcomes for traumatic brain injury (TBI) patients. By analyzing clinical data from 232 patients, the research identifies the random forest and ridge regression models as the most effective for predicting poor outcomes and mortality. Key factors influencing predictions include age, the Glasgow Coma Scale, and specific blood markers. The findings suggest that modern machine learning techniques hold promise for improving TBI prognosis, though further validation is needed. Bruschetta et al. compared traditional regression models with various machine learning (ML) algorithms for predicting clinical outcomes in traumatic brain injury (TBI) patients [5]. The study evaluates the performance of a Classical Linear Regression Model (LM) against a Support Vector Machine (SVM), k-nearest Neighbors (k-NN), Naïve Bayes (NB), Decision Tree (DT), and an ensemble ML approach. The findings indicate that ML algorithms, including ensemble methods, do not significantly outperform traditional regression models when using a limited set of predictor variables. Naïve Bayes showed the poorest performance among the models tested. The study underscores the importance of rigorous comparisons between traditional and ML approaches in TBI prognosis and suggests further validation on more diverse datasets incorporating dynamic predictors like neuroimaging data.

2. Deep learning for TBI outcome prediction

The study by Adil et al. (2020) focuses on developing a deep learning model to predict outcomes after TBI, particularly in low- and middle-income countries (LMICs) [6]. Using clinical data from Uganda, the authors compared the performance of deep neural networks, shallow neural networks, and elastic-net logistic regression. Their results showed that deep learning outperformed traditional methods in specific metrics like AUROC and partial AUPRC, highlighting the potential of advanced algorithms in resource-limited settings, though optimal algorithm choice depends on the clinical context. Matthew et al. proposed a prognostic model combining deep learning from head CT scans with clinical data to predict long-term outcomes in patients with severe traumatic brain injury (sTBI) [7]. In a cohort of 537 patients, the fusion model demonstrated superior performance over the established IMPACT model in predicting mortality and unfavorable outcomes on internal

data. However, in external validation with 220 patients, the fusion model did not significantly outperform the IMPACT model. Despite the mixed results, the study underscores the promise of integrating imaging and clinical information for sTBI prognostication. Another study employs a deep learning framework to evaluate the sensitivity of a CNN model in detecting anatomical changes in brain MRI scans related to traumatic brain injury (TBI) outcomes [2]. The objective is to validate the model's clinical relevance using a dataset from the TRACK-TBI study, which includes 203 patients, and to analyze the model's performance across varying severity subgroups. Results indicate that the model can identify latent predictive information from MR images that is not captured in ground truth labels, suggesting promise in classifying TBI-related MR images.

3. Language models for disease prediction

Large language models (LLMs) have demonstrated impressive capabilities not only in the domain of natural language processing but also in various other tasks. In medical analysis, some early works have applied LLMs to enhance the performance of existing deep-learning models within a single modality. We present recent significant works that apply LLMs to address medical-related problems.

In [8], raw clinical texts have been refined and augmented by an LLM then combined with embedding computed from tabular data from the Electrical Medical Report of Patients. The final model, namely called CKLE, has been trained for two health event prediction tasks in the field of cardiology, heart failure prediction and hypertension prediction. Health-LLM incorporated health records and medical knowledge into a large model to ask relevant questions for a given disease [9]. Med-BERT is an adaptation of the BERT framework originally developed for the text domain to the structured Electrical Health Records domain [10]. Hegselmann et al. proposed TabLLM for the classification task of tabular data [11]. Each feature in the tabular data was converted into a natural language string using a manual template, table-to-text, or LLM. Li et al. introduced LViT that combined text description with images as input for segmentation task [12]. In LViT, medical text annotation is incorporated to compensate for the quality deficiency in image data. Le et al. conducted an evaluation of TBI 6-month outcome prediction using ChatGPT4 model [13] and showed that it was less performant than the conventional IMPACT model and that physicians.

Although there have been some attempts to use large language models (LLMs) for disease prediction, to the best of our knowledge, no studies have yet explored the use of LLMs with medical notes from CT images of TBI patients. In our work, we will utilize both pre- and post-CT medical notes from doctors, as well as a combination of these notes, to predict the outcomes of TBI. We believe this is the first initiative to explore the potential of LLMs for TBI prediction using medical text.

III. PROPOSED METHOD FOR TBI PROGNOSIS

1. General framework

As aforementioned, our work focuses on analyzing the medical notes of doctors from CT scans of TBI patients and outputs one among four severity levels of TBI: mild, moderate, severe, and critical. In this study, we propose a framework for the prognosis of TBI using the medical text of the doctors from CT scans as illustrated in Fig. 1. It is composed of the following steps:

- *Data splitting*: First, raw data (i.e. medical notes of physicians from CT scans of TBI patients) are split into two subsets, one for training and one for testing. In our work, we employ K folds (with K = 10) so nine folds will be used for training and one fold will be used for testing.
- *Pre-processing*: The raw medical notes must be pre-processed before being input into the classification model. We deploy two techniques: acronym normalization and special character removal to convert raw text to normalized text. This step is applied for both training and testing datasets.
- *Data augmentation*: To enrich the training set, we implement various techniques to create new medical texts with the same meaning. This step is applied only to the training dataset.
- *Classifier training*: We utilize the pre-train ViHealth-BERT model to fine-tune the viTBI-BERT using an augmented training dataset.
- *Inference*: The test samples are first pre-processed and then inputted into the trained viTBI-BERT to infer the severity level of the TBI.

Subsequently, we provide a detailed explanation of each of these steps, providing a comprehensive view of our methodology and the underlying rationale.

2. Pre-processing

a) Acronym normalization

In medical texts, acronyms are often used to quickly denote specific conditions, symptoms, or medical procedures. In our TBI dataset, a total of 77 acronyms have been found in medical notes. For example, "CTSN" is the acronym for "Chấn thương sọ não" (Traumatic Brain Injury), "MTTN" is the acronym for "Máu tụ trong não" ("intracerebral hemorrhage"), TNGT means "Tai nạn giao

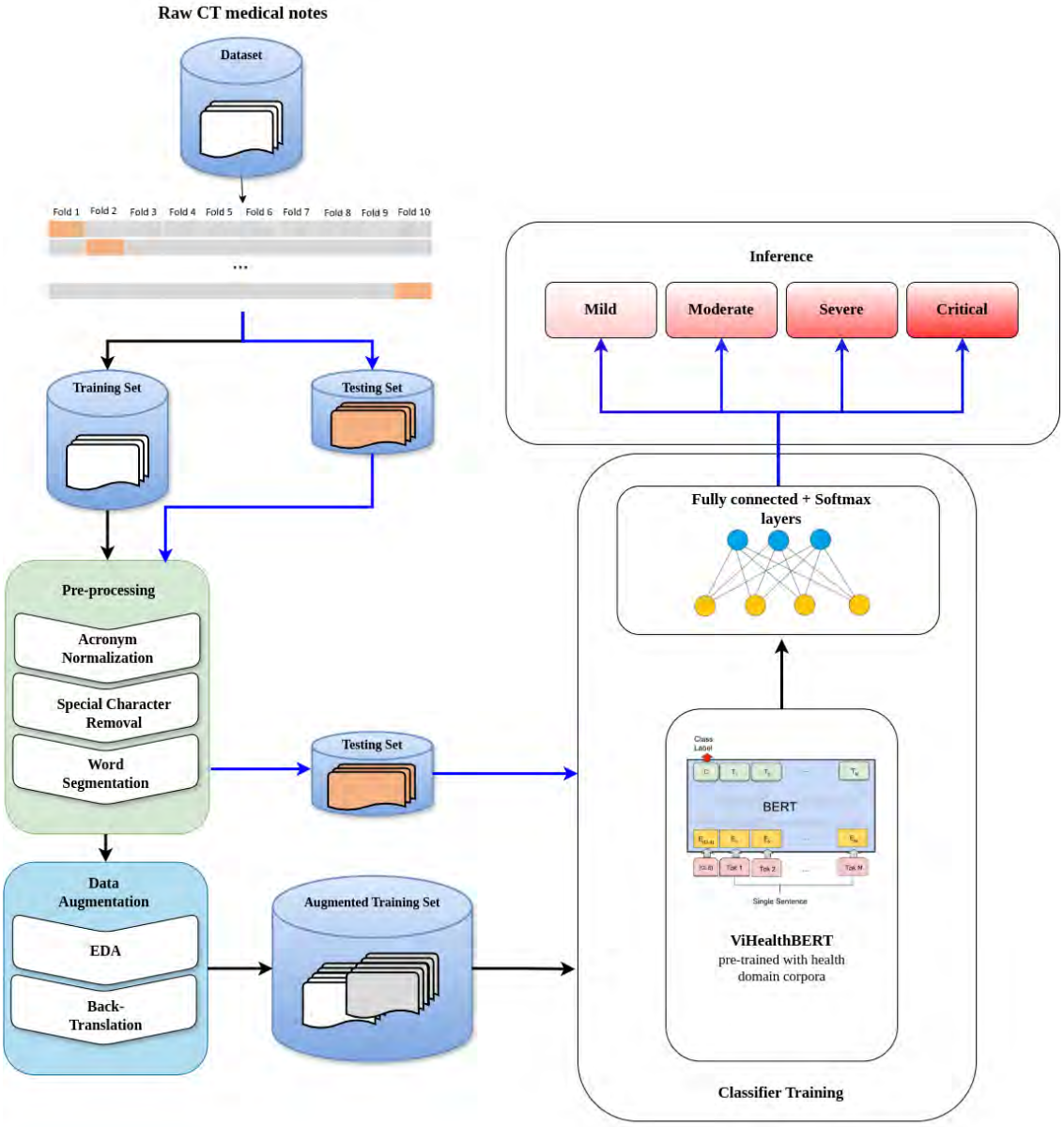

Figure 1. Proposed viTBI-BERT for outcome prediction after Traumatic Brain Injury from Vietnamese medical notes.

thông" ("traffic accident"). To ensure accurate interpretation and improve consistency in the data, all acronyms must be normalized.

In the medical text domain, Vo et al. proposed a method to detect abbreviation in Vietnamese clinical texts [14]. However, this method does not encompass all acronyms present in our dataset related to traumatic brain injury. Following the approach used in [3], we generate a dictionary of 77 acronyms. This involves extracting all noun phrases (NP) from the dataset using the *underthesea* POS tagger. Consequently, we manually decrypt the data by clarifying their meanings provided by expert doctors at 103 Military Hospital. For example, acronyms such as "CTSN" are expanded to "Chấn thương sọ não", and "MTMNC" is expanded to "Máu tụ ngoài màng cứng" (epidural hematoma).

b) Special characters removal

Machine learning models often struggle to interpret punctuation or special characters. In our dataset, the medical notes frequently contain various punctuation marks such as exclamation points, apostrophes, and commas [15]. To address this, we preprocess the medical notes by removing all special characters and punctuation, including ":", ",", ".", "/", "(", ")", "-", and ";".

c) Word segmentation

Word segmentation is a crucial pre-processing step in natural language processing (NLP), especially for languages like Vietnamese, which lack explicit word boundaries. In our workflow, we perform word segmentation using the RDRSegmenter, a component of the VNCoreNLP toolkit. RDRSegmenter, developed by [16], is a Vietnamese word segmentation tool that uses a rule-based approach to accurately separate text into meaningful units. This tool is part of VNCoreNLP, a comprehensive suite of NLP tools for Vietnamese developed by Vu et al. [17]. Since Minh et al. [3] used this method to pre-process the pre-training data and also recommended it for fine-tuning tasks, we use the same word segmenter for our downstream applications. By incorporating these methods, we enhance the quality and precision of our text processing pipeline, making it well-suited for classification tasks. In our experiment, we directly utilize the pre-trained RDRSegmenter for word segmentation.

Figure 2 illustrates an example of pre-precessing raw data "CĐ:CTSN:MTNMC+TKNS do tai nạn sinh hoạt G7". In this sentence, "CĐ" is an acronym for "Chấn động" ("concussion"), "CTSN" is an acronym for "Chấn thương sọ não" ("traumatic brain injury"), "MTNMC" is an acronym for "Máu tụ ngoài màng cứng" ("epidural hematoma"). "TKNS" means "Tụ khí nội sọ" ("intracranial pneumatoma"). "G7" means "giờ thứ 7" ("7th hour"). After normalizing all acronyms, we obtain "chấn động: chấn thương sọ não:máu tụ ngoài màng cứng+tụ khí nội sọ do tai nạn sinh hoạt giờ thứ 7" ("concussion: traumatic brain injury: epidural hematoma+intracranial pneumatoma due to domestic accident 7th hour"). This new data contains special characters and punctuation. We remove all of these and apply RDRSegmenter for word segmentation to finally achieve: "chấn_động chấn_thương sọ não máu tụ ngoài màng_cứng tụ khí nội sọ do tai_nạn sinh_hoạt giờ thứ 7".

3. Data augmentation

a) Augmentation methods

In image classification tasks, data augmentation enhances data diversity through techniques such as rotation, translation, scaling, and noise addition. Similarly, text classification employs augmentation to enrich text data, with numerous studies developing automated methods to generate and improve data quality. This approach improves model accuracy, particularly for small datasets such as ours.

We will first present the EDA (Easy Data Augmentation) techniques introduced by Wei and Zou [18] and apply them to Vietnamese medical texts. The techniques include:

- *Synonym Replacement (SR)*: This technique involves replacing non-stop words with randomly chosen synonyms to create a new sentence. For our experiments, we used the Vietnamese WordNet from [19] to find synonyms and a Vietnamese stopword dictionary to exclude stop words from the sentences.
- *Random Swap (RS)*: Swapping two non-stop words randomly n times.
- *Random Insert (RI*: Finding a random synonym of a non-stop word and inserting it randomly n times in the sentence.
- *Random Delete (RD)*: Removing each word in the sentence with a probability p.

The number of augmentations should be calibrated according to the dataset size, as an excessive amount of augmented data can result in overfitting. A notable limitation of these methods is their inability to maintain the contextual meaning of sentences. To overcome this, we investigate more advanced techniques that preserve the original sentence's meaning.

- *Back Translation (BT)*: This technique generates additional training samples by utilizing translators. It has been used by various research teams to improve translation models [20][21]. The process involves translating the original text into another language and then translating it back to the original language using a different translator. In this work, we employ the *mtranslate* library to implement the back translation method, utilizing English, French, and Russian as intermediate languages.

Table I presents the results of applying the Easy Data Augmentation (EDA) method to the sentence: "chấn_động não do tai_nạn giao_thông" (brain_injury due to traffic_accident). Similarly, Table II shows the results of back-translation applied to the sentence "chấn động: chấn thương sọ não máu tụ ngoài màng cứng tụ khí nội sọ do tai nạn sinh hoạt giờ thứ 7" using three different intermediate languages.

Table I
EXAMPLES OF EASY DATA AUGMENTATION METHODS

Augmented Sentence	Type
chấn_động do tai_nạn giao_thông (deleted "não")	RD
chấn_động **đầu_óc** do tai_nạn giao_thông	SR
não chấn_động do tai_nạn giao_thông	RS
chấn_động não do **đầu** tai_nạn giao_thông	RI

4. Classifier training

a) Model architecture

In our experiments, we employed monolingual domain-specific pre-trained BERT models, designed specifically

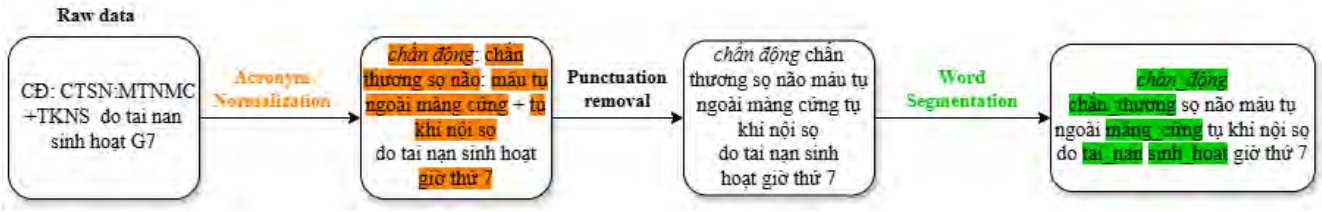

Figure 2. Example of pre-processing steps.

Table II
EXAMPLE OF DATA AUGMENTATION USING BACK-TRANSLATION WITH THREE LANGUAGES: ENGLISH, FRENCH AND RUSSIAN. WE ALSO ADDED THE TRANSLATION OF THE AUGMENTED DATA FROM VIETNAMESE TO ENGLISH USING GOOGLE TRANSLATOR.

Language	Back-Translated Sentence
English	Chấn động não: chấn thương sọ não, tụ máu ngoài màng cứng, tràn khí nội sọ do tai nạn gia đình, giờ thứ 7 (Concussion: traumatic brain injury, epidural hematoma, intracranial pneumothorax due to domestic accident, 7th hour)
French	Chấn động: chấn thương đầu, tụ máu ngoài màng cứng, tụ khí nội sọ do tai nạn hàng ngày vào giờ thứ 7 (Concussion: head trauma, epidural hematoma, intracranial emphysema due to daily accidents at 7 o'clock)
Russian	Chấn động: chấn thương sọ não, tụ máu ngoài màng cứng, tích tụ khí nội sọ do hậu quả của một sự kiện trong đời vào giờ thứ 7 (Concussion: traumatic brain injury, epidural hematoma, intracranial gas accumulation as a result of a life event in the 7th hour).

for Vietnamese healthcare text for the TBI outcome classification task based on medical notes. ViHealthBERT, a highly effective baseline model for this domain, was rigorously evaluated using various training strategies. It achieved state-of-the-art (SOTA) results across three downstream tasks: Named Entity Recognition (NER) for COVID-19 and ViMQ datasets, Acronym Disambiguation, and Summarization [3].

In this study, we used the word-level version of ViHealthBERT, which contains 135 million parameters and utilizes a word-level tokenizer. This model follows the same architecture as $\text{BERT}_{\text{base}}$ [22]. One of the key advantages of using BERT-based Transformers for sequence classification is their ability to capture the meaning and context of the input sequence more effectively than traditional neural networks. This capability stems from BERT's pretraining on a vast corpus of text, where it learns to predict missing words or sentences based on their surrounding context. This pretraining enables BERT to understand the relationships between different words and phrases within the input, resulting in a more informative representation of the sequence.

The pre-trained model consists of 12 encoder-only Transformer layers [23], with 768 hidden units and 12 attention heads. Each encoder layer has two main components: a Multi-Head Self-Attention mechanism and a Feedforward Neural Network. The attention mechanism allows the model to weigh the importance of different words in the sequence, capturing long-range dependencies. Mathematically, the output of the Multi-Head Attention for a single head can be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. In practice, the model employs multiple heads, each with different learned parameters, allowing it to focus on various parts of the input.

After the attention mechanism, the output passes through a feedforward network, which typically consists of two linear transformations with a ReLU activation in between:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \qquad (2)$$

where W_1, W_2, b_1, and b_2 are learned parameters.

As shown in Figure 1, we added an additional classifier head. This classifier head operates on the [CLS] token (Classifier token) in the last hidden layer output by the pretrained BERT model. The [CLS] token is a special token prepended to every input sequence, representing the entire sequence for classification tasks. The classifier head consists of a fully connected layer followed by a softmax activation function, which produces probabilities for each class label.

5. Inference

To gain deeper insights into the model's decisions, we employed Integrated Gradients (IG) [24], an explainable AI technique. This method computes attribution scores for each word in a clinical note, revealing which words most significantly influenced the model's predictions.

Formally, let $F : \mathbb{R}^n \to \mathbb{R}$ represent a deep neural network, where $F(x)$ is the output for input x. The attribution of each input feature x_i is given by:

$$\text{IG}_i(x) = (x_i - x_i') \times \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (3)$$

Here, x is the input, x' is the baseline (neutral or zero input), and α interpolates between x' and x. The gradient $\frac{\partial F(x)}{\partial x_i}$ is computed at points along this path. The integral can be approximated as:

$$\text{IG}_i(x) \approx (x_i - x_i') \times \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m}(x - x'))}{\partial x_i} \quad (4)$$

IG satisfies two properties: **Sensitivity**, ensuring non-zero attributions when output changes are due to a single feature, and **Implementation Invariance**, guaranteeing identical attributions for functionally equivalent models.

IV. EXPERIMENTS

1. Dataset

In the literature, there is no dataset of medical notes of traumatic brain injury. As a result, to validate our proposed framework, we collected a dataset consisting of 503 patients treated at the Military Hospital 103 in Hanoi, Vietnam, in 2023. This dataset, referred to as **TBI_MH** (**T**raumatic **B**rain **I**njury collected at **M**ilitary **H**ospital), was carefully curated to exclude records with incomplete information, ensuring data accuracy. The patients had a mean age of 43.89 years with a standard deviation of 21.3 years. A significant portion of the cohort, approximately 76.09%, were male, as described in Table III and Figure 3.

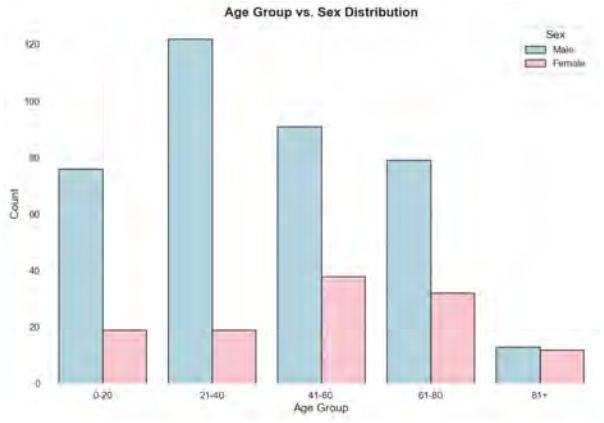

Figure 3. Age groups and sex distribution of patients in the TBI_MH dataset

Table III
AGE AND SEX DISTRIBUTION IN TBI_MH DATASET

Category	Subcategory	Percentage (%)
Age Group	0-20	18.92
	21-40	28.29
	41-60	25.70
	61-80	22.11
	81+	4.98
Sex	Male	76.10
	Female	23.90

The dataset includes a comprehensive range of patient data, covering demographic details, clinical indicators, laboratory results, and CT (computed tomography) imaging information. The demographic data encompasses age, sex, and key patient identifiers. Clinical data provides insights into health status, blood pressure, and temperature, along with details on traumatic brain injuries as documented in medical histories. The Glasgow Coma Scale (GCS) [25] was assessed during physical examinations, and the dataset also records related conditions such as stroke, diabetes, hypertension, and cardiovascular diseases. CT scan data plays a critical role in diagnosing traumatic brain injuries, particularly through the use of the Rotterdam score [26]. Additionally, the dataset includes laboratory test results, such as blood testing, serum biochemistry, electrolyte levels, and electrocardiogram readings.

In this research, our focus is on analyzing the pre-discharge and post-discharge medical notes provided by doctors to assess patient severity using AI-based natural language processing (NLP) techniques. Patient outcomes were categorized into four severity levels (1 = Mild, 2 = Moderate, 3 = Severe, 4 = Critical) by experienced clinicians and physicians, as illustrated in Table IV and Figure 4. Table V illustrates some examples of medical notes at pre-discharge and post-discharge for four severity levels.

Table IV
OUTCOMES OF THE TBI_MH DATASET

Severity Level	Description	Number of Patients
1	Mild	39
2	Moderate	265
3	Severe	155
4	Critical	44

Before using this dataset for research, we obtained the necessary approvals from the relevant organizational authorities to ensure compliance with established data usage protocols. The study's methodology strictly adhered to the guidelines and regulations governing research practices, and ethical clearance was obtained from the hospital for access to the dataset.

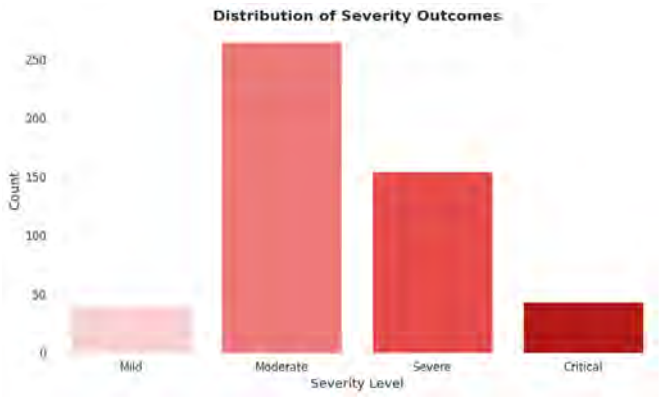

Figure 4. Outcomes distribution of TBI_MH Dataset

We consider three sets of medical notes extracted from the raw **TBI_MH** dataset as follows:

1) **PreMN_TBI_MH**: The first dataset includes the medical notes of doctors before the patient was admitted to the hospital.
2) **PostMN_TBI_MH**: The second dataset includes the medical notes of doctors after the patient was discharged from the hospital.
3) **MN_TBI_MH**: The third dataset is a combination of both the pre-admission and post-discharge medical notes.

Table VI provides a comprehensive summary of the statistics related to the length of notes across three datasets PostMN_TBI_MH, PreMN_TBI_MH, and MN_TBI_MH by severity levels (Mild, Moderate, Severe, and Critical). The metrics include the number of medical notes, along with descriptive statistics about their lengths: mean, standard deviation (Std), minimum (Min), 25th percentile (25%), median (50%), 75th percentile (75%), and maximum (Max). It is noted that the lengths of medical notes at pre-discharge and post-discharge are similar on average. However, the more severe the TBI, the longer the medical notes tend to be.

We separate the dataset into the training set and the testing set using K-Fold. Then the training set is augmented to re-balance the samples among classes. Figure 5 shows the distribution of original and augmented training sets.

2. Implementation details

a) Data Augmentation

According to Wei and Zou (2019) [18], for the SR, RI, and RS methods, the number of changed words n is calculated as $n = \alpha \times l$, where α is the percentage of words to be replaced and l is the sentence length. For the RD method, the deletion probability p equals α, with α defined by the user.

Table V
EXAMPLE OF PRE- AND POST-MN_TBI_MH SET WITH SEVERITY LEVELS

PreMN_TBI_MH	PostMN_TBI_MH	Severity
Xuất huyết dưới nhện liềm đại não gãy hở 13 dưới xương mác trái tụ máu dưới da quanh mi mắt phải do tai nạn giao thông (Subarachnoid hemorrhage, open fracture of the left fibula, subcutaneous hematoma around the right eyelid due to traffic accident)	Xuất huyết dưới nhện liềm đại não gãy kín 13d xương mác và vết thương phần mềm cẳng chân bên trái do tai nạn giao thông giờ thứ 3 đã khâu vết thương ngày thứ nhất (Subarachnoid hemorrhage, closed fracture of the 13d fibula and soft tissue injury of the left leg due to a traffic accident. The wound was sutured on the first day.)	Mild
Máu tụ dưới màng cứng trán đỉnh phải nghiện rượu (Subdural hematoma right frontal parietal alcoholic)	Máu tụ dưới màng cứng trán đỉnh phải đám mờ không thuần nhất ngoại vi 13 dưới phổi phải teo não tuổi già nghiện rượu (Right frontal subdural hematoma peripheral heterogeneous opacity 13 right lower lung cerebral atrophy senile alcoholism)	Moderate
Máu tụ dưới màng cứng mạn tính bán cầu phải (Chronic subdural hematoma right hemisphere)	Máu tụ dưới màng cứng mạn tính bán cầu phải đã phẫu thuật bơm rửa di lệch ổ máu tụ ngày thứ nhất (Chronic subdural hematoma of the right hemisphere was surgically discharged and displaced on the first day)	Severe
Chấn thương sọ não nặng máu tụ dưới màng cứng bán cầu trái xuất huyết dưới nhện rải rác 2 bán cầu vỡ xương đá phải do tai nạn giao thông g3 (Severe traumatic brain injury, left hemisphere subdural hematoma, scattered subarachnoid hemorrhage, bilateral hemisphere fracture of right petrous bone due to traffic accident g3)	Chấn thương sọ não nặng máu tụ dưới màng cứng bán cầu trái dập não xuất huyết trán và thái dương và đỉnh trái xuất huyết dưới nhện rải rác 2 bán cầu máu tụ ngoài màng cứng và vỡ xương thái dương đỉnh phải vỡ xương thái dương bên trái xương đá phải thân xương bướm do tai nạn giao thông viêm phổi do a baumanii toàn kháng (Severe traumatic brain injury, left hemisphere subdural hematoma, brain contusion, left frontal and temporal and parietal hemorrhage, scattered subarachnoid hemorrhage in both hemispheres, epidural hematoma and right temporal-parietal bone fracture, left temporal bone fracture, right petrous bone, sphenoid bone body due to traffic accident, pan-resistant A baumannii pneumonia)	Critical

Table VI
STATISTICS OF DIAGNOSIS LENGTH ACROSS DATASETS AND SEVERITY LEVELS

Dataset	Severity Level	Count	Mean	Std	Min	25%	50%	75%	Max
PostMN_TBI_MH	Mild	39	16.59	9.45	5	10	14	22.50	48
	Moderate	265	28.17	14.70	3	18	25	37	85
	Severe	155	35.37	18.34	7	22	30	48	141
	Critical	44	49.45	24.91	17	29.75	47	57.25	140
PreMN_TBI_MH	Mild	39	18.33	8.41	7	11	17	26	37
	Moderate	265	29.09	14.30	5	19	26	36	86
	Severe	155	34.47	14.96	6	22.50	31	45	79
	Critical	44	45.07	24.18	13	29.75	40.50	53.25	121
MN_TBI_MH	Mild	39	34.92	16.86	12	22	32	46	75
	Moderate	265	57.26	28.30	11	36	52	74	171
	Severe	155	69.85	32.29	22	44.50	61	94.50	220
	Critical	44	94.52	47.45	35	62.75	82.50	111.50	261

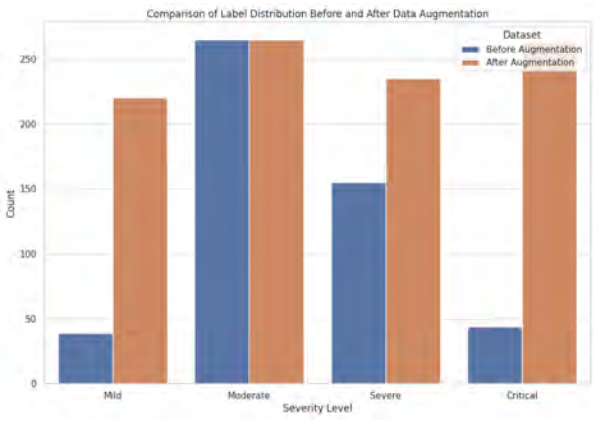

Figure 5. Distribution of original and augmented training dataset

In this experiment, we used the number of augmentations equal to 2 for each sentence. The augmentation techniques applied include synonym replacement, random insertion, random swap, and random deletion, with the following parameters: $\alpha_{\text{sr}} = 0.1$, $\alpha_{\text{ri}} = 0.1$, $\alpha_{\text{rs}} = 0.1$, and $\alpha_{\text{rd}} = 0.1$. Additionally, back-translation is applied using intermediate languages from a set of {English, French, Russian}. Figure 5 illustrates the distribution of a training fold before and after applying the aforementioned augmentation methods. It is noted that the original data across classes is quite imbalanced, with the first and fourth classes having fewer samples. After applying data augmentation, the number of samples in each class becomes more balanced.

b) Classifier Training

viTBI-BERT is implemented in Python. To track and fine-tune the model for optimal hyperparameters. After all, we use the hyperparameters set as described in Table VII below. The optimizer is AdamW, and the method to find hyperparameters is Bayesian. The model training process spans 40 epochs.

Table VII
HYPERPARAMETERS SETTING IN THE EXPERIMENTS.

Hyperparameter	Value
Batch Size	16
Dropout Rate	0.2
Learning Rate	0.00005
Epoch	40
Number layers	1
Pretrained Model	word-level

c) Inference

We applied the Integrated Gradients (IG) method to evaluate the contribution of individual features to the model's predictions. As the baseline input, we used a zero embedding vector, consistent with the approach described in [24], which is widely adopted in text classification tasks. For calculating and visualizing the results, we leveraged the Captum library [27].

3. Evaluation Metrics

To evaluate the performance of the viTBI-BERT, we use various metrics derived from the confusion matrix. This matrix is essential for assessing the performance of classifiers by including counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The following formulas define the evaluation metrics, which help in determining the most suitable classification method.

*a) **Accuracy***

Accuracy measures the proportion of correct predictions among the total number of cases examined. It is given by (5):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

*b) F_1 **score***

The F_1 score is a harmonic mean of precision and recall, providing a single metric for a model's performance. It is defined as (6):

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (6)$$

c) *Precision*

Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is calculated using (7):

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

d) *Recall*

Recall, also known as sensitivity or the true positive rate, measures the model's ability to correctly identify true positives. It is given by (8):

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

4. Experimental results

a) *Results on three datasets*

Table VIII illustrates the average performance of our proposed method on three sets: PreMN_TBI_MH, PostMN_TBI_MH, and MN_TBI_MH. In terms of Sensitivity (or Recall), viTBI-BERT achieved scores of 0.67, 0.68, and 0.71 on the three sets on average, respectively. It is reasonable that the sensitivity on the PostMN_TBI_MH set is higher than on the PreMN_TBI_MH set because the information about the disease is more detailed and precise when recorded after discharge, especially in cases involving severe and critical patients. When combining both pre-discharge and post-discharge medical notes, the sensitivity increases to 0.71, which is 4% higher than using only pre-discharge notes and 3% higher than using only post-discharge notes.

Table VIII
COMPARISON OF PERFORMANCE METRICS ACROSS THREE DATASETS.

Set of data	Precision	Recall	F_1 score	Accuracy
PreMN_TBI_MH	0.72	0.67	0.67	0.68
PostMN_TBI_MH	0.71	0.68	0.67	0.68
MN_TBI_MH	0.71	0.71	0.74	0.71

To delve deeper, we extracted the model with the highest accuracy, trained on the third fold using the *PostMN_TBI_MH* dataset, as shown in Table IX. Figure 6 displays the confusion matrix, offering a comprehensive view of the model's classification performance. Figure 7 illustrates the training and testing accuracy over epochs, clearly demonstrating the model's learning progression. Lastly, we utilized the model to create visualizations using the Integrated Gradients method, emphasizing its interpretability in practical applications.

Table IX
PERFORMANCE METRICS FOR THE THIRD FOLD OF THE *PostMN_TBI_MH* DATASET

Metric	Value
Precision	0.7861
Recall	0.8007
F1	0.7861
Accuracy	0.7736

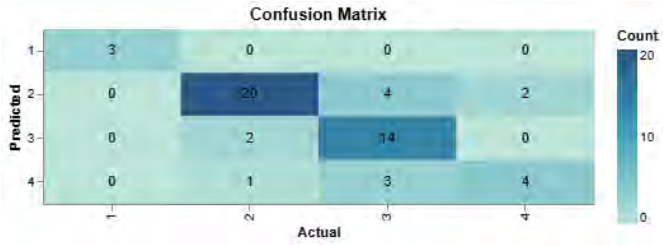

Figure 6. Confusion matrix for third fold of the applied method on PostMN_TBI_MH.

Figure 8 illustrates the application of the Integrated Gradients (IG) method to the medical note "chấn_động não vết_thương phần_mềm đỉnh trái do ngã" (concussion soft tissue injury, left parietal, due to fall). In this case, the true label is 1 (moderate), and the model correctly predicted the outcome. The words contributing most significantly to the model's decision are "chấn_động" (concussion), "vết_thương" (injury), and "do_ngã" (due to fall). These insights can potentially aid doctors in identifying critical terms relevant to patient diagnosis.

Comparision with conventional machine learning methods We applied several traditional machine learning methods as described in [28], along with additional machine learning models:

- **Multinomial Naive Bayes:** A probabilistic classifier based on Bayes' theorem, commonly used for text classification tasks.
- **Decision Tree Classifier:** A simple, interpretable model that splits data into branches based on feature

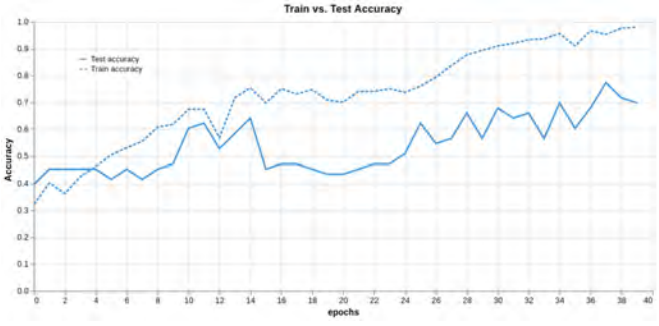

Figure 7. Training and testing progress of Fold 3

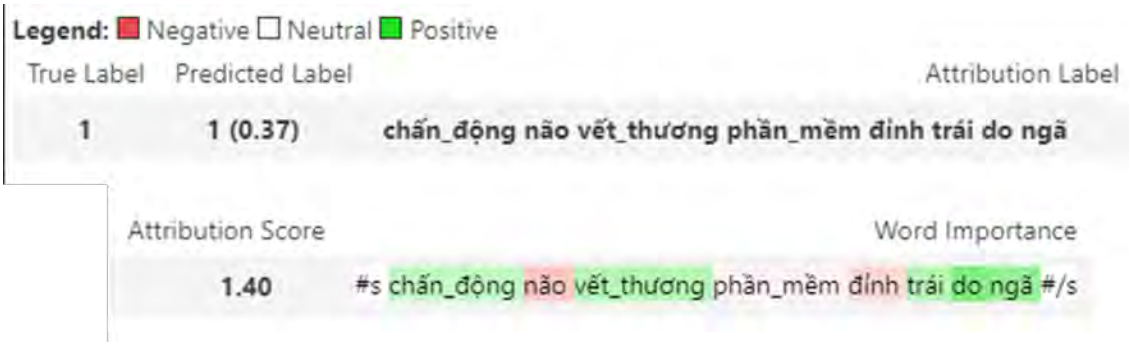

Figure 8. Intepreation of the result using IG method.

values to make predictions.

- **Logistic Regression:** A linear model for binary or multi-class classification that predicts probabilities using the logistic function.
- **Stochastic Gradient Descent Classifier:** A linear classifier optimized using stochastic gradient descent, suitable for large-scale and sparse data.
- **Linear Support Vector Classifier:** A linear SVM implementation configured with Cramer-Singer multi-class support ($C = 1$) to handle multi-class classification tasks.
- **XGBoost Classifier:** A high-performance gradient boosting method.

These traditional models provide a baseline for comparison with our proposed approach, with the results summarized in Table X. The results demonstrate that our method outperformed the others, while the Stochastic Gradient Descent Classifier also achieved relatively high performance.

5. Ablation study

a) Impact of pre-trained models

We compared the performance of viTBI-BERT using three baseline pretrained BERT models: PhoBERTbase [29], trained on general Vietnamese text, to evaluate the impact of general linguistic knowledge, ViPubmedDeBERTaxsmall [30] and ViHealthBERT [3], a domain-specific biomedical model to assess the benefits of specialized pretraining. Using the same hyperparameters in Table VII, we found that ViHealthBERT achieved the highest accuracy. The results are presented in Figure 9. Across all experimented datasets, ViHealthBERT consistently enabled viTBI-BERT to achieve higher accuracy.

b) Impact of pre-processing steps

We have conducted additional experiments along with four different strategies to assess the impact of various pre-processing techniques on model performance. These

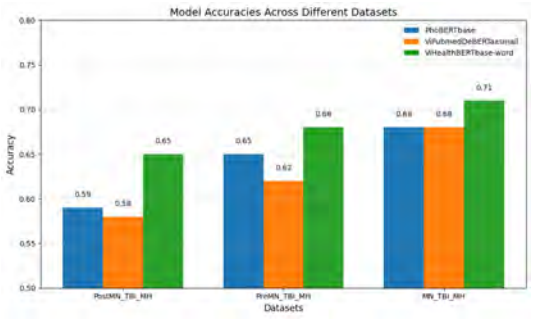

Figure 9. Performance comparison of various pre-trained on our three datasets.

strategies are outlined below and their responding results are indicated by Figures 10, 11, and 12.

- **Strategy 1: Raw data without augmentation:** In this strategy, we use the original, unprocessed text data without applying any additional preprocessing or data augmentation. This serves as a baseline to assess the effectiveness of subsequent preprocessing techniques.
- **Strategy 2: Punctuation removal without augmentation:** In this strategy, we remove punctuation marks from the text, keeping the data otherwise unchanged, to assess how simplifying the text in this way impacts the model's performance. No data augmentation is applied at this stage.
- **Strategy 3: Acronym normalization without augmentation:** In this strategy, we replace acronyms in the text with their full forms to enhance the model's ability to understand medical terminology in its complete form. Again, no data augmentation is introduced at this stage.
- **Strategy 4: Full pre-processing without data augmentation:** This approach includes the complete set of preprocessing techniques—punctuation removal, acronym normalization, and other necessary steps—without applying any data augmentation. This allows us to evaluate the effects of the entire prepro-

Table X
COMPARISION WITH CONVENTIONAL MACHINE LEARNING METHODS

Classifier	PostMN_TBI_MH				PreMN_TBI_MH				MN_TBI_MH			
	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy
MultinomialNB	0.46	0.34	0.34	0.56	0.46	0.34	0.34	0.56	0.46	0.34	0.34	0.56
DecisionTreeClassifier	0.44	0.47	0.42	0.51	0.40	0.43	0.41	0.46	0.36	0.41	0.36	0.46
LogisticRegression	0.31	0.34	0.32	0.52	0.31	0.35	0.33	0.52	0.31	0.35	0.33	0.51
SGDClassifier	0.61	0.59	0.56	0.57	0.62	0.48	0.51	0.56	0.58	0.45	0.41	0.50
LinearSVC	0.64	0.56	0.52	0.57	0.35	0.42	0.38	0.51	0.46	0.50	0.51	0.56
XGBClassifier	0.33	0.32	0.31	0.47	0.44	0.44	0.42	0.50	0.38	0.36	0.33	0.40
viTBI-BERT (our)	**0.71**	**0.68**	**0.67**	**0.68**	**0.72**	**0.67**	**0.67**	**0.68**	**0.71**	**0.71**	**0.74**	**0.71**

cessing pipeline in isolation.

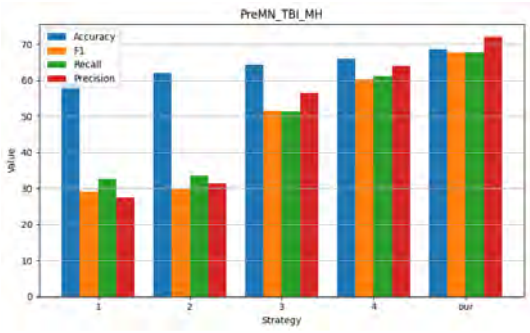

Figure 10. Results of applying several strategies on PreMN_TBI_MH dataset.

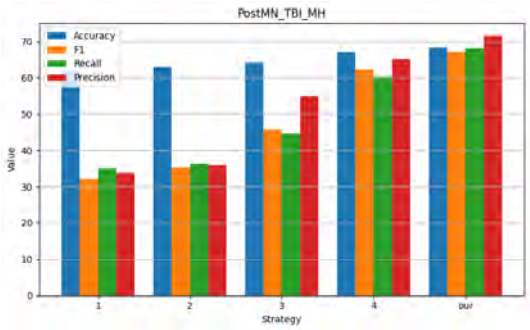

Figure 11. Results of applying several strategies on PostMN_TBI_MH dataset.

We observed an improvement across four evaluation metrics on all three datasets after applying each strategy. This demonstrates the positive impact of the updated preprocessing steps on the model's performance. Compared to the our final framework, the four strategies achieved lower performance.

V. CONCLUSIONS

In this paper, we presented a new framework, viTBI-BERT, for outcome prediction after traumatic brain injury from medical notes of CT scan using a large language model. Our framework consists of two main stages. The preprocessing stage involves expanding all acronyms, removing

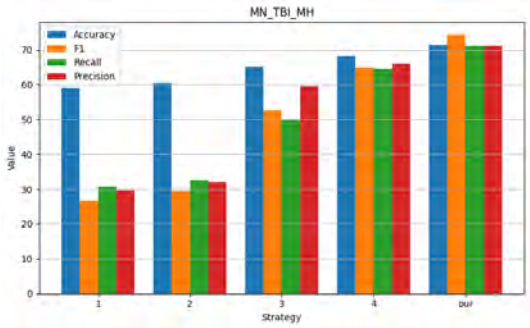

Figure 12. Results of applying several strategies on MN_TBI_MH dataset.

special characters, augmenting data, and performing word segmentation. This stage was executed using dictionary mapping, EDA, back-translation, and RDRSegmenter techniques. The processed data is then fed into the pre-trained ViHealthBERT model, enhanced with an additional fully connected layer and a softmax layer for classification. Our framework was validated on three self-collected medical note sets, comprising 503 patients across four severity levels. Experimental results showed that the sensitivity ranged from 0.67 to 0.68 and 0.71 on medical notes collected before admission, after discharge, and from both periods, respectively. Despite the limited amount of medical notes, the framework was able to effectively distinguish the severity of TBI. In the future, we plan to combine these results with clinical tests and CT scan data for further improvement of the framework.

ACKNOWLEDGEMENTS

This research is funded by the Ministry of Education and Training (MOET) under grant number B2023.BKA.09, titled "*Research and development of supporting tools for prognosis of traumatic brain injury using multi-modal information and artificial intelligence.*"

REFERENCES

[1] H. Khalili, M. Rismani, M. A. Nematollahi, M. S. Masoudi, A. Asadollahi, R. Taheri, H. Pourmontaseri, A. Valibeygi, M. Roshanzamir, R. Alizadehsani *et al.*, "Prognosis prediction in traumatic brain injury patients using machine learning algorithms," *Scientific reports*, vol. 13, no. 1, p. 960, 2023.

[2] D. Yeboah, H. Nguyen, D. B. Hier, G. R. Olbricht, and T. Obafemi-Ajayi, "A deep learning model to predict traumatic brain injury severity and outcome from mr images," in *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Oct 2021, pp. 1–6.

[3] M. P. Nguyen, V. H. Tran, V. Hoang, T. D. Huy, T. H. Bui, and S. Q. H. Truong, "ViHealthBERT: Pre-trained Language Models for Vietnamese in Health Text Mining," *13th Edition of its Language Resources and Evaluation Conference*, 2022.

[4] K. Matsuo, H. Aihara, T. Nakai, A. Morishita, Y. Tohma, and E. Kohmura, "Machine learning to predict in-hospital morbidity and mortality after traumatic brain injury," *Journal of Neurotrauma*, vol. 37, no. 1, pp. 202–210, 2020. [Online]. Available: https://doi.org/10.1089/neu.2018.6276

[5] R. Bruschetta, G. Tartarisco, L. F. Lucca, E. Leto, M. Ursino, P. Tonin, G. Pioggia, and A. Cerasa, "Predicting outcome of traumatic brain injury: Is machine learning the best way?" *Biomedicines*, vol. 10, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2227-9059/10/3/686

[6] S. M. Adil, C. Elahi, D. N. Patel, A. Seas, P. I. Warman, A. T. Fuller, M. M. Haglund, and T. W. Dunn, "Deep learning to predict traumatic brain injury outcomes in the low-resource setting," *World Neurosurgery*, vol. 164, pp. e8–e16, 2022.

[7] M. Pease, D. Arefan, J. Barber, E. Yuh, A. Puccio, K. Hochberger, E. Nwachuku, S. Roy, S. Casillo, N. Temkin, D. O. Okonkwo, S. Wu, , N. Badjatia, Y. Bodien, A.-C. Duhaime, V. R. Feeser, A. R. Ferguson, B. Foreman, R. Gardner, S. Gopinath, C. D. Keene, C. Madden, M. McCrea, P. Mukherjee, L. B. Ngwenya, D. Schnyer, S. Taylor, and J. K. Yue, "Outcome prediction in patients with severe traumatic brain injury using deep learning from head ct scans," *Radiology*, vol. 304, no. 2, pp. 385–394, 2022. [Online]. Available: https://doi.org/10.1148/radiol.212181

[8] S. Ding, J. Ye, X. Hu, and N. Zou, "Distilling the knowledge from large-language model for health event prediction," *medRxiv*, pp. 2024–06, 2024.

[9] M. Jin, Q. Yu, D. Shu, C. Zhang, L. Fan, W. Hua, S. Zhu, Y. Meng, Z. Wang, M. Du *et al.*, "Health-llm: Personalized retrieval-augmented disease prediction system," *arXiv preprint arXiv:2402.00746*, 2024.

[10] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ digital medicine*, vol. 4, no. 1, p. 86, 2021.

[11] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, "Tabllm: Few-shot classification of tabular data with large language models," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 5549–5581.

[12] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, D. Jin, Y. Zhang, and Q. Hong, "Lvit: language meets vision transformer in medical image segmentation," *IEEE transactions on medical imaging*, 2023.

[13] C. Le Barbey, "Evaluation of artificial language model chatgpt4 to predict 6-month outcome after traumatic brain injury," Ph.D. dissertation, 2023.

[14] C. Vo, T. Cao, and B. Ho, "Abbreviation detection in vietnamese clinical texts," *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 34, no. 2, 2018. [Online]. Available: //jcsce.vnu.edu.vn/index.php/jcsce/article/view/211

[15] A. Tabassum and D. R. R. Patil, "A survey on text pre-processing & feature extraction techniques in natural language processing," 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:235211496

[16] D. Q. Nguyen, D. Q. Nguyen, T. Vu, M. Dras, and M. Johnson, "A fast and accurate vietnamese word segmenter," *CoRR*, vol. abs/1709.06307, 2017. [Online]. Available: http://arxiv.org/abs/1709.06307

[17] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese natural language processing toolkit," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, 2018, pp. 56–60. [Online]. Available: https://aclanthology.org/N18-5012

[18] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, nov 2019, pp. 6382–6388. [Online]. Available: https://aclanthology.org/D19-1670

[19] T. P. Nguyen, V.-L. Pham, H.-A. Nguyen, H.-H. Vu, N.-A. Tran, and T.-T.-H. Truong, "A two-phase approach for building Vietnamese WordNet," in *Proceedings of the 8th Global WordNet Conference (GWC)*. Global Wordnet Association, 27–30 2016, pp. 261–266. [Online]. Available: https://aclanthology.org/2016.gwc-1.38

[20] M. Xia, X. Kong, A. Anastasopoulos, and G. Neubig, "Generalized data augmentation for low-resource translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, jul 2019, pp. 5786–5796. [Online]. Available: https://aclanthology.org/P19-1579

[21] A. Sugiyama and N. Yoshinaga, "Data augmentation using back-translation for context-aware neural machine translation," in *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*. Association for Computational Linguistics, nov 2019, pp. 35–44. [Online]. Available: https://aclanthology.org/D19-6504

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[24] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[25] Z. Zhao, R. Anand, and M. Wang, "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform," in *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2019, pp. 442–452.

[26] F. Gaillard, "Rotterdam CT score of traumatic brain injury | Radiology Reference Article | Radiopaedia.org — radiopaedia.org," https://radiopaedia.org/articles/rotterdam-ct-score-of-traumatic-brain-injury.

[27] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020.

[28] Q. T. Nguyen, T. L. Nguyen, N. H. Luong, and Q. H. Ngo, "Fine-tuning BERT for sentiment analysis of vietnamese reviews," *CoRR*, vol. abs/2011.10426, 2020. [Online]. Available: https://arxiv.org/abs/2011.10426

[29] D. Q. Nguyen and A. T. Nguyen, "Phobert: Pre-trained language models for vietnamese," *CoRR*, vol. abs/2003.00744, 2020. [Online]. Available: https://arxiv.org/abs/2003.00744

[30] L. Phan, T. Dang, H. Tran, T. H. Trinh, V. Phan, L. D. Chau, and M.-T. Luong, "Enriching biomedical knowledge for low-resource language through large-scale translation," 2022. [Online]. Available: https://arxiv.org/abs/2210.05598

Duc-Khiem Doan is currently a final-year undergraduate student in Biomedical Engineering at Hanoi University of Science and Technology. His main research interests include medical data processing and the application of machine learning and deep learning models to biomedical data.
Email: khiem.DD210483@sis.hust.edu.vn

Thanh-Hai Tran graduated with an engineer's degree in Information Technology from Hanoi University of Science and Technology (HUST) in 2001. She holds an M.S. degree and a Ph.D. degree in Imagery Vision Robotics from Grenoble INP, France, in 2002 and 2006 respectively. Currently, she is a lecturer/researcher at the School of Electrical and Electronics Engineering and International Research Institute in Multimedia, Information, Communication, and Application, HUST. Her main research interests are visual object recognition, video understanding, human-robot interaction, and text detection for applications in Computer Vision.
Email: hai.tranthithanh1@hust.edu.vnnd

Trung-Kien Tran graduated from Hanoi University of Science and Technology in 2004 and worked at the Military Information Technology Institute, AMST, as a researcher in the field of cybersecurity until 2014. After that, he began pursuing a Masterś degree and Ph.D. in Machine Learning at the Artificial Intelligence Laboratory, National Defense Academy of Japan, and graduated in March 2020. Currently, he is an AI specialist at the Military Information Technology Institute, AMST, with primary research interests in Edge Artificial intelligence, Human-Robot Interaction, and Unmanned Vehicle.
Email: t2kien@gmail.com

Thi-Lan Le graduated in Information Technology from Hanoi University of Science and Technology (HUST), Vietnam. She obtained an MS. degree in Signal Processing and Communication from HUST, Vietnam. In 2009, she received her Ph.D. degree at INRIA Sophia Antipolis, France in video retrieval. She is currently an associate professor at the School of Electrical and Electronic Engineering (SEEE), HUST, Vietnam. Her research interests include image processing, computer vision, content-based indexing and, retrieval, video understanding, and human-robot interaction.
Email: lan.lethi1@hust.edu.vn

Hai Vu received a PhD in Computer Science from Osaka University, Japan, in 2009. Currently, the author is a lecturer at the School of Electrical and Electronic Engineering (SEEE), HUST, Vietnam. His research interests include computer vision-based techniques in smart-surveillance camera networks, medical image analysis for diagnostic assistance and Human-Machine Interaction.
Email: hai.vu@hust.edu.vn

Huu-Khanh Nguyen is a neurosurgery resident at 103 Hospital in Hanoi, Vietnam. He graduated from Vietnam Military Medical University in 2022, where he completed seven years of comprehensive medical training. Since then, he has been specializing in neurosurgery, developing his skills and knowledge at one of Vietnam's leading medical institutions.
Email: HuukhanhHVQY@gmail.com

Van Mao Can graduated from the Military Academy of Medicine, Hanoi, Vietnam, in 2000. He earned his Master's Degree in Medicine in 2003 and a Doctoral Degree in System Emotional Science from Toyama University, Japan in 2009. He was appointed Associate Professor in Medicine in March 2018. Since 2018, he has been the Head of the Pathophysiology Department at the Military Medical University, contributing significantly to education, research, and medical advancements.
Email: canvanmao@vmmu.edu.vn

Thanh Bac Nguyen graduated from the Military Academy of Medicine, Hanoi, in 2000 and completed his residency in neurosurgery at Military Hospital 103 in 2005. He earned his Ph.D. in 2020. Since then, he has been the Head of the Department of Neurosurgery at Military Hospital 103. In December 2024, he was appointed Associate Professor. He is dedicated to advancing neurosurgery and mentoring the next generation of medical professionals in Vietnam.
Email: bacnt103@gmail.com