Detecting of a Patient's Condition From Clinical Narratives Using Natural Language Representation

Thanh-Dung Le, *Member, IEEE*, Rita Noumeir, *Member, IEEE*, Jérôme Rambaud, Guillaume Sans, and Philippe Jouvet

***Abstract*—The rapid progress in clinical data management systems and artificial intelligence approaches enable the era of personalized medicine. Intensive care units (ICUs) are ideal clinical research environments for such development because they collect many clinical data and are highly computerized. *Goal:* We designed a retrospective clinical study on a prospective ICU database using clinical natural language to help in the early diagnosis of heart failure in critically ill children. *Methods:* The methodology consisted of empirical experiments of a learning algorithm to learn the hidden interpretation and presentation of the French clinical note data. This study included 1386 patients' clinical notes with 5444 single lines of notes. There were 1941 positive cases (36% of total) and 3503 negative cases classified by two independent physicians using a standardized approach. *Results:* The multilayer perceptron neural network outperforms other discriminative and generative classifiers. Consequently, the proposed framework yields an overall classification performance with 89% accuracy, 88% recall, and 89% precision. *Conclusions:* This study successfully applied learning representation and machine learning algorithms to detect heart failure in a single French institution from clinical natural language. Further work is needed to use the same methodology in other languages and institutions.**

***Index Terms*—Clinical natural language processing, cardiac failure, machine learning, imbalance learning, feature selection.**

***Impact Statement*—The study is a showcase to confirm that, dealing with a small dataset of clinical notes, a multilayer perceptron neural network classifier is a better approach compared to conventional classifiers, especially, pretrained-based deep learning models. Additionally, instead of losing information from numeric values, they can be retained and encoded for the representation learning. Consequently, it achieves better results for the classification task.**

I. Introduction

CURRENTLY, clinical narratives are continuously provided and stored in electronic medical records (EMR), but they are underutilized in clinical decision support systems. The limitation comes from their unstructured or semi-structured format. Besides, another problem with clinical narratives is that they are written in incomplete sentences but in an information-dense way for communication between clinicians [1]. Because of the two reasons, clinical narrative sources impose constraints in an actual application for clinical outcome prediction.

Since 2013, the Pediatric Critical Care Unit at CHU Sainte-Justine (CHUSJ) has used an EMR. The patients' information, including vital signs, laboratory results, and ventilator parameters are updated every 5 minutes to 1 h [2]. Primarily, a significant data source of French clinical notes is currently stored. There are seven caregiver notes/patient/day from 1386 patients (containing a dataset of more than 2.5×10^7 words). These notes are scribed extensively from admission notes and evaluation notes. Admission notes outline reasons for admission to intensive care units, historical progress of the disease, medication, surgery, and the patient's baseline status. Daily ailments and test results are described in evaluation notes, from which patient condition is evaluated and diagnosed later by doctors. However, these information sources are being used as documentation for reporting and billing instead of clinical knowledge for predicting conditions or decision support.

A. Problem Statement

The diagnosis of acute respiratory distress syndrome (ARDS) is frequently delayed or even not diagnosed in intensive care

Manuscript received 27 April 2022; revised 27 June 2022, 29 July 2022, and 22 September 2022; accepted 22 September 2022. Date of publication 26 September 2022; date of current version 13 December 2022. This work was supported in part by the Natural Sciences and Engineering Research Council, in part by the Institut de Valorisation des données de l'Université de Montréal, in part by the Fonds de la recherche en sante du Quebec, and in part by the Fonds de recherche du Québec – Nature et technologies. The review of this article was arranged by Associate Editor Toshiyo Tamura. *(Corresponding author: Thanh-Dung Le.)*

Thanh-Dung Le is with the Biomedical Information Processing Lab, École de Technologie Supérieure, University of Québec, Montreal, QB H3G 1M8, Canada, and also with the Research Center at CHU Sainte-Justine Hospital, University of Montreal, Montreal, QB H3T 1J4, Canada (e-mail: thanh-dung.le.1@etsmtl.net).

Rita Noumeir is with the Biomedical Information Processing Lab, École de Technologie Supérieure, University of Québec, Montreal, QB H3G 1M8, Canada (e-mail: rita.noumeir@etsmtl.ca).

Jérôme Rambaud, Guillaume Sans, and Philippe Jouvet are with the Research Center at CHU Sainte-Justine Hospital, University of Montreal, Montreal, QB H3T 1J4, Canada (e-mail: jerome.rambaud-ext@aphp.fr; guillaumesans@orange.fr; philippe.jouvet@umontreal.ca).

This article has supplementary downloadable material available at https://doi.org/10.1109/OJEMB.2022.3209900, provided by the authors.

Digital Object Identifier 10.1109/OJEMB.2022.3209900

units. In the largest international cohort of patients with ARDS, the diagnosis of ARDS was delayed or missed in two-thirds of patients, with the diagnosis missed entirely in 40% of patients [3]. To make the diagnosis of ARDS, three main conditions need to be detected: hypoxemia (low blood oxygenation), presence of infiltrates on chest X Ray and absence of cardiac failure [4]. The development of a clinical decision support system (CDSS) in real time that automatically screen the EMR data, chest X Rays and other data sources (medical devices collecting vital signs, ventilator settings) has the potential to increase diagnosis rate and then improve the management of this syndrome [4]. Our research team has developed the first two algorithms for hypoxemia [5] and chest X Ray analysis [6]. This work contributes to the third algorithm development i.e. identifying the absence of cardiac failure.

Cardiac failure is clinically suspected and the test that confirms its absence or presence is ususally an echocardiography. This echocardiography could have been performed prior to PICU admission, even in another institution and could not be digitally available for analysis. However, when an echocardiography has been performed, physicians report its result in the notes. It is the reason why, using notes to exclude or confimed a cardiac failure was assumed to be the best way to electronically collect as soon as possible the information.

Generally, there is a list of golden indicators to classify cardiac failure patients. Those indicators could be either from the medical history, clinical exam, chest X-Ray interpretation, recent cardiovascular performance evaluation, or laboratory test results. Medication, such as Levosimendan, Milrinone, Dobutamine, is a surrogate to the gold standard. Its list can be retrieved from syringe pump data, prescriptions, and notes. If any medication from the three is present, there is certainly a cardiac failure. Besides, cardiovascular performance evaluation also contributes to indicate the cardiac failure diagnosis. One of the evaluations is ejection fraction (EF) $< 50\%$. EF refers to the percentage of blood pumped (or ejected) out of the ventricles with each contraction. It is a surrogate for left ventricular global systolic function, defined as the left ventricular stroke volume divided by the end-diastolic volume. The other indicator for cardiovascular performance evaluation is shortening fraction (SF) $< 25\%$. FR is the length of the left ventricle during diastole and systole. It measures diastolic/systolic changes for inter-ventricular septal and posterior wall dimensions. Finally, brain natriuretic peptide, known as pro-BNP ng/L > 1000, comes from laboratory test results being useful in the acute settings for differentiation of cardiac failure from pulmonary causes of respiratory distress. Pro-BNP is continually produced in small quantities in the heart and released in more substantial quantities when the heart needs to work harder.

Consequently, the clinical knowledge representation will summarize detailed attributes that are essential to detecting cardiac failure. All notes are taken into account if they are encompassed by the information of the prescription history of Milrinone (mcg/kg/min), measurement notes of pro-BNP (ng/L), dilated cardiomyopathy, acute left cardiac failure, chronic cardiac failure, postoperative cardiac failure, coronary microvascular disorder history notes, notes of a measurement result of either EF (%) or SF (%). As a result, a patient is considered to have a cardiac failure if he/she has one of the criteria. Unfortunately, as all the mentioned information above that helps diagnose cardiac failure is not readily available electronically, we will develop a machine learning algorithm based on natural language processing (NLP) that automatically detects this desired concept label from clinical notes. The algorithm can automatically see whether a patient has a cardiac failure or a healthy condition lacking gold indicators from the notes. In such a situation, the proposed algorithm can effectively learn a latent representation of clinical notes, which traditionally rule-based approaches cannot depict.

B. Motivation

The recent study [7] extensively analyzed and confirmed the feasibility of employing machine learning for cardiac failure. However, we are dealing with two challenges from clinical notes in French and a limited amount of dataset size in our case. We will examine data retrospectively to validate the diagnosis. And the main objective of this study consists of two sub-objectives that overcome the mentioned limitations, as follows:

- Which representation learning approach should be used? The representation learning approach, which can retain the words' semantic and syntactic analysis in critical care data, enriches the mutual information for the word representation by capturing word-to-word correlation.
- Which machine learning classifier should be employed? The classifier can avoid the overfitting associated with the machine learning rule by marginalizing over the model parameters instead of making point estimates of its values.

II. Materials and Methods

A. Clinical Narrative Data At CHUSJ

Fig. 1 illustrates the conceptual framework for conducting the experiments. First, the data integration process has been completed at the Pediatric Intensive Care Unit, CHUSJ, for more than 1300 patients. After the research protocol was approved by the research ethics board from Research Center of the Sainte-Justine University Hospital. We only took information from two types of notes, including admission and evaluation notes. Since, these notes documented the reasons why a patient was admitted to the hospital by the physician in charge. And, the notes also provided the initial instruction for that patient's care based on the patient's health status. Primarily, we focused on medical background, history of the disease to admission, and cardiovascular evaluation. Furthermore, we only used notes for each patient's first stay within the first 24 h since the admission. If a patient had more than one ICU stay, we only analyzed the first one. We did not have any missing notes but we can not exclude some information that were not collected by physicians and then not reported in the note. However, the data fully reflect real clinical practice. Then, two doctors from the CHUSJ (Dr. Jérôme Rambaud and Dr. Guillaume Sans), who did not compose the notes at the first hand, separately reviewed each patient's notes; each note was manually labeled "YES" or "NO" for positively cardiac

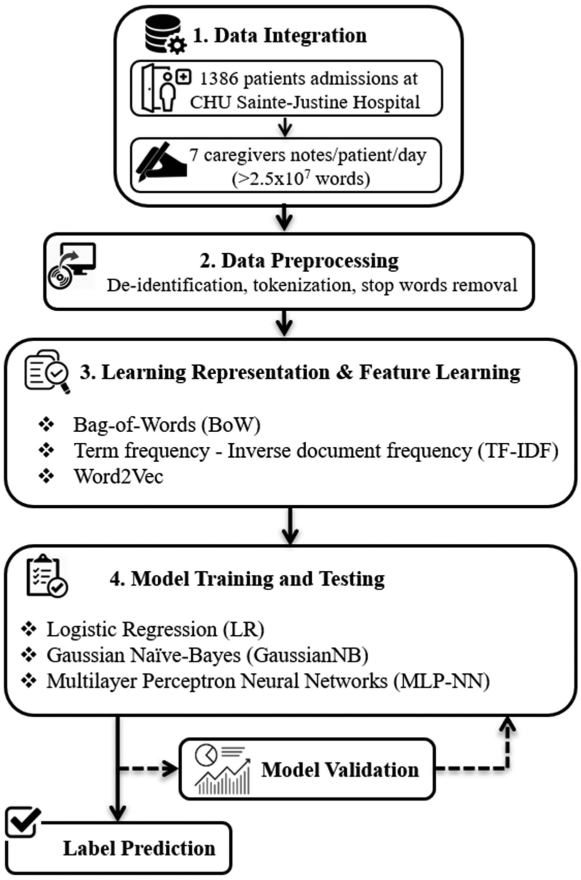

1. Data Integration
1386 patients admissions at CHU Sainte-Justine Hospital
7 caregivers notes/patient/day (>2.5x10^7 words)
2. Data Preprocessing
De-identification, tokenization, stop words removal
3. Learning Representation & Feature Learning
❖ Bag-of-Words (BoW)
❖ Term frequency - Inverse document frequency (TF-IDF)
❖ Word2Vec
4. Model Training and Testing
❖ Logistic Regression (LR)
❖ Gaussian Naïve-Bayes (GaussianNB)
❖ Multilayer Perceptron Neural Networks (MLP-NN)
Model Validation
Label Prediction

Fig. 1. An overview of the proposed methodology to detect cardiac failure from clinical notes at CHUSJ.

failure or under a healthy condition, respectively. By doing so, we could double-check that missing data was not problematic. To avoid data contamination, we checked both the "patientID" and "careproviderID" to ensure no notes were simultaneously present in the training and testing cohort. Finally, we have 5444 line of notes with 1941 positive cases (36% of total) and 3503 negative cases. The average length of the number of characters is 601 and 704. The average length of the number of digits is 25 and 26 for the positive and negative cases, respectively.

B. Data Pre-Processing

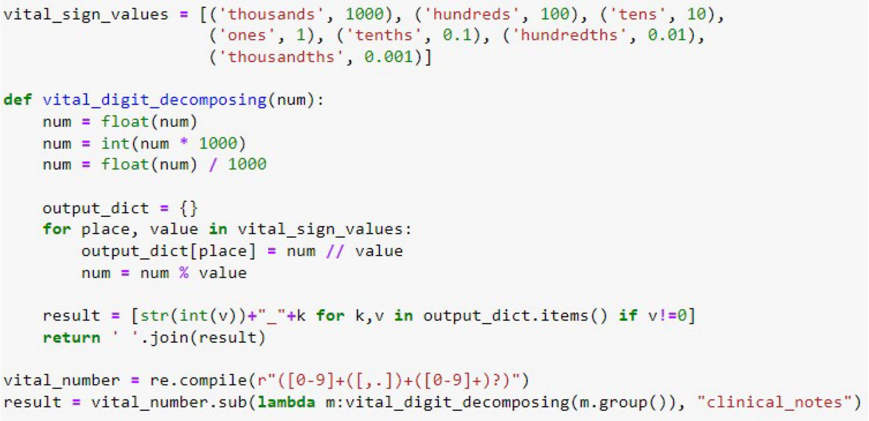
Generally, it is proven that if the preprocessing steps are well prepared, the result for the end-task will be improved [8]. Therefore, there are steps that were used as case lowering, and stop words removing. From the list, all these words are the definition in French; so, they do not contribute to the learning representation. Besides, we did not consider any French linguistic feature as our method is based on uncorrelated words. The longest n-gram is over 400 words, but most of the n-gram length distribution is between 50 and 125 words.

In addition, it is essential to pay attention to negation in medical expression. First, the negation criteria from the study [9] were used for detecting the negative meaning from French notes. Then, a negation technique is applied [10]: a term "neg_" is added as a prefix for a term. An example note is "Patient explique qu'à ce moment là, il *n'était pas* capable de parler et l'air *ne passait pas* au niveau de sa gorge. Respiration plus rapide, mais état général préservé, parents *n'étaient pas* inquiets. (Patient explained at that time, he was not able to speak and the air did not pass at the level of his throat. Breathing faster, but general condition preserved, parents were not worried)". The negation will be tagged as: "Patient explique qu'à ce moment là, il *neg_était* capable de parler et l'air *neg_passait* au niveau de sa gorge. Respiration plus rapide, mais état général préservé, parents *neg_étaient* inquiets."

For the vital numeric values (heart rate, blood pressure, etc.), most of the NLP representation learnings cannot accommodate the numeric values effectively. Most NLP models treat numeric values in the text the same way as other tokens. It has been proven that the pre-trained token representations (word2vec) can naturally encode the numeric values [11]. Unfortunately, it required a large amount of data with specific labeling progress for this task. At the same time, the state-of-the-art for numerical reasoning results is much less good (47%) compared with the expert human performance (96.4%) in the f1 score metric [12]. Another study only focuses on how to extract the number, not dealing with representation learning [13]. Even, study [14] proposes an alternative approach to deal with both large and small datasets. However, the authors either removed all of the vital sign numeric values or did not mention how to deal with numeric values. Because we have limited data, we decide to keep all numeric values for vital sign values (nearly 4% of the notes) and apply the decoding for those number values. In fact, a numeric value consists in a numerical measurement value and a measurement unit as ruled by Digital Imaging and Communication in Medicine standard for report document [15]. Therefore, we performed four experiments to evaluate the contribution from the numeric value to the classifiers. Fig. 2 shows an example of code snippet in Python, which help us conducting the decomposing the numerical measurement value. Finally, Table I summarizes the four different approaches to decode the numeric values, including (i) keeping all of the original numeric values and their units, (ii) removing all of the numeric values and their units, (iii) encoding the decimal into a string named dot, and (iv) decomposing into digits.

C. Clinical Natural Language Representation Learning

There is no doubt about the effectiveness of neural word embedding. The study [16] confirms that word2vec representation has been successfully used for various disease classifications from medical notes. Especially for the French clinical notes, the study [17] shows that word2vec and GloVec effectively embed the clinical notes. And, the word2vec had the highest score on 3 out of 4 rated tasks (analogy-based operations, odd one similarity, and human validation). In addition, studies [18], [19], [20], [21] confirm that conventional approaches bag-of-words (BoW), term frequency-inverse document frequency (TF-IDF) have better performance than other deep learning techniques on a smaller corpus with long texts in clinical note corpus.



```python
vital_sign_values = [('thousands', 1000), ('hundreds', 100), ('tens', 10),
                     ('ones', 1), ('tenths', 0.1), ('hundredths', 0.01),
                     ('thousandths', 0.001)]

def vital_digit_decomposing(num):
    num = float(num)
    num = int(num * 1000)
    num = float(num) / 1000

    output_dict = {}
    for place, value in vital_sign_values:
        output_dict[place] = num // value
        num = num % value

    result = [str(int(v))+"_"+k for k,v in output_dict.items() if v!=0]
    return ' '.join(result)

vital_number = re.compile(r"([0-9]+([,.])+([0-9]+)?)")
result = vital_number.sub(lambda m:vital_digit_decomposing(m.group()), "clinical_notes")
```

Fig. 2. An example of code snippet in Python for decomposing numeric values (Example 4).

TABLE I
A SUMMARY OF EXPERIMENTS DEALING WITH VITAL SIGN NUMERIC VALUES

Experiment	Description	Illustration*
Exp_1	Keep all of the numeric values and units	[vg, sévèrement, dilate, 64.8, mm, diastole, 58.3, mm, systole]
Exp_2	Remove all of the numeric values and units	[vg, sévèrement, dilate, diastole, systole]
Exp_3	Encoding the decimal point into string (DOT)	[vg, sévèrement, dilate, 64, dot, 8, mm, diastole, 58, dot, 3, mm, systole]
Exp_4	Decomposing numeric values into digits	[vg, sévèrement, dilate, 6_tens, 4_ones, 8_tenths, mm, diastole, 5_tens, 8_ones, 3_tenths, mm, systole]

*The original notes are "VG sévèrement dilaté (64.8mm en diastole et 58.3mm en systole) - Severely dilated LV (64.8mm in diastole and 58.3mm in systole)"

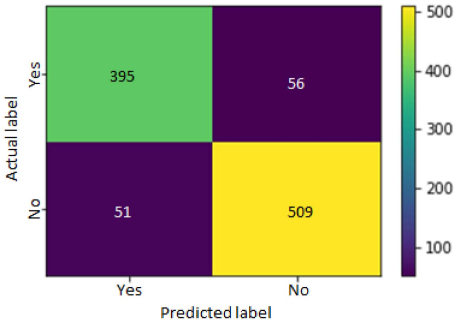


Fig. 3. Confusion matrix of the MLP-NN classifier, showing the classification of positive (Yes) and negative (No) between predicted and actual labels.

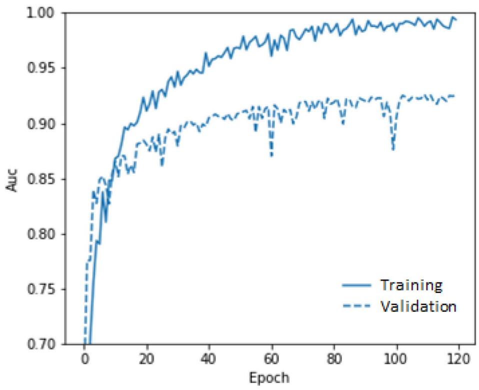


Fig. 4. Area Under the Curve (AUC) performance of MLP-NN.

Therefore, we will evaluate the effectiveness of two conventional representation approaches, including BoW, TF-IDF, and the word2vec neural embedding model.

D. Machine Learning Classifiers

The state-of-the-art machine learning-based NLP currently focuses on deep learning for clinical notes [22], [23], [24], [25], [26]. For example, to predict cardiac failure, deep learning (Convolution Neural Network-based) shows its exceptional performance, F1 score of 0.756, to the conventional approach Random Forest (RF) with an F1 score of 0.674 [27]. And, study [16] shows the best performance to predict multiple chronic diseases (cerebral infraction, pulmonary infection and coronary atherosclerotic heart disease) by combining of word2vec and deep learning with the average accuracy and F1 score exceeded 90%.

However, a large enough amount of data is needed to have a good generalization capability of deep learning, while this data availability requirement is not always provided [28]. Especially, clinical notes in a language other than English, the challenge is more difficult to mitigate [29]. Deep learning architectures generally work well for large scale data sets with short texts while do not outperform conventional approaches (BoW) on a smaller corpus with long texts in clinical note corpus [19]. Automatic methods to extract New York heart association classification from clinical notes [20] confirm that the machine learning method, support vector machines (SVM) with n-gram features, achieves the best performance at 93% F-measure. Also, study [18] proved the achievement by combining the BoW and Naïve Bayes classifier on clinical notes for accessing hospital readmission offering an area under the curve (AUC) of 0.690.

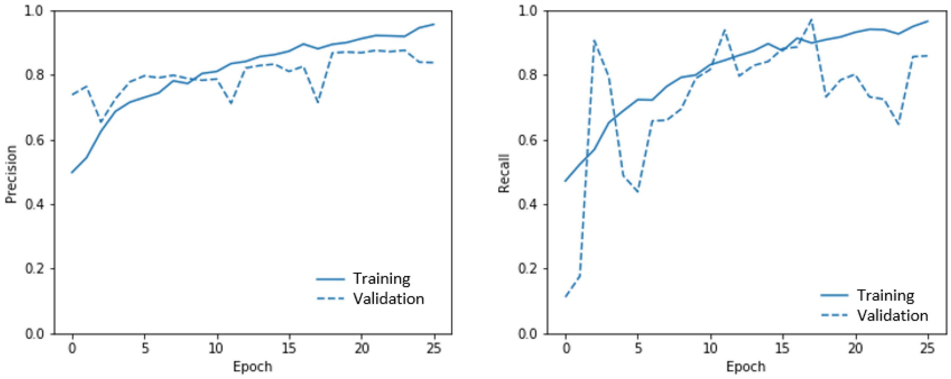


Fig. 5. Precision (left) and recall (right) performance based on the Transformer configuration.

This study confirms that, with the small dataset, TF-IDF and BoW have better performance than other techniques on coronary microvascular classification [21].

Besides, logistic regression (LR) and generative Naïve Bayes perform better than the other classifiers, particularly for small datasets. Several classifiers have been trained for short text classification; it includes RF, Gaussian Naïve Bayes (GaussianNB), Multinomial Naïve Bayes (MultimonalNB), LR, SVM and K-nearest neighbour. The experimental results from [30], [31] confirm that LR, and GaussianNB perform much the better than the other classifiers. Moreover, study [32] evaluated different classifiers' performance, including discriminative and generative learning approaches. And, it also confirms that the discriminative LR algorithm has a lower asymptotic error, while the generative Naïve Bayes classifier converges quickly.

Additionally, when the ratio value for the number of samples/number of words per sample is small (<1500), a small multilayer perceptron neural network (MLP-NN) that takes n-grams as input performs better or at least as well as deep learning models [33]. Besides, an MLP-NN is simple to define and understand, and it takes less computation time than sequence models. A detailed explanation of using an MLP-NN in medical analysis can be seen from [34].

Consequently, we implemented and compared all the above mentioned methods; the result of RF, MultimonalNB, and SVM was less than 75% for accuracy. Again, the result shows that only LR, GaussianNB and MLP-NN are comparable, and perform better than RF, MultimonalNB, and SVM classifer. Therefore, in this study, we focus on three different machine learning classifiers, including LR, GaussianNB, and MLP-NN.

III. Results

We did the analysis to select of proper neural network sizes and architectures [35]. We have used the structure of an MLP-NN that consists of $L = 3$ layers, where layer 1 is the input layer, layer 3 is the output layer, and layer 2 is the hidden layer. The total number of neurons in the hidden layer is $N_t = 100$ neurons. To prevent the neural network from overfitting, we applied the dropout [36] with the probability of dropping out rate p = 0.25, and GlorotNormal kernel initializer [37].

We used the scikit-learn library [38] and Keras [39] in Python to implement our model. No preprocessing was required to deal with missing data. The data was divided into 60% training, 20% validation, and 20% testing. To make our results more consistent, we used the k-fold cross validation ($k = 5$) [40]; each dataset was divided into k subsets called folds, the model was trained on $k - 1$ of them and tested on the left out. This process was repeated k times, and the results were averaged to get the final one. Furthermore, we also employed the univariate feature selection with sparse data from the learning representation feature space. This selection process works by selecting the best features based on univariate statistical tests named SelectKBest algorithms, which removes all but the K highest scoring features (K = 20000).

To effectively assess the performance of our method, metrics including accuracy, precision, recall (or sensitivity), and F1 score were used [41]. These metrics are defined as follows:

$$\begin{aligned}
\text{Accuracy (acc)} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\
\text{Precision (pre)} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
\text{Recall/Sensitivity (rec)} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{F1-Score (f1)} &= \frac{2^{\star}\text{Precision}^{\star}\text{Recall}}{\text{Precision } + \text{Recall}}
\end{aligned}$$

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

IV. Discussion

Table II presents the results of our method. First, among four experiments for dealing with numeric values, experiment 3 yields the best performance. Encoding the decimal point into a string "DOT" has helped the learning representation process retain the information from numeric values. It is also interesting to mention that when we keep all numeric values and do nothing (experiment 1), the results are worse than if we remove all the

TABLE II
SUMMARIZATION OF EXPERIMENTS PERFORMANCE EVALUATION

Representation		ML	Exp_1				Exp_2				Exp_3				Exp_4			
			acc	pre	rec	f1	acc	pre	rec	f1	acc	pre	rec	f1	acc	pre	rec	f1
W/o Feature Selection	BoW	LR	0.80	0.77	0.81	0.79	0.81	0.80	0.82	0.81	0.82	0.80	0.83	0.81	0.82	0.80	0.81	0.8
		GaussianNB	0.77	0.72	0.80	0.76	0.78	0.76	0.81	0.78	0.79	0.78	0.81	0.79	0.79	0.76	0.80	0.78
		MLP-NN	0.81	0.78	0.81	0.79	0.81	0.80	0.82	0.81	0.81	0.81	0.84	0.82	0.81	0.81	0.82	0.81
	TF-IDF	LR	0.81	0.79	0.82	0.8	0.78	0.76	0.79	0.77	0.81	0.78	0.81	0.79	0.77	0.75	0.77	0.76
		GaussianNB	0.79	0.75	0.81	0.78	0.77	0.74	0.80	0.77	0.78	0.75	0.81	0.78	0.76	0.74	0.80	0.77
		MLP-NN	0.81	0.80	0.82	0.81	0.84	0.82	0.85	0.83	0.85	0.84	0.85	0.84	0.82	0.81	0.81	0.81
	Embedding	LR	0.74	0.72	0.79	0.75	0.76	0.74	0.79	0.76	0.78	0.75	0.82	0.78	0.76	0.73	0.77	0.75
		GaussianNB	0.72	0.71	0.77	0.74	0.76	0.71	0.79	0.75	0.76	0.73	0.80	0.76	0.75	0.72	0.72	0.72
		MLP-NN	0.74	0.74	0.76	0.75	0.77	0.76	0.78	0.77	0.79	0.77	0.80	0.78	0.77	0.73	0.78	0.75
W/ Feature Selection	BoW	LR	0.80	0.81	0.78	0.79	0.81	0.81	0.79	0.80	0.78	0.78	0.77	0.77	0.80	0.80	0.79	0.79
		GaussianNB	0.80	0.81	0.78	0.79	0.80	0.78	0.79	0.78	0.78	0.79	0.77	0.78	0.80	0.81	0.78	0.79
		MLP-NN	0.80	0.79	0.80	0.79	0.82	0.82	0.81	0.81	0.83	0.82	0.83	0.82	0.84	0.83	0.84	0.83
	TF-IDF	LR	0.76	0.71	0.79	0.75	0.82	0.81	0.83	0.82	0.83	0.82	0.83	0.82	0.78	0.78	0.80	0.79
		GaussianNB	0.80	0.78	0.80	0.79	0.81	0.82	0.79	0.80	0.81	0.81	0.82	0.81	0.79	0.78	0.79	0.78
		MLP-NN	**0.84**	**0.84**	**0.85**	**0.84**	**0.87**	**0.86**	**0.88**	**0.87**	***0.89***	***0.89***	***0.88***	***0.88***	**0.85**	**0.84**	**0.84**	**0.84**
	Embedding	LR	0.80	0.78	0.80	0.79	0.80	0.79	0.80	0.79	0.82	0.82	0.83	0.82	0.81	0.78	0.79	0.78
		GaussianNB	0.77	0.76	0.78	0.77	0.79	0.79	0.79	0.79	0.81	0.81	0.80	0.8	0.79	0.78	0.78	0.78
		MLP-NN	0.80	0.79	0.80	0.79	0.80	0.80	0.80	0.80	0.82	0.81	0.81	0.81	0.80	0.79	0.80	0.79

numbers and their units (experiment 2). Experiment 4 confirms that if the numbers are extensively encoded, it will negatively affect the result, lowering the performance.

The combination of TF-IDF and MLP-NN consistently outperforms other combinations with overall performance and is the most stable in all circumstances. Without any feature selection, the proposed framework yielded an overall classification performance with acc, pre, rec, and f1 of 85% and 84%, 85%, and 84%, respectively. Also, the representation matrix from the TF-IDF above is sparse because every word is treated separately. Hence, the semantic relationship between separated entities is ignored, which would cause information loss. Therefore, if the feature selection (SelectKBest) was well applied and tuned, it could improve up to 3–4% for each evaluation in the overall performance. Consequently, it achieves the best performance with 89%, 89%, 88%, and 88% for acc, pre, rec, and f1, respectively. And, the detailed confusion matrix showing the classification of positive cases (1) and negative cases (0) is shown in Fig. 3.

Furthermore, with limited data, the BoW and TF-IDF have proven their capacity to better retain information from the notes representation. It has been shown in [31] that the TF-IDF has the highest accuracy compared to neural word embeddings in short text classification (less than 20 words per sample). In our study, we could not increase our samples beyond 80 words per sample. However, our results show that the TF-IDF performs better than the neural word embedding when used on short narratives (approximately 80 words per example in our case). It is in agreement with the comparison discussed in [31]. The difference in performance was less significant in our case. One can expect the neural word embeddings to outperform others approaches, when the word number increases as shown in [42].

Besides, with the same learning presentation approach (BoW, TF-IDF, or neural word embeddings), the LR classifiers had better performance than GaussianNB classifiers. The results align with the theoretical and experimental analysis from [32], [43]. LR performs better with smaller data sizes because it effectively approaches its lower asymptotic error from the initial learning steps. However, MLP-NN models always dominated with their best generalization. They have achieved their generalization capacity because the misclassification probability can be reduced and trained closer to optimal points that cannot be achieved with simple algorithms [44].

By applying the dropout (p = 0.25) [36], GlorotNormal initializer [37], and balancing the classes by using the Bayes Imbalance Impact Index [45], the classifier was successful in avoiding the overfitting. Primarily, Fig. 4 represents the Area Under the Curve (AUC) with respect to the epoch for the training and validation. We can see that the classifier can achieve nearly 100% of separability of the two classes during the training. The classier can achieve almost 90% of the separability during the validation. The distance between the two curves does not change with the increasing epoch number. And, the validation curve does not drop out to the growing epoch number. This indicates the algorithm does not overfit.

We also tested with the model CamemBERT, which is specifically a transformer-based language model for the french language [46]. It is motivated by the success of a Bidirectional Encoder Representations from Transformers (BERT) for natural language understanding [47]. Unfortunately, the result was not as good as expected; we could only achieve less than 60% accuracy, even though we applied the drop-out technique as recommended from the study [48]. We continued investigating with the simpler Transformer, which is solely based on attention mechanisms through the connection of the encoder and decoder [49], and it is implemented by Keras [50]. The result has achieved a decent performance compared to advanced and complicated BERT-based models. However, it is still far below the performance from the simple MLP-NN, where the highest precision and recall are continually fluctuating at around 80% as shown in Fig. 5. Moreover, from the result of Fig. 5, we can conclude that the transformer-based model underperforms in classification tasks for a small sample size, short of clinical NLP. This conclusion is in agreement with the limitations identified and discussed in [51]; the authors have proved that the transformer-based model was well suited for understanding the contextual meaning of a long sequence rather than understanding key words or phrases.

V. Conclusion

We have employed both learning representation and machine learning algorithms to tackle the French clinical natural language processing for detecting cardiac failure in children at CHUSJ. We have extensively conducted and analyzed a conceptual framework to detect a patient's health condition from the contextual input to the contextual output. Our numerical results have confirmed the feasibility of the proposed design by combining TF-IDF and MLP-NN; the proposed mechanism could also be improved with the feature selection from the learning representation vector space. Consequently, the proposed framework yields an overall classification performance with 89% accuracy, 88% recall, and 89% precision.

Secondly, we assumed that the numeric values significantly contribute to the classifier. Instead of losing them, we addressed different decoding approaches for numeric values in our work. In our case study, encoding the decimal point into a string "DOT" has helped the learning representation process retain the information from the numerical values in clinical notes. Otherwise, it is better to remove the numeric values rather than keep them without any encoding, or extensive encoding.

Finally, with the MLP-NN learning algorithm, we can train closer to optimal architectures, which cannot be trained with simple algorithms (LR, GaussianNB, RF, MultinomialNB, and SVM). Although BERT-based models are currently known as the state-of-the-art in natural language processing tasks, the final results suggest that these Transformer-based methods perform less effectively than existing alternatives.

One of the limitations is that the CDSS is still under development (in process currently). The next step of our project is to create the CDSS to diagnose ARDS early by integrating this NLP algorithm with the other algorithms on hypoxemia and chest X-Ray analysis. When the integration is done in the PICU electronic medical infrastructure, we will validate the CDSS's ability to screen ARDS prospectively. Furthermore, future research should carefully consider the potential effects of numerical values alongside unstructured notes. Ideally, an algorithm, which can automatically extract and represent the numerical values from the clinical notes, should be investigated for further validation. This may be a promising aspect of using a semantic neural network to determine the boundaries and extract the numerical values from the text. And, generative learning has a great potential for an evaluation [12].

Acknowledgment

The clinical data were provided by the Research Center at CHU Sainte-Justine hospital. The authors thank Dr. Sally Al Omar, Dr. Rambaud Jérôme and Dr. Sans Guillaume for their data support of this research. This work was supported by a scholarship from Fonds de recherche du Québec – Nature et technologies (FRQNT) to Thanh-Dung Le, the funds from Natural Sciences and Engineering Research Council (NSERC) and Institut de valorisation des donnÃ©es (IVADO), and the Fonds de la recherche en sante du Quebec (FRQS).

V. Supplementary Materials

The additional exploratory data analysis from the data preprocessing step is presented in supplementary materials. We also provide an overview theoretical explanation of employed methods for clinical note representation learning, machine learning classifiers, and imbalance learning in that document.

References

[1] A. E. W. Johnson et al., "Machine learning and decision support in critical care," *Proc. IEEE*, vol. 104, no. 2, pp. 444–466, Feb. 2016.

[2] M.-P. Matton et al., "Databases and computerized systems in picu: Electronic medical record in pediatric intensive care: Implementation process assessment," *J. Pediatr. Intensive Care*, vol. 5, pp. 129–138, 2016.

[3] 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.

[4] P. A. L. I. C. C. Group et al., "Pediatric acute respiratory distress syndrome: Consensus recommendations from the pediatric acute lung injury consensus conference," *Pediatr. Crit. Care Med.: A J. Soc. Crit. Care Med. World Federation Pediatr. Intensive Crit. Care Societies*, vol. 16, no. 5, pp. 428–439, 2015.

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

[6] N. Zaglam et al., "Computer-aided diagnosis system for the acute respiratory distress syndrome from chest radiographs," *Comput. Biol. Med.*, vol. 52, pp. 41–48, 2014.

[7] C. R. Olsen, R. J. Mentz, K. J. Anstrom, D. Page, and P. A. Patel, "Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure," *Amer. Heart J.*, vol. 229, pp. 1–17, 2020.

[8] S. Kannan et al., "Preprocessing techniques for text mining," *Int. J. Comput. Sci. Commun. Netw.*, vol. 5, no. 1, pp. 7–16, 2014.

[9] L. Deléger et al., "Detecting negation of medical problems in French clinical notes," in *Proc. 2nd ACM SIGHIT Int. Health Informat. Symp.*, 2012, pp. 697–702.

[10] S. Dubois et al., "Learning effective representations from clinical notes," 2017, *arXiv:1705.07025*.

[11] E. Wallace et al., "Do NLP models know numbers? probing numeracy in embeddings," in *Proc. Conf. Empirical Methods NLP*, 2019, pp. 5310–5318.

[12] D. Dua et al., "DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, vol. 1, pp. 2368–2378.

[13] T. Cai et al., "EXTraction of EMR numerical data: An efficient and generalizable tool to EXTEND clinical research," *BMC Med. Informat. Decis. Mak.*, vol. 19, no. 1, 2019, Art. no. 226.

[14] V. Kumar, D. R. Recupero, D. Riboni, and R. Helaoui, "Ensembling classical ML and DL approaches for morbidity identification from clinical notes," *IEEE Access*, vol. 9, pp. 7107–7126, 2021.

[15] R. Noumeir, "DICOM structured report document type definition," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 4, pp. 318–328, Dec. 2003.

[16] X. Shi et al., "Multiple disease risk assessment with uniform model based on medical clinical notes," *IEEE Access*, vol. 4, 2016.

[17] E. Dynomant et al., "Word embedding for the French natural language in health care: Comparative study," *JMIR Med. Informat.*, vol. 7, no. 3, 2019, Art. no. e12310.

[18] A. Agarwal, C. Baechle, R. Behara, and X. Zhu, "A natural language processing framework for assessing hospital readmissions for patients with COPD," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 588–596, Mar. 2018.

[19] Y. Li, L. Yao, C. Mao, A. Srivastava, X. Jiang, and Y. Luo, "Early prediction of acute kidney injury in critical care setting using clinical notes," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2018, pp. 683–686.

[20] R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, "Automatic methods to extract New York Heart Association classification from clinical notes," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2017, pp. 1296–1299.

[21] S. J. Fodeh,, T. Li, H. Jarad, and B. Safdar, "Classification of patients with coronary microvascular dysfunction," *IEEE Trans. Comput. Biol. Bioinform.*, vol. 17, no. 2, pp. 704–711, Mar./Apr. 2020.

[22] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *J. Biomed. Informat.*, vol. 69, pp. 218–229, 2017.

[23] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, 2018, Art. no. 18.

[24] D. W. Otter,, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for NLP," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604–624, Feb. 2021.

[25] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[26] S. Sheikhalishahi et al., "Nlp of clinical notes on chronic diseases: Systematic review," *JMIR Med. Informat.*, vol. 7, 2019, Art. no. e12239.

[27] X. Liu et al., "Predicting heart failure readmission from clinical notes using deep learning," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2019, pp. 2642–2648.

[28] A. Paleyes et al., "Challenges in deploying machine learning: A survey of case studies," *ACM Comput. Surv.*, to be published, doi: 10.1145/3533378.

[29] A. Névéol et al., "Clinical natural language processing in languages other than english: Opportunities and challenges," *J. Biomed. Semantics*, vol. 9, no. 1, pp. 1–13, 2018.

[30] O. Z. Maimon et al., *Data Mining With Decision Trees: Theory and Applications*, vol. 81. Singapore: World Scientific, 2014.

[31] Y. Wang et al., "Comparisons and selections of features and classifiers for short text classification," in *Proc. IOP Conf. series: Mater. Sci. Eng.*, vol. 261, 2017, Art. no. 012018.

[32] A. Ng and M. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, vol. 14, pp. 841–848.

[33] Google, "Machine learning guides text classification," Accessed: Sep. 30, 2020. [Online]. Available: https://developers.google.com/machine-learning/guides/text-classification/step- 2-5, 2019-10-21.

[34] A. Pasini, "Artificial neural networks for small dataset analysis," *J. Thoracic Dis.*, vol. 7, no. 5, p. 953, 2015.

[35] D. Hunter, H. Yu, M. S. Pukish III, J. Kolbusz, and B. M. Wilamowski, "Selection of proper neural network sizes and architectures: A comparative study," *IEEE Trans. Ind. Inform.*, vol. 8, no. 2, pp. 228–240, May 2012.

[36] N. Srivastava et al., "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.'

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

[38] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[39] F. Chollet et al., "Keras," 2015. [Online]. Available: https://keras.io/

[40] R. Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 1995, vol. 14, pp. 1137–1145.

[41] C. Goutte et al., "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proc. Eur. Conf. Inf. Retrieval*, 2005, pp. 345–359.

[42] M. Sahlgren et al., "The effects of data size and frequency range on distributional semantic models," in *Proc. Conf. Empirical Methods NLP*, 2016, pp. 975–980.

[43] C. Perlich et al., "Tree induction vs. logistic regression: A learning-curve analysis," *J. Mach. Learn. Res.*, vol. 4, pp. 211–255, 2003.

[44] P. L. Bartlett, "The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network," *IEEE Trans. Inf. Theory*, vol. vol. 44, no. 2, pp. 525–536, Mar. 1998.

[45] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, "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, Sep. 2020.

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

[47] J. Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, vol. 1, pp. 4171–4186.

[48] K. Pasupa et al., "A comparison between shallow and deep architecture classifiers on small dataset," in *Proc. 8th Int. Conf. Inf. Technol. Elect. Eng.*, 2016, pp. 1–6.

[49] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[50] A. Nandan, "Text classification with transformer," Accessed: Sep. 30, 2020. [Online]. Available: https://keras.io/examples/nlp/text_classification_with_transformer/, 2020-05-10.

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