# Improving Transformer Performance for French Clinical Notes Classification Using Mixture of Experts on a Limited Dataset

## THANH-DUNG LE [1,2], (Senior Member, IEEE), PHILIPPE JOUVET [3], AND RITA NOUMEIR [1], (Member, IEEE)

[1]Biomedical Information Processing Laboratory, École de Technologie Supérieure, University of Quebec, Quebec City, QC G1K 9H6, Canada
[2]Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg
[3]Research Center at CHU Sainte-Justine, University of Montreal, Montreal, QC H3T 1J4, Canada
CORRESPONDING AUTHOR: T.-D. LE (thanh-dung.le@uni.lu)

**ABSTRACT** Transformer-based models have shown outstanding results in natural language processing but face challenges in applications like classifying small-scale clinical texts, especially with constrained computational resources. This study presents a customized Mixture of Expert (MoE) Transformer models for classifying small-scale French clinical texts at CHU Sainte-Justine Hospital. The MoE-Transformer addresses the dual challenges of effective training with limited data and low-resource computation suitable for in-house hospital use. Despite the success of biomedical pre-trained models such as CamemBERT-bio, DrBERT, and AliBERT, their high computational demands make them impractical for many clinical settings. Our MoE-Transformer model not only outperforms DistillBERT, CamemBERT, FlauBERT, and Transformer models on the same dataset but also achieves impressive results: an accuracy of 87%, precision of 87%, recall of 85%, and F1-score of 86%. While the MoE-Transformer does not surpass the performance of biomedical pre-trained BERT models, it can be trained at least 190 times faster, offering a viable alternative for settings with limited data and computational resources. Although the MoE-Transformer addresses challenges of generalization gaps and sharp minima, demonstrating some limitations for efficient and accurate clinical text classification, this model still represents a significant advancement in the field. It is particularly valuable for classifying small French clinical narratives within the privacy and constraints of hospital-based computational resources.

**INDEX TERMS** Clinical natural language processing, cardiac failure, BERT, Transformer, Mixture of Expert.

*Clinical and Translational Impact Statement*—This study highlights the potential of customized MoE-Transformers in enhancing clinical text classification, particularly for small-scale datasets like French clinical narratives. The MoE-Transformer's ability to outperform several pre-trained BERT models marks a stride in applying NLP techniques to clinical data and integrating into a Clinical Decision Support System in a Pediatric Intensive Care Unit. The study underscores the importance of model selection and customization in achieving optimal performance for specific clinical applications, especially with limited data availability and within the constraints of hospital-based computational resources.

## 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) [6], and medical imaging [7], electroencephalogram [8], to extract 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, Transformer 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 Transformer models, researchers and clinicians can develop algorithms that automatically classify these narratives based on different criteria, such as diagnosis, treatment, or patient outcomes. 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 Transformer models for clinical narrative classification include identifying clinical coding [9], diagnosing health conditions [10], and detecting clinical events [11]. As such, Transformers-based models have become an increasingly important tool in clinical NLP and are likely to continue playing a significant role.

Despite their many benefits, Transformer 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 [12]. 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 [13]. There is a risk of bias in the data used to train these models, leading to errors or disparities in the predictions made [14]. 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 [15]. 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 [16]. While Transformer models have great potential for clinical text classification, they also require careful attention to their limitations and the potential biases that can arise.

Additionally, engaging with real clinical data presents various challenges and constraints. A primary issue is the limited availability of clinical data, compounded by its inherently confidential nature. In addition, computational resources are constrained as we rely on a shared cloud computing infrastructure hosted on the hospital server. The hospital's server is the sole environment permitted for data processing and suffers from limited computational capacity. This presents a significant bottleneck, especially when developing sophisticated machine learning algorithms that demand high computational power to provide real-time analysis and

predictive insights. Another challenge is that clinical datasets are often small and imbalanced, making it difficult to train accurate models using Transformer [17]. 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 [18]. Overall, while Transformer 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 abovementioned challenges using Transformer-based models for clinical text classification for a small French clinical note and constrained computation resources. Motivated by the recent Switch Transformer model developed by Google [19], we will adapt the Mixture of Experts (MoE) mechanism to improve the performance of the Transformer model, namely the MoE-Transformer. A MoE-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 MoE-Transformers with an MoE mechanism and Transformers with self-attention is how they model complex input-output relationships. An example of the effectiveness of MoE has been proven by [20]; 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 layer. This allows for better modeling capacity and potentially better performance.

Additionally, the study [21] 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, MoE-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 mechanism, MoE-Transformers employ multiple local attention mechanisms focusing on different input aspects. Depending on the context, the gating mechanism used in MoE-Transformers selects which expert model to use for a given input. 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 MoE-Transformer model, and trained from scratch. This would allow other researchers to understand and replicate the methodology used in the study, which is essential for advancing this work.
- Second, our study provides experimental evidence showing the limitations of Transformer-based models regarding generalization gap and sharp minima. This highlights the importance of carefully training 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) [22]. 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.

The main contribution of this work is the introduction of a carefully curated, small-scale French clinical text dataset from CHU Sainte-Justine (CHUSJ) Hospital, which can serve as a valuable benchmark for resource-constrained NLP research in healthcare. Moreover, our findings demonstrate the feasibility of deploying the proposed MoE-Transformer in a practical clinical decision support system (CDSS), even when both the available data and computational resources are minimal.

## II. MATERIALS AND METHODS
### A. FRENCH CLINICAL DATA at CHUSJ
The CDSS system in the 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 [23], emphasizing the need for more effective diagnostic tools. To diagnose ARDS, three main conditions must be detected: hypoxemia, chest X-ray infiltrates, and absence of cardiac failure [24]. The research team at CHUSJ has developed algorithms for detecting hypoxemia [25], analyzing chest X-rays [26], [27], 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 [28], [29]. Implementing these algorithms could increase ARDS diagnosis rates and improve patient outcomes.
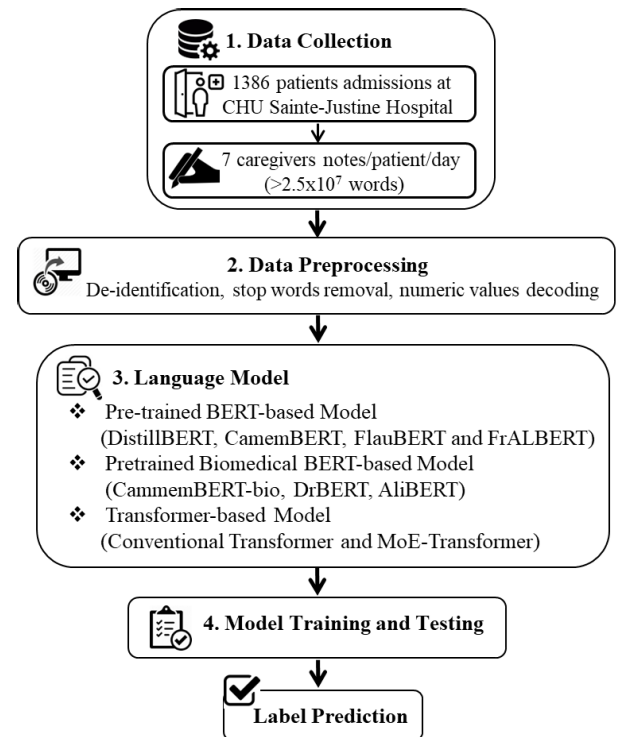


**FIGURE 1.** Workflow demonstration of the proposed methodology to classify French clinical narratives at CHUSJ hospital.

This study was conducted following ethical approval from the research ethics board at CHUSJ (protocol number: 2020-2253), 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. 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. 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 containing these values. All the notes are short narratives; detailed characteristics for the notes at CHUSJ can be found in the Supplementary Materials from the study [28].

### 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. 1. Moreover, conventional and MoE-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. 2 (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 [31]. Given a sequence of input embeddings $x_1, \ldots, x_n$, the self-attention mechanism computes a set of context-aware embeddings $h_1, \ldots, h_n$ 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:

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

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. FNNs 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.

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 quadratic computation time concerning the input sequence length, making it difficult to scale the model to very long sequences [32], and lower generalizability for a short sequence [12]. 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 great performance, it can still struggle to capture complex input-output relationships requiring more specialized models.

Motivated by study [19], the MoE structure attempts to address these limitations. The MoE mechanisms decompose the problem into smaller, simpler sub-problems, allowing the model to handle sequences and complex input-output relationships better. As mentioned earlier, 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 [12]. The MoE mechanisms allow the model to divide the sequence into smaller, more manageable segments and apply different experts to each segment [33]. Hence, it can overcome the limitations by increasing the capacity of the FFNs of conventional Transformers [34]. Consequently, this approach has improved the model's sequence task classification performance and achieved state-of-the-art results on several benchmarks [20], [21], [35].

The critical difference in the mathematical equation of the MoE-Transformer compared to the conventional Transformer is replacing the FFN with the MoE mechanism, shown in Fig. 2. In the conventional Transformer (left), the FFN consists of two linear layers with a ReLU activation function in between. On the other hand, the MoE mechanism (right) 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 MoE-Transformer can be represented by the following equation:

$$z_t = \sum_{i \in \mathcal{T}} g_i(x)E_i(x) \quad (6)$$

where $g_i(x_t)$ is a gating function that determines the importance of expert module i for input $x_t$, and $e_i(x_t)$ is the output of expert module i for input $x_t$. The top-$k$ gate values are selected for routing the token $x$. If $\mathcal{T}$ is the set of selected top-$k$ indices, then the output computation of the layer is the
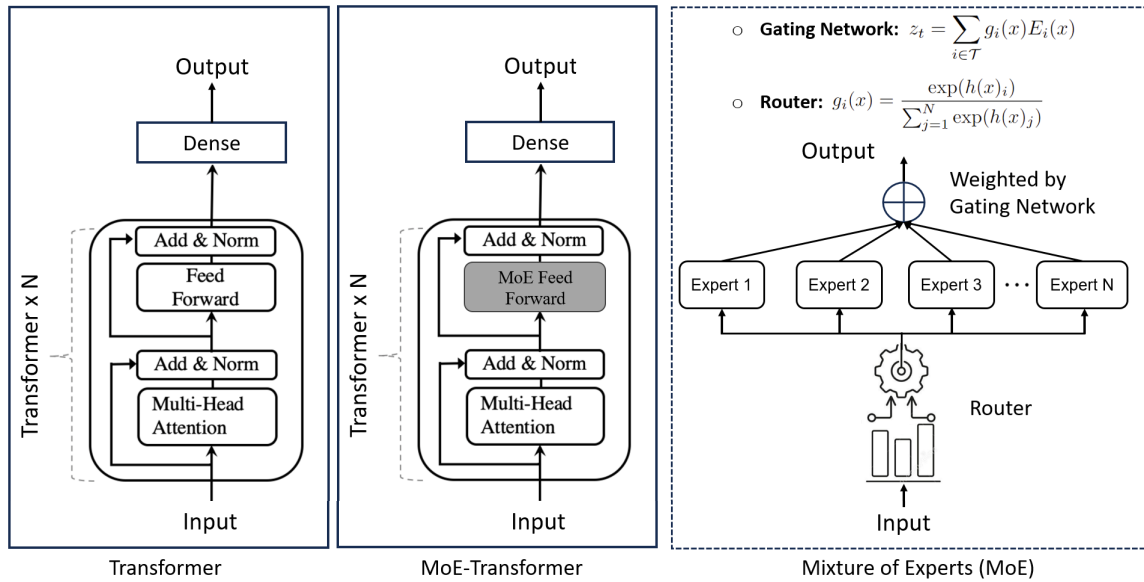
**FIGURE 2.** Illustration of a conventional transformer [30] (left), an MoE-Transformer (middle), and the detailed MoE block (right).

linearly weighted combination of each expert's computation on the token by the gate value.

Next the MoE layer will take as an input a token representation $x$ and then routes this to the best-determined top-$k$ experts, selected from a set $\{E_i(x)\}_{i=1}^{N}$ of $N$ experts. The router variable $W_r$ produces logits $h(x) = W_r \cdot x$ which are normalized via a softmax distribution over the available $N$ experts at that layer. The gate-value for expert $i$ is given by,

$$g_i(x) = \frac{\exp(h(x)_i)}{\sum_{j=1}^{N} \exp(h(x)_j)} \quad (7)$$

The gating network 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 MoE-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 MoE-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, this 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 [36] 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, including CamemBERT [37], FlauBERT [38], FrALBERT [39], and DistillBERT [40].

### 3) PRE-TRAINED BIOMEDICAL BERT-BASED MODELS for FRENCH

The biomedical pre-trained BERT-based models, including CamemBERT-bio [41], DrBERT [42], and AliBERT [43], are specifically designed for processing and understanding biomedical text. CamemBERT-bio is tailored for French biomedical data, leveraging the strengths of the CamemBERT architecture to provide robust performance in this domain. DrBERT and AliBERT also contribute to the landscape of specialized models, offering high accuracy and efficiency in various biomedical NLP tasks. These models are particularly well-suited for French clinical notes classification, as they

**TABLE 1.** Total parameters of the fine-tuned models.

| Models | Hidden Layers | Attention Heads | Total Parameters (Millions) |
|---|---|---|---|
| DistillBERT | 6 | 12 | 67.2 |
| CamemBERT | 12 | 12 | 111.4 |
| FlauBERT | 6 | 8 | 54.9 |
| FrALBERT | 12 | 12 | 12.7 |
| CamemBERT-bio | 12 | 12 | 111.4 |
| DrBERT | 12 | 12 | 111.4 |
| AliBERT | 12 | 12 | 117.6 |
| Tranformer | 4 | 4 | 2.9 |
| MoE-Transformer | 4 | 4 | 5.7 |

have been trained on extensive French biomedical corpora, ensuring they capture the nuances and specific terminology used in French medical practice.

## III. EXPERIMENTAL IMPLEMENTATION

Table 1 shows the total parameters of different Transformer-based models used in this study, including CamemBERT, DistillBERT, FlauBERT, FrALBERT, Transformer, and MoE-Transformer. The parameters compared include hidden layers, attention heads and total parameters. Regarding total parameters, CamemBERT, CamemBERT-bio, DrBERT and AliBERT have the highest number of parameters, with more than 110 million. While Transformer, and MoE-Transformer have significantly fewer parameters, with 2.3 million, and 5.7 million, respectively. The variation in parameters 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.

Defining the hyperparameters during the training process of Transformers is a critical step in achieving good performance. 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 [44].
- **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 that drop out during each iteration [45].
- **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) [46], AdamW (Adam with weight decay) [47] were used.

- **Learning rate**: Consine 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 [48].
- **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. In implementing the MoE, we followed the guidelines from [49] and [50], and an example pseudocode snippet in Python for implementing the MoE layer shown in Fig. 3.

Choosing appropriate values for these hyperparameters requires careful experimentation and tuning to achieve the best possible results. Only a few parameters require careful tuning. As reported in [51], the model size, learning rate, batch size, and maximum sequence length are the critical hyperparameters 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. The combination with the highest estimated performance was considered the optimal solution, and this approach balances computational efficiency and models' accuracy.

Finally, table 2 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 0.5 and a cosine decay learning rate. The Transformer and MoE-Transformer models use an AdamW optimizer with a dropout rate of 0.35 and a cosine decay learning rate. The Adam$_\epsilon$ is only specified for the Transformer and MoE-Transformer models and is set to

```python
class Router(layers.Layer):
    def __init__(self, num_experts, expert_capacity):
        self.num_experts = num_experts
        self.route = layers.Dense(units=num_experts)
        self.expert_capacity = expert_capacity
        super().__init__()


class MoE(layers.Layer):
    def __init__(self, num_experts, embed_dim, num_tokens_per_batch, capacity_factor=1):
        self.num_experts = num_experts
        self.embed_dim = embed_dim
        self.experts = [
            create_feedforward_network(embed_dim) for _ in range(num_experts)
        ]

        self.expert_capacity = num_tokens_per_batch // self.num_experts
        self.router = Router(self.num_experts, self.expert_capacity)
        super().__init__()
```

**FIGURE 3.** Pseudocode snippet in Python for the implementation of MoE layer.

**TABLE 2.** Hyperparameters of the fine-tuned models.

| Hyperparameters | Pretrained BERT-based | Transformer | MoE-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_$\epsilon$ | N/A | 5*1e-06 | 5*1e-06 |
| Maximum sequence length | 256 | 256 | 256 |

```
+-------------------------------------------------------------------------------+
| NVIDIA-SMI 528.02        Driver Version: 528.02        CUDA Version: 12.0      |
|-------------------------------+----------------------+------------------------|
| GPU   Name           TCC/WDDM | Bus-Id        Disp.A | Volatile Uncorr. ECC   |
| Fan   Temp  Perf  Pwr:Usage/Cap|             Memory-Usage | GPU-Util  Compute M. |
|                               |                      |                 MIG M. |
|===============================+======================+========================|
|   0   Quadro P620        WDDM | 00000000:01:00.0  On |                   N/A  |
| 50%   62C    P0    N/A /  40W |   1950MiB /  2048MiB |     88%      Default   |
|                               |                      |                   N/A  |
+-------------------------------+----------------------+------------------------+
```

**FIGURE 4.** GPU utilization and resource usage during model training on NVIDIA Quadro P620.

5*1e-06. The maximum sequence length for all models is set to 256. Additionally, the GlorotNormal initializer [52], batch normalization [53] are employed for models' stability. 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 [54]. These metrics are defined as follows.

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

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall/Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}}$$

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

First, all models were trained using an NVIDIA Quadro P620 GPU. As shown in Fig. 4, this GPU has a total memory capacity of 2048 MB, of which 1950 MB was actively utilized during training. The GPU utilization rate reached 88%, indicating a high level of resource engagement. The Quadro P620, while robust, operates within relatively modest
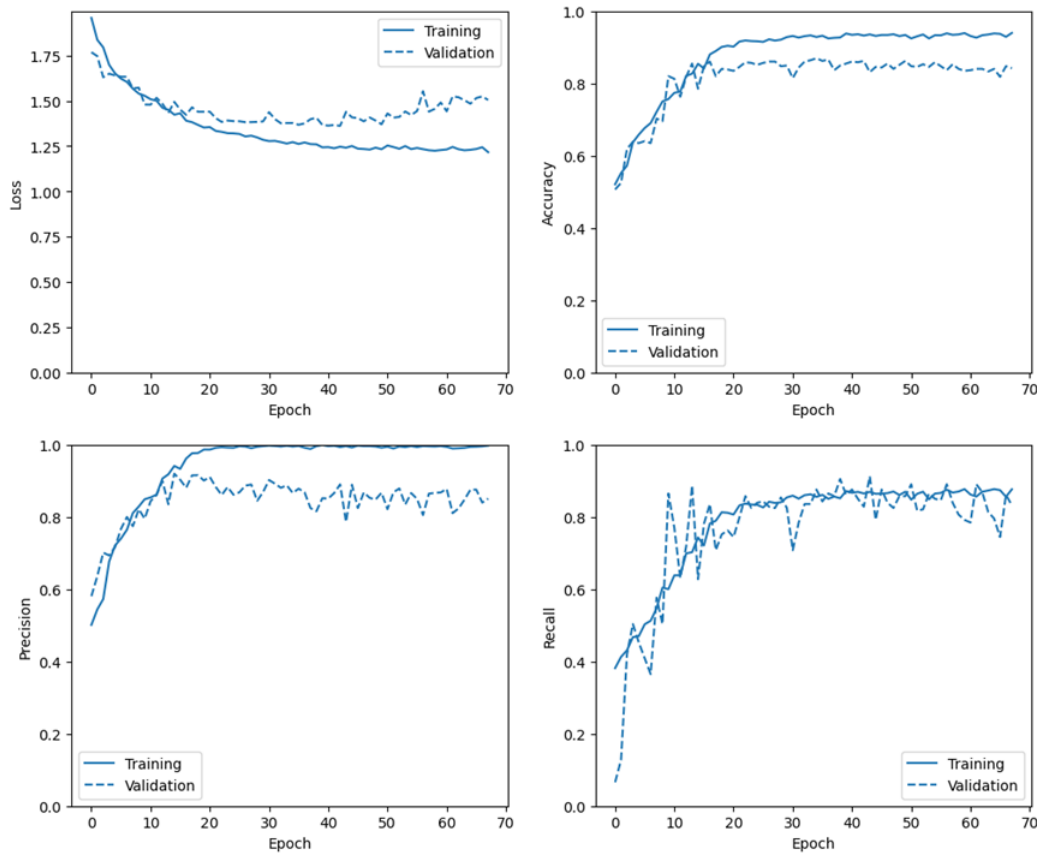
**FIGURE 5.** Training and validation performance results from MoE-Transformer model.

resource constraints compared to more powerful GPUs typically used for deep learning tasks. Training pre-trained models under such limited computational resources poses significant challenges. Efficient resource management and optimization techniques are crucial to achieve competitive performance without overloading the system.

During training and validation shown in Fig. 5, the MoE-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 fluctuated quite a bit. The model's overall performance was good, with high accuracy, 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.

Based on the performance comparison presented in Table 3, the MoE-Transformer demonstrates a significant advantage in computational efficiency while maintaining competitive performance metrics compared to pre-trained BERT-based and biomedical BERT-based models. Specifically, in Fig. 6, the MoE-Transformer achieves high scores
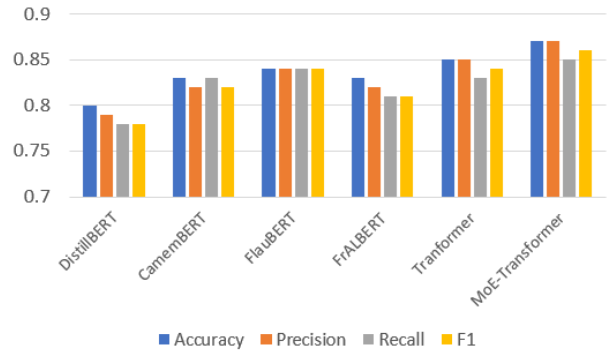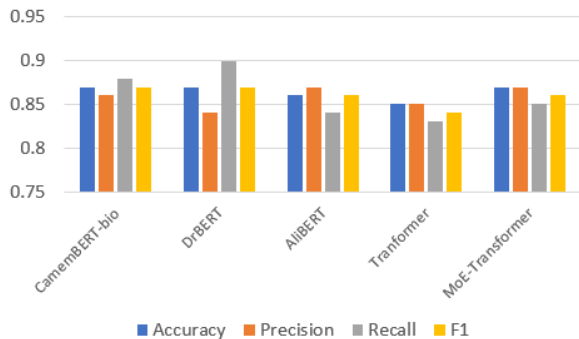


**FIGURE 6.** Performance comparison between MoE-Transformer vs. pre-trained BERT-based models, including DistillBERT, CamemBERT, FlauBERT, and FrALBERT.

across all performance metrics: accuracy (0.87), precision (0.87), recall (0.85), and F1 score (0.86). These results are comparable to or better than those of pre-trained BERT-based models such as DistillBERT, CamemBERT, FlauBERT, and FrALBERT. Notably, the MoE-Transformer outperforms CamemBERT and FrALBERT in all metrics. Fig. 7 demonstrates that the MoE-Transformer performs competitively with the biomedical pre-trained models. Specifically, MoE-Transformer achieves an accuracy of 0.87, which is on par

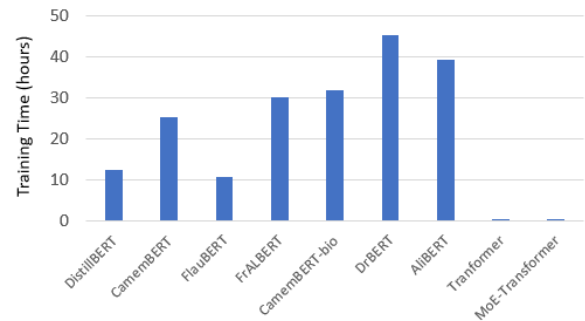**TABLE 3.** A comparison performance of different classifiers.

| Models | Accuracy | Precision | Recall | F1 | Training Time (hours) | Inference Time (s) |
|---|---|---|---|---|---|---|
| DistillBERT | 0.80 | 0.79 | 0.78 | 0.78 | 12.4 | 62 |
| CamemBERT | 0.83 | 0.82 | 0.83 | 0.82 | 25.1 | 132 |
| FlauBERT | 0.84 | 0.84 | 0.84 | 0.84 | 10.6 | 39 |
| FrALBERT | 0.83 | 0.82 | 0.81 | 0.81 | 30.2 | 131 |
| CamemBERT-bio | **0.87** | 0.86 | 0.88 | **0.87** | 31.7 | 142 |
| DrBERT | **0.87** | 0.84 | **0.90** | **0.87** | 45.2 | 133 |
| AliBERT | 0.86 | **0.87** | 0.84 | 0.86 | 39.3 | 128 |
| Transformer | 0.85 | 0.85 | 0.83 | 0.84 | 0.11 | 3 |
| MoETransformer | **0.87** | **0.87** | 0.85 | 0.86 | 0.17 | 4 |



**FIGURE 7.** Performance comparison between MoE-Transformer vs. pre-trained biomedical BERT-based models, including CamemBERT-bio, Dr.BERT, and AliBERT.



**FIGURE 8.** Training times comparison for fine-tuning the pre-trained biomedical BERT-based models vs training MoE-Transformer model.

**TABLE 4.** Confusion matrix comparison for all classifiers.

| Models | TN ↑ | TP ↑ | FP ↓ | FN ↓ |
|---|---|---|---|---|
| DistillBERT | 233 | 201 | 54 | 56 |
| CamemBERT | 239 | 214 | 48 | 43 |
| FlauBERT | 246 | 215 | 41 | 42 |
| FrALBERT | 241 | 209 | 46 | 48 |
| CamemBERT-bio | 228 | 243 | 40 | 33 |
| DrBERT | 238 | 235 | 46 | 25 |
| AliBERT | 235 | 231 | 44 | 44 |
| Transformer | 250 | 213 | 37 | 44 |
| MoE-Transformer | **253** | 219 | **34** | 38 |

with CamemBERT-bio and AliBERT and only slightly lower than DrBERT. Its precision and F1 score are both 0.87, matching the top-performing models. Although its recall is slightly lower than DrBERT, it still achieves a comparable score of 0.85.

Fig. 8 highlights the significant computational efficiency of the MoE-Transformer. The training time for the MoE-Transformer is a mere 0.17 hours, which is drastically lower than any of the pre-trained biomedical models. In contrast, DrBERT, which achieves the highest recall, requires 45.2 hours of training time (266 times longer), while CamemBERT-bio and AliBERT need 31.7 (186 times longer) and 39.3 hours (231 times longer), respectively. This apparent difference underscores the efficiency of the MoE-Transformer in terms of computational resources. It makes the MoE-Transformer highly suitable for environments with limited computational resources like hospitals. In summary, the MoE-Transformer offers a compelling balance of performance and efficiency. It achieves results on par with or better than several pre-trained and biomedical-specific models while requiring a fraction of the training time. This makes it an excellent choice for clinical applications where computational resources are constrained.

Table 4 compares the confusion matrices obtained from 9 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. The DrBERT model achieved the highest number of correct classifications with TP+TN = 473 and the lowest number of misclassifications with FN+FP = 71. Meanwhile, the MoE-Transformer model obtained the highest number of TN (253) and the smallest number of FP (34), making it the second-best model overall. The MoE-Transformer achieved a high number of correct classifications with TP+TN = 472 and a low number of misclassifications with FN+FP = 72. This suggests that the simpler MoE-Transformer model, regarding the number of parameters, may perform comparably or even better than larger pre-trained models for a limited clinical narrative dataset.

Although the MoE-Transformer outperforms several other models and the conventional Transformer model, its performance falls short when compared to two of our previous studies [28], [29] that extensively analyzed a conceptual framework for detecting a patient's health condition from contextual input to output. On the same dataset, the proposed framework in those studies utilized a combination of TF-IDF (term frequency-inverse document frequency) and MLP-NN
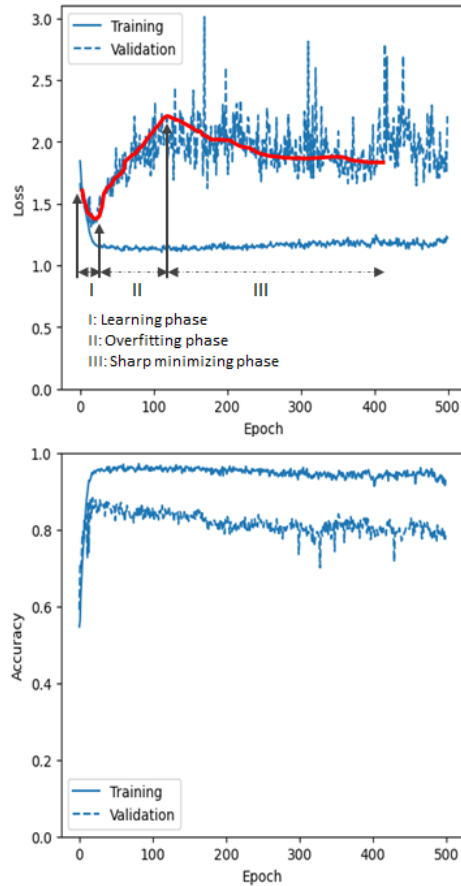
**FIGURE 9.** Generalization gap and sharp minima during training MoE-Transformer without early stopping.

(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 MoE-Transformer model.

While the MoE-Transformer model has demonstrated promising results in clinical text classification, there is still room for further improvement in its performance. One possible area of investigation is the training methodology, as suggested by previous research [55], [56]. 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. 9. Initially, the model underwent the learning phase, where the loss gradually decreased and 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 a 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 [57].

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 MoE-Transformer embedding for each layer (from 1 to 4) in Figure 10. 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 MoE-Transformer model by visualizing the results from the misclassification. There are 72 cases of misclassification from the results of the MoE-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 better understand the reasons behind these misclassifications. The clinician analyzes and confirms which information was inferred to label the data.

Technically, Integrated Gradients (IG) [22] 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
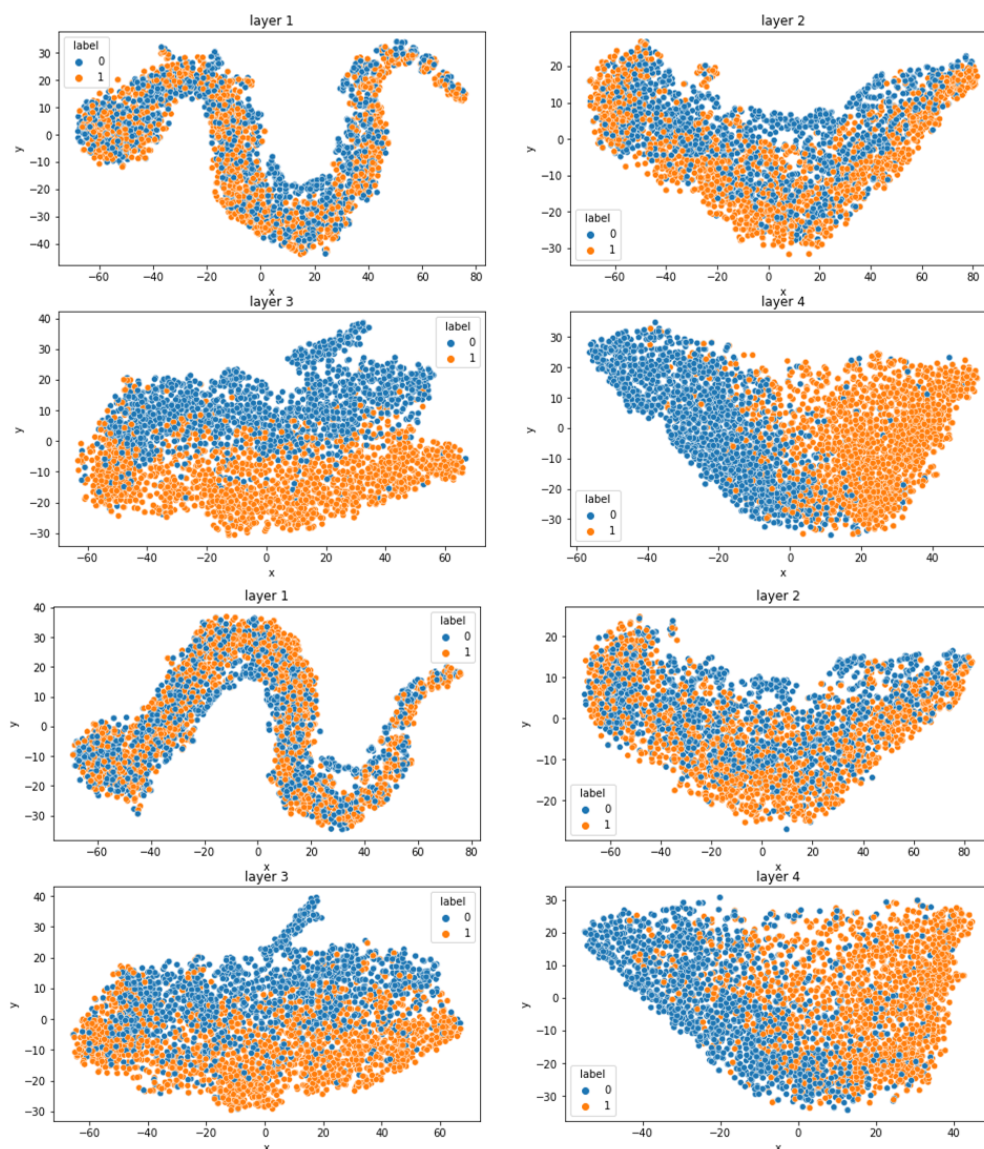
**FIGURE 10.** Hidden embedding visualization during training (top 4 figures) and validation (bottom 4 figures) for the MoE-Transformer at the 30th epoch.

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. 11 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 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 [13]. 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

```
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 néonatologie."
True Label: 1
Predicted Label: 0
Predicted Probability: 0.528659999370575
Attribution Score: -2.07
```

**FIGURE 11.** The highlighted misclassification cases from the MoE-Transformer model.

for the classification task and may shift attention away from the critical keywords [12].

## VI. LIMITATIONS

While large language models (LLMs) have undoubtedly advanced numerous NLP tasks [58], their deployment in specialized clinical environments remains impractical. In our setting, stringent data privacy regulations prohibit transferring sensitive patient information to external servers or APIs, precluding fine-tuning and secure deployment of large off-the-shelf LLMs. Additionally, these models demand extensive computational resources—typically high-memory GPUs or multi-GPU clusters—far exceeding the capacity of our NVIDIA Quadro P620 (2 GB GPU memory). Furthermore, because LLMs are generally trained on broad-domain corpora, they often struggle to generalize to pediatric clinical narratives without significant retraining, amplifying data, and computational requirements. Given that our dataset comprises only 5,444 short clinical narratives, any attempt to fine-tune such large models would likely lead to overfitting and suboptimal performance. For example, a recent study [59] confirms that employing Low-Rank Adaptation during fine-tuning has proven ineffective in overcoming these limitations. Hence, a more focused and computationally efficient model, such as our MoE-Transformer, is a more viable alternative under these constraints.

Similarly, although state-of-the-art LLMs like Mistral [60] and LLaMA 3.21b [61] have demonstrated impressive capabilities in general NLP tasks, several fundamental limitations hinder their integration into our CDSS at CHUSJ. First, even the smallest LLMs typically contain no fewer than 5 billion parameters [62], making them more than 40 times larger than our domain-specific models, specifically trained on French biomedical clinical texts. This vast difference in scale implies that, even in zero-shot settings, general-purpose LLMs are prone to hallucinations and unreliable predictions when applied to pediatric clinical narratives [63]. Second, while these models are trained on massive datasets ranging from 825 billion to 6.3 trillion tokens, our dataset

comprises only 580,000 tokens, further questioning their effectiveness without extensive domain-specific fine-tuning. Third, the computational demands of even the smallest LLMs (e.g., requiring a minimum of 1024-core NVIDIA Ampere architecture GPU with 32 tensor cores, 16G DRAM) are incompatible with our hardware capabilities. Finally, strict hospital privacy regulations preclude cloud-based inference, ruling out the possibility of relying on externally hosted models or resource-intensive fine-tuning methods. Despite their success in broader NLP applications, LLMs are not a feasible solution for our specialized clinical setting, thereby underscoring the advantages of our MoE-Transformer for local, real-time clinical text classification.

## VII. CONCLUSION

We compared the performance of 9 classifiers on a binary classification task. The results indicated that MoE can improve the performance of Transformer models over pre-trained BERT-based models. The MoE-Transformer model performed comparable to biomedical pre-trained models but with 100 times less computation resources. These results confirm that the MoE-Transformer is particularly valuable for classifying small French clinical narratives within the privacy and constraints of hospital-based computational resources.

The study used attribution scores to demonstrate the MoE-Transformer model's ability to predict output based on input features. However, the model did not perform 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 MoE-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 MoE-Transformer model's performance in clinical settings.

These findings suggest that carefully training the Transformer-based models from scratch can significantly improve the performance of clinical narrative classification tasks. The CDSS at CHUSJ is especially 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 [25] and chest, X-ray analysis [26], [27], 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 verify the CDSS's ability to detect ARDS prospectively once the integration with the PICU e-medical infrastructure is completed.

Future work will further explore enhancements to the evaluation methodology to gain deeper insights into the model's performance and fairness in clinical text classification. Specifically, plans include incorporating class-wise performance metrics [64] to reveal detailed strengths and weaknesses across different clinical categories, thereby identifying potential biases and underperforming segments.

Complementing this, a qualitative evaluation framework employing text perturbation experiments—such as synonym replacements, misspellings, and grammatical variations—will be developed to assess the robustness and generalization capabilities of the MoE-Transformer compared to large language model baselines. Additionally, drawing on insights from recent work on extracting social determinants of health from clinical texts [65], comparative analyses will be conducted to benchmark this approach against state-of-the-art methodologies. Recognizing the critical importance of fairness, future studies will also systematically examine potential biases in data curation, compilation, and model development by leveraging established frameworks like the AI Risk Repository, as highlighted in [66]. These explorations are expected to validate the approach's practicality and robustness while contributing to developing fair, responsible, and practical AI applications in healthcare.

## ACKNOWLEDGMENT

## REFERENCES

[1] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[2] J. K. Tripathy et al., "Comprehensive analysis of embeddings and pre-training in NLP," *Comput. Sci. Rev.*, vol. 42, Nov. 2021, Art. no. 100433.

[3] Y.-J. Huang et al., "Assessing schizophrenia patients through linguistic and acoustic features using deep learning techniques," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 947–956, 2022.

[4] L. Ilias and D. Askounis, "Explainable identification of dementia from transcripts using transformer networks," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 4153–4164, Aug. 2022.

[5] Y. Meng, W. Speier, M. K. Ong, and C. W. Arnold, "Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 8, pp. 3121–3129, Aug. 2021.

[6] A. Blanco, A. Pérez, and A. Casillas, "Exploiting ICD hierarchy for classification of EHRs in Spanish through multi-task transformers," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 1374–1383, Mar. 2022.

[7] Z. Deng et al., "RFormer: Transformer-based generative adversarial network for real fundus image restoration on a new clinical benchmark," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 9, pp. 4645–4655, Sep. 2022.

[8] H. Phan, K. Mikkelsen, O. Y. Chen, P. Koch, A. Mertins, and M. De Vos, "SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 8, pp. 2456–2467, Aug. 2022.

[9] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, and F. J. Veredas, "Transformers for clinical coding in Spanish," *IEEE Access*, vol. 9, pp. 72387–72397, 2021.

[10] M. Rizwan, M. F. Mushtaq, U. Akram, A. Mehmood, I. Ashraf, and B. Sahelices, "Depression classification from tweets using small deep transfer learning language models," *IEEE Access*, vol. 10, pp. 129176–129189, 2022.

[11] H. K. Kim et al., "Identifying alcohol-related information from unstructured bilingual clinical notes with multilingual transformers," *IEEE Access*, vol. 11, pp. 16066–16075, 2023.

[12] S. Gao et al., "Limitations of transformers on clinical text classification," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 9, pp. 3596–3607, Sep. 2021.

[13] O. J. Bear Don't Walk IV, T. Sun, A. Perotte, and N. Elhadad, "Clinically relevant pretraining is all you need," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 9, pp. 1970–1976, Aug. 2021.

[14] I. Alimova, E. Tutubalina, and S. I. Nikolenko, "Cross-domain limitations of neural models on biomedical relation classification," *IEEE Access*, vol. 10, pp. 1432–1439, 2022.

[15] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the transformer-based models for NLP tasks," in *Proc. 15th Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, Sep. 2020, pp. 179–183.

[16] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[17] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than English: Opportunities and challenges," *J. Biomed. Semantics*, vol. 9, no. 1, pp. 1–13, Dec. 2018.

[18] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Aug. 2021, pp. 12116–12128.

[19] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res*, vol. 23, pp. 1–40, Jan. 2021.

[20] F. Xue, Z. Shi, F. Wei, Y. Lou, Y. Liu, and Y. You, "Go wider instead of deeper," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, Jun. 2022, pp. 8779–8787.

[21] A. Lazaridou et al., "Mind the gap: Assessing temporal generalization in neural language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Jan. 2021, pp. 29348–29363.

[22] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.

[23] G. Bellani 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.

[24] Pediatric Acute Lung Injury Consensus Conference Group, "Pediatric acute respiratory distress syndrome: Consensus recommendations from the pediatric acute lung injury consensus conference," *Pediatric Crit. Care Med.*, vol. 5, pp. 428–439, Jun. 2015.

[25] M. Sauthier et al., "Estimated Pao$_2$: A continuous and noninvasive method to estimate Pao$_2$ and oxygenation index," *Crit. Care Explor.*, vol. 3, no. 10, 2021, Art. no. e0546.

[26] N. Zaglam, P. Jouvet, O. Flechelles, G. Emeriaud, and F. Cheriet, "Computer-aided diagnosis system for the acute respiratory distress syndrome from chest radiographs," *Comput. Biol. Med.*, vol. 52, pp. 41–48, Sep. 2014.

[27] M. Yahyatabar, P. Jouvet, and F. Cheriet, "Dense-Unet: A light model for lung fields segmentation in chest X-ray images," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1242–1245.

[28] T.-D. Le, R. Noumeir, J. Rambaud, G. Sans, and P. Jouvet, "Detecting of a patient's condition from clinical narratives using natural language representation," *IEEE Open J. Eng. Med. Biol.*, vol. 3, pp. 142–149, 2022.

[29] T.-D. Le, R. Noumeir, J. Rambaud, G. Sans, and P. Jouvet, "Adaptation of autoencoder for sparsity reduction from clinical notes representation learning," *IEEE J. Translational Eng. Health Med.*, vol. 11, pp. 469–478, 2023.

[30] J. Alammar, "The illustrated transformer," in *The Illustrated Transformer–Jay Alammar–Visualizing Machine Learning One Concept At a Time*, vol. 27, 2018, pp. 1–2. [Online]. Available: https://jalammar.github.io/illustrated-transformer/

[31] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, Jan. 2022.

[32] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[33] N. Dikkala, N. Ghosh, R. Meka, R. Panigrahy, N. Vyas, and X. Wang, "On the benefits of learning to route in mixture-of-experts models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Dec. 2023, pp. 9376–9396.

[34] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li, "Towards understanding mixture of experts in deep learning," 2022, *arXiv:2208.02813*.

[35] A. Fan et al., "Beyond English-centric multilingual machine translation," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 4839–4886, 2021.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[37] L. Martin et al., "CamemBERT: A tasty French language model," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7203–7219.

[38] H. Le et al., "FlauBERT: Unsupervised language model pre-training for French," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 2479–2490.

[39] O. Cattan, C. Servan, and S. Rosset, "On the usability of transformers-based models for a French question-answering task," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2021, pp. 244–255.

[40] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[41] R. Touchent et al., "Camembert-bio: A tasty French language model better for your health," 2023, *arXiv:2306.15550*.

[42] Y. Labrak et al., "DrBERT: A robust pre-trained model in French for biomedical and clinical domains," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, Toronto, ON, Canada, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds., Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 16207–16221, doi: 10.18653/v1/2023.acl-long.896.

[43] A. Berhe, G. Draznieks, V. Martenot, V. Masdeu, L. Davy, and J.-D. Zucker, "AliBERT: A pre-trained language model for French biomedical text," in *Proc. 22nd Workshop Biomed. Natural Lang. Process. BioNLP Shared Tasks*, Toronto, ON, Canada, D. Demner-Fushman, S. Ananiadou, and K. Cohen, Eds., 2023, pp. 223–236, doi: 10.18653/v1/2023.bionlp-1.19.

[44] À. R. Atrio and A. Popescu-Belis, "Small batch sizes improve training of low-resource neural MT," in *Proc. 18th Int. Conf. Natural Lang. Process. (ICON)*, Jan. 2022, pp. 18–24.

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[46] O. Konur, D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[47] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, 2019.

[48] A. Gotmare, N. S. Keskar, C. Xiong, and R. Socher, "A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation," in *Proc. 7th Int. Conf. Learn. Represent.*, New Orleans, LA, USA, Jan. 2018.

[49] D. Fan, B. Messmer, and M. Jaggi, "Towards an empirical understanding of MoE design choices," 2024, *arXiv:2402.13089*.

[50] J. Zhao, P. Wang, and Z. Wang, "Generalization error analysis for sparse mixture-of-experts: A preliminary study," 2024, *arXiv:2403.17404*.

[51] M. Popel and O. Bojar, "Training tips for the transformer model," 2018, *arXiv:1804.00247*.

[52] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[53] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[54] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proc. Eur. Conf. Inf. Retr.* Berlin, Germany: Springer, Mar. 2005, pp. 345–359.

[55] E. Hoffer, I. Hubara, and D. Soudry, "Train longer, generalize better: Closing the generalization gap in large batch training of neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2017.

[56] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *J. Stat. Mech., Theory Exp.*, vol. 2021, no. 12, Dec. 2021, Art. no. 124003.

[57] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *Proc. 5th Int. Conf. Learn. Represent.*, Toulon, France, Sep. 2017.

[58] W. Khan, S. Leem, K. B. See, J. K. Wong, S. Zhang, and R. Fang, "A comprehensive survey of foundation models in medicine," *IEEE Rev. Biomed. Eng.*, early access, Jan. 20, 2025, doi: 10.1109/RBME.2025.3531360.

[59] T.-D. Le, T. T. Nguyen, V. N. Ha, S. Chatzinotas, P. Jouvet, and R. Noumeir, "The impact of LoRA adapters for LLMs on clinical NLP classification under data limitations," 2024, *arXiv:2407.19299*.

[60] A. Q. Jiang et al., "Mistral 7B," 2023, *arXiv:2310.06825*.

[61] A. Grattafiori et al., "The llama 3 herd of models," 2024, *arXiv:2407.21783*.

[62] Z. Lu et al., "Small language models: Survey, measurements, and insights," 2024, *arXiv:2409.15790*.

[63] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou, "Large language models in medicine: The potentials and pitfalls: A narrative review," *Ann. Internal Med.*, vol. 177, no. 2, pp. 210–220, Feb. 2024.

[64] P. Benz, C. Zhang, A. Karjauv, and I. S. Kweon, "Robustness may be at odds with fairness: An empirical study on class-wise accuracy," in *Proc. NeurIPS Workshop Pre-Registration Mach. Learn.*, Jan. 2020, pp. 325–342.

[65] S. Raza, E. Dolatabadi, N. Ondrusek, L. Rosella, and B. Schwartz, "Discovering social determinants of health from case reports using natural language processing: Algorithmic development and validation," *BMC Digit. Health*, vol. 1, no. 1, p. 35, Sep. 2023.

[66] P. Slattery et al., "The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence," 2024, *arXiv:2408.12622*.

• • •