

# A Small-Scale Switch Transformer and NLP-based Model for Clinical Narratives Classification

Thanh-Dung Le, *Member, IEEE*, Philippe Jouvét M.D., Ph.D., and Rita Noumeir Ph.D., *Member, IEEE*

**Abstract**—In recent years, Transformer-based models such as the Switch Transformer have achieved remarkable results in natural language processing tasks. However, these models are often too complex and require extensive pre-training, which limits their effectiveness for small clinical text classification tasks with limited data. In this study, we propose a simplified Switch Transformer framework and train it from scratch on a small French clinical text classification dataset at CHU Sainte-Justine hospital. Our results demonstrate that the simplified small-scale Transformer models outperform pre-trained BERT-based models, including DistillBERT, CamemBERT, FlauBERT, and FrALBERT. Additionally, using a mixture of expert mechanisms from the Switch Transformer helps capture diverse patterns; hence, the proposed approach achieves better results than a conventional Transformer with the self-attention mechanism. Finally, our proposed framework achieves an accuracy of 87%, precision at 87%, and recall at 85%, compared to the third-best pre-trained BERT-based model, FlauBERT, which achieved an accuracy of 84%, precision at 84%, and recall at 84%. However, Switch Transformers have limitations, including a generalization gap and sharp minima. We compare it with a multi-layer perceptron neural network for small French clinical narratives classification and show that the latter outperforms all other models.

**Index Terms**—Clinical natural language processing, cardiac failure, BERT, Transformer.

**Clinical and Translational Impact Statement**— The application of Switch Transformer to clinical text classification represents a promising avenue for improved performance over pre-trained BERT-based models. While it does not outperform a small MLP-NN neural network, the framework has the potential to enhance accuracy on small French clinical narrative classification.

## I. INTRODUCTION

Recent advancements in deep learning have led to the development of Transformer models [1], which have shown remarkable performance in various natural language processing (NLP) tasks [2]. As a result, there is a growing interest in applying Transformer-based models to clinical applications, such as predicting disease risk [3], identifying disease [4], and improving clinical decision-making [5]. These models can be trained on various data sources, including electronic health records (EHRs) [5]–[7], medical imaging [8]–[10], electrogram [11], [12], and genome [13], [14] to extract clinically relevant information and provide accurate predictions. Overall, Transformer models present a powerful tool for clinical applications and can potentially play an increasingly important role in healthcare.

In clinical NLP, Transformers-based models have shown great promise in clinical narrative classification. In this context, clinical narrative refers to patient encounters in EHRs or other clinical documentation. Using Transformers-based models, researchers and clinicians can develop algorithms that automatically classify these narratives based on different criteria, such as diagnosis, treatment, or patient outcomes.

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC), in part by the Institut de Valorisation des données de l'Université de Montréal (IVADO), in part by the Fonds de la recherche en santé du Québec (FRQS), and in part by the Fonds de recherche du Québec-Nature et technologies (FRQNT).

Thanh-Dung Le is with the Biomedical Information Processing Lab, École de Technologie Supérieure, University of Québec, Canada, and also with the Research Center at CHU Sainte-Justine, University of Montreal, Canada (Email: thanh-dung.le.1@ens.etsmtl.ca).

Rita Noumeir is with the Biomedical Information Processing Lab, École de Technologie Supérieure, University of Québec, Canada.

Philippe Jouvét are with the Research Center at CHU Sainte-Justine, University of Montreal, Canada.

This can help streamline clinical workflows and improve patient care by providing more accurate and efficient clinical data processing. Some examples of successful applications of Transformers-based models for clinical narrative classification include identifying clinical coding [15], [16], diagnosing health conditions [17]–[20], and detecting clinical events [21]–[23]. As such, Transformers-based models have become an increasingly important tool in clinical NLP and are likely to continue playing a significant role in this field [24].

Despite their many benefits, Transformers-based models for clinical text classification have some limitations that must be considered. One major challenge is the need for large amounts of annotated clinical data to train these models effectively. Clinical data is often scarce and sensitive, which makes it challenging to obtain and annotate in a way that preserves patient privacy [25]. Additionally, clinical language is highly specialized and can vary significantly across different specialties and regions, making it difficult to develop models that generalize well across different contexts [26]. There is a risk of bias in the data used to train these models, leading to errors or disparities in the predictions made [27]. Furthermore, the computational requirements of Transformer-based models can be pretty high, which can limit their use in resource-constrained settings where computational resources are limited [28]. Finally, the interpretability of these models can be limited, making it difficult for clinicians to understand how they make their predictions and trust their outputs [29], [30]. While Transformers-based models have great potential for clinical text classification, they also require careful attention to their limitations and the potential biases that can arise.

- Computational requirements: If a model lacks the necessary computational capacity, its training efforts will fail, regardless of the learning algorithm's sophistication or the

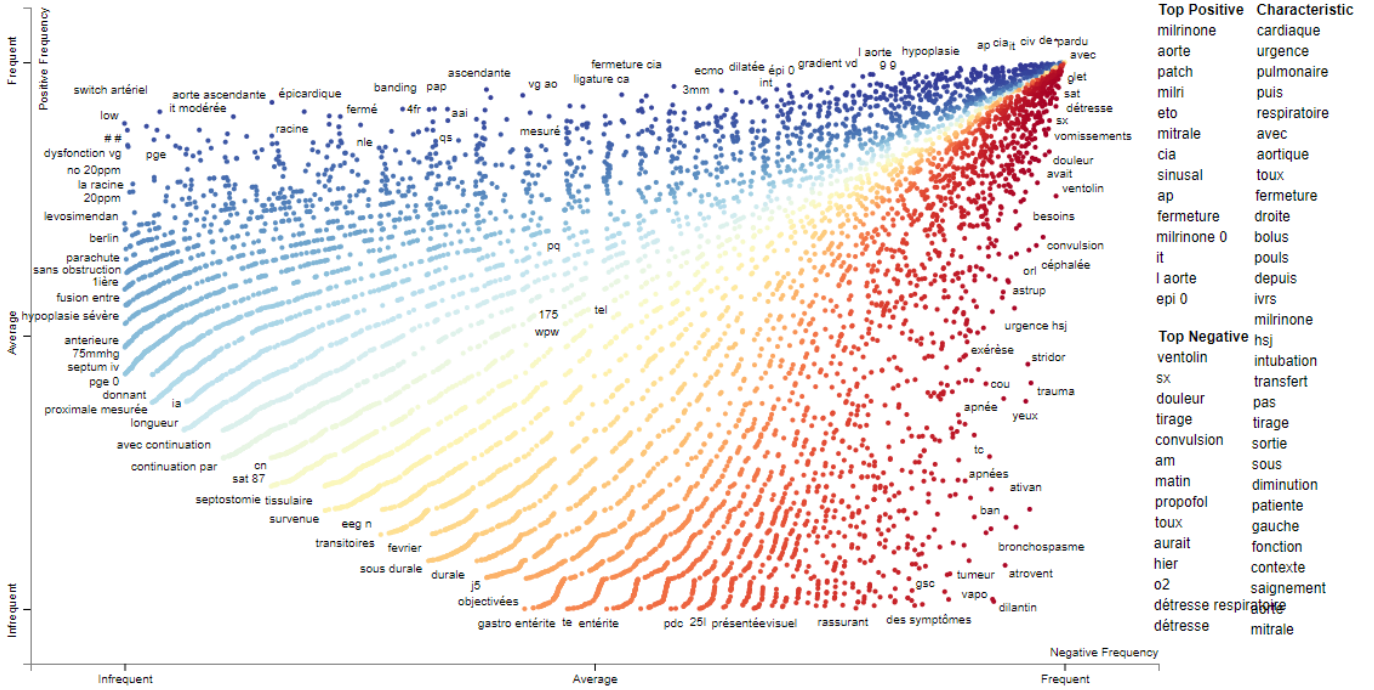


Fig. 1. French clinical note at CHUSJ illustration by using Scattertext visualization.

training data's quality [31]. This can be a limiting factor for smaller clinical text or resource-constrained settings.

- Data requirements: Transformer-based models require large amounts of labeled data for training, which may not be available for some clinical text classification tasks, especially for rare or low-frequency conditions [32].
- Domain-specific language: Clinical text is highly domain-specific and contains jargon and abbreviations that may not be covered by general language models such as Transformers. This can lead to suboptimal performance on clinical text classification tasks [33].
- Interpretability: Transformer-based models are highly complex and difficult to interpret, making it challenging to understand how the model makes predictions, which is essential for clinical decision-making [34].

Another significant limitation of using Transformer-based models for clinical text classification is that they may not perform as well for languages other than English and that are in limited availability. Most Transformer-based models have been developed and trained on English-language text, and their performance may suffer when applied to other languages [35]. This is particularly important in the clinical context, where patient data can be collected in many languages. Another challenge is that clinical datasets are often small and imbalanced, making it difficult to train accurate models using Transformer-based [36]. Small datasets can also lead to overfitting, where the model performs well on the training data but fails to generalize to new data. When there is insufficient data, the Transformer model does not learn to focus on local features in the lower layers of the network. This may result in reduced model performance, as it cannot effectively capture relevant information from the input data [37]. Overall, while Transformer-based models offer many advantages for clinical

text classification, their effectiveness is influenced by the data's language and the training dataset's size and quality.

This study aims to overcome the challenges of using Transformer-based models for clinical text classification for a small French clinical note by employing the Mixture-of-expert (MoE) framework from the recent Switch Transformer model developed by Google [38]. Switch Transformer is an extension of the Transformer architecture motivated by the original model's self-attention mechanisms. Still, it uses an MoE mechanism to address the limitations of the conventional Transformer [1]. A key technical difference between Switch Transformers with an MoE mechanism and Transformers with self-attention is how they handle the modeling of complex input-output relationships. An example of the effectiveness of MoE has been proven by [39]; that study shows that the approach of using parameter sharing to compress along the depth of the model, which is used in existing works, is limited in terms of performance. To improve the model's capacity, the authors propose scaling along the model's width by replacing the feed-forward network with an MoE. This allows for better modeling capacity and potentially better performance.

Additionally, the study [40] suggests that simply increasing the model's size is insufficient to address the issue of performance degradation over time from neural language models. However, the researchers found that using models that continuously update their knowledge with new information can help alleviate this problem. While Transformers with self-attention model these relationships through a single attention mechanism that captures dependencies between all input and output positions, Switch Transformers with an MoE mechanism decompose the problem into smaller, simpler sub-problems, each handled by a different "expert" model. In other words, instead of using a single global attention mech-

anism, Switch Transformers employ multiple local attention mechanisms focusing on different input aspects. The gating mechanism used in Switch Transformers selects which expert model to use for a given input, depending on the context. Therefore, this approach can potentially improve the modeling of complex input-output relationships and increase the model's efficiency, especially when dealing with complex data from the clinical domain. This is particularly important in clinical data, where information is often conveyed through complex and nuanced language. By employing this approach, our study aims to improve the accuracy and generalizability of clinical text classification models for small datasets in languages other than English. We have made several significant contributions to clinical text classification using Transformer-based models.

- First, our study demonstrates a comprehensive implementation of a simplified Switch Transformer model from scratch. This would allow other researchers to understand and replicate the methodology used in the study, which is essential for building on and advancing this work.
- Second, our study provides experimental evidence showing the limitations of Transformer-based models in terms of generalization gap and sharp minima. This highlights the importance of carefully selecting and preprocessing the data used to train these models to avoid overfitting and improve generalization performance.
- Finally, our study illustrates the interpretable output of the model by adapting the Integrated Gradients (IG) [41]. It provides a way to attribute importance to the input features of a model, allowing clinicians and researchers to gain insight into how the model is making its predictions.

This study significantly contributes to developing accurate and interpretable clinical text classification models and sheds light on the limitations and challenges of using Transformer-based models in this context. By leveraging the MoE technique, this approach offers a promising solution to the problem of small datasets in clinical text classification, enabling the practical adaptation of Transformer-based models to real-world clinical data. The MoE allows the model to learn from multiple experts, each specialized in different aspects of the data, and to combine their outputs to achieve improved performance. Furthermore, a Transformer-based model provides a powerful tool for capturing the complex relationships between words and phrases in clinical text. However, our proposed method underperforms compared to a smaller and simpler framework that combines statistical representation learning with term frequency-inverse document frequency and multilayer perceptron network. Despite this limitation, our work demonstrates the potential of combining MoE with Transformer-based models to overcome data limitations and improve the accuracy and interpretability of clinical text classification models, which could have a significant impact on clinical decision-making.

This paper is organized as follows. Section II will discuss the materials and methods. Then, the experimental results and discussion will be discussed in section III, and IV, respectively. Misclassification cases will be discussed in section V. Finally, section VI provides concluding remarks.

## II. MATERIALS AND METHODS

### A. French Clinical Data at CHUSJ

The clinical decision support system (CDSS) system in the CHU Sainte Justine (CHUSJ) hospital aims to improve the diagnosis and management of acute respiratory distress syndromes (ARDS) in real-time by automatically screening data from electronic medical records, chest X-rays, and other sources. Previous studies have found that the diagnosis of ARDS is often delayed or missed in many patients [42], emphasizing the need for more effective diagnostic tools. Three main conditions must be detected to diagnose ARDS: hypoxemia, chest X-ray infiltrates, and absence of cardiac failure [43]. The research team at CHUSJ has developed algorithms for detecting hypoxemia [44], analyzing chest X-rays [45], [46], and identifying the absence of cardiac failure. In addition, the team has performed extensive analyses of machine learning algorithms for detecting cardiac failure from clinical narratives using natural language processing [47], [48]. Implementing these algorithms could increase ARDS diagnosis rates and improve patient outcomes.

This study was conducted following ethical approval from the research ethics board at CHUSJ; and, the study's design focused on identifying cardiac failure in patients within the first 24 hours of admission by analyzing admission and evolution notes during this initial period. Therefore, we conducted a retrospective analysis of EHRs from the Research Center of CHUSJ in this study. The dataset consisted of 580,000 unigrams extracted from 5,444 single lines of short clinical narratives. Of these, 1,941 cases were positive (36% of the total), and 3,503 cases were negative. ScatterText [49] was utilized to visualize the notes and identified over 580,000 unigrams (n-grams), as depicted in Fig. 1. The visualization showcases the most frequent words for positive cases in the upper right corner, negative cases in the lower-left corner, and less frequent words for both cases in the center. The top terms for positive and negative cases are also presented on the right-hand side. Upon inspection, we observed that most top terms for positive cases were positively related to cardiac malfunction, such as milrinone or milri (milrinone), and aorte or aortique valve (aortic valve). In contrast, terms like respiratoire (respiratory), détresse respiratoire (distress respiratory), and O<sub>2</sub> (oxygen) indicated respiratory syndromes in negative cases. While the longest n-gram was over 400 words, most n-grams had a length distribution between 50 and 125 words. The average length of the number of characters was 601 and 704, and the average size of the number of digits was 25 and 26 for the positive and negative cases, respectively. We pre-processed the data by removing stop-words and accounting for negation in medical expressions. Numeric values for vital signs (heart rate, blood pressure, etc.) were also included and decoded to account for nearly 4% of the notes that contained these values. All the notes are short narratives; detailed characteristics can be found in the Supplementary Materials from [47].

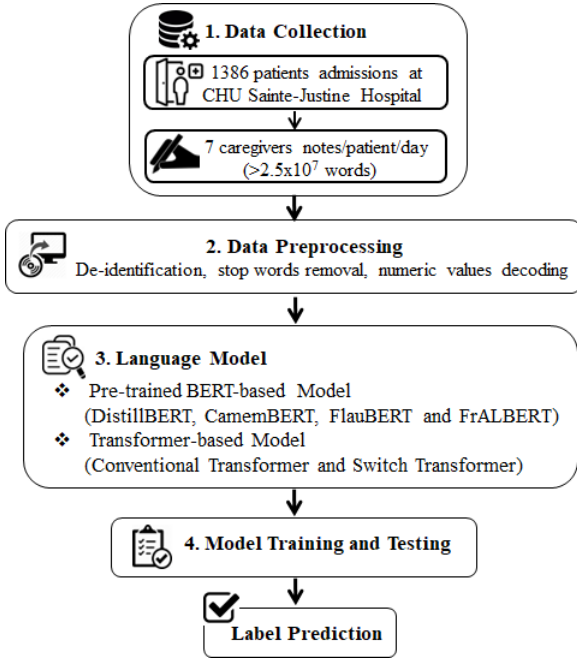


Fig. 2. Workflow demonstration of the proposed methodology to classify French clinical narratives at CHUSJ hospital.

### B. Language Models for Clinical Narratives

This manuscript thoroughly analyzes the present state of pre-trained BERT-based models and Transformer models for clinical narrative classification, with a particular emphasis on limited datasets. Various pre-trained BERT-based models for the French language are leveraged, such as FlauBERT, FrALBERT, CamemBERT, and DistilBERT, as depicted in Fig. 2. Moreover, conventional and Switch Transformer models are constructed from scratch to perform the same task. Finally, we compare the performance of all models based on various evaluation metrics for binary classification, including accuracy, precision, recall, F1-score, and area under the curve (AUC). This study endeavors to offer insights into the efficacy of these models on limited datasets, which is a critical aspect in real-world clinical settings for non-English notes.

1) *Transformer-based Models*: Transformer-based models have been highly effective for various NLP tasks, including text classification. The conventional Transformer model [1] with multi-head self-attention is a widely used architecture for this task. Shown in Fig. 3 (left), its architecture comprises an encoder consisting of multiple layers of multi-head self-attention and feedforward neural networks (FFN). The multi-head self-attention mechanism allows the model to weigh the importance of different words in a sequence based on their semantic relationships, while the FFNs transform the output of the self-attention layer into a more helpful representation. The Transformer's core is the self-attention mechanism based on mathematical expressions [51]. Given a sequence of input embeddings  $x_1, \dots, x_n$ , the self-attention mechanism computes a set of context-aware embeddings  $h_1, \dots, h_n$  as follows:

$$h_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (1)$$

where Attention is the scaled dot-product attention function:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Then, the multi-head attention is a concatenation of all head of  $h_i$ , as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(h_1, \dots, h_n)W^O \quad (3)$$

Additionally, the position-wise FFNs are multi-layer perceptrons applied independently to each position in the sequence, which provide a nonlinear transformation of the attention outputs. FFNs are calculated as follows:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \quad (4)$$

For each layer, there is a Layer Normalization which normalizes the inputs to a layer in a neural network to improve training speed and stability.

$$\text{LayerNorm}(x) = \gamma \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \quad (5)$$

where  $Q$ ,  $K$ , and  $V$  are the query, key, and value matrices,  $W_i^Q$ ,  $W_i^K$ , and  $W_i^V$  are the learned weight matrices for the  $i$ -th head of the multi-head attention,  $W_1$  and  $W_2$  are the weight matrices for the position-wise FFNs,  $\gamma$  and  $\beta$  are learned scaling and shifting parameters for layer normalization, and  $\mu$  and  $\sigma$  are the mean and standard deviation of the input feature activations. The working mechanism in the Transformer architecture can be summarized into the following steps:

- 1) **Linear Transformation**: The input sequence is projected into three vectors, query  $Q$ , key  $K$ , and value  $V$ , by applying a linear transformation to the input embedding.
- 2) **Splitting**: The  $Q$ ,  $K$ , and  $V$  vectors are then split into multiple heads  $h_i$ , allowing the model to simultaneously attend to different aspects of the input sequence Eq. 1.
- 3) **Scaled Dot-Product Attention**: For each  $h_i$ , the model calculates the attention weights between the  $Q$  and  $K$  vectors by scaling their dot product by the square root of the vector dimension. It calculates each  $K$  vector's importance to the corresponding  $Q$  vector.
- 4) **Softmax**: The resulting attention weights are normalized using a softmax function, ensuring that they sum to 1.
- 5) **Weighted Sum**: The attention weights are then used to weigh the  $V$  vectors, producing an attention output for each head  $h_i$  Eq. 2.
- 6) **Concatenation**: The attention outputs from each head are concatenated and projected back to the original vector dimension through another linear transformation Eq. 3.
- 7) **Feed Forward Network**: The resulting output is passed through a feedforward network, which introduces non-linearity and allows the model to capture more complex relationships between the input and output Eq. 4.

By performing these steps for each layer in the encoder and decoder, the multi-head self-attention mechanism allows the Transformer architecture to capture rich semantic relationships between different words in a sequence and is highly effective for a wide range of NLP tasks. However, the conventional Transformer architecture has some limitations. One of the main issues is that the self-attention mechanism requires

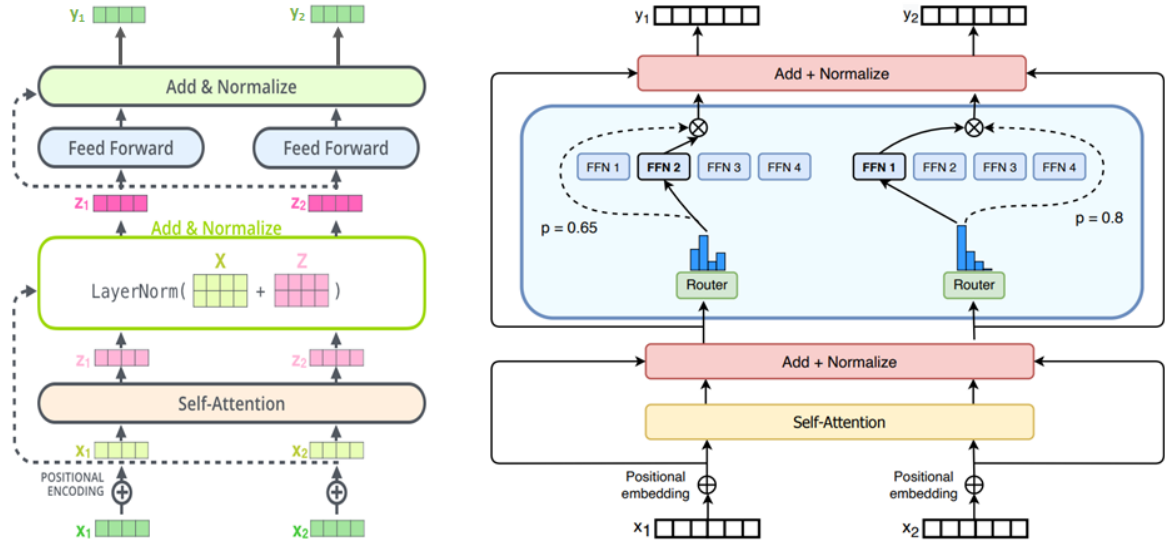


Fig. 3. Illustration of a Conventional Transformer [50] (left), and a Switch Transformer [38] (right) encoder block.

quadratic computation time concerning the input sequence length, making it difficult to scale the model to very long sequences [52], and lower generalizability for a short sequence [25]. Additionally, the self-attention mechanism treats all positions in the input sequence equally, which may not be optimal for certain types of inputs where some positions are more critical than others. While the Transformer model has shown state-of-the-art performance on many NLP tasks, it can still struggle to capture complex input-output relationships requiring more specialized models.

Switch Transformers [38] attempt to address these limitations by introducing a mixture of expert (MoE) mechanisms that decompose the problem into smaller, simpler sub-problems, allowing the model to handle long sequences and complex input-output relationships better. As mentioned above, the multi-head self-attention mechanism in the Transformer model is motivated by the need to capture semantic relationships between words in a sequence, but it has limitations when dealing with short sequences [25]. The MoE mechanisms allow the model to divide the sequence into smaller, more manageable segments and apply different experts to each segment. This approach has improved the model's performance on short sequence tasks and has achieved state-of-the-art results on several benchmarks [39], [40], [53].

The critical difference in the mathematical equation of the Switch Transformer compared to the conventional Transformer is replacing the FFN with the MoE mechanism, shown in Fig. 3 (right). In the conventional Transformer, the FFN consists of two linear layers with a ReLU activation function in between. The MoE mechanism, on the other hand, uses a set of expert networks to learn different aspects of the input data and then combines their outputs with a gating network. It allows the model to dynamically choose between multiple sets of parameters (i.e., expert modules) based on the input. This contrasts the original Transformer model in Eq. 4, which uses a fixed set of parameters for all inputs. Formally, the MoE mechanism in the Switch Transformer can be represented by

the following equation:

$$z_t = \sum_j g_j(x_t) * e_j(x_t) \quad (6)$$

where  $g_j(x_t)$  is a gating function that determines the importance of expert module  $j$  for input  $x_t$ , and  $e_j(x_t)$  is the output of expert module  $j$  for input  $x_t$ . The switch mechanism is implemented by learning the parameters of the gating functions, which are used to select the expert modules dynamically. This allows the model to adapt to different input distributions and perform better on various tasks. Here is a summary of how the MoE mechanism works in the Switch Transformer:

- 1) The input is split into multiple subspaces, and each subspace is processed by a separate expert. Each expert is a separate neural network trained to specialize in a specific subset of the input space.
- 2) The output of each expert is a vector that represents its prediction for the given input subspace.
- 3) A gating mechanism selects the most relevant expert for a given input. This gating mechanism takes the input and produces a set of weights that determine the importance of each expert's prediction.
- 4) The final output is a weighted combination of the experts' predictions. The weights used in the combination are determined by the gating mechanism.

Overall, the MoE allows the Switch Transformer to learn complex patterns in the input space by leveraging the specialized knowledge of multiple experts. The MoE framework enables the model to learn from multiple experts, each specialized in different aspects of the data, and combine their outputs to achieve better performance. This can lead to better performance on tasks requiring understanding inputs and offers a promising solution to the challenge of small datasets in clinical text classification. Consequently, the study uses its ability to capture the complex relationships between words and phrases in the clinical text.



2) *Pre-trained BERT-based Models for French*: Pre-trained BERT-based models have become increasingly popular, enabling researchers and practitioners to perform various language-processing tasks with unprecedented accuracy. While BERT [54] was initially developed for English language processing, it has since been adapted to several other languages, including French. In this context, we will explore some of the most popular pre-trained BERT-based models for French language processing available from Huggingface.

**CamemBERT [55]**: This is a pre-trained Transformer-based language model designed explicitly for processing French text. It is based on the Roberta architecture and was trained on a large corpus of French text that was filtered and pre-processed to improve the data quality. Its pre-training objective is a masked language model, where some input tokens are masked, and the model is trained to predict the missing tokens. Overall, CamemBERT is a highly effective tool for processing French language text and can be fine-tuned for specific downstream tasks or used for transfer learning in multilingual settings.

**FlauBERT [56]**: It is based on the original BERT architecture and was trained on a large corpus of the French text. It has been shown to perform strongly on several natural language processing tasks in French, including named entity recognition and sentiment analysis. It also performs well on tasks related to French morphosyntaxes, such as part-of-speech tagging and dependency parsing. It was trained using a masked language model objective, where a portion of the input tokens are masked, and the model is trained to predict the missing tokens. FlauBERT is a powerful language model for processing French text that can be fine-tuned for specific downstream tasks.

**FrALBERT [57]** is a Transformer-based language model designed explicitly for text classification tasks in French. It is based on the ALBERT architecture and was trained on a large corpus of the French text. It has achieved state-of-the-art performance on several text classification tasks in French, including sentiment analysis, news categorization, and toxic comment classification. The model was fine-tuned using a supervised learning approach, where the model was trained on labeled data to predict the correct class label for a given input text. FrALBERT is available for download and can be fine-tuned on specific text classification tasks in French or used for transfer learning in multilingual settings.

**DistillBERT [58]** is a smaller and more efficient version of the BERT architecture designed to reduce the computational and storage requirements of the model while maintaining its performance. It was trained on a large corpus of French text and has been shown to perform strongly on various natural language processing tasks, including text classification. It is particularly useful for text classification tasks in French, such as sentiment analysis and news categorization. DistillBERT is much smaller than the original BERT model, making it more suitable for deployment on resource-constrained devices or in applications where speed and efficiency are a concern.

### III. EXPERIMENTAL IMPLEMENTATION

Table I shows the hyperparameters of different Transformer-based models used in this study, including CamemBERT,

DistillBERT, FlauBERT, FrALBERT, Transformer, and Switch Transformer. The hyperparameters compared include hidden layers and total parameters. CamemBERT and FrALBERT have 12 hidden layers, whereas DistillBERT, FlauBERT, Transformer, and Switch Transformer have 6, 6, 4, and 4 hidden layers, respectively. Regarding to total parameters, CamemBERT has the highest number of parameters, with 111 million, followed by DistillBERT with 66.7 million parameters, and FlauBERT with 54.6 million parameters. FrALBERT, Transformer, and Switch Transformer have significantly fewer parameters, with 12.3 million, 2.3 million, and 5.7 million, respectively. The variation in hyperparameters across different models reflects the differences in the architecture and design of the models. This information is crucial for understanding each model's computational complexity and efficiency and helps select the most suitable model.

When training a machine learning model, the hardware and software specifications used for the training process can significantly impact the model's performance and efficiency. In this case, the models were trained on a local machine with a Quadro P620 GPU and CUDA library version 12. Including these specifications when describing the trained models can provide important context for others looking to replicate or build upon the work.

Defining the hyperparameters during the training process of Transformers is a critical step in achieving good performance. Hyperparameters are the settings that control the behavior of the training algorithm, and they can significantly impact the final performance of the model. Here are some of the critical hyperparameters that are tuned during the training process of BERT-based and Transformer models in this study:

- **Maximum sequence length**: This is the maximum number of tokens that can be inputted into the model simultaneously. Setting an appropriate maximum sequence length can affect the performance and memory usage of the model. Due to computational constraints, the maximum sequence length varies from 128 to 256.
- **Batch size**: Choosing an appropriate batch size can affect the speed and stability of the training process. We varied the training batch size for each trial, ranging from 4 to 32 (with gradient accumulation as 4), based on the knowledge that training with smaller batches is more effective for highly low-resource language training [59].
- **Drop-out**: This regularization technique randomly drops out some of the neurons during training to prevent overfitting. The dropout rate determines the proportion of neurons to drop out during each iteration [60].
- **Optimizers**: These algorithms update the model weights during training to minimize the loss function. Different optimizers have different strengths and weaknesses, and choosing the right one can impact the final performance of the model. Adaptive Moment Estimation (Adam) [61], AdamW (Adam with weight decay) [62] were used.
- **Learning rate**: Cosine annealed learning rate with warmup can help prevent training instability in the deeper layers of a neural network; its primary purpose is to help the model converge more quickly and effectively to a better solution overall [63].

TABLE I  
MODELS HYPERPARAMETERS

Hyperparameters	CamemBERT	DistilBERT	FlauBERT	FrALBERT	Transformer	Switch Transformer
Hidden Layers	12	6	6	12	4	4
Total Parameters	111 M	66.7 M	54.6 M	12.3 M	2.3 M	5.7 M

TABLE II  
HYPERPARAMETERS OF THE FINE-TUNED MODELS

Hyperparameters	Pretrained BERT-based	Transformer	Switch Transformer
Number of multi-head attention	N/A	4	4
Number of Experts	N/A	N/A	4
Batch size	16	16	16
Dropout	0.5	0.35	0.35
Learning rate	Cosine annealed	Cosine annealed	Cosine annealed
Optimizer	Adam	AdamW	AdamW
Adam_ε	N/A	5*1e-06	5*1e-06
Maximum sequence length	256	256	256

- **Number of multi-head attention:** This determines the number of attention heads used in the multi-head attention layer of the Transformer. Increasing the number of attention heads can improve the model's ability to attend to different input parts.
- **Number of experts:** This determines the number of experts used in the MoE layer of the Transformer. Increasing the number of experts can improve the model's ability to handle diverse inputs.

Choosing appropriate values for these hyperparameters requires careful experimentation and tuning to achieve the best possible results. Additionally, optimizing hyper-parameters is essential for achieving high performance in machine learning models, but this process comes with a tradeoff between the quality of the final solution and the time required for computation. However, not all hyperparameters significantly impact model accuracy, and only a few parameters require careful tuning. As reported in [64], the model size, learning rate, batch size, and maximum sequence length are the three critical hyper-parameters for Transformer model training. For this reason, grid search can be an efficient approach for optimizing these parameters by simultaneously exploring all possible combinations of intervals. Compared to Bayesian optimization, grid search has advantages in parallelization and flexibility of resource allocation [65]. In this study, we used grid search to optimize hyper-parameters for model training. The combination with the highest estimated performance was considered the optimal solution, and this approach balances computational efficiency and models' accuracy.

Finally, table II presents the hyperparameters used to fine-tune three models. For the pre-trained BERT-based model, the number of multi-head attention and the number of experts are not applicable (N/A), as this model is already trained and does not require further customization. The batch size, epochs, dropout rate, learning rate, and optimizer for all models are specified. The trained BERT-based model uses an Adam optimizer with a dropout rate of 0.5 and a cosine decay learning rate. The Transformer and Switch Transformer models use an AdamW optimizer with a dropout rate of 0.35 and a cosine decay learning rate. The Adam\_ε is only specified

for the Transformer and Switch Transformer models and is set to 5\*1e-06. The maximum sequence length for all models is set to 256. The fine-tuning process for the pre-trained BERT-based model was performed for 40 epochs, while the Transformer and Switch Transformer models were fine-tuned for 70 epochs. Additionally, the GlorotNormal initializer [66], batch normalization [67], [68] are employed for models' stability, and balancing the classes by using the Bayes Imbalance Impact Index [69] to deal with the imbalanced classes. Then, these hyperparameters were carefully chosen to achieve optimal performance and prevent overfitting.

The data was divided into 80% training and 10% validation and 10% testing. To assess the performance of our method, metrics including accuracy, precision, recall, and F1 score were used [70]. These metrics are defined as follows.

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall/Sensitivity} &= \frac{TP}{TP + FN} \\ \text{F1-Score} &= \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned}$$

where TN and TP stand for true negative and true positive, respectively, and are the number of negative and positive patients correctly classified. FP and FN represent false positives and false negatives and the number of incorrectly predicted positive and negative patients.

#### IV. RESULTS AND DISCUSSION

During training and validation shown in Fig. 4, the Switch Transformer model showed a gradual decrease in loss with increasing epochs. The loss started to converge after around 20 epochs and reached its minimum at the 30th epoch. Applying the early stopping at this point helped prevent the model's overfitting. The accuracy and precision of the model showed a smooth convergence to their optimal values for both the training and validation phases. However, the recall values for the two phases were observed to be quite fluctuating. The model's overall performance was good, with high accuracy,

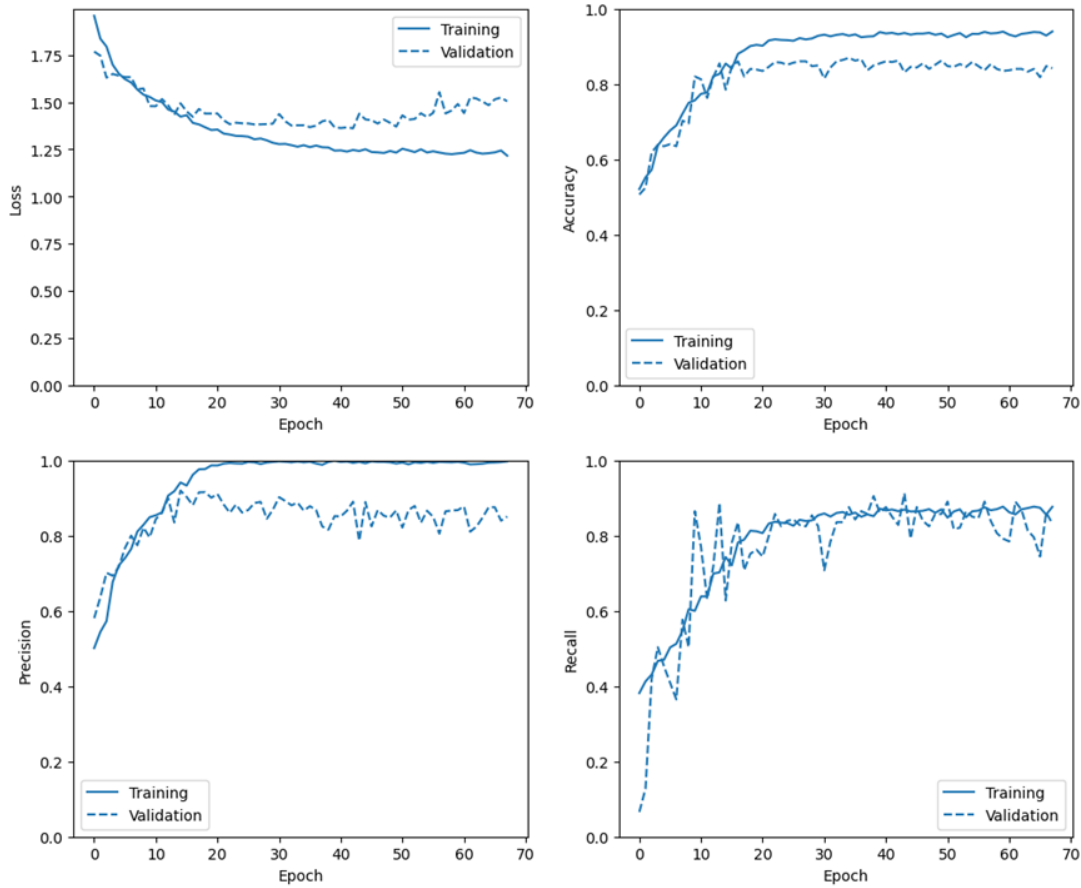


Fig. 4. Training and validation performance results from Switch Transformer model.

precision, and recall. The model's ability to reach its optimal values with smooth convergence and with the help of early stopping indicates the model's effectiveness in the given task.

The results presented in the table III indicate that careful hyperparameter tuning can result in better performance of Transformer models over pre-trained BERT-based models for the given task. The table compares the performance of six classifiers with metrics such as accuracy, precision, recall, F1, and AUC. The classifiers include DistillBERT, CamemBERT, FlauBERT, FrALBERT, Transformer, and Switch Transformer. The results show that the best-performing classifier in accuracy, precision, recall, F1, and AUC is Switch Transformer, with an accuracy score of 0.87, precision of 0.87, recall of 0.85, F1 score of 0.86, and AUC of 0.92. The Transformer model has the second-best performance with an accuracy score of 0.85. DistillBERT, CamemBERT, and FrALBERT perform comparably well, with accuracy scores ranging from 0.80 to 0.83. The Switch Transformer and Transformer models achieved the best accuracy, precision, recall, F1 score, and AUC. These models demonstrated faster training and evaluation times than others, making them the most suitable options for the given task. However, it is essential to note that FlauBERT achieved the best precision, recall, F1 score, and AUC among all models, although it required longer training and evaluation times. Compared to other methods (excluding fine-tuning), mixture-of-experts (MoEs) is more

efficient regarding the computational resources required [71]. The study suggests that Switch Transformer and Transformer models are the most suitable for the given task, given their high performance and faster training and evaluation times. Overall, these findings suggest that careful selection of Transformer-based models and hyperparameter tuning can significantly improve the performance of small clinical narrative classification.

Fig. 5 compares the confusion matrices obtained from six models. Each confusion matrix presents the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for binary classification tasks. This study labels the classes '0' for 'Negative' and '1' for 'Positive.' The Switch Transformer model obtained the highest number of TP and TN, with 253 and 219, respectively. It misclassified 34 instances as false positives and 38 instances as false negatives. DistillBERT, on the other hand, obtained 253 TP and 201 TN, with 54 instances misclassified as false positives and 56 instances as false negatives. FlauBERT and FrALBERT models had similar results with 246 TP and 215 TN and 241 TP and 209 TN, respectively. Both models misclassified around 15% of instances. CamemBERT model obtained 239 TP and 214 TN, with 48 and 43 instances misclassified as false positives and false negatives, respectively. Finally, the Transformer model obtained 250 TP and 213 TN, with 37 and 44 instances misclassified as false positives and false negatives, respectively. In summary, the Switch Transformer



TABLE III  
A COMPARISON PERFORMANCE OF DIFFERENT CLASSIFIERS

Model	Accuracy	Precision	Recall	F1	AUC	Training Time	Evaluation Time
DistilBERT	0.80	0.79	0.78	0.78	0.84	109	5
CamemBERT	0.83	0.82	0.83	0.82	0.89	212	19
FlauBERT	0.84	0.84	0.84	0.84	0.91	51	6
FrALBERT	0.83	0.82	0.81	0.81	0.89	196	19
Transformer	0.85	0.85	0.83	0.84	0.91	<b>4</b>	<b>1</b>
Switch Transformer	<b>0.87</b>	<b>0.87</b>	<b>0.85</b>	<b>0.86</b>	<b>0.92</b>	34	2

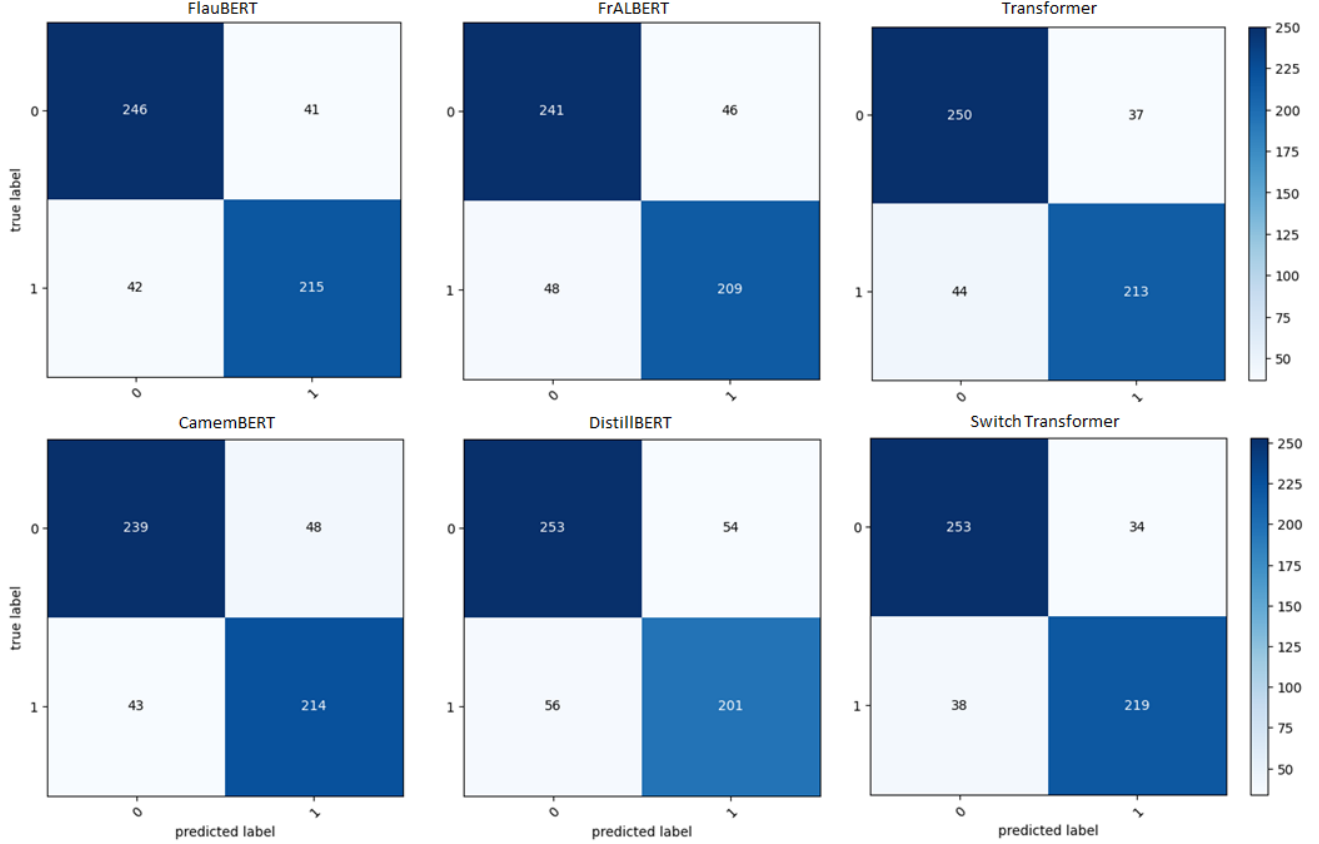


Fig. 5. Confusion matrix comparison for all classifiers.

model achieved the highest number of correct classifications and the lowest number of misclassifications, followed closely by the DistilBERT and Transformer models. The FlauBERT and FrALBERT models performed similarly, with slightly higher misclassifications. However, the CamemBERT model had the lowest number of correct classifications and a relatively high number of misclassifications. These results can guide the selection of models for future classification tasks. Particularly, it suggests that simpler models (in terms of the number of parameters) may perform better for non-English and limited clinical narrative datasets.

Although the Switch Transformer outperforms several other models, including DistilBERT, CamemBERT, FlauBERT, FrALBERT, and the conventional Transformer model, its performance falls short when compared to two of our previous studies [47], [48] that extensively analyzed a conceptual framework for detecting a patient's health condition from contextual input to output. The proposed framework in those studies utilized a combination of TF-IDF (term frequency-inverse doc-

ument frequency) and MLP-NN (multilayer perceptron neural network), achieving an overall classification performance of 89% accuracy, 88% recall, and 89% precision. Moreover, sparsity reduction significantly affected classifier performance in downstream tasks, and a generative AE (autoencoder) learning algorithm effectively leveraged sparsity reduction to help the MLP-NN classifier achieve 92% accuracy, 91% recall, 91% precision, and 91% F1-score. These findings suggest that the simpler frameworks are effective for this specific context and highlight the limitations of the Switch Transformer model.

While the Switch Transformer model has demonstrated promising results in clinical text classification, there is still room for further improvement of its performance. One possible area of investigation is the training methodology, as suggested by previous research [72], [73]. Specifically, the model was trained for 500 epochs without early stopping, which resulted in three distinctive phases in the learning curves of training and validation losses in Fig. 6. Initially, the model underwent the learning phase, where the loss gradually decreased and

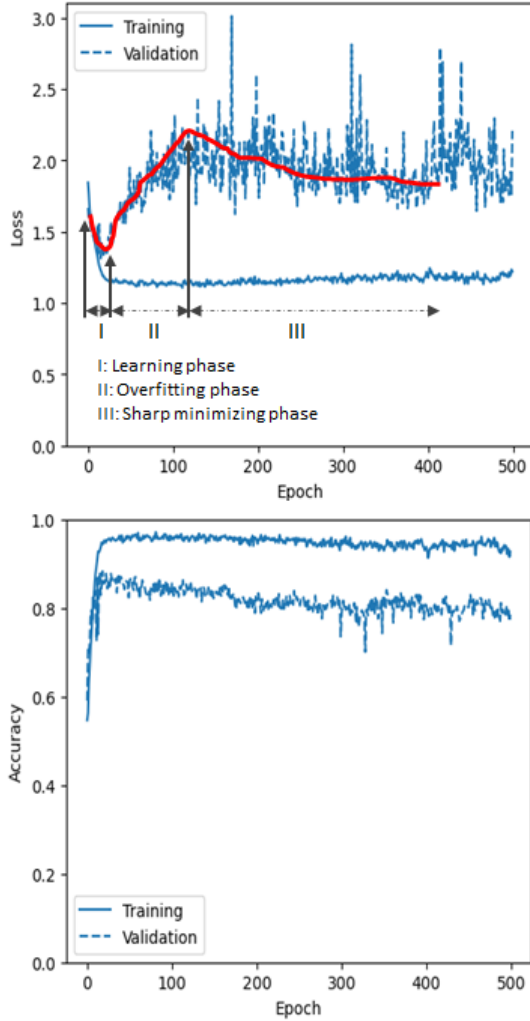


Fig. 6. Generalization gap and sharp minima during training Switch Transformer without early stopping.

reached its minimum at epoch 30. Subsequently, the model entered the second phase, where overfitting occurred, and the loss increased sharply, reaching its maximum at epoch 120. Interestingly, the model experienced double descent, and the loss started decreasing again in the third phase and remained flat until nearly the end of the 400 epochs. During this phase, the classifier was confined to a sharp minimum and failed to improve further. Regarding accuracy, after achieving the optimal value, both learning curves from training and validation remained flat, which is expected. These are typical phenomena in deep learning models trained on small datasets, as the model tends to overfit the data and struggles with generalization. The classifier could not bridge the generalization gap caused by the sharp minima effect due to insufficient data explained in [74].

Furthermore, we propose a novel perspective on this behavior and find a better illustration, viewing them through hidden embedding visualization for each layer during training and validation to explain their behavior. To illustrate this perspective, we present detailed visualizations of the Switch Transformer embedding for each layer (from 1 to 4) in Figure 7. We utilize t-SNE, a nonlinear dimensionality reduction technique

well-suited for embedding high-dimensional data into lower-dimensional data (2 dimensions in our case). By analyzing the hidden embedding from the model, we successfully observe the difference between the training and validation processes. The four top figures illustrate that after the 30th epoch, the model successfully separates the two classes (1: positive, 0: negative) in each hidden layer. Remarkably, the last hidden layer (4th layer) achieves perfect classification accuracy of 98% on the training set. However, this level of performance does not carry over to the validation set at the same epoch. The four bottom figures demonstrate that the two classes overlap, and the model cannot learn a clear boundary between them, resulting in only 87% validation accuracy. Therefore, we observe a generalization gap between the training and validation for a large model with small data.

## V. MISCLASSIFICATION INTERPRETABILITY

Interpretability of misclassifications is essential to model evaluation, particularly in critical applications such as medical diagnosis. In this study, we analyze the misclassification cases of the Switch Transformer model by visualizing the results from the misclassification. Totally, there are 72 cases of misclassification from the results of the Switch Transformer. Our focus has been primarily on the false negatives, where the true label indicates the presence of cardiac failure (True label is 1); however, our classifiers predict the opposite. We have referred to the labeled data to understand the reasons behind these misclassifications better. The clinician analyzes and confirms which information was inferred to label the data.

Technically, Integrated Gradients (IG) [41] are a powerful interpretability technique for explaining the predictions of deep learning models, including the Transformers model used in clinical text classification. IG provides a way to attribute importance to the input features of a model, allowing clinicians and researchers to gain insight into how the model is making its predictions. Then, we compared this information with the information from the classifier based on the IG methods. This helped us identify misclassification sources and improve our classifiers' accuracy in detecting cardiac failure.

The results in Fig. 8 demonstrate the Transformer model's ability to calculate attribution scores to predict output based on input features. The sign of the attribution score indicates the direction of the feature's influence on the output: a positive score means that the feature positively influences the output, while a negative score indicates a negative influence. However, the model did not perform well on the task at hand. The correct labeling of the data requires clinical expertise and professional knowledge. For example, in the first original note, the absence of data on cardiac failure was compensated for by the presence of other clinical signs such as 'Souffle 3/6,' 'très faible pouls fémoral mais pas de pouls pédieux (very weak femoral pulse but no pedal pulse),' and 'Pieds tièdes (warm feet).' Similarly, in the second note, no data on cardiac failure was present, but 'sténose sous pulmonaire et CIV large (subpulmonary stenosis and wide CIV)' suggested its presence. These examples highlight the significant gap in the Transformer model's contextual learning and understanding

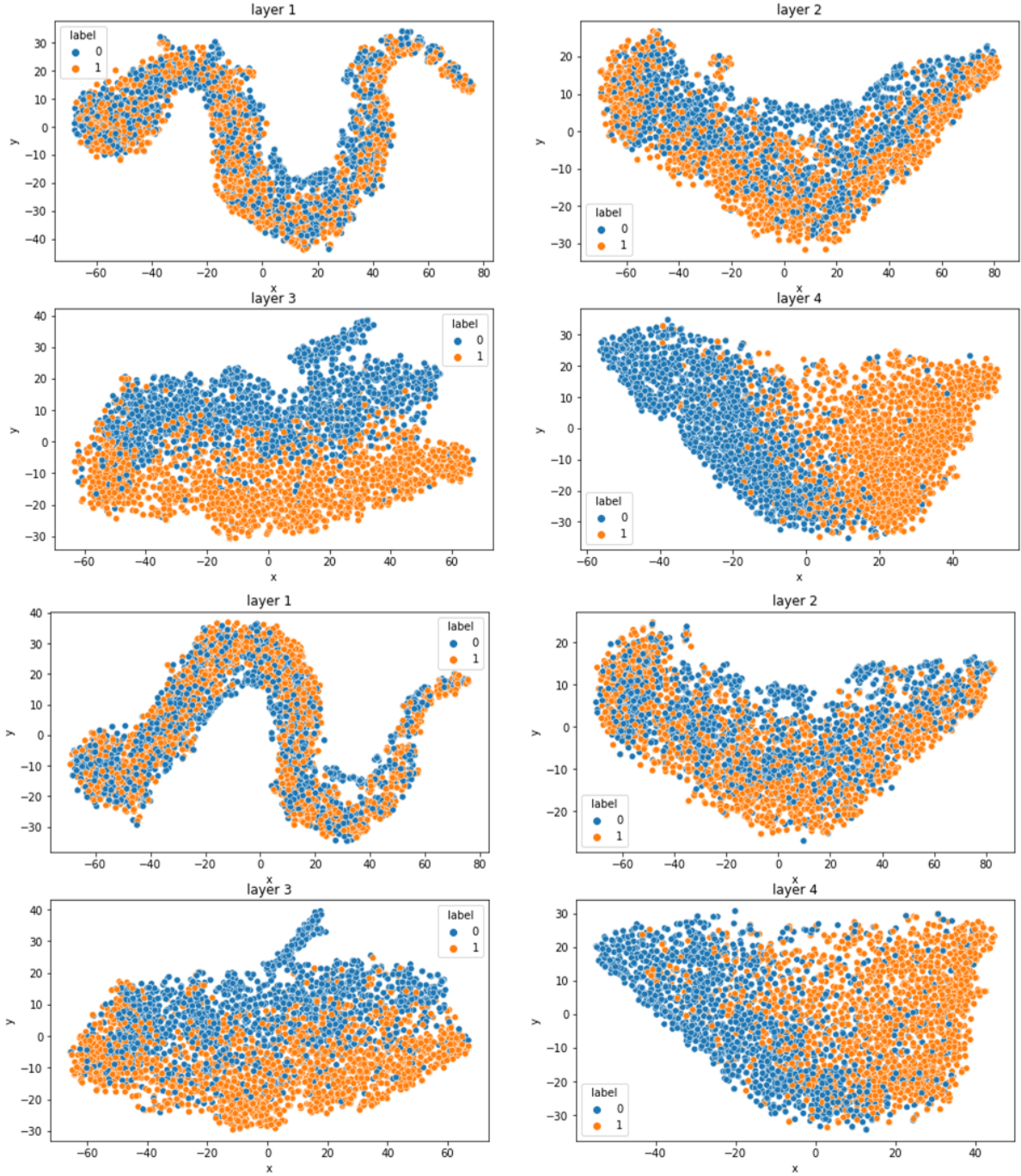


Fig. 7. Hidden embedding visualization during training (top 4 figures) and validation (bottom 4 figures) for the Switch Transformer at the 30th epoch.

of real clinical datasets. There are two possible reasons for this limitation. First, while Transformer models have shown promising performance in new tasks, it remains unclear if they can generalize across the differences in settings within the clinical domain [26]. Second, the tasks in the clinical domain often have a low signal-to-noise ratio, where the presence of a few essential keywords may suffice to determine a specific label. In contrast, Transformer's training process involves learning intricate and nuanced relations between all words in the pretraining corpus, which may not be relevant for the classification task and may shift attention away from the critical keywords [25].

## VI. CONCLUSION

We compared the performance of 6 classifiers on a binary classification task: CamemBERT, DistillBERT, FlauBERT, FrALBERT, Transformer, and Switch Transformer. The results indicated that careful hyperparameter tuning could significantly improve the performance of Transformer models over pre-trained BERT-based models. The Switch Transformer model achieved the highest performance in Accuracy, Precision, Recall, F1, and AUC, with an accuracy score of 0.87, precision of 0.87, recall of 0.85, F1 score of 0.86, and AUC of 0.92. The Transformer model achieved the second-best

**Original note 1:** "Souffle 3/6 PSG irradiant à l'apex. Pouls facilement palpables MS, possible très faible pouls fémoral mais pas de pouls pédieux, pieds tièdes mais bien colorés."

**True Label:** 1  
**Predicted Label:** 0  
**Predicted Probability:** 0.5532299876213074  
**Attribution Score:** 0.42

**Original note 2:** "Grossesse gémellaire naissance à 37+4 D-TGV avec stenose, sous pulmonaire et CIV large Rashkin + prosta en néonatalogie."

**True Label:** 1  
**Predicted Label:** 0  
**Predicted Probability:** 0.528659999370575  
**Attribution Score:** -2.07

Fig. 8. The highlighted misclassification cases from the Switch Transformer model.

performance, with an accuracy score of 0.85.

Furthermore, we presented the confusion matrices obtained from six models. The results indicated that the Switch Transformer model obtained the highest number of correct classifications and the lowest number of misclassifications, followed closely by the DistillBERT and Transformer models. FlauBERT and FrALBERT models performed similarly, with slightly higher misclassifications. Finally, the CamemBERT model obtained the lowest number of correct classifications and a relatively high number of misclassification.

The study used attribution scores to demonstrate the Transformer model's ability to predict output based on input features. However, the model did not perform very well on the clinical dataset due to its inability to contextualize and understand real-world data. The clinical tasks have a low signal-to-noise ratio, and the Transformer's training process may shift attention away from critical keywords. Additionally, it remains unclear whether Transformer models can generalize across different settings in the clinical domain. Overall, the results suggest the need for further research to improve the Transformer model's performance in clinical settings.

These findings suggest that careful selection of Transformer-based models and hyperparameter tuning can significantly improve the performance of clinical narrative classification tasks. Especially the CDSS at CHUSJ is currently under development. By combining this NLP algorithm to detect the absence of heart failure with the two other algorithms already developed on hypoxemia detection [44] and chest, X-ray analysis [45], [46], the next step of our study is to implement the resulting CDSS (integration of the three algorithms) within the cyberinfrastructure of the pediatric intensive care unit (PICU) at Sainte-Justine Hospital to diagnose ARDS early. We will then verify the ability of the CDSS to detect ARDS prospectively once the integration with the PICU e-Medical infrastructure is completed.

## VII. FUTURE WORKS

The study only considers binary classification tasks and does not examine the performance of Transformer-based models on

multiclass classification tasks. The dataset used for the study is relatively small, with almost more than 5000 instances, which may limit the generalizability of the findings to larger datasets. The study did not examine the impact of fine-tuning on the performance of the Transformer-based models. To improve the performance of this study, some potential solutions would be 1) including multiclass classification tasks to examine the performance of Transformer-based models on more complex classification tasks; 2) expanding the dataset to increase the generalizability of the findings. The impact of fine-tuning could be examined to determine if it improves the performance of the Transformer-based models. In summary, potential future directions could be explored as follows:

- 1) Model optimization: Transformer-based models can be optimized to reduce their computational requirements while maintaining accuracy, such as using distillation or pruning methods to reduce the number of parameters.
- 2) Data augmentation: Data augmentation techniques can be used to increase the amount of labeled data available for training Transformer-based models, such as using synthetic data generation methods or unsupervised learning techniques to leverage unlabeled data.
- 3) Domain-specific pre-training: pre-trained Transformer-based models on clinical text data can be employed to improve their understanding of domain-specific language and performance on clinical text classification.

## ACKNOWLEDGMENT

Clinical data were provided by the Research Center of CHU Sainte-Justine hospital, University of Montreal. The authors thank Dr. Sally Al Omar, Dr. Michael Sauthier, Dr. Rambaud Jerome and Dr. Sans Guillaume for their data support of this research. This work was supported by a scholarship from the Fonds de recherche du Quebec-Nature et technologies (FRQNT) to Thanh-Dung Le, and the grants from the Natural Sciences and Engineering Research Council (NSERC), the Institut de valorisation des données (IVADO), and the Fonds de la recherche en santé du Québec (FRQS).



## REFERENCES

- [1] A. Vaswani and et. al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [2] J. K. Tripathy and et. al., "Comprehensive analysis of embeddings and pre-training in nlp," *Computer Science Review*, vol. 42, p. 100433, 2021.
- [3] Y.-J. Huang and et. al., "Assessing schizophrenia patients through linguistic and acoustic features using deep learning techniques," *IEEE Trans. Neural Syst. Rehabilitation Eng.*, vol. 30, pp. 947–956, 2022.
- [4] L. Ilias and et. al., "Explainable identification of dementia from transcripts using transformer networks," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 4153–4164, 2022.
- [5] Y. Meng and et. al., "Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 8, pp. 3121–3129, 2021.
- [6] A. Blanco and et. al., "Exploiting icd hierarchy for classification of ehrs in spanish through multi-task transformers," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 3, pp. 1374–1383, 2021.
- [7] Y. Li and et. al., "Hi-BEHT: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records," *IEEE J. Biomed. Health Inform.*, 2022.
- [8] Z. Deng and et. al., "Rformer: Transformer-based generative adversarial network for real fundus image restoration on a new clinical benchmark," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 9, pp. 4645–4655, 2022.
- [9] R. Li and et. al., "Ddptransformer: Dual-domain with parallel transformer network for sparse view ct image reconstruction," *IEEE Trans Comput Imaging*, vol. 8, pp. 1101–1116, 2022.
- [10] A. K. Mondal and et. al., "xvitcos: explainable vision transformer based covid-19 screening using radiography," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–10, 2021.
- [11] H. Phan and et. al., "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 8, pp. 2456–2467, 2022.
- [12] P. Lu and et. al., "Improving classification of tetanus severity for patients in low-middle income countries wearing ecg sensors by using a cnn-transformer network," *IEEE Trans. Biomed. Eng.*, 2022.
- [13] J. Clauwaert and et. al., "Novel transformer networks for improved sequence labeling in genomics," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 19, no. 1, pp. 97–106, 2020.
- [14] N. Huang and et. al., "Sacall: a neural network basecaller for oxford nanopore sequencing data based on self-attention mechanism," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 19, no. 1, pp. 614–623, 2020.
- [15] G. Lopez-Garcia and et. al., "Transformers for clinical coding in spanish," *IEEE Access*, vol. 9, pp. 72 387–72 397, 2021.
- [16] G. Lopez Garcia and et. al., "Explainable clinical coding with in-domain adapted transformers," *Journal of Biomedical Informatics*, 2023.
- [17] K. Roitero and et. al., "Dilbert: Cheap embeddings for disease related medical nlp," *IEEE Access*, vol. 9, pp. 159 714–159 723, 2021.
- [18] C. Mugisha and et. al., "Comparison of neural language modeling pipelines for outcome prediction from unstructured medical text notes," *IEEE Access*, vol. 10, pp. 16 489–16 498, 2022.
- [19] M. Rizwan and et. al., "Depression classification from tweets using small deep transfer learning language models," *IEEE Access*, vol. 10, pp. 129 176–129 189, 2022.
- [20] O. N. Kjell and et. al., "Natural language analyzed with ai-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy," *Scientific Reports*, vol. 12, no. 1, p. 3918, 2022.
- [21] G. Althari and et. al., "Exploring transformer-based learning for negation detection in biomedical texts," *IEEE Access*, vol. 10, pp. 83 813–83 825, 2022.
- [22] H. K. Kim and et. al., "Identifying alcohol-related information from unstructured bilingual clinical notes with multilingual transformers," *IEEE Access*, 2023.
- [23] X. Yang and et. al., "Clinical concept extraction using transformers," *J Am Med Inform Assoc*, vol. 27, no. 12, pp. 1935–1942, 2020.
- [24] B. Zhou and et. al., "Natural language processing for smart healthcare," *IEEE Rev Biomed Eng*, 2022.
- [25] S. Gao and et. al., "Limitations of transformers on clinical text classification," *IEEE J. Biomed. Health Inform.*, 2021.
- [26] O. J. and et. al., "Clinically relevant pretraining is all you need," *J Am Med Inform Assoc*, vol. 28, no. 9, pp. 1970–1976, 2021.
- [27] I. Alimova and et. al., "Cross-domain limitations of neural models on biomedical relation classification," *IEEE Access*, vol. 10, pp. 1432–1439, 2021.
- [28] A. Gillioz and et. al., "Overview of the transformer-based models for nlp tasks," in *15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020, pp. 179–183.
- [29] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
- [30] S. Tonekaboni and et. al., "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 359–380.
- [31] S. Bhattamishra and et. al., "On the computational power of transformers and its implications in sequence modeling," in *Proceedings of the 24th Conference on Computational Natural Language Learning*, 2020, pp. 455–475.
- [32] X. Zeng and et. al., "Pretrained transformer framework on pediatric claims data for population specific tasks," *Scientific Reports*, vol. 12, no. 1, p. 3651, 2022.
- [33] Y. Gu and et. al., "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.
- [34] M. B. Zafar and et. al., "On the lack of robust interpretability of neural text classifiers," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 3730–3740.
- [35] M. AlShuweih and et. al., "Biomedical corpora and natural language processing on clinical text in languages other than english: a systematic review," *Recent Advances in Intelligent Systems and Smart Applications*, pp. 491–509, 2021.
- [36] A. Névéol and et. al., "Clinical natural language processing in languages other than english: opportunities and challenges," *Journal of biomedical semantics*, vol. 9, no. 1, pp. 1–13, 2018.
- [37] M. Raghu and et. al., "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021.
- [38] W. Fedus and et. al., "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, pp. 1–40, 2021.
- [39] F. Xue and et. al., "Go wider instead of deeper," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8779–8787.
- [40] A. Lazaridou and et. al., "Mind the gap: Assessing temporal generalization in neural language models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 348–29 363, 2021.
- [41] M. Sundararajan and et. al., "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3319–3328.
- [42] G. Bellani and et. al., "Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries," *JAMA*, vol. 315, no. 8, pp. 788–800, 2016.
- [43] P. A. L. I. C. C. Group et al., "Pediatric acute respiratory distress syndrome: consensus recommendations from the pediatric acute lung injury consensus conference," *Pediatric critical care medicine: a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies*, p. 428, 2015.
- [44] M. Sauthier and et. al., "Estimated pao2: A continuous and noninvasive method to estimate pao2 and oxygenation index," *Critical care explorations*, vol. 3, no. 10, 2021.
- [45] N. Zaglam and et. al., "Computer-aided diagnosis system for the acute respiratory distress syndrome from chest radiographs," *Computers in biology and medicine*, vol. 52, pp. 41–48, 2014.
- [46] M. Yahyatabar, P. Jouvett, and F. Chérict, "Dense-unet: a light model for lung fields segmentation in chest x-ray images," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 1242–1245.
- [47] T. D. Le and et. al., "Detecting of a patient's condition from clinical narratives using natural language representation," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 3, pp. 142–149, 2022.
- [48] T.-D. Le and et. al., "Adaptation of autoencoder for sparsity reduction from clinical notes representation learning," *IEEE Journal of Translational Engineering in Health and Medicine*, 2023.
- [49] J. Kessler, "Scattertext: a browser-based tool for visualizing how corpora differ," in *Proceedings of ACL 2017, System Demonstrations*, 2017, pp. 85–90.
- [50] J. Alammam, "The illustrated transformer," *The Illustrated Transformer–Jay Alammam–Visualizing Machine Learning One Concept at a Time*, vol. 27, 2018.

- [51] T. Lin and et. al., "A survey of transformers," *AI Open*, 2022.
- [52] C. Raffel and et. al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [53] A. Fan and et. al., "Beyond english-centric multilingual machine translation," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 4839–4886, 2021.
- [54] J. Devlin and et. al., "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
- [55] L. Martin and et. al., "Camembert: a tasty french language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7203–7219.
- [56] H. Le and et. al., "Flaubert: Unsupervised language model pre-training for french," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 2479–2490.
- [57] O. Cattan and et. al., "On the usability of transformers-based models for a french question-answering task," in *International Conference on Recent Advances in Natural Language Processing*, 2021, pp. 244–255.
- [58] V. Sanh and et. al., "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [59] A. Atrio and et. al., "Small batch sizes improve training of low-resource neural mt," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, 2021, pp. 18–24.
- [60] N. Srivastava and et. al., "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.
- [62] I. L. et. al., "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.
- [63] A. Gotmare and et. al., "A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.
- [64] M. Popel and et. al., "Training tips for the transformer model," *arXiv preprint arXiv:1804.00247*, 2018.
- [65] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," *arXiv preprint arXiv:2003.05689*, 2020.
- [66] X. Glorot and et. al., "Understanding the difficulty of training deep feed-forward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings*, 2010, pp. 249–256.
- [67] S. Ioffe and et. al., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*. PMLR, 2015, pp. 448–456.
- [68] N. Bjorck and et. al., "Understanding batch normalization," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [69] Y. Lu and et. al., "Bayes imbalance impact index: A measure of class imbalanced data set for classification problem," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3525–3539, 2019.
- [70] C. Goutte and et. al., "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *European Conference on Information Retrieval*. Springer, 2005, pp. 345–359.
- [71] M. Artetxe and et. al., "Efficient large scale language modeling with mixtures of experts," *arXiv preprint arXiv:2112.10684*, 2021.
- [72] E. Hoffer and et. al., "Train longer, generalize better: closing the generalization gap in large batch training of neural networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [73] P. Nakkiran and et. al., "Deep double descent: Where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, 2021.
- [74] N. S. Keskar and et. al., "On large-batch training for deep learning: Generalization gap and sharp minima," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.