Development of machine-learning-based natural language processing to detect concept labels in clinical narratives

by

Thanh-Dung LE

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, SEPTEMBER 12, 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mrs. Rita Noumeir, Thesis supervisor
Department of Electrical Engineering, École de Technologie Supérieure

Mr. Philippe Jouvet, Thesis Co-Supervisor
Pediatric Intensivist - Ste. Justine Hospital Montréal, Université de Montréal

Mr. Mohamed Cheriet, Chair, Board of Examiners
Department of Electrical Engineering, École de Technologie Supérieure

Mrs. Rachel Bouserhal, Member of the Jury
Department of Electrical Engineering, École de Technologie Supérieure

Mrs. Sabine Bergler, External Examiner
Department of Computer Science and Software Engineering, Concordia University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON AUGUST 14, 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

My journey toward a Ph.D. has been challenging yet rewarding, and I am grateful for the unwavering support and encouragement I received from my supervisors, colleagues, friends, and family. While it's difficult to express my sincere gratitude to them in words, I would like to take this opportunity to acknowledge their contributions.

First and foremost, I express my deepest gratitude to my principal supervisor, Professor Rita Noumeir, for her continuous guidance, motivation, and inspiration throughout my Ph.D. studies. Her invaluable advice and discussions during weekly meetings have positively impacted my research and career trajectory. I am also grateful for her support during the COVID-19 pandemic, where her instructions and supervision helped me adapt to remote working tools and maintain my research progress. I also thank my co-supervisors, Professor Phillipe Jouvet, for his valuable comments and feedback on my research manuscripts. I would also like to thank my Ph.D. assessment jury for their constructive feedback and support.

I thank Fonds de Recherche du Québec Nature et Technologies for awarding me the merit scholarship that enabled me to pursue my Ph.D. project. I also want to thank Pervasive and Smart Wireless Applications for the Digital Economy (PERSWADE) for their support during my research at The Énergie, Matériaux et Télécommunications center, Institut National de la Recherche Scientifique (INRS). Additionally, I would like to thank the anonymous reviewers who provided valuable feedback on my research papers and helped improve my work's quality.

I acknowledge my former supervisors and colleagues at the NECHPY-Lab, INRS: Professor Long Le, Dr. Ha Nguyen Vu, Dr. Hoang Duc Tuong, Dr. Tran Duy Thinh, Dr. Nguyen Ti Ti, Dr. Nguyen Minh Tri, Vu Huy Hoang, Phan Thanh Tung, Dr. Nguyen Minh Dat. Furthermore, I express gratitude to my colleagues at LATIS, ETS: Dr. Georges Matar, Jihad El Tannoury, Gloria Huong, Oussema, Haythem, Mario, Asheok, Wajahat, Khalil, Toufik, Clara, Emily, and many others for their support, encouragement, and the memories we have shared.

Finally, I am indebted to my parents (Lam-Bup, Luyen-La), sisters (Hong Diem, Ngoc Trinh, Le Huyen), brothers (Quoc Vuong, Xuan Son), niece (Hong Vy), nephews (Gia Bao, Thien Phuc) for their unconditional love, support, and sacrifices. Their emotional and financial support have been invaluable during my Ph.D. journey. Special thanks to my wife and our little princess Nha Lam Cecilia, who has been with me every step of the way, providing unwavering support and encouragement throughout my Ph.D. journey. Without their love, dedication, this achievement would not have been possible, and I hope my accomplishments will make them proud.

I would like to thank everyone who made my Ph.D. journey a fulfilling and memorable experience. Thank you all for your support.

Développement d'algorithmes d'apprentissage machine pour le traitement du langage naturel afin de détecter certains concepts dans les récits cliniques

Thanh-Dung LE

RÉSUMÉ

Une abondance de données et d'information est disponible dans le domaine clinique. Les cliniciens ont réussi à combiner les données informatives et structurées, qui comprennent les résultats de tests de laboratoire, l'imagerie médicale et les données de capteurs portables avec de nouveaux algorithmes analytiques pour offrir des soins de santé omniprésents et personnalisés. Cependant, les sources narratives cliniques, qui sont de courtes notes sur les patients, écrites par des médecins, posent des contraintes considérables. Bien que les notes sont fournies continuellement et stockées dans les entrepôts de données cliniques, elles sont peu utilisées en pratique réelle. La limitation provient principalement de leur format non structuré ou semi-structuré. Heureusement, le déploiement de l'apprentissage en profondeur au cours des dernières années aide à capturer efficacement la représentation cachée des récits cliniques, en raison de sa grande capacité de calcul. En particulier, l'amélioration des performances de l'apprentissage en profondeur sur les notes cliniques est continuellement renforcée par l'usage de techniques de traitement de langage naturel (NLP) lors du prétraitement des données. Le NLP devient une approche nécessaire pour surmonter les défis présents dans les notes de texte clinique non structurés, car cette étape peut efficacement mapper les mots dans les données non structurées dans un espace de dimension inférieure.

Heureusement, une grande source de données de notes cliniques est actuellement stockée dans l'entrepôt de données de recherche au CHU Sainte-Justine (CHUSJ). Il y a 7 notes/patient/jour pour 1386 patients (contenant un ensemble de données de plus de 2,5 x 107 mots). Ces notes sont extraites de notes d'admission, notes d'évaluation et notes de synthèse. Les notes d'admission décrivent les raisons pour l'admission aux unités de soins intensifs, le progrès historique de la maladie, les médicaments donnés, la chirurgie et toutes autres données de base supplémentaires du patient. Les affections quotidiennes et les résultats des tests de laboratoire sont décrits dans les notes d'évaluation, desquels l'état du patient sera évalué et diagnostiqué plus tard par des médecins. Tous ces détails, de l'admission à la sortie d'un patient, sont résumés dans les notes de synthèse. Cependant, ces sources d'information sont utilisées comme documentation clinique pour les rapports et la facturation plutôt que servir comme connaissances cliniques antérieures pour prédire la progression de la maladie. Pour éviter la perte d'information scientifique contenue dans ces points de données, un algorithme de NLP basé sur l'apprentissage machine sera développé pour prédire l'état du patient en utilisant des notes cliniques stockées dans l'entrepôt de données de recherche à CHUSJ. L'algorithme proposé peut effectivement apprendre une représentation latente de notes cliniques pour en tirer une conclusion sur l'insuffisance cardiaque du patient, qui ne peut pas être décrite par une approche traditionnelle.

Premièrement, notre étude fournit des informations importantes sur l'utilisation de modèles d'apprentissage automatique dans des ensembles de données limités. Plus précisément, nous

avons constaté que des modèles plus petits et plus simples peuvent mieux fonctionner dans de tels contextes. À cette fin, notre cadre combine TF-IDF et MLP-NN, et nous démontrons que la sélection de caractéristiques à partir de l'espace vectoriel de représentation d'apprentissage peut encore améliorer les performances. Notre algorithme proposé apprend efficacement une représentation latente de notes cliniques pour conclure l'état d'insuffisance cardiaque d'un patient, que les approches traditionnelles ne peuvent pas décrire. Nous avons atteint une performance de classification globale avec une précision de 89%, un rappel de 88% et une précision de 89%. De plus, nous avons constaté que l'encodage des points décimaux sous forme de chaîne "DOT" aide à conserver les informations des valeurs numériques dans les notes cliniques, ce qui peut améliorer les performances du modèle.

De plus, la thèse souligne qu'un facteur critique pour améliorer les performances des classificateurs d'apprentissage automatique dans le traitement clinique du langage naturel est le traitement approprié de la caractéristique de l'espace de représentation. Plus précisément, l'étude démontre que l'incorporation d'un auto-encodeur (AE) pendant la formation peut effectivement compresser l'espace des caractéristiques du modèle terme fréquence-fréquence de document inverse (TF-IDF), ce qui en fait un mécanisme efficace pour l'interprétabilité et la transparence dans le système CDSS. La deuxième étape consiste à utiliser un MLP-NN pour prédire l'état de santé en fonction de l'espace de fonctions compressé. Le modèle d'ensemble efficace atteint une précision de 92%, un rappel de 91%, une précision de 91% et un score f1 de 91%, surpassant toutes les approches alternatives.

Enfin, bien que Transformer ait été largement reconnu comme l'approche de pointe en matière de traitement du langage naturel, il est toujours confronté à des limites lorsqu'il est appliqué à une PNL clinique courte et limitée. Nous proposons un cadre simplifié Switch Transformer que nous formons à partir de zéro sur un petit ensemble de données de classification de textes cliniques en français à l'hôpital CHU Sainte-Justine. Nos résultats montrent que les modèles simplifiés de transformateurs à petite échelle fonctionnent mieux que les modèles pré-formés basés sur BERT, tels que DisstillBERT, CamemBERT, FlauBERT et FrALBERT. Le cadre proposé atteint une précision de 87%, une précision de 87% et un rappel de 85%, ce qui surpasse le troisième meilleur modèle basé sur BERT pré-formé, FlauBERT, qui a atteint une précision de 84%, précision à 84 % et rappel à 84 %. Cependant, les transformateurs de commutation ont des limites, telles qu'un écart de généralisation et des minima nets. Pour répondre à ces limitations, nous le comparons à un réseau de neurones perceptrons multicouches pour la classification des petits récits cliniques français et montrons que ce dernier surpasse tous les autres modèles.

Dans l'ensemble, l'étude démontre l'efficacité du cadre proposé et fournit des informations précieuses pour le développement de techniques de PNL en milieu clinique. Il améliore le processus long et coûteux de traitement des maladies, les interventions de santé et la gestion de la prévention à l'unité de soins intensifs pédiatriques de l'hôpital CHUSJ.

Mots-clés: traitement clinique du langage naturel, insuffisance cardiaque, apprentissage automatique, apprentissage par déséquilibre, sélection de fonctionnalités

Development of machine-learning-based natural language processing to detect concept labels in clinical narratives

Thanh-Dung LE

ABSTRACT

Currently, an abundance of data and information are available in the clinical domain. Grasping this opportunity, clinicians have been successfully combining the informative and structured data, which includes laboratory test results, medical imaging, and wearable sensor data, with novel data analytic algorithms to provide pervasive and personalized healthcare. However, considerable constraints are imposed by clinical narrative sources, which are short notes on patients originally written by doctors and physicians. Although the notes are continuously provided and plentifully stored in clinical data warehouses, they are underutilized in practice. The limitation mainly comes from their unstructured or semi-structured format. Fortunately, the deployment of machine learning algorithms in recent years helps to effectively capture the hidden representation of clinical narratives because of its high computational capacity. In particular, the improvement of machine learning performance on clinical notes is continually reinforced by employing natural language processing (NLP) techniques as a data preprocessing step in advance. NLP becomes a necessary approach to overcome the existing challenges of unstructured clinical text notes because it effectively maps the words in unstructured data into a continuously-valued lower dimensional space.

Fortunately, a large data source of clinical notes is currently stored in our Research Data Warehouse at CHU Sainte-Justine (CHUSJ) hospital. There are 7 caregiver notes/patient/day from 1386 patients (containing a dataset of more than 2.5x107 words). These notes are scribed extensively from admission notes, evaluation notes and summary notes. Admission notes outline reasons for admission to intensive care units, historical progress of disease, medication, surgery and additional baseline status of the patient. Daily ailments and laboratory test results are described in evaluation notes, from which patient condition is evaluated and diagnosed later by doctors. All these details from admission to discharge of a patient are outlined in summary notes. However, these information sources are being used as clinical documentation for reporting and billing instead of prior clinical knowledge for predicting disease condition. To prevent the loss of scientific information from these beneficial data points, a machine-learning-powered NLP method is developed to predict patient condition by using clinical notes stored at the Research Data Warehouse at CHUSJ hospital. The proposed algorithm can effectively learn a latent representation of clinical notes to draw a conclusion about a patient's cardiac failure condition which cannot be depicted by traditional approaches.

First, our study provides important insights into using machine learning models in limited datasets. Specifically, we found that smaller and simpler models can work better in such contexts. To this end, our framework combines TF-IDF and MLP-NN, and we demonstrate that feature selection from the learning representation vector space can further improve performance. Our proposed algorithm effectively learns a latent representation of clinical notes to conclude a patient's

cardiac failure condition, which traditional approaches cannot depict. We achieved an overall classification performance with 89% accuracy, 88% recall, and 89% precision. Furthermore, we found that encoding decimal points as a string "DOT" helps retain the information from numerical values in clinical notes, which can improve model performance.

Furthermore, the thesis highlights that a critical factor for improving the performance of machine learning classifiers in clinical natural language processing is the appropriate treatment of the representation space feature. Specifically, the study demonstrates that incorporating an autoencoder (AE) during training can effectively compress the feature space of the term frequency-inverse document frequency (TF-IDF) model, making it an effective mechanism for interpretability and transparency in the CDSS system. The second step involves using an MLP-NN to predict the health status based on the compressed feature space. The efficient ensemble model achieves 92% accuracy, 91% recall, 91% precision, and 91% f1-score, outperforming all alternative approaches.

Finally, while Transformer has been widely recognized as the state-of-the-art approach in natural language processing, it still faces limitations when applied to short and limited clinical NLP. We propose a simplified Switch Transformer framework that we train from scratch on a small French clinical text classification dataset at CHU Sainte-Justine hospital. Our results show that the simplified small-scale Transformer models perform better than pre-trained BERT-based models, such as DistillBERT, CamemBERT, FlauBERT, and FrALBERT. The proposed framework achieves an accuracy of 87%, precision at 87%, and recall at 85%, which outperforms the third-best pre-trained BERT-based model, FlauBERT, which achieved an accuracy of 84%, precision at 84%, and recall at 84%. However, Switch Transformers have some limitations, such as a generalization gap and sharp minima. To address these limitations, we compare it with a multi-layer perceptron neural network for small French clinical narratives classification and show that the latter outperforms all other models.

Overall, the study demonstrates the effectiveness of the proposed framework and provides valuable insights for developing NLP techniques in clinical settings. It improves the time-consuming and costly disease treatment process, health interventions, and prevention management at the Pediatric Critical Care Unit of CHUSJ hospital.

Keywords: clinical natural language processing, cardiac failure, machine learning, imbalance learning, feature selection.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

Acc	Accuracy
Adam	Adaptive Moment Estimation
AE	Autoencoder
ARDS	acute respiratory distress syndromes
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Word
CDSS	clinical decision support system
CEC	Circulation extracorporelle
CHUSJ	CHU Sainte Justine
CIA	Communication intraauriculaire
CIV	Communication intraventriculaire
EMR	electronic medical records
FC	Fréquence cardiaque
FFN	Feed Forward Network
FN	False Negative
FP	False Positive
GaussianNB	Gaussian Naïve Bayes
ICU	Intensive care units

IG	Integrated Gradients
IVRS	Infection des voies respiratoires supérieures
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MLP	Multilayer Perceptron
MLP-NN	Multilayer Perceptron Neural Network
MoE	Mixture-of-exper
MultimonalNB	Multinomial Naïve Bayes
NCA	Neighborhood Component Analysis
NLP	natural language processin
PCA	Principal Component Analysis
PICU	Pediatric Intensive Care Unit
PO	Per os (by mouth)
Pre	Precision
Rec	Recall
RF	Random Forest
SGD	Stochastic Gradient Descent
SOP	Salle d'opération
SVM	Support Vector Machine
TF-IDF	term frequency-inverse document frequency

TN	True Negative
TP	True Positive

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

EF	Ejection Fraction (%)
pro-BNP	Brain Natriuretic Peptide Test (ng/L)
SF	Shortening Ratio (%)

INTRODUCTION

A clinical decision support system focuses on the real-time analysis of the diagnosis and management of patient condition (Berner, 2007). It has been developing and providing a crucial promotion in the personalized healthcare system because more available data are continuously collected and stored. These data sources are decisive points to advance and enhance the efficiency and effectiveness of clinical decision support systems' operations. Consequently, predictive models currently result in preventive treatment and patient diagnosis for healthcare improvement in an intelligent, precise, yet timely manner.

Unfortunately, limitations for the data collection process of the proposed clinical decision support system remain. One of the reasons is that data collection has been designed to document clinical activity for reporting and billing reasons instead of developing new algorithms and/or knowledge. Therefore, many challenges are being faced in critical care data analysis, such as compartmentalization, corruption, and complexity, as described in (Johnson *et al.*, 2016). These challenges arise from the fragmented nature of data sources, leading to difficulties in integrating diverse data streams and data corruption issues due to inaccuracies and noise. Moreover, the inherent complexity of critical care data, characterized by multifaceted variables and intricate relationships, necessitates advanced analytical approaches. Successfully addressing these challenges mandates interdisciplinary collaboration and innovative methodologies, promising advancements in patient care and critical care practices through informed decision-making and tailored interventions. To overcome these challenges, data validation processes of clinical variables must be effectively elaborated in clinical data management of the clinical decision-support system whose data are continuously collected in the critical care unit.

Following the mentioned achievements, a clinical decision-support system at CHU Sainte-Justine Research Center (CHUSJ) is being developed. Two fundamental processes in the workflow of a clinical decision-support system, which involves the collection and process of critical care data,

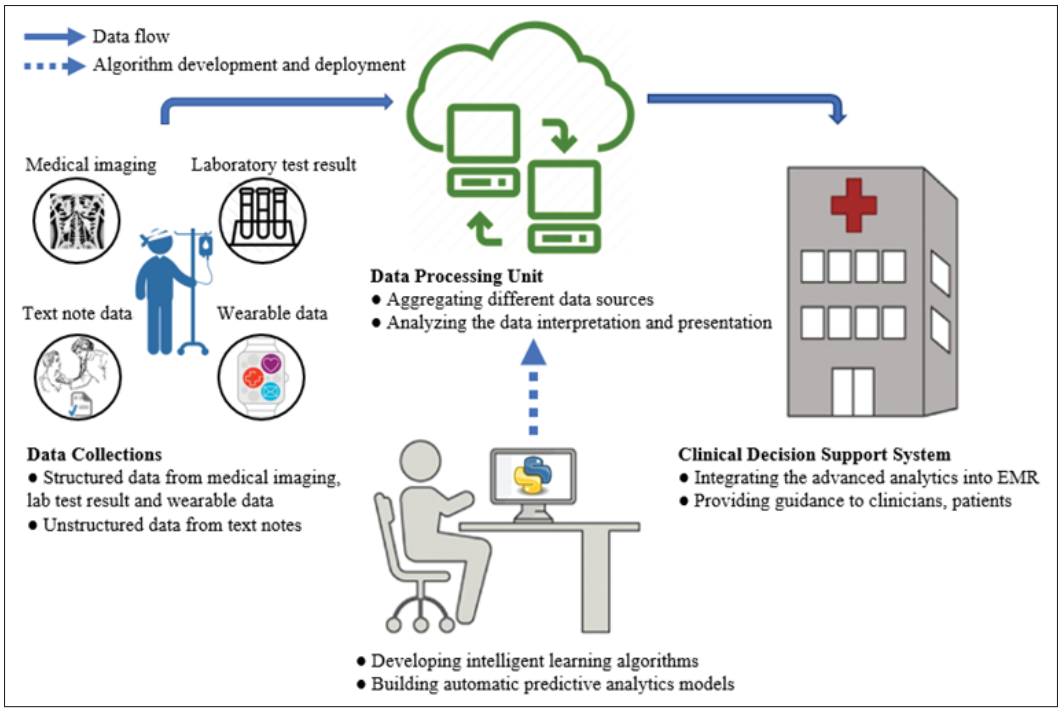

Figure 0.1 Workflow demonstration of a clinical decision-support system at CHUSJ hospital

are shown in Fig. 0.1. First, clinical data is collected and stored in a clinical data warehouse. Second, in the data processing unit, the data are systematically aggregated and processed to convert raw data to machine-readable data. This process helps to analyze the unknown data interpretation and presentation. Consequently, the clinical decision-support system can integrate the advanced analytic result from the data-processing unit and learning algorithms, and clinicians adequately utilize the clinical decision-support system as guidance in early intervention and prevention for healthcare management.

The project points to an approach that could improve and accelerate healthcare adoption and use of information and communication technology. Specifically, one of the targets of the clinical decision support system in CHUSJ is too early to diagnose acute respiratory distress syndromes (ARDS). The primary role of the respiratory system is to facilitate the exchange of gases in our bloodstream. This process hinges on two main actions: inhaling and exhaling. When inhales,

Table 0.1 Details the proposed definitions for those children at risk for pediatric ARDS (Group *et al.*, 2015)

Age	Exclude patients with peri-natal related lung disease			
Timing	Within 7 days of known clinical insult			
Origin of Edema	Respiratory failure not fully explained by cardiac failure or fluid overload			
Chest Imaging	Chest imaging findings of new infiltrate(s) consistent with acute pulmonary parenchymal disease			
Oxygenation	**Non Invasive mechanical ventilation**	**Invasive mechanical ventilation**		
	PARDS (No severity stratification)	Mild	Moderate	Severe
	Full face-mask bi-level ventilation or CPAP ≥5 cm H_20 [2] PF ratio ≤ 300 SF ratio ≤ 264 [1]	4 ≤ OI < 8 5 ≤ OSI < 7.5[1]	8 ≤ OI < 16 7.5 ≤ OSI < 12.3[1]	OI ≥ 16 OSI ≥ 12.3[1]
Special Populations				
Cyanotic Heart Disease	Standard Criteria above for age, timing, origin of edema and chest imaging with an acute deterioration in oxygenation not explained by underlying cardiac disease. [3]			
Chronic Lung Disease	Standard Criteria above for age, timing, and origin of edema with chest imaging consistent with new infiltrate and acute deterioration in oxygenation from baseline which meet oxygenation criteria above.[3]			
Left Ventricular dysfunction	Standard Criteria for age, timing and origin of edema with chest imaging changes consistent with new infiltrate and acute deterioration in oxygenation which meet criteria above not explained by left ventricular dysfunction.			

the diaphragm and intercostal muscles contract, expanding the chest cavity; this causes a drop in lung pressure, drawing air in from our surroundings. This inhaled air brings oxygen, which enters the bloodstream, while carbon dioxide is moved from the blood to the lungs. Upon exhaling, we release this carbon dioxide into the environment. Usually, our lungs supply oxygen to our bloodstream, delivering it to essential organs and removing carbon dioxide (Ware & Matthay, 2000).

However, complications arise with lung injuries or certain viral infections. In such situations, the lungs might not supply enough oxygen to vital organs or effectively remove carbon dioxide from the blood. To compensate, the brain signals additional respiratory muscles to assist. This strain on the respiratory system is called respiratory distress. It's a severe condition; over-relying on these accessory muscles can eventually lead to cardiopulmonary arrest. Recognizing and diagnosing acute respiratory distress early is crucial. Timely detection allows immediate medical

intervention, dramatically increasing the chances of recovery and preventing lasting damage to essential organs (Matthay *et al.*, 2019).

Additionally, ARDS and cardiac failure often present with similar symptoms, making early and accurate diagnosis essential for effective treatment strategies, particularly in critical care units like the Pediatric Intensive Care Unit (PICU). Accurately distinguishing between these conditions can significantly influence patient outcomes, potentially saving lives.

An expert panel on pediatric acute lung injury has posited that a pivotal step in diagnosing respiratory diseases in children is to determine the absence of "cardiac failure." As detailed in Table 0.1, evidence suggests that children exhibiting heart dysfunction still meet all the criteria for ARDS. This overlap in symptoms and criteria, encompassing factors like age, onset timing, and edema origin, makes distinguishing between heart failure and ARDS challenging (Group *et al.*, 2015).

Given the complex interplay between these conditions, a thorough diagnostic approach is indispensable. This often involves a combination of clinical data, laboratory tests, and echocardiography assessments. Pinpointing the absence of cardiac failure becomes a pivotal diagnostic criterion within this framework. By reliably identifying when cardiac failure isn't present, clinicians can refine their ARDS diagnostic accuracy, streamlining the diagnostic process as depicted in 0.2. This clarity can then expedite the right treatment choices, ensuring timely and appropriate medical interventions, ultimately elevating the quality of patient care.

Advantageously, an imaginative data resource of clinical notes is currently stored in our Research Data Warehouse at CHUSJ. As shown in Fig. 0.3, there are 7 caregivers notes/patient/day from 1386 patients (containing a dataset of more than $2.5x10^7$ words). These notes are scribed extensively from admission, evaluation, and summary notes. Expressly, admission notes enfold delineation of indisposition being admitted to intensive care units, historical information of

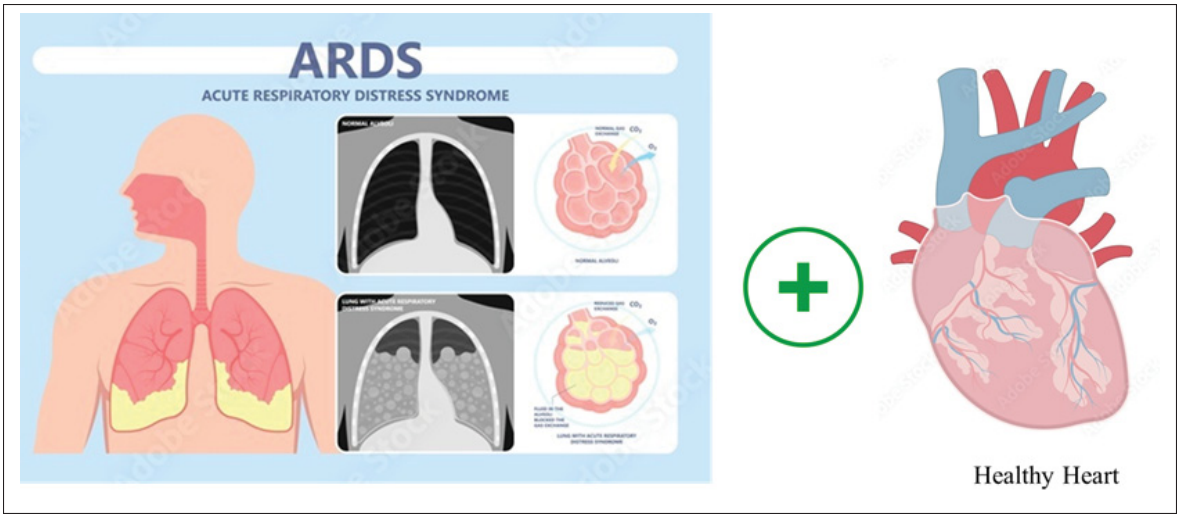

Figure 0.2 Key identification for Acute Respiratory Distress Syndrome

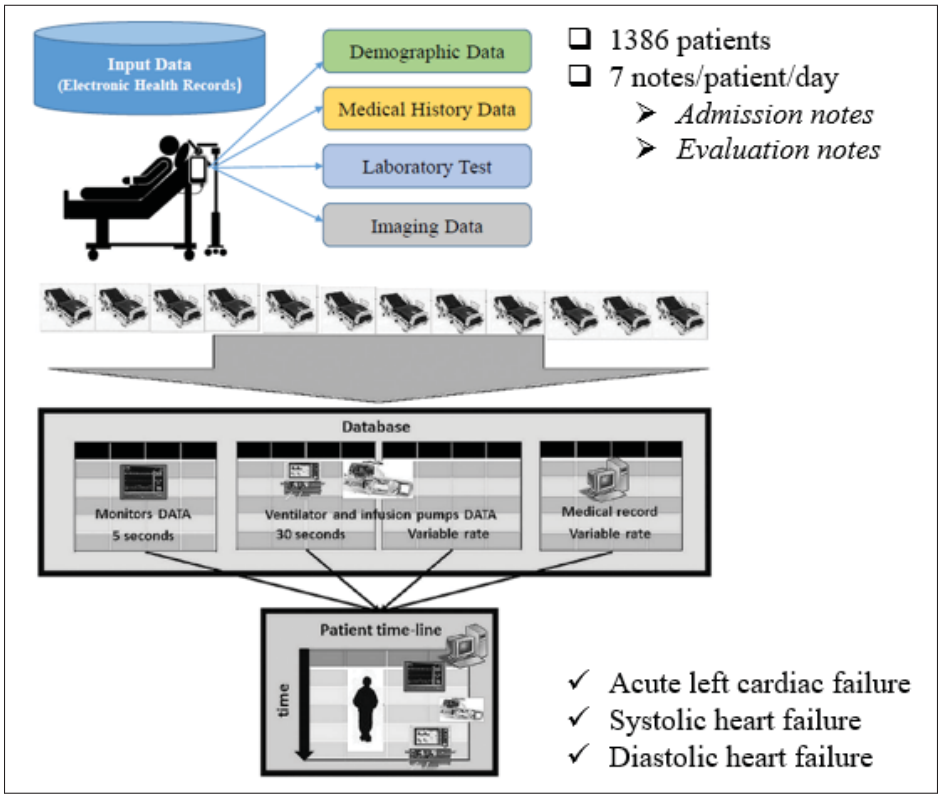

Figure 0.3 Data collection process at CHUSJ

disease, medication, surgery, and the additional baseline status of the patient. Daily ailments and laboratory test results are described in evaluation notes, from whose patient conditions are evaluated and diagnosed later by doctors. Such detailed afflictions from admission to discharge of a patient are adumbrated in summary notes. However, that information source is used as clinical activities for reporting and billing reasons instead of prior clinical knowledge for predicting disease conditions. To prevent the continuous loss of scientific information from these beneficial data points, a machine learning algorithm powering natural language processing inception is developed to prognosticate patient conditions using clinical notes stored at the Research Data Warehouse, CHUSJ. The proposed algorithm can effectively learn a latent representation of clinical notes to conclude patients' cardiac failure condition, which cannot be depicted by the traditional approaches. Consequently, it bounteously extricates from the time-consuming and costly disease treatment process but increases the management improvement for health intervention and prevention at the PICU, CHUSJ.

Based on the stored clinical notes that synthesized all this information, the patient's cardiac condition can be effectively found. As per the recommendations of CHUSJ's medical professionals, the selection of clinical notes for our study is focused on two crucial types: Admission and Evaluation notes. This strategic emphasis stems from recognizing that these note categories, explicitly delving into the medical background, pre-admission disease history, and cardiovascular evaluations, encompass vital information essential for timely cardiac failure detection. The rationale is that Admission notes comprehensively encapsulate the patient's condition upon entry into the PICU, providing an initial overview. Given the context of patients entering the PICU, if we can effectively classify ARDS and cardiac failure, a prompt and precise treatment path can be determined, optimizing the critical time window for medical intervention. This approach holds the potential for real-time decision-making for medical practitioners and saves lives by swiftly and accurately guiding the course of treatment for patients entering the PICU. Then, based on those above inclusion criteria taken from clinical notes in the PICU database, clinical

knowledge representation in detecting cardiac failure is set, as shown in Table 0.2. The clinical knowledge representation summarizes detailed attributes vital to detecting heart failure. As a result, a patient is considered to have a cardiac failure if he/she takes one of the criteria.

Table 0.2 The clinical knowledge representation in detecting cardiac failure

Label
Cardiac failure
Attributes
1. Admission notes: a. **Medical history**: Levosimendan, Milrionone or Dobutamine. b. **History of diagnosed terms**: cardiomyopathie dilatee, choc cardiogenique, defaillance cardiaque gauche aigue, defaillance cardiaque gauche chronique, defaillance cardiaque post-operatorie (LCOS), surcharge liquidienne (hypervolemie), myocardite. 2. Evaluation notes: a. **Evolution par systeme** (Cardiovascular): FE (ejection fraction $<$50%) and/or FR (shortening ratio $<$25%). b. **Laboratory test result**: pro-BNP ng/L ($>$ 1000)

Table 0.2 shows the list of golden indicators to classify the patient with cardiac failure. Technically, in the medical history, we will extract the information for *Levosimendan, Milrinone, Dobutamine*. That medication information is a surrogate to the gold standard because the medication list can be retrieved from syringe pump data, prescriptions, and notes. If any listed medication is present, there is a cardiac failure. However, existing information that helps diagnose cardiac failure is not always readily available electronically.

Besides, we also base on the diagnosed terms to detect the patient with cardiac failure. There are totally six terms that are selected including cardiomypathie dilatee, cho caridogenique, defailance cardiaque gauche aigue, defailance cardiaque gauche chronique, delaillance cardiaque post-operatorie (LCOS), surcharge liquidienne (hypervolemie).

Furthermore, we will take the value of ejection fraction (FE) and shortening fraction (FR) from the cardiovascular evolution notes. The EF ($< 50\%$) refers to the amount, or percentage, of blood that is 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. While the FR ($< 25\%$) 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, we concentrate on the pro-BNP ng/L ($>$1000) from laboratory test results. The brain natriuretic peptides (BNP) are peptides (small proteins) that are either hormones or part of the peptide that contained the hormone at one time. They are continually produced in small quantities in the heart and released in larger quantities when the heart senses that it needs to work harder. This supports fluid retention and volume expansion in the arteries and veins. They are useful in acute settings for differentiating HF from pulmonary causes of respiratory distress.

Consequently, Table 0.3 shows an example of real data and labels from CHU Sainte Justine, with critical indicators for cardiac failure diagnosis. However, as all the information that helps diagnose cardiac failure is not readily available electronically, we will develop a machine learning algorithm based on NLP that automatically detects the desired concept label from clinical notes. Specifically, the algorithm can detect whether a patient has a cardiac failure or healthy condition in terms of lacking gold indicators from the notes. Technically, in such a condition, the proposed algorithm can effectively learn a latent representation of clinical notes, which traditionally rule-based approaches cannot depict.

Continuing the trajectory of refining ADRS diagnosis at CHUSJ, a series of investigations have been undertaken, encompassing the evaluation of chest X-ray infiltrations (Yahyatabar *et al.*, 2023) and the estimation of the oxygenation index through diverse statistical analyses and neural network methodologies (Sauthier, Tuli, Jouvet, Brownstein & Randolph, 2021) shown in Fig.

Table 0.3 An example of patient with cardiac failure from CHUSJ

Attributes	Clinical Findings
Medical background	# **Cardiomyopathie dilatée** Suivi en cardio HSJ - en attente ***de greffe cardiaque*** Dernière écho 20/08/2018: VD taille normale. VG sévèrement dilaté (64.8mm en diastole et 58.3mm en systole) et hypokinésie. ***FR 10% et FE 21%***. # **Retard de la motricité grossière** Carvedilol 3.2 mg BID Lasix 7 mg BID Captopril 7 mg TID Lansoprazole 7.5 mg die Aldactone 5 mg TID (essai récent en cardio)
History of disease	Patiente suivi pour une ***CMD en attente d'une greffe cardiaque.*** Depuis plusieurs mois, stagnation pondérale et diminution de l'énergie. Vu en cardiologie 27/08/2018 et malgré une thérapie pharmacologique aggressive pour optimiser sa FE, elle stagne autours de 20-21%. Risques expliqués aux parents relatifs à la possibilité de mort subite et de détérioration sévère et suggestion d'initier une perfusion de milrinone comme traitement d'une ***insuffisance cardiaque severe due a la CMD.*** Parents initialement hésitant. Retour en clinique 28/08/2018 et acceptent admission à USIP pour initier perfusion de milrinone. Enfant malgré tout en bon état général. Pas de fièvre, pas de Sx IVRS, GI ou GU. 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.
Cardiovascular evaluation	# **Cardio-vasculary evolution note 1** ***Milrinone 1 mcg/kg/min*** ajustée pour son poids 5 decembre Carvedilol 4 mg BID (dose max) pas ajusté pour son poids Ivabradine (x 31-10) 1.2 mg bid - aucune augmentation prévue pour le moment pas ajuster pour son poids Pas d'arythmie depuis 14/11 ETT du 12/12: FEVG stable à 15%, IM modérée idem. Pas IT exploitable. pas thrombus, pas épanchement # **Cardio-vasculary evolution note 2** ***Milrinone 0.7 mcg/kg/min*** Carvedilol 3.2 mg q12h FC 125-150 Chaude, pouls 2+ femoraux B/L Refil $<$2 sec ECG 29/08/2018: tachycardie sinusale avec évidence de dilatation du VG.

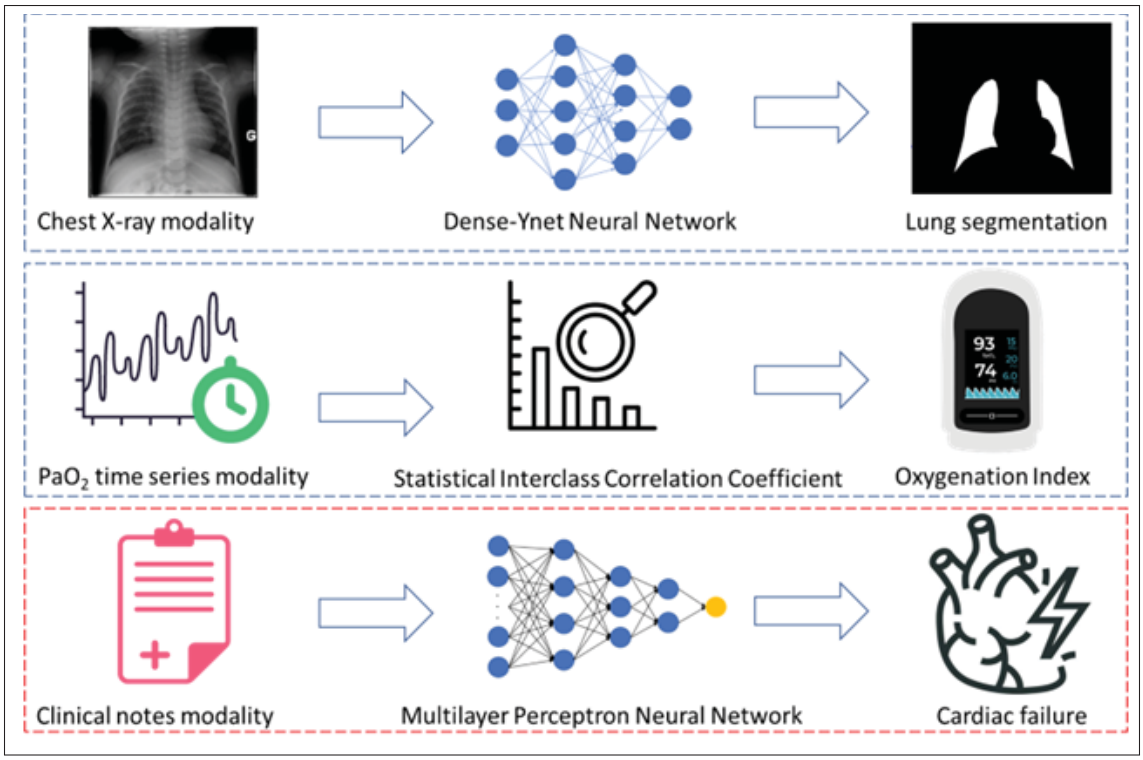

Figure 0.4 Projects help detect ARDS from the CDSS at CHUSJ

0.4. Building upon this foundation, the primary objective of the current study is to forge ahead in the realm of advancement by formulating a machine learning algorithm rooted in natural language processing. This algorithm's core functionality is to autonomously discern whether a patient's condition leans towards "cardiac failure" or "healthy," effectively leveraging physician notes stored within the Research Data Warehouse of CHUSJ. A holistic framework emerges by accomplishing this, revolutionizing real-time ADRS diagnosis at CHUSJ. This transformative approach can significantly enhance clinical decision-making processes, offering timely and precise insights into patient conditions and ultimately contributing to optimizing patient care within the critical care domain. Technically, the main objective consists of two sub-objectives, 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 the model parameters instead of making point estimates of its values.

This thesis comprises four chapters. Chapter 1 offers a comprehensive literature review on clinical natural language representation learning, machine learning classifiers, and their application in detecting cardiac failures. This chapter provides an overview of the latest clinical narrative classifications, which serve as the primary focus of this thesis. The subsequent three chapters delve into each of the previously mentioned objectives, with their corresponding literature reviews presented within.

The following section presents the four primary contributions of this thesis, organized by their respective chapters and peer-reviewed publications:

Contribution 1: One of the main contributions of this thesis is demonstrating the effectiveness of using a multilayer perceptron neural network classifier for small clinical narrative datasets compared to conventional classifiers and pretrained-based deep learning models. This study shows that by retaining and encoding numeric values, the MLP classifier achieves better results for the classification task without losing any valuable information. This work highlights the advantages of using MLP classifiers in clinical narrative classification, mainly when dealing with limited data.

In our experiments, we assessed three learning representations for short-text classification with limited data: Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings. BoW and TF-IDF demonstrated superior capabilities in retaining relevant information from the notes compared to word embeddings. Notably, TF-IDF showcased

the highest classification accuracy, particularly for very short texts (less than 20 words per sample). Although our study's scope was limited to samples of around 80 words each, TF-IDF outperformed neural word embeddings, underscoring its robustness in handling short text data.

We further evaluated several classifiers for their effectiveness in short text classification, including Random Forest (RF), Gaussian Naïve Bayes (GaussianNB), Multinomial Naïve Bayes (MultimonalNB), Logistic Regression (LR), Support Vector Machines (SVM), and K-nearest neighbor. Our comparative analysis revealed that RF, MultimonalNB, and SVM all yielded accuracy rates below 75%. On the other hand, LR, GaussianNB, and Multi-layer Perceptron Neural Network (MLP-NN) exhibited superior performance, making them more suitable choices for short-text classification than the aforementioned classifiers.

This work has been published in the following peer-reviewed journal papers:

Thanh-Dung Le, Rita Noumeir, Jérôme Rambaud, Guillaume Sans, and Philippe Jouvet, "Detecting of a Patient's Condition From Clinical Narratives Using Natural Language Representation," *IEEE Open Journal of Engineering in Medicine and Biology, 3 (2022): 142-149*.

Contribution 2: Another significant contribution of this thesis is the development of an autoencoder learning algorithm that addresses the issue of sparsity in the feature space representation of a small clinical narrative dataset. Unlike other approaches, the algorithm's lossless compression capacity enables it to learn the most optimal representation of the training data. This feature leads to a considerable improvement in its downstream classification performance, which is not possible with deep learning models. The proposed method offers a promising alternative for improving the classification accuracy of small clinical narrative datasets.

Addressing the challenge of sparsity often leads researchers to dimension-reduction techniques. Two popular methods are Principal Component Analysis (PCA) and Neighborhood Component Analysis (NCA), favored for their simplicity among various dimensionality reduction techniques.

We investigated the potential of PCA and NCA to mitigate sparsity due to the aforementioned advantages. However, neither approach significantly enhanced classification performance. This outcome underscores the inherent limitation of these methods, which attempt to approximate a feature subspace to optimize class separability linearly.

On the other hand, autoencoders (AE) with non-linear activation functions demonstrated superior efficacy in compressing the sparse TF-IDF representation space. We further evaluated this compressed representation's effectiveness for reconstruction.

Several machine learning classifiers, including MLP-NN, LR, GaussianNB, RF, Multinomial Naive Bayes, and SVM, were tested for classification tasks. Among these, the MLP-NN classifier showcased the best performance, recording 92% accuracy, 91% precision, 91% recall, and 91% F1 score. This is a notable 2-3% improvement across each evaluation metric compared to the general classification performance achieved in a sparse TF-IDF feature space (89% accuracy, 89% precision, 88% recall, and 88% F1 score). These results validate that the AE methodology effectively addresses sparsity by compressing the TF-IDF feature space, boosting the MLP-NN classifier's performance and making it more resilient than other techniques.

Furthermore, the behavior of AEs with limited data is consistent even with larger datasets, per the information-theoretic framework. This framework provides insights into the workings of AEs and pinpoints scenarios where they achieve optimal compression. Our study also delves into understanding the behavior of AEs during contraction by examining the mutual information across each hidden layer in both the encoder and decoder segments.

This work has been published in the following peer-reviewed journal paper:

Thanh-Dung Le, Rita Noumeir, Jérôme Rambaud, Guillaume Sans, and Philippe Jouvet, "Adaptation of Autoencoder for Sparsity Reduction From Clinical Notes Representation Learning," *IEEE Journal of Translational Engineering in Health and Medicine,* 11 (2023): 469-478.

Contribution 3: The third contribution of this Ph.D. thesis is the exploration of the Switch Transformer model for clinical text classification. Our findings demonstrate that this model shows promising results in improving performance over pre-trained BERT-based models. Although it did not outperform a small MLP-NN neural network, we believe that this framework has the potential to enhance accuracy on small French clinical narrative datasets. Our work serves as a proof-of-concept for the application of Switch Transformer in clinical natural language processing and highlights its potential for future research.

In our study, we evaluated the performance of six classifiers for a binary classification task: CamemBERT, DistillBERT, FlauBERT, FrALBERT, Transformer, and Switch Transformer. Our findings revealed that diligent hyperparameter optimization could lead the Transformer models to outperform the pre-trained BERT-based counterparts. Given that Transformer models typically require substantial amounts of data for practical training, in our study, we divided the dataset into 80% for training, 10% for validation, and the remaining 10% for testing. To evaluate the effectiveness of our approach, we utilized several metrics such as accuracy, precision, recall, and the F1 score. When trained from scratch, the Switch Transformer model emerged as the top performer, boasting an accuracy score of 87%, precision and recall rates of 87% and 85%, respectively, an F1 score of 86%, and an AUC of 92%.

However, even these commendable metrics could not surpass the results of a meticulously engineered MLP-NN classifier. This classifier, designed with specialized techniques like numerical decoding, negation tagging, and sparsity reduction, outdid the Transformer-based models. One contributing factor is the discernible generalization gap observed in Transformer models during training and validation, especially for longer sequences. We further deduced that these models struggled on the clinical dataset because of their limitations in accurately contextualizing and interpreting real-world data. Clinical tasks often present a low signal-to-noise ratio, and during the training phase, Transformers may divert their attention from

pivotal keywords. Consequently, it remains an open question whether Transformer models can consistently perform across diverse scenarios within the clinical domain.

This work was submitted in the following peer-reviewed journal paper (Under review):

Thanh-Dung Le, Philippe Jouvet, and Rita Noumeir, "A Small-Scale Switch Transformer and NLP-based Model for Clinical Narratives Classification" submitted to *IEEE Journal of Biomedical and Health Informatics in March 2023.*

In this pioneering Ph.D. thesis, the experiments in this study systematically progress from more straightforward methodologies to advanced deep learning techniques. Surprisingly, given the constraints of our dataset (limited in size and highly specialized in the medical domain), pre-trained language models did not offer significant performance enhancements. This thesis tackles a challenging real-world dataset for which there is no benchmark data or prior published comparisons.

Technically, a comprehensive end-to-end framework has been meticulously developed to revolutionize the detection of patient conditions from clinical notes at CHUSJ. The culmination of this research journey has illuminated the most promising candidate for statistical learning representation, namely TF-IDF, coupled with scalable machine learning in the form of MLP-NN. This amalgamation has demonstrated unparalleled potential in transforming the landscape of patient condition detection.

The significance of the framework is further augmented by the inclusion of adept engineering strategies in data preprocessing, elevating the performance of the classification end task. The adept handling of various approaches for encoding vital sign numeric values, coupled with the strategic implementation of autoencoders, has been pivotal in enhancing the accuracy and efficiency of the overall system. These data preprocessing techniques have proven instrumental in mitigating challenges associated with feature sparsity in the representation feature space.

While this thesis has unveiled remarkable achievements, it has also shed light on the potential of Transformers as a promising solution for patient condition detection. Nonetheless, formidable challenges and limitations associated with Transformers have surfaced. The identified generalization gap and the Transformer's inherent limitation in effectively comprehending shorter texts are pivotal areas that warrant further exploration and refinement.

In essence, this Ph.D. thesis marks a significant milestone in healthcare informatics. The novel end-to-end framework, powered by the symbiotic synergy of statistical learning representation and scalable machine learning, presents a transformative approach to patient condition detection. As the healthcare landscape evolves, the insights gleaned from this research serve as a springboard for future endeavors to advance patient care's accuracy, scalability, and efficacy through innovative data-driven methodologies.

CHAPTER 1

LITERATURE REVIEW

Recent advancements have seen a convergence of machine learning and natural language processing (NLP) to enhance the interpretation of clinical notes, notably through temporal and analytical reasoning (Sheikhalishahi *et al.*, 2019). This synergy has applications across various clinical domains, such as early diagnosis, treatment intervention, and predicting patient readmissions.

The combination of machine learning (ML) and NLP has become increasingly popular in learning from clinical notes by facilitating temporal and analytical reasoning (Sheikhalishahi *et al.*, 2019). This combination has been applied to a wide range of clinical applications, including early diagnosis (Shi *et al.*, 2016; Huddar *et al.*, 2016; Soguero-Ruiz *et al.*, 2014), treatment intervention (Liu *et al.*, 2019a; Suresh *et al.*, 2017), and readmission prediction (Rumshisky *et al.*, 2016; Curto, Carvalho, Salgado, Vieira & Sousa, 2016; Agarwal, Baechle, Behara & Zhu, 2017). Compared to conventional statistical learning models for clinical text processing, machine learning-based natural language processing has proven to be a dominant method because of its applicability in real practice. For instance, a system for diagnosing common diseases such as hypertension, diabetes, and chronic obstructive pulmonary was developed and found to be feasible and effective for use in Huangshi Central Hospital (Yang *et al.*, 2018). Additionally, Mayo Clinic Research Center developed an automated system for identifying peripheral arterial disease cases from clinical narratives (Afzal *et al.*, 2017), while Massachusetts General Hospital successfully predicted early readmission by using a neural language model (Rumshisky *et al.*, 2016). At Harvard Medical School, a machine learning-based natural language processing classifier was developed to classify medical subdomains, which integrated deep learning algorithms and distributed word representation (Weng, Wagholikar, McCray, Szolovits & Chueh, 2017).

These advancements underscore that ML-integrated NLP has become instrumental in knowledge discovery within the clinical sector, especially when processing clinical text notes. Alternative

approaches pale in comparison to the results achieved with this combined methodology. This evolution addresses the decade-old question, "What can natural language processing do for clinical decision support?" (Demner-Fushman, Chapman & McDonald, 2009), by delivering effective solutions and insightful answers. Two primary factors have catalyzed the aforementioned advancements in clinical note interpretation: the capability of NLP for feature extraction and the representation learning abilities of neural networks.

First, NLP and Feature Extraction: NLP can quickly learn feature extraction from unstructured notes. The key to understanding clinical note features is word representation, which represents words quantitatively and then transforms those features into machine-readable or structured data. There are many ways to represent words quantitatively, such as fixed word representation that does not assume semantics and similarity of words (one-hot representation, co-occurrence matrix representation), and word embedding through distributional word representation that incorporates semantics and similarity information of words into embedding (Word2Vec (Mikolov, Sutskever, Chen, Corrado & Dean, 2013), GloVe (Pennington, Socher & Manning, 2014)). These word representation techniques are popularly used for clinical notes feature extraction (Liu *et al.*, 2019a; Weng *et al.*, 2017; Fan & Zhang, 2018). Consequently, one of NLP's strengths is its ability to extract features from unstructured notes. The crux of comprehending clinical note features revolves around word representation — a technique that quantitatively defines words and transforms them into structured, machine-readable data. There are several approaches to achieve this:

1. Fixed Word Representations: These methods, like one-hot representation and co-occurrence matrix representation, represent words in a quantitative manner but don't inherently capture semantic nuances or similarities between words.
2. Distributional Word Representations: Techniques like Word2Vec (Mikolov *et al.*, 2013) and GloVe (Pennington *et al.*, 2014) go beyond mere quantification, embedding semantic and similarity information into the representation. Such word representation methods have been widely adopted for feature extraction from clinical notes.

Second, Neural Networks and Representation Learning: ML can represent high-level abstraction by learning complicated functions (Bengio, Courville & Vincent, 2013). Hence, the interdisciplinary field of machine learning for biomedical data has been accelerating due to the availability of well-annotated data and greater clinician involvement. The achievements of ML in healthcare have been extensively analyzed from clinical collaboration, clinical data availability, clinical conditions and tasks, and ML methods, as summarized in (Beaulieu-Jones *et al.*, 2019). Moreover, ML is encouraged to be adopted in healthcare because ML methodologies computationally learn the best treatment decisions and provide guidelines to build a successful end-to-end smart healthcare system (Kreimeyer *et al.*, 2017; Young, Hazarika, Poria & Cambria, 2018). The push towards incorporating ML in healthcare is not just due to its computational accomplishments in determining optimal treatment decisions but also because it provides a framework for building holistic, smart healthcare systems.

In case of resourceful data availability, the state-of-the-art ML-based NLP is focused on using deep learning (DL) for clinical notes to overcome the abovementioned limitations (Pham, Tran, Phung & Venkatesh, 2017; Rajkomar *et al.*, 2018). For instance, deep learning models like Convolutional Neural Networks (CNN) have shown exceptional performance in predicting cardiac failure, achieving an F1 score of 0.756 compared to the conventional approach of Random Forest, which achieved an F1 score of 0.674 (Liu *et al.*, 2019b). Similarly, combining word2vec and deep learning has yielded the best performance for predicting multiple chronic diseases, such as cerebral infarction, pulmonary infection, and coronary atherosclerotic heart disease, with an average accuracy and F1 score exceeding 90% (Shi *et al.*, 2016).

However, while DL architectures generally perform well on large-scale datasets with short texts, they do not necessarily outperform conventional approaches like Bag-of-Words (BoW) on smaller datasets with longer clinical notes (Li *et al.*, 2018). For instance, automatic methods for extracting the New York Heart Association classification from clinical notes (Zhang *et al.*, 2017) have found that the support vector machine (SVM) with n-gram features achieved the best performance, with an F-measure of 93%. Similarly, a study by Agarwal et al. (Agarwal *et al.*, 2017) showed that combining BoW and the Na"ive Bayes classifier on clinical notes

for predicting hospital readmission yielded an area under the curve (AUC) of 0.690. Finally, Fodeh et al. (Fodeh, Li, Jarad & Safdar, 2019) found that with a small dataset, TF-IDF and BoW techniques performed better than other techniques for classifying coronary microvascular dysfunction.

Despite the significant achievements in ML-based natural language processing for clinical text knowledge extraction, concerns remain regarding the trajectory of clinical natural language processing research (Young *et al.*, 2018; Sheikhalishahi *et al.*, 2019). One limitation is learning the semantic and syntactic structure of feature extraction for clinical texts. Combining semantic and syntactic information can lead to conflation deficiencies when a word level is tightened to the semantic and syntactic level. This challenge is even more difficult to mitigate in clinical notes written in languages other than English (Névéol, Dalianis, Velupillai, Savova & Zweigenbaum, 2018). In short, ML-enhanced NLP has made significant strides in extracting knowledge from clinical texts, but critical concerns remain about the trajectory of clinical NLP research, as highlighted in (Kreimeyer *et al.*, 2017).

1. **Feature Learning from Clinical Texts**: One main challenge is extracting semantic and syntactic structures from clinical text representations. Traditional text feature extraction methods often employ count-based strategies, such as one-hot representations or co-occurrence matrices. These methods can require extensive manual effort to produce meaningful representations. Moreover, it's been established that not all textual data equally contribute to the meaningfulness of clinical notes. However, when the dataset is small, word embedding techniques like Word2Vec or GloVe struggle to generalize in the unique language landscape of clinical notes.
2. **Deep-Supervised Learning Limitations**: Deep-supervised learning models have limitations, particularly when used for classification. In large datasets, these models learn differences between actual data instances and synthesized instances produced by the learning rules. In scenarios with limited data, clinical text's actual underlying probability distribution remains obscured. Consequently, models trained under these conditions can have a deficiency in discriminative learning. This is of particular concern in the medical

field, where machine-learning algorithms must align with expert medical insights. It's crucial to articulate why a specific learning algorithm is trustworthy, especially when most discriminative learning is centered on curve fitting.

As mentioned, DL requires a large amount of data to achieve good generalization capability, which may not always be available (Paleyes, Urma & Lawrence, 2020). Study (Kumar, Recupero, Riboni & Helaoui, 2020) proposes an alternative approach to address the issue of small datasets, but it either removes vital sign numeric values or does not provide information on how to handle them. To improve the effectiveness of neural network-based natural language processing, the semantic enrichment of clinical notes and the classification deficiency of supervised learning must be significantly strengthened. Therefore, the project will adopt two strategies: 1) focus on learning the underlying structure of clinical text by adapting data engineering and 2) improve the interpretability of learning representation using a simpler neural network. Recent studies (Wang, Zhou, Jin, Liu & Lu, 2017; Fodeh *et al.*, 2019) have demonstrated that combining statistical learning representations and traditional ML techniques yields superior results for classifying clinical notes for smaller datasets. In light of this evidence, this study will concentrate on the following statistical learning representations and conventional ML classifiers:

1.1 Clinical Natural Language Representation Learning

There is no doubt about the effectiveness of neural word embedding. The study (Shi *et al.*, 2016) confirms that word2vec representation has been successfully used for various disease classifications from medical notes. Especially for the French clinical notes, the study (Dynomant *et al.*, 2019) shows that word2vec and GloVec effectively embed the clinical notes. 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 (Agarwal *et al.*, 2017; Li *et al.*, 2018; Zhang *et al.*, 2017; Fodeh *et al.*, 2019) 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. Therefore, we

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

1.1.1 Bag-of-Words

A bag-of-words model (Joachims, 1998), or BoW for short, represents text that describes the occurrence of words within a document. Technically, if documents have similar content, they will be similar. It extracts features from the text by considering each word count as a feature from a vocabulary of known words. There is a theoretical analysis for understanding the BoW model (Zhang, Jin & Zhou, 2010), which proves that the success of BoW representation is not by using a heuristic clustering process but by a statistical approach based on statistical consistency. However, in the BoW representation, any two different words drawn from the vocabulary are treated equally if they are assigned the same topic; in reality, it neglects much correlation information among words.

1.1.2 TF-IDF

The TF-IDF, first introduced in (Salton & Yang, 1973), stands for term frequency (TF) × inverse document frequency (IDF). TF-IDF weighting is commonly used in information retrieval for text mining. The intuition is that term importance increases with the term's frequency in the text, but its frequency neutralizes it in the domain of interest. Given a collection of terms $t \in T$ that appear in a set of N documents $d \in D$, each of length n_d, tf-idf weighting is computed as:

$$tf_{t,d} = \frac{f_{t,d}}{n_d} \tag{1.1}$$

$$idf_t = \log \frac{N}{df_t} \tag{1.2}$$

$$W_{t,d} = tf_{t,d} \times idf_t \tag{1.3}$$

where $f_{t,d}$ is the frequency of term t in document d, and df_t is the document frequency of term t, the number of documents in which term t appears. Several variations were offered, including normalizing $tf_{t,d}$, and optional weighting schemes.

TF-IDF and its variations do not only focus on the fundamental statistical relation of note representation. But, its strong theoretical arguments are explained for its heuristic (Robertson, 2004), and as a probabilistic theoretical explanation (Havrlant & Kreinovich, 2017). And, it is shown that if TF-IDF is treated carefully with the bias, it will manifest better performance in representation learning.

1.1.3 Neural Word Embeddings

The neural word embedding was introduced in (Mikolov *et al.*, 2013), and named the word2vec model. Each word will be represented by two sets of vectors, u_w, and v_w. u_w is used when word w is the context word, and v_w is used when word w is the center word. Using these two vectors, the probability for the central target word w_o and context word w_c for each word w will look like this:

$$pw_o \mid w_c = \frac{\exp u_o^\top v_c}{\sum_{i \in V} \exp u_i^\top v_c} \tag{1.4}$$

The key to the success of word2vec is that it can compute the logarithmic conditional probability for the central word vector and the context word vector. Then, its computation obtains the conditional likelihood for all the words in the dictionary given the context word w_c, and is trained by a neural network.

$$\log pw_o \mid w_c = u_o^\top v_c - \log \left(\sum_{i \in V} \exp u_i^\top v_c \right) \tag{1.5}$$

where $V = \{0, 1, \ldots, |V| - 1\}$ is the vocabulary index. After the training, for any word in the dictionary with index w_i, we will get its two-word vector sets v_i and u_i . In applications of NLP, the central target word vector in the skip-gram model is generally used as the representation vector of a word.

In this study, we adapted the instruction of how to generate a good neural word embedding (Lai, Liu, He & Zhao, 2016). Significantly, the three critical components in training word embeddings, including the model, corpus, and training parameters, are well customized. In addition, hyper-parameter choices are significant in neural word embedding systems; therefore, we also controlled the effects of data size and frequency range on distributional semantic models based on the recommendations from (Levy, Goldberg & Dagan, 2015).

The neural word embedding only focuses on one-hot representation and ignores the morphological knowledge. However, morphological knowledge can help decrease the training time, as shown in (Santos, Macedo, Bispo & Zanchettin, 2020) on a dataset with 1 billion tokens.

1.2 Machine Learning Classifiers

Deep learning has proven its superiority to representation learning and classification problems from various data structures such as medical imaging, time-series data, and clinical natural language (Otter, Medina & Kalita, 2020). Unfortunately, not all cases can apply deep learning, especially with limited data.

When the ratio value for the number of samples/number of words per sample is small ($<$ 1500), a small MLP-NN that takes n-grams as input performs better or at least as well as deep learning-based sequence models. Besides, an MLP-NN is simple to define and understand and takes less computation time than sequence models. A detailed explanation of using an MLP-NN in medical analysis can be seen from (Pasini, 2015).

Several classifiers have been trained for short text classification; this includes Random Forest, Gaussian Naïve Bayes, Multinomial Naïve Bayes, logistic regression, support vector machines,

and K-nearest neighbour. The experimental results from (Maimon & Rokach, 2014; Wang *et al.*, 2017) confirm that logistic regression, and generative Naïve Bayes perform much better than the other classifiers.

Furthermore, we implemented and compared all the above-mentioned methods; the result of Random Forest, MultimonalNB, and support vector machine was less than 75% for accuracy. Again, the result shows that only logistic regression, Gaussian Naïve Bayes and neural network multilayer perceptron are comparable and perform better than Random forest, multimodal, and support vector machine classifiers.

Moreover, study (Ng & Jordan, 2002) evaluated different classifiers' performance, including discriminative and generative learning approaches, particularly for small datasets. And, it confirms that the discriminative logistic regression algorithm has a lower asymptotic error, while the generative Naïve Bayes classifier converges more quickly. Therefore, in this study, we choose with three different machine learning classifiers logistic regression, Gaussian Naïve Bayes, and MLP-NN.

Here we consider a binary classification problem. Given n training samples $D = x_1, y_1 ..., x_n, y_n$ where $x_i \in \mathcal{R}^p$ is a p-dimensional column vector and label $y_i \in \{0, 1\}$.

1.2.1 Logistic Regression (LR)

LR uses a logistic function to model the probability of a binary dependent variable (particular class or event) such as unhealthy/healthy. Therefore, it is widely used in most medical fields (Tolles & Meurer, 2016). We first write the logistic function as follows:

$$px; w = py = 1 \mid x; w = \frac{1}{1 \ \exp\left(-w^T x\right)} \tag{1.6}$$

where w is the weight vector of coefficients, and p is a sigmoid function. Here, we assume that the n training examples are generated independently. Finally, we thus can obtain the following log-likelihood.

$$\begin{aligned} \ell w &= \sum_{i=1}^{n} \log p\left(y_i \mid x_i; w\right) \\ &= \sum_{i=1}^{n} \left\{ y_i w^T x_i - \log\left(1 \ \exp\left(w^T x_i\right)\right\}\right. \end{aligned} \tag{1.7}$$

1.2.2 Gaussian Naïve Bayes (GaussianNB)

GaussianNB algorithm for classification is a set of supervised learning algorithms based on applying Bayes' theorem with the "Naïve" assumption of conditional independence between every pair of features given the value of the class variable. Bayes' theorem states the following relationship, given class variable Y and dependent feature vector x_1 through x_n:

$$py \mid x_1, \ldots, x_n = \frac{pypx_1, \ldots, x_n \mid y}{px_1, \ldots, x_n} \tag{1.8}$$

Using the Naïve conditional independence assumption that

$$px_i|y, x_1, \ldots, x_{i-1}, x_{i1}, \ldots, x_n = px_i|y, \tag{1.9}$$

for all i, this relationship is simplified to

$$py \mid x_1, \ldots, x_n = \frac{py \prod_{i=1}^{n} px_i \mid y}{px_1, \ldots, x_n} \tag{1.10}$$

Since $px_1, \ldots, x_n$ is constant given the input, we can use the following classification rule: $Py \mid x_1, \ldots, x_n \propto py \prod_{i=1}^{n} px_i \mid y$. Then

$$\hat{y} = \arg\max_{y} py \prod_{i=1}^{n} Px_i \mid y \tag{1.11}$$

and we can use the maximum a posteriori estimation to estimate py and $px_i \mid y$; the former is the relative frequency of class y in the training set. However, the difference between GaussianNB and Naïve Bayes classifiers is mainly based on $px_i \mid y$ distribution. The likelihood of the features is assumed to be Gaussian with mean μ and variance σ^2 for GaussianNB (Hastie, Tibshirani, Friedman & Friedman, 2009):

$$px_i \mid y = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{x_i - \mu_y{}^2}{2\sigma_y^2}\right) \tag{1.12}$$

1.2.3 Multilayer Perceptron Neural Network (MLP-NN)

The earliest work in the field of “neural network" is attempted to understand, model, and emulate neurological function and learning in brains (McCulloch & Pitts, 1943). Since then, a commonly used neural network has been the MLP-NN. In an MLP-NN, the neurons are structured into layers, consisting of at least three layers: the input layer, hidden layer or layers, and an output layer (Demuth, 2014). Typical MLP-NN networks are feedforward neural networks where the computation is carried out in a single direction from input to output.

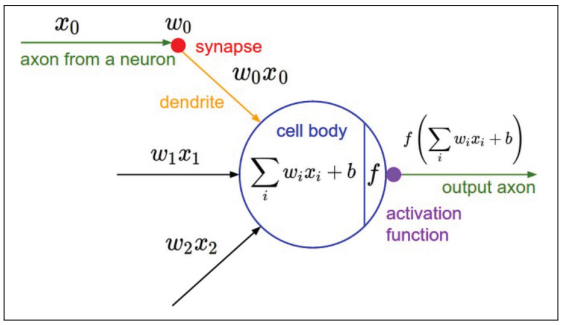

Figure 1.1 An mathematical model for a biologicallyinspired neural network

As shown in Fig. 1.1 (Karparthy, 2020), x_i is the i^{th} input (axon from a neuron) to an input neuron, and a weight w_i is the effect of the i^{th} synapse on the neural network. Then, the total impact of the input on the synapse from the cell body is:

$$\sum_{i}^{n} w_i x_i \;\; b \tag{1.13}$$

Following this affine transformation of the weighted sum of its input, a nonlinear activation function $f\cdot$ is defined for the cell output. This nonlinear activation function $f\cdot$ generally enables the MLP-NN to learn and solve nonlinear problems. It is proven that ReLU activation in combination with stochastic gradient descent optimization algorithm shows a better convergence and gives a better optimization in small MLP-NN (Li & Yuan, 2017). Therefore, we apply the ReLU activation for the hidden layer as shown below:

$$f\cdot = ReLU\alpha = \max\alpha, 0 \tag{1.14}$$

CHAPTER 2

DETECTING OF A PATIENT'S CONDITION FROM CLINICAL NARRATIVES USING NATURAL LANGUAGE REPRESENTATION

Thanh-Dung Le[1,2] , Rita Noumeir[1] , Jérôme Rambaud[2] , Guillaume Sans[2] , Philippe Jouvet[2]

[1] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Research Center at CHU Sainte-Justine Hospital, University of Montreal,
3175 Chem. de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 1C5

Article published in « IEEE Open Journal of Engineering in Medicine and Biology » in September 2022.

2.1 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.

2.2 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 (Johnson *et al.*, 2016). 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 hour (Matton *et al.*, 2016). 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.

2.2.1 Problem Statement

The diagnosis of acute respiratory distress syndrome (ARDS) is frequently delayed or even not diagnosed in intensive care 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 (Bellani *et al.*, 2016). 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 (Group *et al.*, 2015). 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 (Group *et al.*, 2015). Our research team has developed the first two algorithms for hypoxemia (Sauthier *et al.*, 2021) and chest X Ray analysis (Zaglam, Jouvet, Flechelles, Emeriaud & Cheriet, 2014). 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.

2.2.2 Motivation

The recent study (Olsen, Mentz, Anstrom, Page & Patel, 2020) 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.

2.3 Materials and Methods

2.3.1 Clinical Narrative Data at CHUSJ

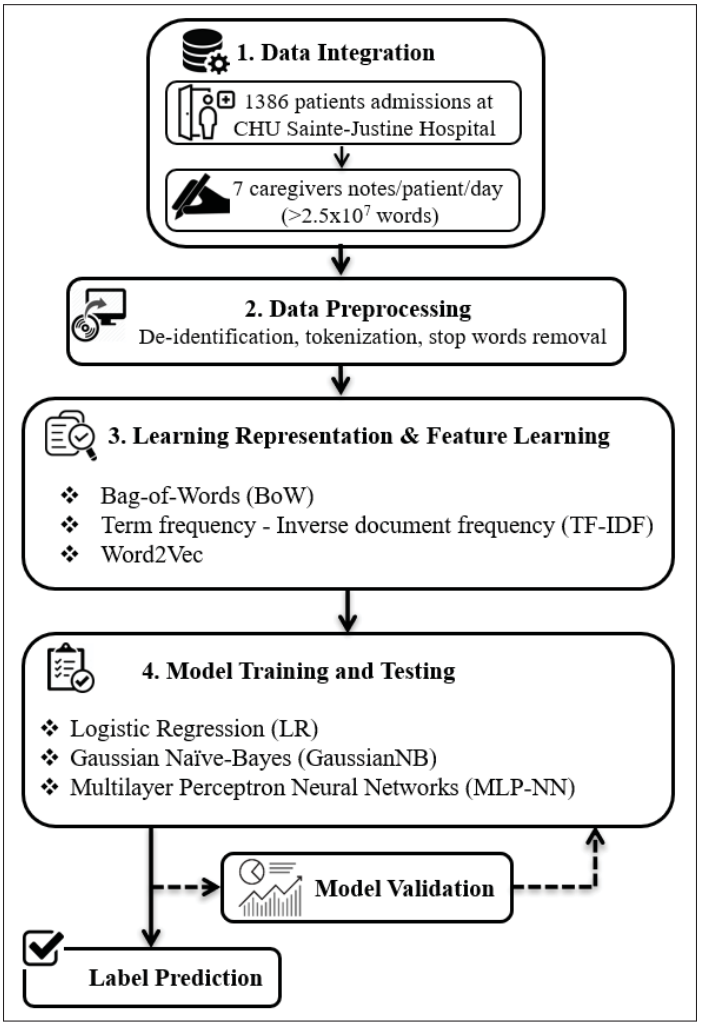

Figure 2.1 An overview of the proposed methodology to detect cardiac failure from clinical notes at CHUSJ

Fig. 2.1 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 24h 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 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. Fig. 2.2 shows an example of clinical notes with labels. Besides, 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.

2.3.2 Data Pre-Processing

Generally, it is proven that if the preprocessing steps are well prepared, the result for the end-task will be improved (Kannan *et al.*, 2014). Therefore, there are steps that were used as case lowering, and stop words removing. Fig. 2.3 shows the statistics for the list of stop words removed. From the list, all these words are the definition in French; therefore, they do not

	notes	label	len_notes
0	CIV par malalignement, CIA type II large, hypo...	no	100
1	Cardiopathie congénitale complexe cyanogène: o...	no	494
2	Née à terme, Grossess et accouchement sans com...	yes	1056
3	Grossesse normale, aucun Rx Né à terme 38 sem ...	yes	135
4	CARDIOMYOPATHIE 1) S'est présenté le 28.04 à ...	yes	1412
5	Enfant jumeau, né à 35 semaines. Période néona...	no	1312
6	Naissance terme 40+6 semaines Grossesse s/p, A...	no	576
7	Né à 41 semaines, 1ère grossesse, pas complica...	no	779
8	Grossesse : découverte masse intracardiaque VD...	yes	1053
9	Naissance à terme Développement de Sx cardiaqu...	yes	392

Figure 2.2 An example of clinical notes from CHUSJ

contribute to the learning representation. Besides, we did not consider any French linguistic feature as our method is based on uncorrelated words. All the notes are short narratives, and the n-gram length distribution is shown in Fig. 2.5. The longest n-gram is over 400 words, but most of the n-gram length distribution is between 50 and 125 words. Then, the ratio of the number of samples/words per sample is much smaller than 1500, as given by a tutorial for small text classification (Google, 2019).

term	Positive freq	Negative freq	pos_precision	pos_freq_pct	pos_hmean
de	11460	13500	0.459135	0.043645	0.079712
et	5180	5519	0.484157	0.019728	0.037911
à	4437	6803	0.394751	0.016898	0.032409
avec	3818	4680	0.449282	0.014541	0.028169
pas	3449	5335	0.392646	0.013135	0.025420
d	3381	4191	0.446513	0.012876	0.025031
en	2957	3250	0.476398	0.011262	0.022003
la	2577	3404	0.430864	0.009814	0.019191
l	2425	3655	0.398849	0.009235	0.018053

Figure 2.3 Clinical notes analyzing for stop words removing

In addition, it is essential to pay attention to negation in medical expression. First, the negation criteria from the study (Deléger & Grouin, 2012) were used for detecting the negative meaning from French notes. Then, a negation technique is applied (Dubois, Romano, Kale, Shah & Jung, 2017): 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."

Table 2.1 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)”		

```
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")
```

Figure 2.4 An example of code snippet in Python for decomposing numeric values (Example 4)

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 (Wallace, Wang, Li, Singh & Gardner, 2019). 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 (Dua *et al.*, 2019). Another study only focuses on how to extract the number, not dealing with representation learning (Cai *et al.*, 2019). Even, study (Kumar *et al.*, 2020) 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 (Noumeir, 2003). Therefore, we performed four experiments to evaluate the contribution from the numeric value to the classifiers. Fig. 2.4 shows an example of code snippet in Python, which help us conducting the decomposing the numerical measurement value. Finally, Table 2.1 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.

For the note visualization, we apply the ScatterText (Kessler, 2017). We have more than 580000 (n-grams) word count from the data shown in Fig. 2.6. The figure shows the most frequent words for the positive case in the upper right corner; the most frequent words for the negative cases in the lower-left corner; and, all less frequent words for both cases are in the center. Besides, the top 20 terms from the positive and negative cases are presented on the right-hand side. Their frequency distribution is illustrated in Fig. 2.7 and Fig. 2.8, respectively. There is an abbreviation for medical terms, whose descriptions and characteristics are summarized in

Table 2.2. For example, there are CIV (Communication intraventriculaire), CEC (Circulation extra-corporelle), CIA (Communication intra-auriculaire), FC (fréquence cardiaque), and IVRS (Infection des voies respiratoires supérieures). For more specific, Fig. 2.9 shows the top 30 n-grams that frequently appear for both cases, and Fig. 2.10 also visualizes the terms distributed for both cases. In positive cases, we quickly see that most of these terms are positively related to cardiac malfunction: milrinone (milri), aorte, aortique, and valve. In contrast, terms such as *urgence, ad, respiratoire, vers, toux, and ivrs* indicate respiratory syndromes. From Fig. 2.9, we can see that the overlapping: terms that strongly indicate one class also appear to a lesser degree in the other class.

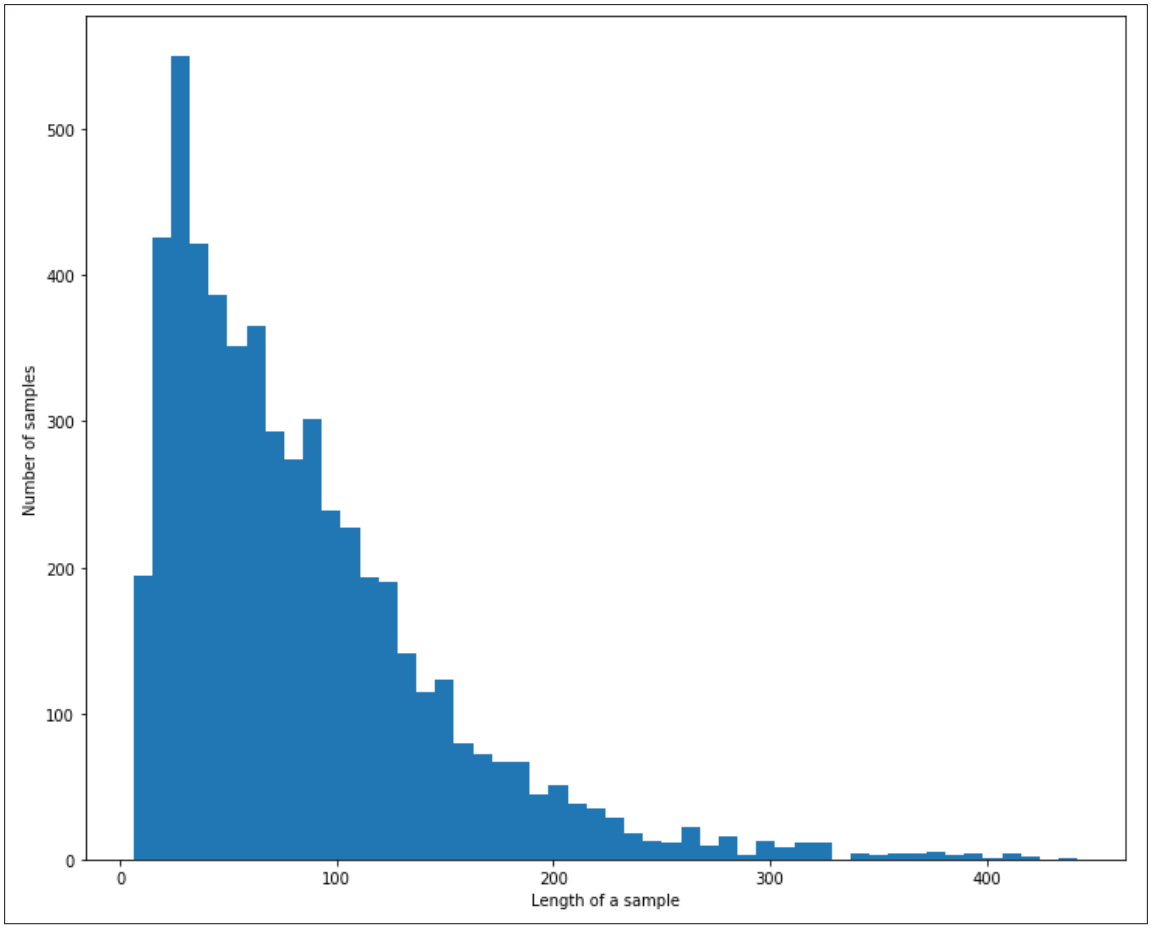

Figure 2.5 The distribution of length of notes in the CHUSJ dataset

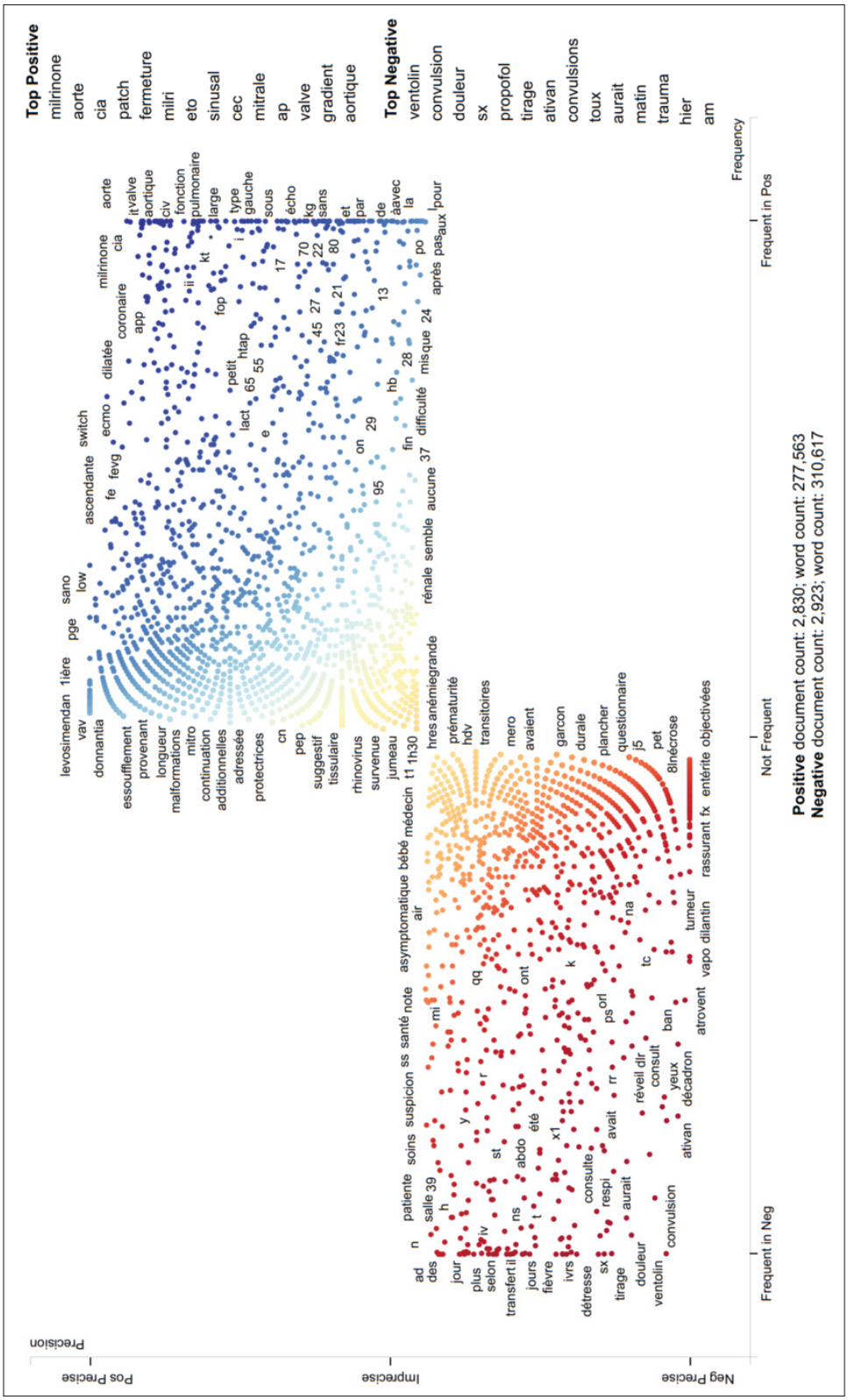

Figure 2.6 Clinical note illustration by using Scattertext visualization

Table 2.2 Important Abbreviations for Medical Terms

Abbreviations	Descriptions (In French)	Characteristics
CIV	Communication intraventriculaire	Cardiac malformation
CEC	Circulation extracorporelle	Treatment for cardiac failure
CIA	Communication intraauriculaire	Cardiac malformation
FC	Fréquence cardiaque	Cardiac frequency
IVRS	Infection des voies respiratoires supérieures	Virus responsible for respiratory distress
SOP	Salle d'opération	Operations
PO	Per os (by mouth)	Feeding

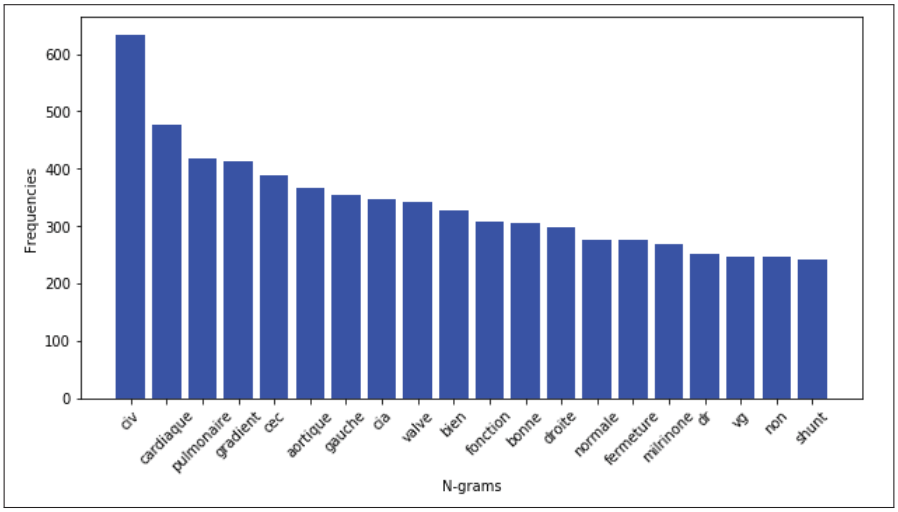

Figure 2.7 N-grams's frequency distribution for positive cases (Top 20 n-grams)

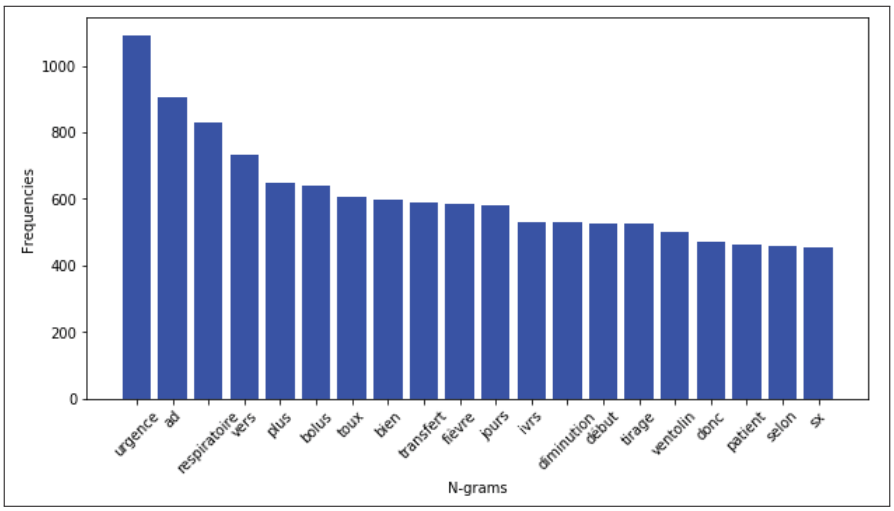

Figure 2.8 N-grams's frequency distribution for negative cases (Top 20 n-grams)

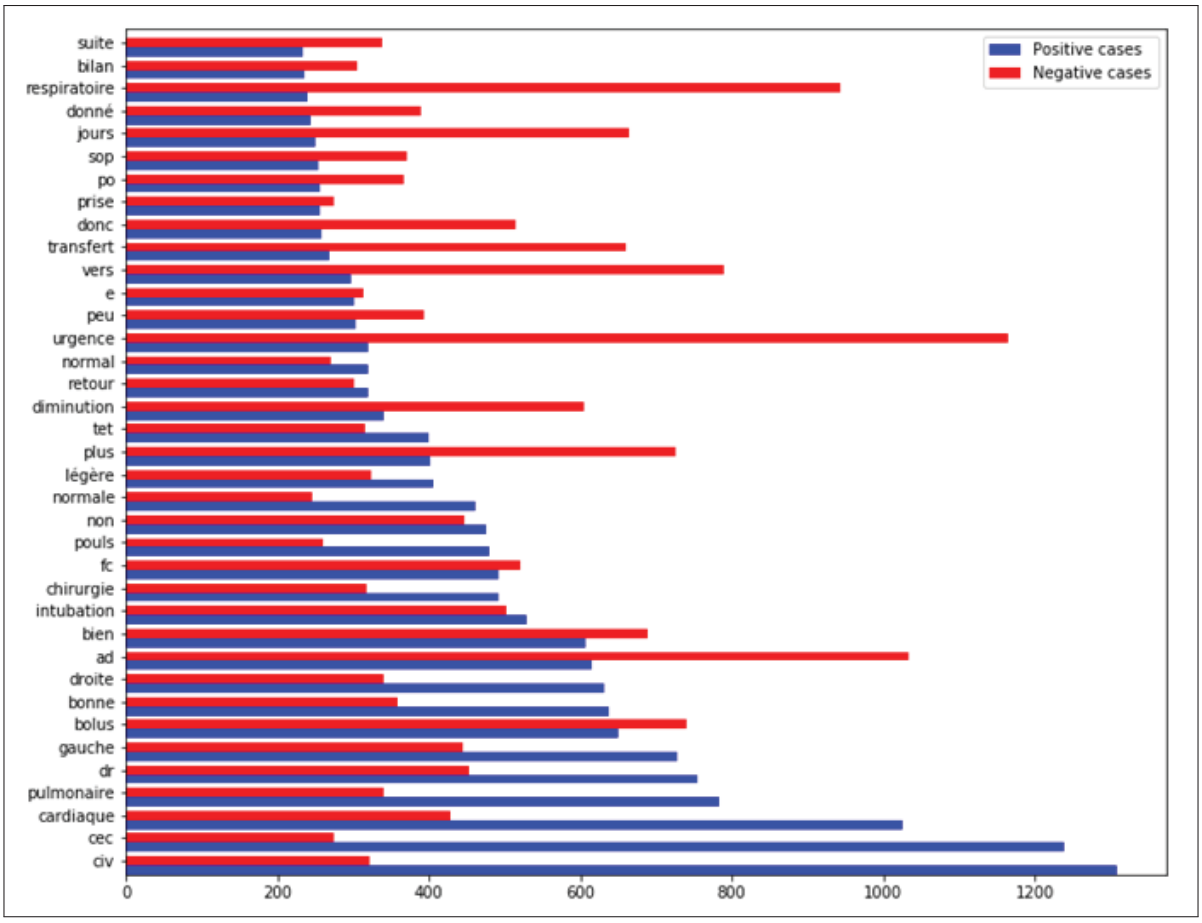

Figure 2.9 Top 30 frequent n-grams overlapping respecting to both two classes distribution.

2.3.3 Clinical Natural Language Representation Learning

There is no doubt about the effectiveness of neural word embedding. The study (Shi *et al.*, 2016) confirms that word2vec representation has been successfully used for various disease classifications from medical notes. Especially for the French clinical notes, the study (Dynomant *et al.*, 2019) 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 (Agarwal *et al.*, 2017; Li *et al.*, 2018; Zhang *et al.*, 2017; Fodeh *et al.*, 2019) 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. Therefore, we will evaluate the effectiveness of two conventional representation approaches, including BoW, TF-IDF, and the word2vec neural embedding model.

2.3.4 Machine Learning Classifiers

The state-of-the-art machine learning-based NLP currently focuses on deep learning for clinical notes (Pham *et al.*, 2017; Rajkomar *et al.*, 2018; Otter *et al.*, 2020; Young *et al.*, 2018; Sheikhalishahi *et al.*, 2019). 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 (Liu *et al.*, 2019b). And, study (Shi *et al.*, 2016) 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 (Paleyes *et al.*, 2020). Especially, clinical notes in a language other than English, the challenge is more difficult to mitigate (Névéol *et al.*, 2018). 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 (Li *et al.*, 2018). Automatic methods to extract New York heart association classification from clinical notes (Zhang *et al.*, 2017) confirm that the machine learning method, support vector machines (SVM) with n-gram features, achieves the best performance at 93% F-measure. Also, study (Agarwal *et al.*, 2017) 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. This study confirms that, with the small dataset, TF-IDF and BoW have better performance than other techniques on coronary microvascular classification (Fodeh *et al.*, 2019).

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 (Maimon & Rokach, 2014; Wang *et al.*, 2017) confirm that LR, and GaussianNB perform much the better than the other classifiers. Moreover, study (Ng & Jordan, 2002) 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 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 (Pasini, 2015).

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.

2.4 Results

We did the analysis to select of proper neural network sizes and architectures (Hunter, Yu, Pukish III, Kolbusz & Wilamowski, 2012). 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 (Srivastava, Hinton,

Krizhevsky, Sutskever & Salakhutdinov, 2014) with the probability of dropping out rate p=0.25, and GlorotNormal kernel initializer (Glorot & Bengio, 2010).

We used the scikit-learn library (Pedregosa *et al.*, 2011b) and Keras (Chollet, 2015) 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 ($k = 5$) (Kohavi, 1995); 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 (Goutte & Gaussier, 2005). These metrics are defined as follows:

$$\text{Accuracy (acc)} = \frac{\text{TP TN}}{\text{TP TN FP FN}}$$

$$\text{Precision (pre)} = \frac{\text{TP}}{\text{TP FP}}$$

$$\text{Recall/Sensitivity (rec)} = \frac{\text{TP}}{\text{TP FN}}$$

$$\text{F1-Score (f1)} = \frac{2^{\star}\text{Precision}^{\star}\text{Recall}}{\text{Precision Recall}}$$

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.

2.5 Discussion

Table 2.3 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 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. 2.11.

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 (Wang *et al.*, 2017) 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 (Wang *et al.*, 2017). 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 (Sahlgren & Lenci, 2016).

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 (Ng & Jordan, 2002; Perlich, Provost & Simonoff, 2003). 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 (Bartlett, 1998).

By applying the dropout (p=0.25) (Srivastava *et al.*, 2014), GlorotNormal initializer (Glorot & Bengio, 2010), and balancing the classes by using the Bayes Imbalance Impact Index (Lu, Cheung & Tang, 2019), the classifier was successful in avoiding the overfitting. Primarily, Fig. 2.12 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 (Martin *et al.*, 2020). It is motivated by the success of a Bidirectional Encoder Representations from Transformers (BERT) for natural language understanding (Kenton & Toutanova, 2019). Unfortunately, the result was not as good as expected; we could only achieve less than 60% accuracy, even though we applied the dropout technique as recommended from the study (Pasupa & Sunhem, 2016). We continued investigating with the simpler Transformer, which is solely based on attention mechanisms

through the connection of the encoder and decoder (Vaswani *et al.*, 2017), and it is implemented by Keras

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. 2.13. Moreover, from the result of Fig. 2.13, 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 (Gao *et al.*, 2021); 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.

2.6 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 (Dua *et al.*, 2019).

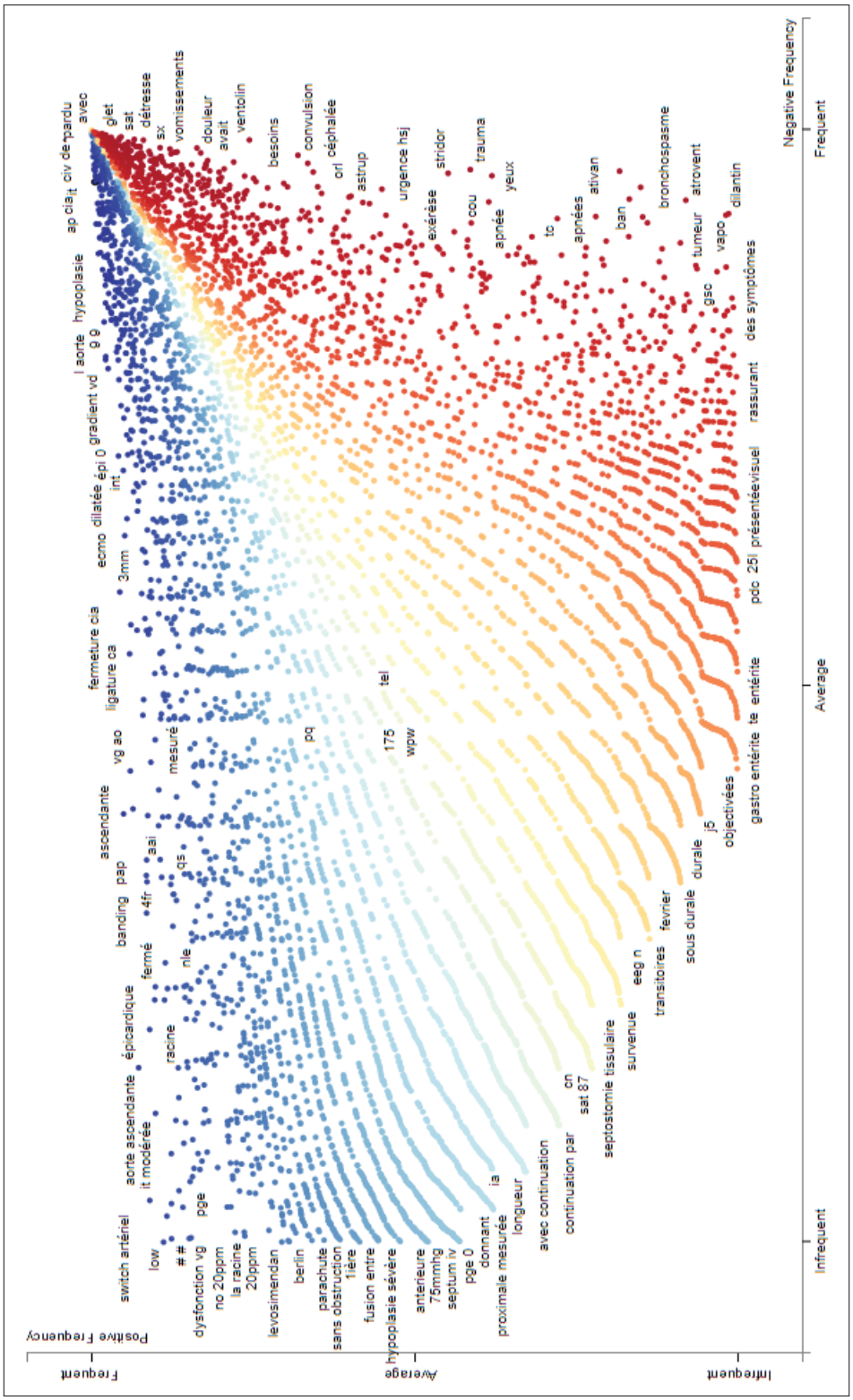

Figure 2.10 Visualization of terms distribution for both classes

Table 2.3 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

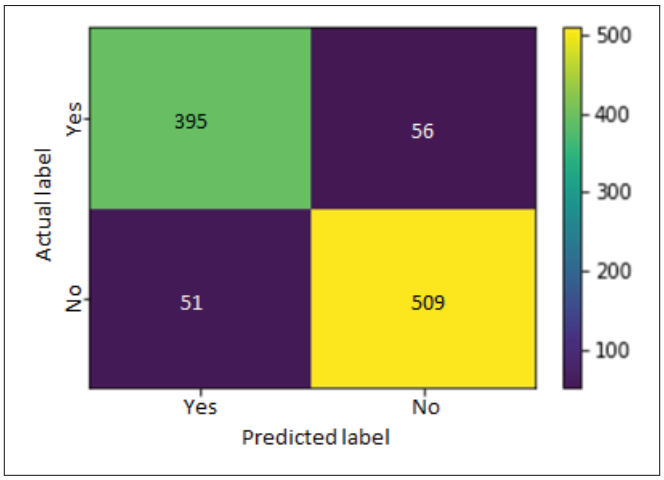

Figure 2.11 Confusion matrix of the MLP-NN classifier, showing the classification of positive (Yes) and negative (No) between predicted and actual labels

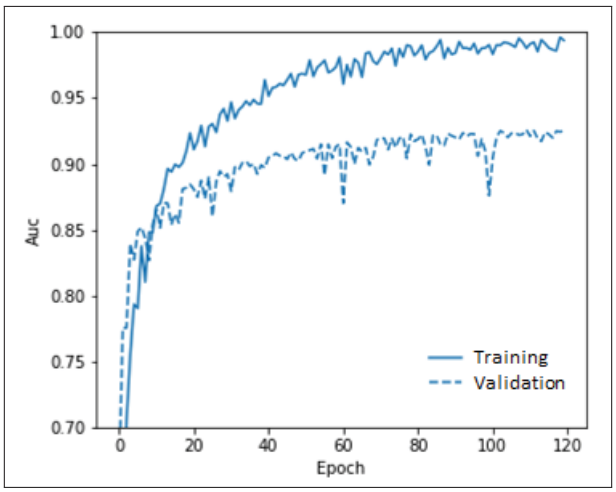

Figure 2.12 Area Under the Curve (AUC) performance of MLP-NN

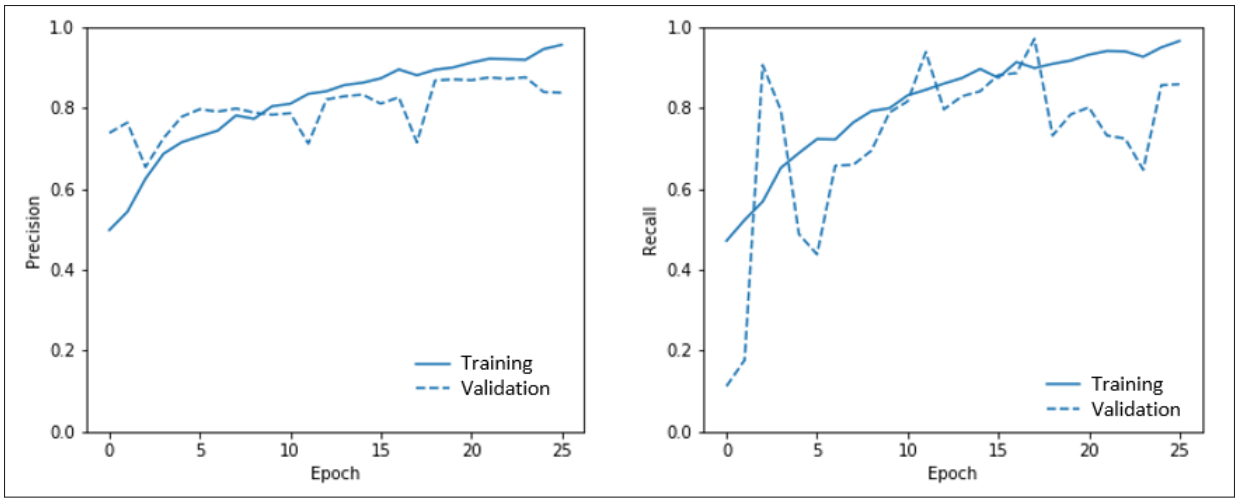

Figure 2.13 Precision (left) and recall (right) performance based on the Transformer configuration

CHAPTER 3

ADAPTATION OF AUTOENCODER FOR SPARSITY REDUCTION FROM CLINICAL NOTES REPRESENTATION LEARNING

Thanh-Dung Le[1,2] , Rita Noumeir[1] , Jérôme Rambaud[2] , Guillaume Sans[2] , Philippe Jouvet[2]

[1] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Research Center at CHU Sainte-Justine Hospital, University of Montreal,
3175 Chem. de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 1C5

Article published in « IEEE Journal of Translational Engineering in Health and Medicine » in February 2023.

3.1 Abstract

When dealing with clinical text classification on a small dataset, recent studies have confirmed that a well-tuned multilayer perceptron outperforms other generative classifiers, including deep learning ones. To increase the performance of the neural network classifier, feature selection for the learning representation can effectively be used. However, most feature selection methods only estimate the degree of linear dependency between variables and select the best features based on univariate statistical tests. Furthermore, the sparsity of the feature space involved in the learning representation is ignored. *Goal*: Our aim is, therefore, to access an alternative approach to tackle the sparsity by compressing the clinical representation feature space, where limited French clinical notes can also be dealt with effectively. *Methods*: This study proposed an autoencoder learning algorithm to take advantage of sparsity reduction in clinical note representation. The motivation was to determine how to compress sparse, high-dimensional data by reducing the dimension of the clinical note representation feature space. The classification performance of the classifiers was then evaluated in the trained and compressed feature space. Results: The proposed approach provided overall performance gains of up to 3% for each test set evaluation. Finally, the classifier achieved 92% accuracy, 91% recall, 91% precision, and 91% f1-score in detecting the patient's condition. Furthermore, the compression working mechanism and

the autoencoder prediction process were demonstrated by applying the theoretic information bottleneck framework.

3.2 Introduction

Clinical decision support systems (CDSS) are continuously being developed and play a crucial role in promoting a personalized healthcare system, as more and more data are collected and stored continuously (Musen, Middleton & Greenes, 2021). These data represent decisive points in advancing and enhancing the efficiency and effectiveness of CDSS operations. Predictive models have been developed based on the latter for preventive treatment and patient diagnosis, culminating in intelligent, precise, and timely healthcare improvement (Sutton *et al.*, 2020). In one notable example, a recent study (Gold *et al.*, 2022) analyzed the effect of CDSS on cardiovascular risk in 18,578 patients in 70 community health centers. In that case, CDSS significantly reduced the risk of cardiovascular disease among vulnerable high-risk patients.

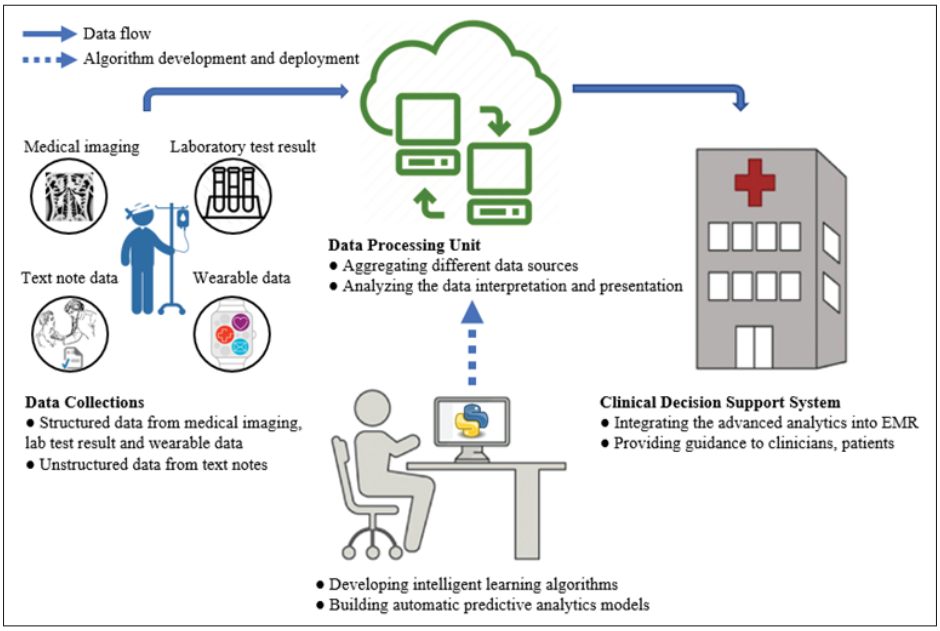

Figure 3.1 Workflow demonstration of a clinical decision-support system at CHUSJ hospital

Following the above successes, a CDSS was developed at CHU Sainte-Justine Research Center (CHUSJ). The system monitors pediatric intensive care management for all patients ranging in age from 0 to 18 years old. Fig. 3.1 illustrates two fundamental processes in the CDSS workflow at CHUSJ, which involve collecting and processing critical care data. First, clinical data are collected and stored in a clinical data warehouse. The data processing unit is then systematically aggregated and processed to convert raw data to a machine-readable form in the data processing unit. This process helps analyze the unknown data interpretation and presentation. The CDSS can thus integrate the advanced analytic result of the data processing unit and learning algorithms; then, clinicians can adequately use the CDSS to guide early intervention and prevention for healthcare management.

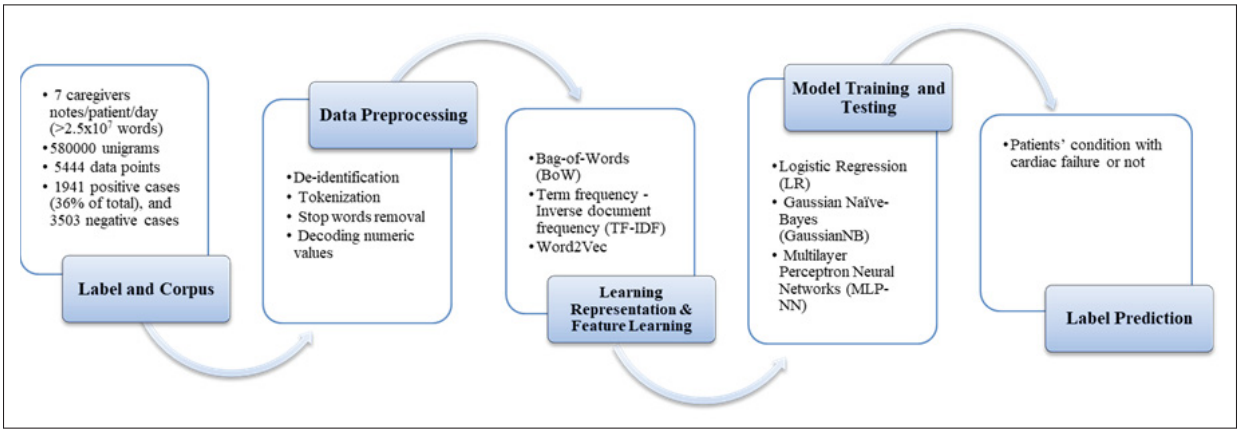

Figure 3.2 The clinical NLP based on machine learning for patients' condition prediction at CHUSJ hospital

One of the goals of the CDSS system in CHUSJ is automatically screening the data from electronic medical records, chest X-rays, and other data sources, which can increase the diagnosis rate and improve the management of acute respiratory distress syndromes (ARDS) in real time. Usually, the diagnosis of ARDS was delayed or missed in two-thirds of patients, and the diagnosis was missed completely in 40% of patients (Bellani *et al.*, 2016). Three main conditions need to be detected to diagnose ARDS: hypoxemia (low blood oxygenation), presence of infiltrates on chest X-Ray and absence of cardiac failure (Group *et al.*, 2015). Our research team has developed algorithms for hypoxemia (Sauthier *et al.*, 2021), chest X-ray analysis (Zaglam *et al.*,

2014; Yahyatabar, Jouvet & Cheriet, 2020), and identification of the absence of cardiac failure (Le, Noumeir, Rambaud, Sans & Jouvet, 2022, 2021). Technically, it successfully carried out extensive analyzes of machine learning algorithms (ML) aimed at detecting cardiac failure from clinical narratives using natural language processing (NLP) based on such algorithms (Le *et al.*, 2022). The study's design was to detect a cardiac failure in a patient's first 24 hours of admission using admission notes and evolution notes within the first 24 h. As summarized in Fig. 3.2, the study included the clinical notes of 1386 patients classified by two independent physicians using a standardized approach. Then, a comparative analysis was performed to discover the effective combination of various representation learning techniques with different machine learning classifiers. Consequently, it confirmed that the framework proposed herein outperforms other combinations with an overall classification performance of 89% accuracy, 88% recall, and 89% precision by applying a multilayer perceptron neural network (MLP-NN) classifier in combination with a term frequency x inverse document frequency (TF-IDF) learning representation.

These results were made possible by the contributions of the feature selection process, also known as SelectKBest. The advantage of the process was proven for supervised models as the classifier performance brought overall improvements of up to 3-4% over the case without the feature selection. It is obvious to understand because there are fewer misleading features; the classifier accuracy is improved after selecting the best K features. Unfortunately, the SelectKBest feature selection continues to have certain limitations in the proposed framework. One reason is that the feature selection method is based on a statistical test that estimates the degree of linear dependency between random variables. Then, it removes irrelevant features and ignores the correlation between data elements. As a result, more samples are required for an accurate estimation and avoidance of overfitting, which is not possible in our case (Jain & Singh, 2018). Furthermore, SelectKBest does not deal mainly with the sparsity of the feature space in the note representation matrix (Forman, 2003). Consequently, the sparsity that characterizes the learning representation space is ignored.

In healthcare, the autoencoder algorithm (AE) has lived up to its promises and has shown its effectiveness in improving outcomes for efficient clinical decision-making. AE can find informative transformed feature vectors through the compressed latent representation. For example, a study (Zhou, Jia & Motani, 2018) demonstrates an efficient framework for automatically learning compact representations from heterogeneous raw data sources from patient health data. In addition, AE can improve the predictability of the six different learning models to detect Parkinson's classification (Xiong & Lu, 2020). Another study (Kolyvakis, Kalousis, Smith & Kiritsis, 2018) shows that AE improved the performance of a novel outlier detection mechanism by retrofitting word vectors for the biomedical ontology matching task. In addition, having rich and accurate clinical data is very challenging (Quiroz *et al.*, 2019) because the acquisition and sharing of medical data face a significant obstacle in the form of privacy issues and the sensitive nature of the data. AE can be applied for sparsity reduction in clinical representation feature to tackle problems related to limited data availability. It could effectively discover the low dimensional embeddings and reveal the underlying effective manifold structure from a sparse high dimensional document-term matrix (Leyli-Abadi, Labiod & Nadif, 2017).

Therefore, the present study examines alternatives to feature selection and focuses mainly on compressing data without loss of information by employing an AE algorithm. First, the study aims to achieve a better feature space without sparsity. The authors are interested in compressing the sparse TF-IDF matrix and reducing its dimensions to improve the efficiency of the feature space representation. Notably, a neural network is incorporated to learn efficient codings of unlabeled data to address the issues caused by sparse vectors generated from the TF-IDF representation feature space for clinical notes. Then, the compressed vector space from the TF-IDF matrix is fed into the classifiers as a refined input. Finally, ML classifiers conduct the learning process to draw comparative results, which are then used to evaluate the classification performance.

Our study confirms that AE effectively compresses the vector space of the TF-IDF representation for clinical narratives into a lower dimension. The proposed approach can retain the critical feature by capturing the correlation between attributes during the training process, hence; the

downstream classification task can generally be increased to 2-3% for each evaluation criterion. Furthermore, the value of AE behaviors in a limited data set is also highlighted. The working mechanism of the AE is analyzed and explained how the AE works to compress data through the encoder and decoder. Based on the information-theoretic framework, the working mechanism of the AE is to optimize the information bottleneck during the compression and prediction process, respectively. As a result, the behavior of AE in limited data is exactly in harmony with such cases where there is much larger data availability.

Section 3.3 will discuss the materials and methods. The experimental results and discussion then will be discussed in section 3.4, 3.5. Finally, section 3.6 provides concluding remarks.

3.3 Materials and Methods

3.3.1 Data Sparsity Challenges

In numerical analysis, a sparse matrix or array is a matrix in which most elements are zero (Hurley & Rickard, 2009). The number of zero-valued elements divided by the total number of elements (e.g., $m \times n$ for a $m \times n$ matrix) is called the matrix sparsity (equal to 1 minus the density of the matrix). Using these definitions, a matrix will be sparse when its sparsity is more significant than 0.5. In our case, after the research ethics board approved the research protocol from the Research Center of the Sainte-Justine Hospital, the data were retrospectively extracted from the electronic medical record. There are more than 580000 (unigrams) word count from 5444 single lines of notes with 1941 positive cases (36% of total) and 3503 negative cases. All the notes are short narratives, and detailed description characteristics can be found in the Supplementary Materials from (Le *et al.*, 2022). The longest n-gram is over 400 words, but most n-gram length distribution is between 50 and 125 words. The average length of the number of characters is 601 and 704. And the average size of the number of digits is 25 and 26 for the positive and negative cases, respectively. Then, the data was pre-processed by applying the stop-word removal to exclude the minor information. In addition, the negation in medical expression was used to add the negative meaning from French notes. For the vital numeric

values (heart rate, blood pressure, etc.), all numeric values for vital sign values were kept (nearly 4% of the notes), and the decoding for those number values was used to decode the numeric values. Finally, the feature selection, SelectKBest, was used to select the top best 'k=20000' of the vectorized features for the TF-IDF representation learning feature space. Hence, there is a matrix of features of 5444×20000. It is calculated by the Eq. 3.1, and the sparsity of the matrix is greater than 0.9.

It confirms that the representation matrix from the TF-IDF is sparse because every word is treated separately. Hence, the semantic relationship between separated entities is ignored, which would cause information loss. Although the combination of TF-IDF and MLP-NN consistently outperformed other combinations with overall performance and was the most stable under all circumstances (Le *et al.*, 2022), the sparsity remains. Therefore, the motivation is to compress the sparse, high-dimensional data by reducing the dimension from the TF-IDF feature space of clinical notes representation

$$\text{sparsity} = 1 - \frac{\text{count\_nonzero(TF-IDF)}}{\text{total\_elements\_of\_(TF-IDF)}} \tag{3.1}$$

3.3.2 Autoencoder Learning Algorithm

An AE was originated by (Kramer, 1991) to solve a nonlinear dimensional reduction; later, AE was famously promoted by training an MLP-NN with a small central layer to reconstruct high-dimensional input vectors (Hinton & Salakhutdinov, 2006; Wang, Yao & Zhao, 2016). Technically, AE takes an input $X \in \mathcal{R}^{N \times D}$ and maps it to a latent representation $Z \in \mathcal{R}^{N \times M}$ via a nonlinear mapping. Let us call $x \in X$, and $z \in Z$, then it will be as:

$$z = gWx \;\; b \tag{3.2}$$

W is a weight matrix during training, b is a bias vector, and $g\cdot$ stands for a nonlinear function, such as the logistic sigmoid function or a hyperbolic tangent function. The encoded feature representation x is then used to reconstruct the input x by reverse mapping, leading to the reconstructed input x':

$$x' = fW'z \;\; b' \tag{3.3}$$

where W' is usually limited to the form of $W' = W^T$, i.e., the same weight is used to encode the input and decode the latent representation. $f\cdot$ is also a non-linear function. The AE tries to learn a function $f_{W',b'}x \approx x'$. In other words, it is trying to learn an approximation of the identity function for the output x' that is similar to x. Still, by placing constraints on the network, such as limiting the number of hidden units, interesting data structures can be discovered. Then, the reconstruction error is defined as the Euclidean distance between x and x' that is constrained to approximate the input data x (that is, minimizing $||x - x'||^2$).

$$\begin{aligned} \mathcal{L}\left(x, x'\right) &= \left\|x - x'\right\|^2 \\ &= \left\|x - fW'\left(gWx \;\; b\right) \;\; b'\right\|^2 \end{aligned} \tag{3.4}$$

For the reconstruction evaluation between the original data x, and the reconstructed output x', the statistical measure R_i^2 will be applied for the i^{th} variable of x_i, and it can be computed as:

$$R_i^2 = 1 - \frac{\sum_{j=1}^{m} x_{j,i} - {x'_{j,i}}^2}{\sum_{j=1}^{m} x_{j,i}^2} \tag{3.5}$$

Since $R^2 = 1$ will be a perfect reconstruction. Consequently, the reconstruction will be evaluated by how much the value of R^2 is close to 1.

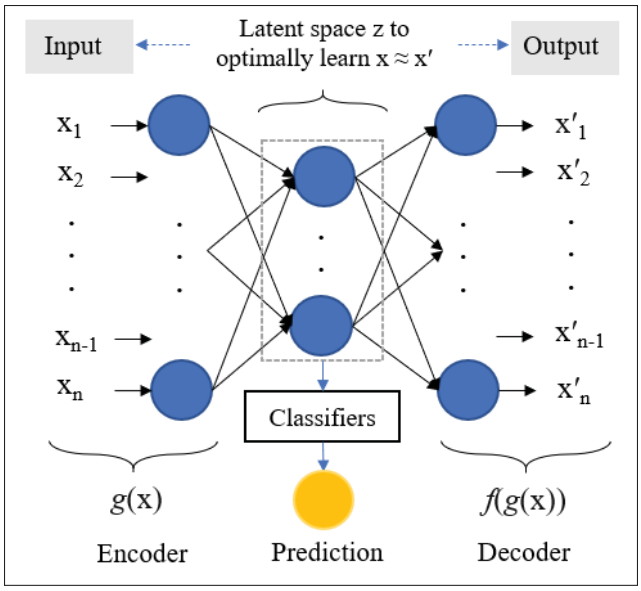

Figure 3.3 Schematic structure of an AE-based for compression and prediction

Ideally, an effective AE can be designed and trained based on the minimization of reconstruction error from Eq. 3.4 and maximization of the reconstructed effectiveness from Eq. 3.5; however, it is substantially based on its width (number of neuron units or latent representation dimension M) and its depth (number of hidden layers). First, conventional AE relies on the dimension of the latent representation z being smaller than that of the input x ($M < D$), which means that it tends to learn a low-dimensional compressed representation. The study (Garg & Liang, 2020) presents methods to learn the decoder function $f\cdot$ as a learnable function through the reconstruction error in Eq. 3.4 in several representation learning approaches. It is concluded that the compression depends on dimension M but less on dimension D. Second, it has been shown that training a neural network-based by increasing the number of hidden layers (in combination with an increase in the number of neuron units per layer) achieves less consistent results (Steinmeyer & Wiese, 2020). Therefore, a small and simple AE will be used in our case. An AE with three layers (one input layer, one hidden layer, and one output layer) is employed. Mainly, to reduce the parameters from the latent space of the AE, the regularization technique is applied from study (Shi, Lei, Ma & Niu, 2019) to remove redundant parameters.

After training, the weight matrix from the hidden layer as a pre-trained tool is used. A classifier subsequently uses this pre-train latent space representation to perform the binary classification, as shown in Fig. 3.3. For the classifiers, it is essential to have consistency in evaluating the proposed approach's performance. Then, six different ML classifiers, including Random Forest (RF), Multinomial Naive Bayes (MultinomialNB), Logistic Regression (LR), Support Vector Machine (SVC), Gaussian Naive Bayes (GaussianNB), and Multilayer Perceptron Neural Network (MLP-NN) are used.

Furthermore, to understand the dynamics of learning and the behavior of AE, particularly in our case with limited data, the behavior of AE during the training process from the encoder and decoder is analyzed. Technically, it is captured to understand how the AE can retain the information during the compression process. To do that, the information-theoretic quantities and their estimators are applied. The technique is based on information-theoretic learning, which computes and optimizes information-theoretic descriptors named mutual information. The information-theoretic framework (Yu & Principe, 2019; Tapia & Estévez, 2020; Lee & Jo, 2021) has been utilized for a detailed theoretical explanation of an AE. These studies rely on the "information bottleneck" (Tishby, Pereira & Bialek, 2000; Shwartz-Ziv & Tishby, 2017) to understand and estimate how the AE works by quantifying its information plane coordinates. The information bottleneck can be used as an optimal bound that maximally compresses the input x, for a given mutual information on the desired output x'. There are comprehensive overviews of recent studies (Geiger, 2021; Geiger & Kubin, 2020; Alomrani, 2021). Technically, the output activation is firstly binned as stated in (Shwartz-Ziv & Tishby, 2017), and each hidden layer i ($1 \leq i \leq K$) is treated as a single variable T_i. Then it will be able to estimate the mutual information between all the hidden layers and the input/output layers by estimating the joint distribution PX, T_i and PT_i, X', and use them to calculate the mutual information of the encoder (between the input X and the hidden layer T_i), and the mutual information of the decoder (between the hidden layer T_i and the desired output X') using the following equations Eq. 3.6, 3.7. Finally, the good representation TX can be learned, which is characterized by its

encoder and decoder distribution $PT|X$, and $PX'|T$, respectively, to effectively map the input patterns X to a good prediction of the desired output X'.

$$IX;T_i = \sum_{x\in X, t\in T_i} Px,t \log\left(\frac{Px,t}{PxPt}\right) \tag{3.6}$$

$$IT_i;X' = \sum_{t\in T_i, x'\in X'} Pt,x' \log\left(\frac{Pt,x'}{PtPx'}\right). \tag{3.7}$$

3.4 Experimental Implementation

To assess the performance of our method, metrics including accuracy, precision, recall (or sensitivity), and F1 score were used (Goutte & Gaussier, 2005). These metrics are defined as follows.

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

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

For implementation, the same hyperparameters are used as from the previous study (Le *et al.*, 2022) for all classifiers to have a consistent evaluation of the performance: avoiding overfitting by applying the dropout (p=0.25) (Srivastava *et al.*, 2014), and the GlorotNormal initializer (Glorot & Bengio, 2010); balancing the classes by using the Bayes Imbalance Impact Index (Lu *et al.*, 2019) to deal with the imbalanced classes. The data was also divided into 60% training, 20% validation, and 20% testing. The implementation was done using Python Scikit learn (Pedregosa *et al.*, 2011a) and Keras (Chollet, 2015).

There is a tradeoff between the guarantee to identify the best combination of hyper-parameters and the computation time. And, for training a neural network, usually, only some hyper-parameters matter. The others have little impact on the machine learning model's accuracy. Based on the study (Luo, 2016), there are three essential hyper-parameters, including the number of hidden layers, the number of nodes on each hidden layer, and the learning rate for the backpropagation algorithm. With this limited range of hyper-parameters, the grid search will quickly become feasible to optimize every parameter simultaneously, including the cross-product of all intervals. Then, the models can be trained quickly. Further advantages of grid search include easier parallelization and flexible resource; the equivalent does not hold for Bayesian optimization (Yu & Zhu, 2020). Therefore, this study used grid search for up to three hidden layers and 500 neurons per layer, and other hyperparameters are summarized in Table 3.1 for AE training. For the optimizers, the Stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (ADAM) was used with small scalar ϵ, and the forgetting factors for gradients and second moments of gradients, β_1 and β_2. Then, a combination with the highest estimations was considered the best performance.

3.5 Results and Discussion

To deal with sparsity, many researchers focus on dimension reduction. There are two most popular techniques, namely Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), for their simplicity among other dimension reduction techniques (Anowar, Sadaoui & Selim, 2021), even with a large dataset (Reddy *et al.*, 2020). Especially when the

Table 3.1 Summary of Hyperparameters

Hyperparameter	**Ranges**
Hidden layers	1-3
Neurons	100-500
Activation	Sigmoid
Kernel initializer	GlorotNormal
Optimizers	SGD, ADAM
Learning rate	0.001 - 0.01
β_1	0.9
β_2	0.999
ϵ	e^{-8} - e^{-7}

training data set is small, and the PCA-supervised discriminative approach can outperform, it is also less sensitive to the variability of the training sets (Martinez & Kak, 2001). The study (Gárate-Escamila, El Hassani & Andrès, 2020) shows that PCA can increase the performance of different ML classifiers for predicting cardiac failure.

It can be said that the classifiers performed better after applying LDA to the linear data set. If the classes are non-linearly separable, the LDA cannot effectively discriminate between these classes (Tharwat, Gaber, Ibrahim & Hassanien, 2017). Otherwise, in the case of linear data, LDA can reduce the dimensionality and be used in different classification tasks (Ghosh & Shuvo, 2019). However, the TF-IDF enhanced with the LDA approach did not allow the classifier to score high accuracy compared to the other two methods when smaller datasets were fed (Dzisevič & Šešok, 2019). One of the reasons was explained in (Reddy *et al.*, 2020); the results showed that ML algorithms with PCA produce better results when the dimensionality of the data sets is high. When the dimensionality of datasets is low, the ML algorithms without dimensionality reduction yield better results. Another possible way is using an unsupervised generative Latent Dirichlet allocation to estimate the topic distribution (topics) by using observed variables (words). Latent Dirichlet allocation shows the effectiveness of overcoming the sparsity from the feature space matrix of TF-IDF (Kim & Gil, 2019). It can also help to make texts more semantically focused and reduce sparseness (Chen, Yao & Yang, 2016). However, its selection of characteristics does not improve performance with small data (Fodeh *et al.*, 2019).

Table 3.2 A comparison performance of feature selection approaches

Feature selection	Accuracy	Precision	Recall	F1
SelectKBest (Le *et al.*, 2022)	0.89	0.89	0.88	0.88
PCA	0.88	0.88	0.86	0.87
NCA	0.89	0.88	0.89	0.88

The possibility of PCA for sparsity reduction was explored because of the advantages mentioned above. The training was tuned and performed, and the best performance was achieved by decreasing to 2 principal dimensions. The completed test has an accuracy of 88%, a precision of 88%, a recall of 86%, and an f1-score of 87%. Furthermore, following the recommendation of (Laghmati, Cherradi, Tmiri, Daanouni & Hamida, 2020), a statistical method, Neighborhood Component Analysis (NCA) (Goldberger, Hinton, Roweis & Salakhutdinov, 2005), was also used to reduce the dimensions of the data set. NCA has shown that it works well on a small dataset for the medical domain. However, the result is slightly better than PCA; NCA only achieves an accuracy of 89%, a precision of 88%, a recall of 89%, and an f1-score of 88%. From Fig. 3.4, 3.5, it can be easily seen the features overlap; hence, the classification task hardly separates the boundary for the binary classification. Neither PCA nor NCA can improve classification performance summarized in Table. 3.2. It confirms the limitation of these approaches by linearly approximating a feature subspace to maximize class separability.

Furthermore, the non-linear activation function AE performs best on compression of the sparse TF-IDF representation space. This study compares the effectiveness of reconstruction based on the reconstruction evaluation from Eq. 3.5 between PCA, linear activation function AE (LAE), AE, and stacked AE (SAE) (Gehring, Miao, Metze & Waibel, 2013). The results confirm that the PCA and LAE have the same performance, achieving about 80% of the reconstruction. When the activation of AE is linear, then PCA and LAE are identical. There is no improvement if the SAE is used to extract the features in cases of limited data. Besides, the effectiveness of non-linear activation in AE is proved when it can maximally reconstruct up to 86% compared to the original spare data. It is one of the advantages of nonlinear transformation from AE, trained by a neural network, which is superior to the linear transformation from other approaches.

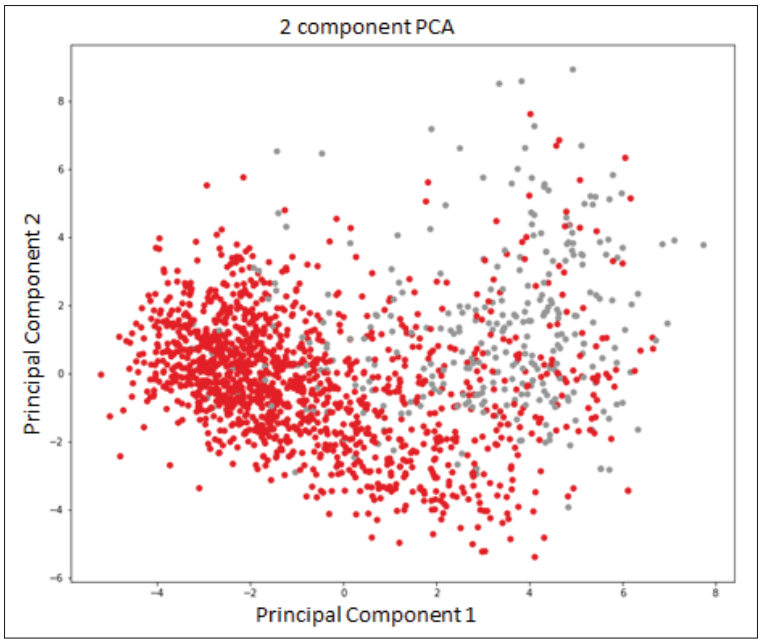

Figure 3.4 Visualization of the representation space
for 2 components from Principle Component Analysis (PCA)

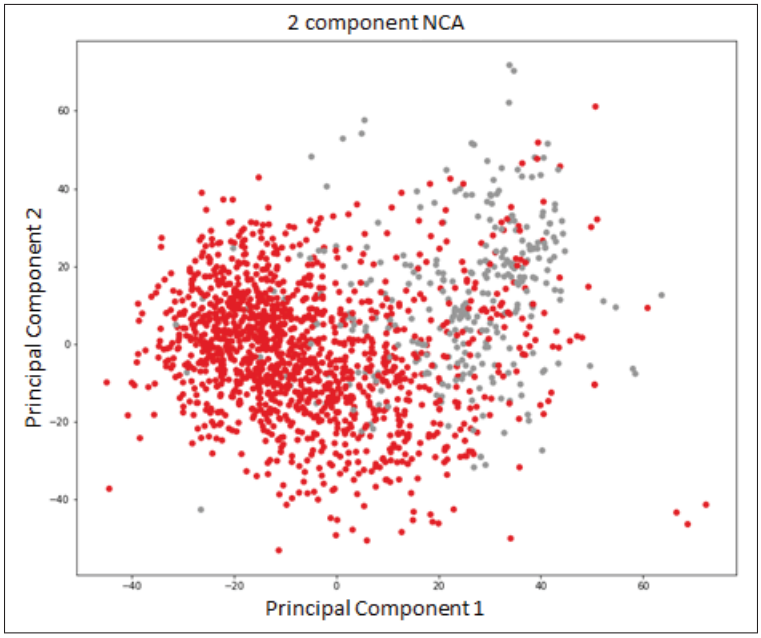

Figure 3.5 Visualization of the representation space
for 2 components from Neighborhood Component Analysis (NCA)

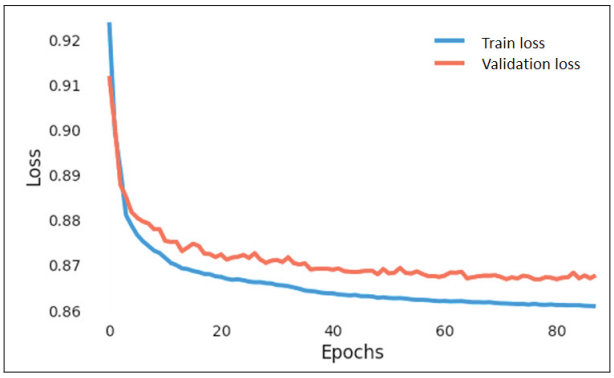

Figure 3.6 Loss for training and validation for the AE algorithm

Overall, the downstream classification performances are effectively improved by feeding the compressed feature space output from the AE to ML classifiers. Fig. 3.6 shows the loss during the training and validation process by optimizing the loss function from Eq. 3.4 for training the AE; both training and validation losses have quite-smooth convergence. After successfully training the AE, there is a pre-trained compressed, low-dimension feature space. Then, machine learning classifiers are employed to perform the classification and evaluate the performance. Instead of performing on MLP-NN, LR, and GaussianNB, it is also tested with other classifiers such as Random Forest (RF), Multinomial Naive Bayes, and Support Vector Machine. The best performance from MLP-NN classifier is achieved at 92%, 91%, 91%, and 91%, respectively, for accuracy, precision, recall, and f1 score. And the detailed confusion matrix showing the classification of positive cases (1) and negative cases (0) between predicted and actual labels for the holdout set is shown in Fig. 3.7. The experimental results are improved to 2-3 % for each evaluation criterion from (Le *et al.*, 2022), which had a general classification performance in a sparse TF-IDF feature space at 89% accuracy, 89% precision, 88% recall, and 88% f1 score. It confirms that the AE method can deal with sparsity by compressing the TF-IDF feature space. Consequently, it improves the downstream task performance of the MLP-NN classifier and is more robust than other methods. Recent work (Mienye, Sun & Wang, 2020) also confirmed a similar effect, but it was applied to a different dataset type and larger data availability. These results confirm the effectiveness of compressing the feature representation learning space into a

low-dimensional representation using the AE algorithm. The robust transformation can outplay the deep learning models with limited data resources.

Cross-validation was further used to accurately estimate the model's predictive performance and determine the reliability of ML algorithms (Arlot & Celisse, 2010). Fig. 3.8 shows the accuracy comparison, using a box plot, of the 5-fold cross-validation. It can be seen that the best three classifiers are MLP-NN, LR, and GaussianNB, respectively. All their median accuracy is over 80%; mainly, the MLP-NN classifier's median accuracy is the highest, over 90%. While there is not much difference between LR and GaussianNB, the median accuracy is around 82-83%. In addition, MultinominalNB, RF, and SVC follow right after as the three most minor performances, respectively, with median accuracy lower than 75%. Second, although the models' performance is assumed that the returns of accuracy follow a normal distribution, in reality, the returns are usually skewed. Notably, there is two skewness of the accuracy distribution for all classifiers. There is a negatively skewed distribution (skewed left) from the MLP-NN, LR, and RF, which may expect frequent smaller accuracy than their median in practice. In contrast, it should be expected to have higher accuracy than the median from the GaussianNB, MultinominalNB, and SVC because they all have positively skewed distribution (skewed right). Lastly, the dispersion distribution for most classifiers' accuracy is quite similar because the variability range contains all the smallest and largest accuracy values at the end of the whiskers. However, there is an exception for the LR and MultinomialNB classifiers, which have values outside the box plot's whiskers. It means that the two classifiers are less stable and reliable. In short, MLP-NN gives the best performances because of its high and stable accuracy for the model generalization validation; GaussianNB follows right after; LR is comparatively similar to GaussianNB. And all other classifiers are less effective.

Furthermore, an important aspect of performance analysis is that the proposed approach still shows its advantageous capacity to increase data availability. The study investigated the effectiveness of AE for compressing feature space and studied how algorithm performance varies with the increasing of training examples from the compressed feature space. The performance of two classifiers, GaussianNB and MLP-NN, was assessed to evaluate their effectiveness. When it

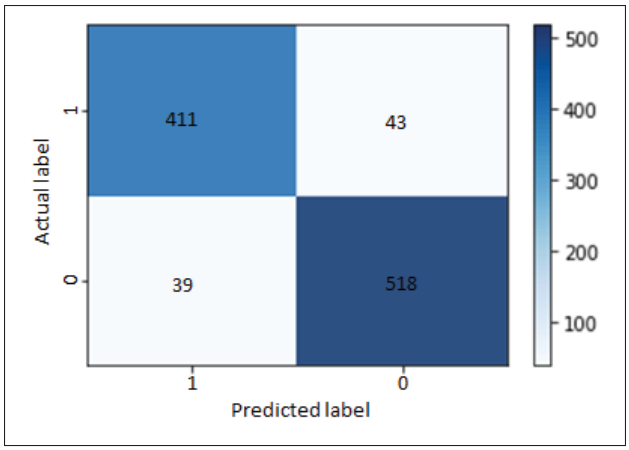

Figure 3.7 Confusion matrix of the MLP-NN classifier, showing the classification of positive (1) and negative (0) between predicted and actual labels

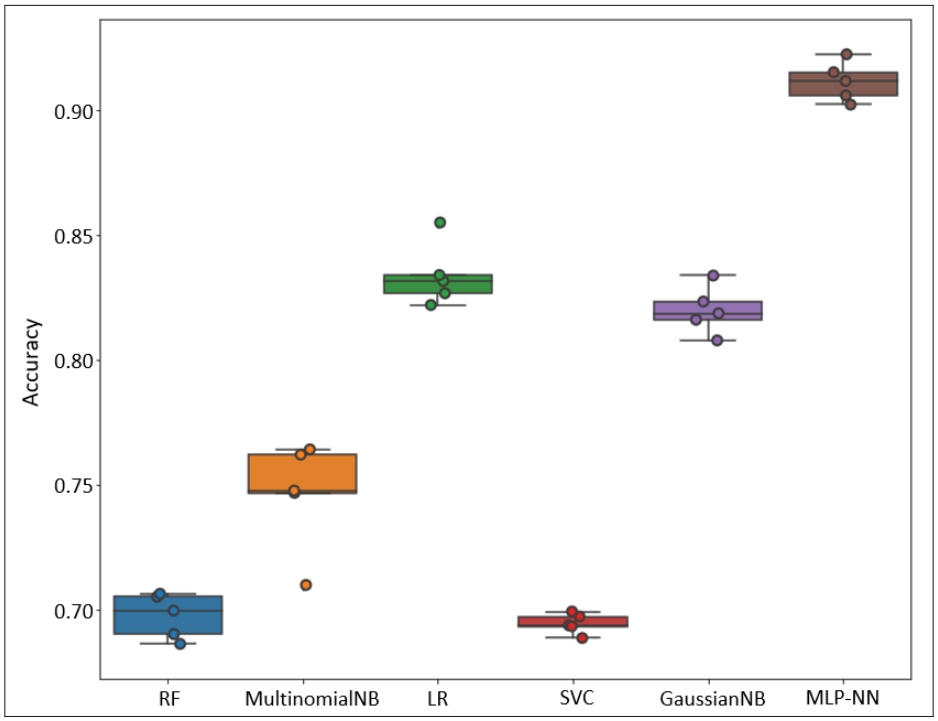

Figure 3.8 A comparison evaluation of the box plot 5-fold cross-validation results for classifiers performance.

possibly increases data availability in the future, whether the classifier improves performance or not. In this case, study (Ng & Jordan, 2002) confirms that when the number of training examples

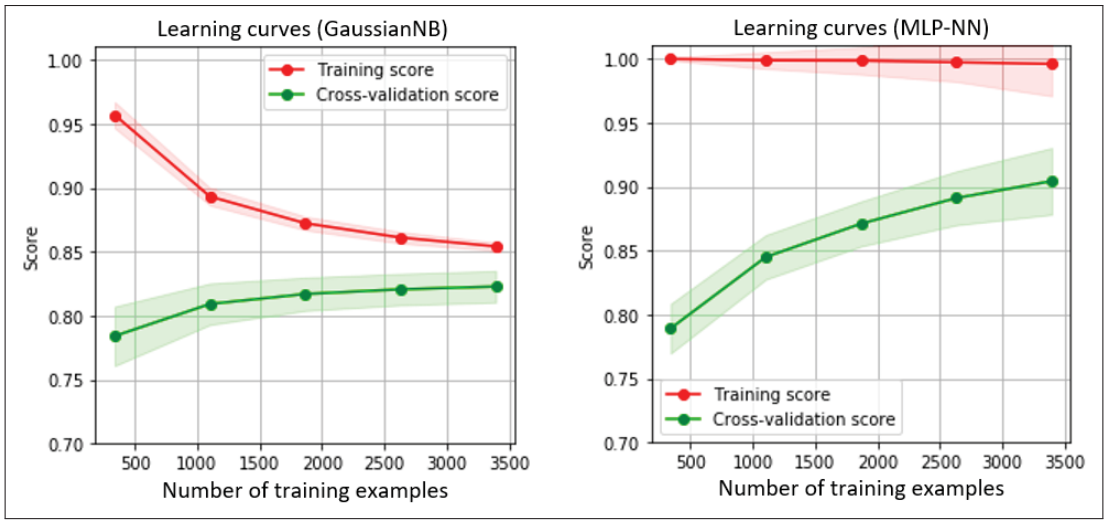

Figure 3.9 Performance of classifiers in case of increasing the training size: GaussianNB (left) and MLP-NN (right)

increases, the generative model based on Naive Bayes would expect to perform better. However, our results are in contrast to that confirmation. Fig. 3.9 shows the GaussianNB (left) and MLP-NN (right) training and validation scores when increasing the number of training examples. Technically, the GaussianNB reaches a plateau of performance after around the 2000^{th} training examples with the same dataset size, and the cross-validation score could not improve. It should be expected that this is one of the limitations of GaussianNB, namely the linear discrimination characteristic for a real-world dataset, discussed in (Xue & Titterington, 2008). In contrast, the MLP-NN shows improvement with the increasing size of the dataset. Its cross-validation score gradually increases from the point at 500^{th} to the 2500^{th} training examples; especially, the slope shows no signs of decreasing after reaching the maximum number of the training example. In short, GaussianNB shows improvement, but not as much as the MLP-NN, and reaches a plateau more quickly. It can be confirmed that our approach with MLP-NN is still applicable when data is possibly increased and continually improves its classification performance.

Moreover, the behavior of AE in limited data is in harmony with more significant data cases based on the information-theoretic framework. The behavior of AE was analyzed, and the technique was based on an information-theoretic framework, as mentioned in Eq. 3.6, and 3.7.

It aims at understanding how the AE behaves during the compression process by analyzing the mutual information of each hidden layer from the encoder and decoder. Generally, this type of analysis has been performed for a larger data set and has mainly focused on other data sources compared to our case; such as computer vision (Viola & Wells III, 1997), medical imaging (Pluim, Maintz & Viergever, 2003), and genetics (Olsen, Meyer & Bontempi, 2008). The analysis for two AE models was performed concerning various hidden layers (three hidden layers and five hidden layers). As shown in Fig. 3.10, there are two phases of the information plane in each hidden layer of the three-layer and five-layer cases. It is noted that from left to right, it illustrates the behavior of each hidden layer. And in each hidden layer, from top to bottom, it captures the mutual information for each training epoch. Finally, all trajectories follow a similar path during the learning process, eventually converging and getting closer to the optimal points in the bottleneck bound.

Specifically, it can be divided into two phases for the working mechanism of AE in Fig. 3.10. The first phase is called the drift phase, where the AE attempts to learn the latent representation TX with a smaller dimension than the original data X. During the compression, there will be information loss, which is why it can be seen the trend of decreasing the mutual information of encoder $IX; T$. At the end of this step, there will be a compressed latent representation TX, and optimal mutual information $IX; T$. Then, the second phase is named the diffusion phase. Within this step, the AE tries to find the reconstructed data X', which is optimally close to the original data X. The AE maps the latent representation TX to the reconstructed data X' by maximizing the mutual information of the decoder $IT; X'$. By doing that, there is an increasing trend of $IT; X'$; until $IT; X'$ reaches its optimal bound for each layer. And the optimal mutual information will get smaller when AE has more hidden layers. In the case of three hidden layers, the optimal mutual information of the encoder IX, T is larger by 6.0 but is maximum at 5.5 for five hidden layers. It is the same for the optimal mutual information of the decoder IT, X' at nearly 7.0 and 6.5 for three and five hidden layers, respectively. These results illustrate the mechanism of an AE is to optimize the information bottleneck trade-off TX during compression and prediction for each layer. Remarkably, it is trained on a small and sparse dataset; still,

it proves its effectiveness by compressing and maximizing the mutual information from the TF-IDF feature space.

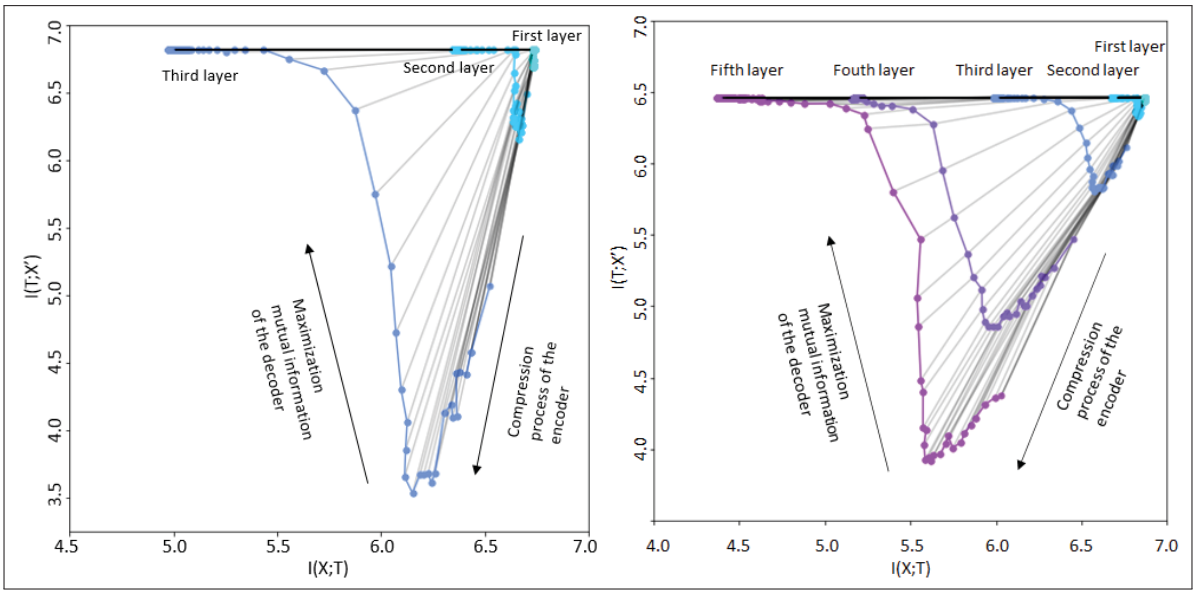

Figure 3.10 The evolution of the layers with epochs in the information plane for three hidden layers (left) and five hidden layers (right)

3.6 Conclusion

First, this study has shown that the participation of an AE in training can effectively compress the feature space of TF-IDF. The AE with a nonlinear activation function can achieve the reconstruction capacity at 86% compared to the original data. It outperforms other approaches such as PCA, NCA, LAE (AE with linear activation function), and stacked AE. It concludes that AE can learn the best representation of the training data due to its lossless compression capacity.

Additionally, the AE also works well with a small clinical dataset, especially in harmony with the information-theoretic mechanism of an AE for a larger dataset and from different data sources. It has two learning phases; the encoder's drift phase by trying to compress the data. The second phase is related to the diffusion phase by maximizing the mutual information process in the decoder. Consequently, it shows the effectiveness of lost information in compressing the data.

By doing so, the interpretability can also be captured as comprehensibility and transparency of the proposed model for decision-making in our CDSS system recommended by (Rudin, 2019).

The second step involves using an MLP-NN to predict the health status based on the compressed feature space. It has been shown that the sparsity reduction for the feature space strongly affects the classifier performance in the downstream task. AE learning algorithm effectively leverages the sparsity reduction. As a result, it helps the MLP-NN classifier achieve 92% accuracy, 91% recall, 91% precision, and 91% f1-score. This efficient ensemble model can outperform all alternative approaches: GaussianNB, LR, RF, MultimonialNB, and SVC.

The proposed approach is still proving successful in cases where data availability is increased. The MLP-NN effectively achieves a better performance after the GaussianNB reaches its maximum capacity. In future work, the optimal parameters will be chosen, and our method will be validated on more datasets. The weak supervision approach will be explored, as it recently proved its effectiveness in 4,000 cardiac magnetic resonance sequences with imperfect labels (Fries *et al.*, 2019); because it can maximize unlabeled data at scale, which is costly to annotate.

Finally, the CDSS is still under development. By combining this NLP algorithm to detect the absence of heart failure with the two other algorithms already developed on hypoxemia detection (Sauthier *et al.*, 2021) and chest X-ray analysis (Zaglam *et al.*, 2014; Yahyatabar *et al.*, 2020), the next step of our study is to implement the resulting CDSS (integration of the three algorithms) within the cyberinfrastructure of the pediatric intensive care unit (PICU) at Sainte-Justine Hospital to diagnose ARDS early. We will then verify the ability of the CDSS to detect ARDS prospectively once the integration with the PICU e-Medical infrastructure will be completed.

CHAPTER 4

A SMALL-SCALE SWITCH TRANSFORMER AND NLP-BASED MODEL FOR CLINICAL NARRATIVES CLASSIFICATION

Thanh-Dung Le[1,2] , Rita Noumeir[1] , Philippe Jouvet[2]

[1] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Research Center at CHU Sainte-Justine Hospital, University of Montreal,
3175 Chem. de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 1C5

Article submitted to « IEEE Journal of Biomedical and Health Informatics » in March 2023.

4.1 Abstract

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

4.2 Introduction

Recent advancements in deep learning have led to the development of Transformer models (Vaswani *et al.*, 2017), which have shown remarkable performance in various natural language processing (NLP) tasks (Tripathy *et al.*, 2021). As a result, there is a growing interest in applying Transformer-based models to clinical applications, such as predicting disease risk (Huang *et al.*, 2022), identifying disease (Ilias & Askounis, 2022), and improving clinical decision-making (Meng, Speier, Ong & Arnold, 2021). These models can be trained on various data sources, including electronic health records (EHRs) (Meng *et al.*, 2021; Blanco, Pérez & Casillas, 2021; Li *et al.*, 2022b), medical imaging (Deng *et al.*, 2022; Li *et al.*, 2022a; Mondal, Bhattacharjee, Singla & Prathosh, 2021), electrogram (Phan *et al.*, 2022; Lu *et al.*, 2022), and genome (Clauwaert & Waegeman, 2020; Huang, Nie, Ni, Luo & Wang, 2020) to extract clinically relevant information and provide accurate predictions. Overall, Transformer models present a powerful tool for clinical applications and can potentially play an increasingly important role in healthcare.

In clinical NLP, Transformers-based models have shown great promise in clinical narrative classification. In this context, clinical narrative refers to patient encounters in EHRs or other clinical documentation. Using Transformers-based models, researchers and clinicians can develop algorithms that automatically classify these narratives based on different criteria, such as diagnosis, treatment, or patient outcomes. This can help streamline clinical workflows and improve patient care by providing more accurate and efficient clinical data processing. Some examples of successful applications of Transformers-based models for clinical narrative classification include identifying clinical coding (López-García, Jerez, Ribelles, Alba & Veredas, 2021, 2023), diagnosing health conditions (Roitero, Portelli, Popescu & Della Mea, 2021; Mugisha & Paik, 2022; Rizwan *et al.*, 2022; Kjell, Sikström, Kjell & Schwartz, 2022), and detecting clinical events (Althari & Alsulmi, 2022; Kim *et al.*, 2023; Yang, Bian, Hogan & Wu, 2020). As such, Transformers-based models have become an increasingly important tool in clinical NLP and are likely to continue playing a significant role in this field (Zhou, Yang, Shi & Ma, 2022).

Despite their many benefits, Transformers-based models for clinical text classification have some limitations that must be considered. One major challenge is the need for large amounts of annotated clinical data to train these models effectively. Clinical data is often scarce and sensitive, which makes it challenging to obtain and annotate in a way that preserves patient privacy (Gao *et al.*, 2021). 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 (Bear Don't Walk IV, Sun, Perotte & Elhadad, 2021). There is a risk of bias in the data used to train these models, leading to errors or disparities in the predictions made (Alimova, Tutubalina & Nikolenko, 2021). 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 (Gillioz, Casas, Mugellini & Abou Khaled, 2020). 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 (Rudin, 2019; Tonekaboni, Joshi, McCradden & Goldenberg, 2019). While Transformers-based models have great potential for clinical text classification, they also require careful attention to their limitations and the potential biases that can arise.

- Computational requirements: If a model lacks the necessary computational capacity, its training efforts will fail, regardless of the learning algorithm's sophistication or the training data's quality (Bhattamishra, Patel & Goyal, 2020). This can be a limiting factor for smaller clinical text or resource-constrained settings.
- Data requirements: Transformer-based models require large amounts of labeled data for training, which may not be available for some clinical text classification tasks, especially for rare or low-frequency conditions (Zeng, Linwood & Liu, 2022).
- Domain-specific language: Clinical text is highly domain-specific and contains jargon and abbreviations that may not be covered by general language models such as Transformers. This can lead to suboptimal performance on clinical text classification tasks (Gu *et al.*, 2021).
- Interpretability: Transformer-based models are highly complex and difficult to interpret, making it challenging to understand how the model makes predictions, which is essential for clinical decision-making (Zafar *et al.*, 2021).

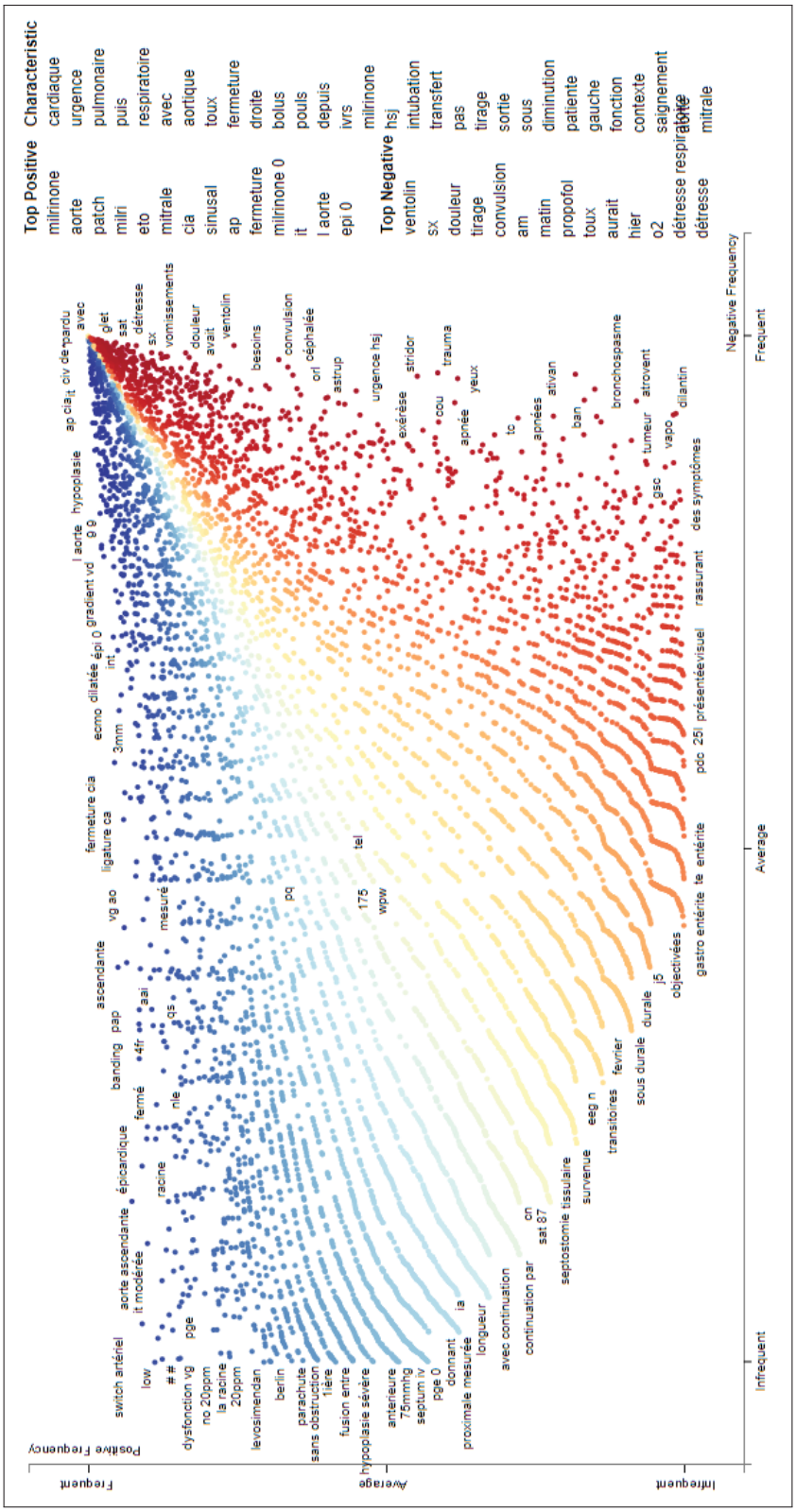

Figure 4.1 French clinical note at CHUSJ illustration by using Scattertext visualization

Another significant limitation of using Transformer-based models for clinical text classification is that they may not perform as well for languages other than English and that are in limited availability. Most Transformer-based models have been developed and trained on English-language text, and their performance may suffer when applied to other languages (AlShuweihi, Salloum & Shaalan, 2021). This is particularly important in the clinical context, where patient data can be collected in many languages. Another challenge is that clinical datasets are often small and imbalanced, making it difficult to train accurate models using Transformer-based (Névéol *et al.*, 2018). 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 (Raghu, Unterthiner, Kornblith, Zhang & Dosovitskiy, 2021). Overall, while Transformer-based models offer many advantages for clinical text classification, their effectiveness is influenced by the data's language and the training dataset's size and quality.

This study aims to overcome the challenges of using Transformer-based models for clinical text classification for a small French clinical note by employing the Mixture-of-expert (MoE) framework from the recent Switch Transformer model developed by Google (Fedus, Zoph & Shazeer, 2021). Switch Transformer is an extension of the Transformer architecture motivated by the original model's self-attention mechanisms. Still, it uses an MoE mechanism to address the limitations of the conventional Transformer (Vaswani *et al.*, 2017). A key technical difference between Switch Transformers with an MoE mechanism and Transformers with self-attention is how they handle the modeling of complex input-output relationships. An example of the effectiveness of MoE has been proven by (Xue *et al.*, 2022); that study shows that the approach of using parameter sharing to compress along the depth of the model, which is used in existing works, is limited in terms of performance. To improve the model's capacity, the authors propose scaling along the model's width by replacing the feed-forward network with an MoE. This allows for better modeling capacity and potentially better performance.

Additionally, the study (Lazaridou *et al.*, 2021) suggests that simply increasing the model's size is insufficient to address the issue of performance degradation over time from neural language models. However, the researchers found that using models that continuously update their knowledge with new information can help alleviate this problem. While Transformers with self-attention model these relationships through a single attention mechanism that captures dependencies between all input and output positions, Switch Transformers with an MoE mechanism decompose the problem into smaller, simpler sub-problems, each handled by a different "expert" model. In other words, instead of using a single global attention mechanism, Switch Transformers employ multiple local attention mechanisms focusing on different input aspects. The gating mechanism used in Switch Transformers selects which expert model to use for a given input, depending on the context. Therefore, this approach can potentially improve the modeling of complex input-output relationships and increase the model's efficiency, especially when dealing with complex data from the clinical domain. This is particularly important in clinical data, where information is often conveyed through complex and nuanced language. By employing this approach, our study aims to improve the accuracy and generalizability of clinical text classification models for small datasets in languages other than English. We have made several significant contributions to clinical text classification using Transformer-based models.

- First, our study demonstrates a comprehensive implementation of a simplified Switch Transformer model from scratch. This would allow other researchers to understand and replicate the methodology used in the study, which is essential for building on and advancing this work.
- Second, our study provides experimental evidence showing the limitations of Transformer-based models in terms of generalization gap and sharp minima. This highlights the importance of carefully selecting and preprocessing the data used to train these models to avoid overfitting and improve generalization performance.
- Finally, our study illustrates the interpretable output of the model by adapting the Integrated Gradients (IG) (Sundararajan, Taly & Yan, 2017). It provides a way to attribute importance to the input features of a model, allowing clinicians and researchers to gain insight into how the model is making its predictions.

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

This paper is organized as follows. Section 4.3 will discuss the materials and methods. Then, the experimental results and discussion will be discussed in section 4.4, and 4.5, respectively. Misclassification cases will be discussed in section 4.6. Finally, section 4.7 provides concluding remarks.

4.3 Materials and Methods

4.3.1 French Clinical Data at CHUSJ

The clinical decision support system (CDSS) system in the CHU Sainte Justine (CHUSJ) hospital aims to improve the diagnosis and management of acute respiratory distress syndromes (ARDS) in real-time by automatically screening data from electronic medical records, chest X-rays, and other sources. Previous studies have found that the diagnosis of ARDS is often delayed or missed in many patients (Bellani *et al.*, 2016), emphasizing the need for more effective diagnostic tools. Three main conditions must be detected to diagnose ARDS: hypoxemia, chest X-ray

infiltrates, and absence of cardiac failure (Group *et al.*, 2015). The research team at CHUSJ has developed algorithms for detecting hypoxemia (Sauthier *et al.*, 2021), analyzing chest X-rays (Zaglam *et al.*, 2014; Yahyatabar *et al.*, 2020), 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 (Le *et al.*, 2022; Le, Noumeir, Rambaud, Sans & Jouvet, 2023c). Implementing these algorithms could increase ARDS diagnosis rates and improve patient outcomes.

This study was conducted following ethical approval from the research ethics board at CHUSJ; and, the study's design focused on identifying cardiac failure in patients within the first 24 hours of admission by analyzing admission and evolution notes during this initial period. Therefore, we conducted a retrospective analysis of EHRs from the Research Center of CHUSJ in this study. The dataset consisted of 580,000 unigrams extracted from 5,444 single lines of short clinical narratives. Of these, 1,941 cases were positive (36% of the total), and 3,503 cases were negative. ScatterText (Kessler, 2017) was utilized to visualize the notes and identified over 580,000 unigrams (n-grams), as depicted in Fig. 4.1. The visualization showcases the most frequent words for positive cases in the upper right corner, negative cases in the lower-left corner, and less frequent words for both cases in the center. The top terms for positive and negative cases are also presented on the right-hand side. Upon inspection, we observed that most top terms for positive cases were positively related to cardiac malfunction, such as milrinone or milri (milrinone), and aorte or aortique valve (aortic valve). In contrast, terms like respiratoire (respiratory), détresse respiratoire (distress respiratory), and 02 (oxygen) indicated respiratory syndromes in negative cases. While the longest n-gram was over 400 words, most n-grams had a length distribution between 50 and 125 words. The average length of the number of characters was 601 and 704, and the average size of the number of digits was 25 and 26 for the positive and negative cases, respectively. We pre-processed the data by removing stop-words and accounting for negation in medical expressions. Numeric values for vital signs (heart rate, blood pressure, etc.) were also included and decoded to account for nearly 4% of the notes that

contained these values. All the notes are short narratives; detailed characteristics can be found in the Supplementary Materials from (Le *et al.*, 2022).

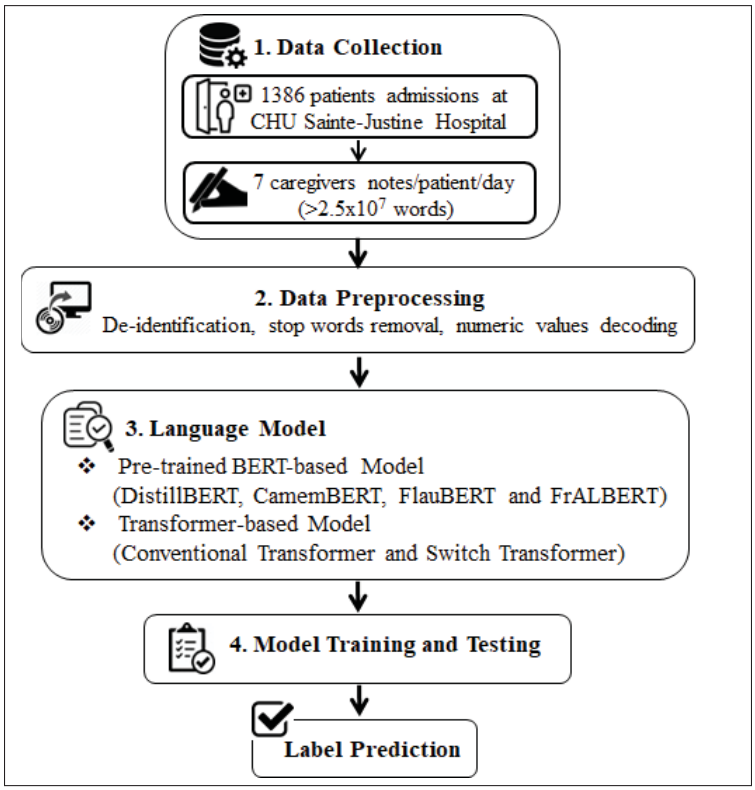

Figure 4.2 Workflow demonstration of the proposed methodology to classify French clinical narratives at CHUSJ hospital

4.3.2 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. 4.2. Moreover, conventional and Switch Transformer models are constructed from scratch to perform the same task. Finally, we compare the performance of all models based on various evaluation metrics for binary classification, including accuracy, precision, recall, F1-score, and area under the

curve (AUC). This study endeavors to offer insights into the efficacy of these models on limited datasets, which is a critical aspect in real-world clinical settings for non-English notes.

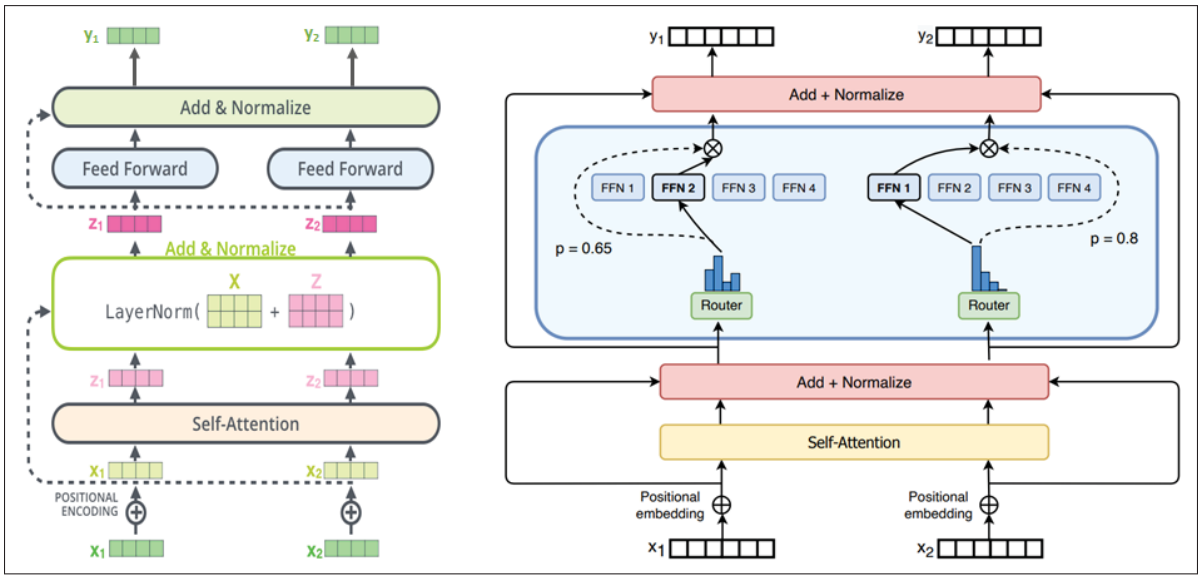

Figure 4.3 Illustration of a Conventional Transformer (Alammar, 2018) (left), and a Switch Transformer (Fedus *et al.*, 2021) (right) encoder block

4.3.2.1 Transformer-based Models

Transformer-based models have been highly effective for various NLP tasks, including text classification. The conventional Transformer model (Vaswani *et al.*, 2017) with multi-head self-attention is a widely used architecture for this task. Shown in Fig. 4.3 (left), its architecture comprises an encoder consisting of multiple layers of multi-head self-attention and feedforward neural networks (FFN). The multi-head self-attention mechanism allows the model to weigh the importance of different words in a sequence based on their semantic relationships, while the FFNs transform the output of the self-attention layer into a more helpful representation. The Transformer's core is the self-attention mechanism based on mathematical expressions (Lin, Wang, Liu & Qiu, 2022). Given a sequence of input embeddings $x_1, ..., x_n$, the self-attention mechanism computes a set of context-aware embeddings $h_1, ..., h_n$ as follows:

$$h_i = \text{Attention} QW_i^Q, KW_i^K, VW_i^V \tag{4.1}$$

where Attention is the scaled dot-product attention function:

$$\text{Attention} Q, K, V = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.2}$$

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

$$\text{MultiHead} Q, K, V = \text{Concat} h_1, \ldots, h_n W^O \tag{4.3}$$

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:

$$\text{FFN} x = \text{ReLU} xW_1 \;\; b_1 W_2 \;\; b_2 \tag{4.4}$$

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.

$$\text{LayerNorm} x = \gamma \frac{x - \mu}{\sqrt{\sigma^2 \;\; \epsilon}} \;\; \beta \tag{4.5}$$

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, γ and β are learned scaling and shifting parameters for layer normalization, and μ and σ are the mean and standard deviation of the input feature activations.

The working mechanism in the Transformer architecture can be summarized into the following steps:

1. Linear Transformation: The input sequence is projected into three vectors, query Q, key K, and value V, by applying a linear transformation to the input embedding.
2. Splitting: The Q, K, and V vectors are then split into multiple heads h_i, allowing the model to simultaneously attend to different aspects of the input sequence Eq. 4.1.
3. Scaled Dot-Product Attention: For each h_i, the model calculates the attention weights between the Q and K vectors by scaling their dot product by the square root of the vector dimension. It calculates each K vector's importance to the corresponding Q vector.
4. Softmax: The resulting attention weights are normalized using a softmax function, ensuring that they sum to 1.
5. Weighted Sum: The attention weights are then used to weigh the V vectors, producing an attention output for each head h_i Eq. 4.2.
6. Concatenation: The attention outputs from each head are concatenated and projected back to the original vector dimension through another linear transformation Eq. 4.3.
7. Feed Forward Network: The resulting output is passed through a feedforward network, which introduces non-linearity and allows the model to capture more complex relationships between the input and output Eq. 4.4.

By performing these steps for each layer in the encoder and decoder, the multi-head self-attention mechanism allows the Transformer architecture to capture rich semantic relationships between different words in a sequence and is highly effective for a wide range of NLP tasks. However, the conventional Transformer architecture has some limitations. One of the main issues is that the self-attention mechanism requires quadratic computation time concerning the input sequence length, making it difficult to scale the model to very long sequences (Raffel *et al.*, 2020), and lower generalizability for a short sequence (Gao *et al.*, 2021). Additionally, the self-attention mechanism treats all positions in the input sequence equally, which may not be optimal for certain types of inputs where some positions are more critical than others. While

the Transformer model has shown state-of-the-art performance on many NLP tasks, it can still struggle to capture complex input-output relationships requiring more specialized models.

Switch Transformers (Fedus *et al.*, 2021) attempt to address these limitations by introducing a mixture of expert (MoE) mechanisms that decompose the problem into smaller, simpler sub-problems, allowing the model to handle long sequences and complex input-output relationships better. As mentioned above, the multi-head self-attention mechanism in the Transformer model is motivated by the need to capture semantic relationships between words in a sequence, but it has limitations when dealing with short sequences (Gao *et al.*, 2021). The MoE mechanisms allow the model to divide the sequence into smaller, more manageable segments and apply different experts to each segment. This approach has improved the model's performance on short sequence tasks and has achieved state-of-the-art results on several benchmarks (Xue *et al.*, 2022; Lazaridou *et al.*, 2021; Fan *et al.*, 2021).

The critical difference in the mathematical equation of the Switch Transformer compared to the conventional Transformer is replacing the FFN with the MoE mechanism, shown in Fig. 4.3 (right). In the conventional Transformer, the FFN consists of two linear layers with a ReLU activation function in between. The MoE mechanism, on the other hand, uses a set of expert networks to learn different aspects of the input data and then combines their outputs with a gating network. It allows the model to dynamically choose between multiple sets of parameters (i.e., expert modules) based on the input. This contrasts the original Transformer model in Eq. 4.4, which uses a fixed set of parameters for all inputs. Formally, the MoE mechanism in the Switch Transformer can be represented by the following equation:

$$z_t = {}_j\, g_j x_t * e_j x_t \tag{4.6}$$

where $g_j x_t$ is a gating function that determines the importance of expert module j for input x_t, and $e_j x_t$ is the output of expert module j for input x_t. The switch mechanism is implemented by

learning the parameters of the gating functions, which are used to select the expert modules dynamically. This allows the model to adapt to different input distributions and perform better on various tasks. Here is a summary of how the MoE mechanism works in the Switch Transformer:

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

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

4.3.2.2 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 (Kenton & Toutanova, 2019) was initially developed for English language processing, it has since been adapted to several other languages, including French. In this context, we will explore some of the most popular pre-trained BERT-based models for French language processing available from Huggingface.

CamemBERT (Martin *et al.*, 2020): This is a pre-trained Transformer-based language model designed explicitly for processing French text. It is based on the Roberta architecture and was trained on a large corpus of French text that was filtered and pre-processed to improve the data quality. Its pre-training objective is a masked language model, where some input tokens are masked, and the model is trained to predict the missing tokens. Overall, CamemBERT is a highly effective tool for processing French language text and can be fine-tuned for specific downstream tasks or used for transfer learning in multilingual settings.

FlauBERT (Le *et al.*, 2020): It is based on the original BERT architecture and was trained on a large corpus of the French text. It has been shown to perform strongly on several natural language processing tasks in French, including named entity recognition and sentiment analysis. It also performs well on tasks related to French morphosyntaxes, such as part-of-speech tagging and dependency parsing. It was trained using a masked language model objective, where a portion of the input tokens are masked, and the model is trained to predict the missing tokens. FlauBERT is a powerful language model for processing French text that can be fine-tuned for specific downstream tasks.

FrALBERT (Cattan, Servan & Rosset, 2021) is a Transformer-based language model designed explicitly for text classification tasks in French. It is based on the ALBERT architecture and was trained on a large corpus of the French text. It has achieved state-of-the-art performance on several text classification tasks in French, including sentiment analysis, news categorization, and toxic comment classification. The model was fine-tuned using a supervised learning approach, where the model was trained on labeled data to predict the correct class label for a given input text. FrALBERT is available for download and can be fine-tuned on specific text classification tasks in French or used for transfer learning in multilingual settings.

DistillBERT (Sanh, Debut, Chaumond & Wolf, 2019) is a smaller and more efficient version of the BERT architecture designed to reduce the computational and storage requirements of the model while maintaining its performance. It was trained on a large corpus of French text and has been shown to perform strongly on various natural language processing tasks, including text

classification. It is particularly useful for text classification tasks in French, such as sentiment analysis and news categorization. DistillBERT is much smaller than the original BERT model, making it more suitable for deployment on resource-constrained devices or in applications where speed and efficiency are a concern.

4.4 Experimental Implementation

Table 4.1 Models Hyperparameters

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

Table 4.1 shows the hyperparameters of different Transformer-based models used in this study, including CamemBERT, DistillBERT, FlauBERT, FrALBERT, Transformer, and Switch Transformer. The hyperparameters compared include hidden layers and total parameters. CamemBERT and FrALBERT have 12 hidden layers, whereas DistillBERT, FlauBERT, Transformer, and Switch Transformer have 6, 6, 4, and 4 hidden layers, respectively. Regarding to total parameters, CamemBERT has the highest number of parameters, with 111 million, followed by DistillBERT with 66.7 million parameters, and FlauBERT with 54.6 million parameters. FrALBERT, Transformer, and Switch Transformer have significantly fewer parameters, with 12.3 million, 2.3 million, and 5.7 million, respectively. The variation in hyperparameters across different models reflects the differences in the architecture and design of the models. This information is crucial for understanding each model's computational complexity and efficiency and helps select the most suitable model.

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

Table 4.2 Hyperparameters of the fine-tuned models

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

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

- **Maximum sequence length**: This is the maximum number of tokens that can be inputted into the model simultaneously. Setting an appropriate maximum sequence length can affect the performance and memory usage of the model. Due to computational constraints, the maximum sequence length varies from 128 to 256.
- **Batch size**: Choosing an appropriate batch size can affect the speed and stability of the training process. We varied the training batch size for each trial, ranging from 4 to 32 (with gradient accumulation as 4), based on the knowledge that training with smaller batches is more effective for highly low-resource language training (Atrio & Popescu-Belis, 2021).
- **Drop-out**: This regularization technique randomly drops out some of the neurons during training to prevent overfitting. The dropout rate determines the proportion of neurons to drop out during each iteration (Srivastava *et al.*, 2014).
- **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) (Kingma & Ba, 2015), AdamW (Adam with weight decay) (Loshchilov & Hutter, 2019) 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 (Gotmare, Keskar, Xiong & Socher, 2019).
- **Number of multi-head attention**: This determines the number of attention heads used in the multi-head attention layer of the Transformer. Increasing the number of attention heads can improve the model's ability to attend to different input parts.
- **Number of experts**: This determines the number of experts used in the MoE layer of the Transformer. Increasing the number of experts can improve the model's ability to handle diverse inputs.

Choosing appropriate values for these hyperparameters requires careful experimentation and tuning to achieve the best possible results. Additionally, optimizing hyper-parameters is essential for achieving high performance in machine learning models, but this process comes with a tradeoff between the quality of the final solution and the time required for computation. However, not all hyperparameters significantly impact model accuracy, and only a few parameters require careful tuning. As reported in (Popel & Bojar, 2018), the model size, learning rate, batch size, and maximum sequence length are the three critical hyper-parameters for Transformer model training. For this reason, grid search can be an efficient approach for optimizing these parameters by simultaneously exploring all possible combinations of intervals. Compared to Bayesian optimization, grid search has advantages in parallelization and flexibility of resource allocation (Yu & Zhu, 2020). In this study, we used grid search to optimize hyper-parameters for model training. The combination with the highest estimated performance was considered the optimal solution, and this approach balances computational efficiency and models' accuracy.

Finally, table 4.2 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 of 0.5 and a cosine decay learning rate. The Transformer and Switch Transformer models use an AdamW optimizer with a dropout rate of 0.35 and a cosine decay learning rate. The Adam_ϵ is only specified for the Transformer and Switch Transformer models and is set to 5*1e-06. The maximum sequence length for all models is set to 256. The fine-tuning process for the pre-trained BERT-based model was performed for 40 epochs, while the Transformer and Switch Transformer models were fine-tuned for 70 epochs. Additionally, the GlorotNormal initializer (Glorot & Bengio, 2010), batch normalization (Ioffe & Szegedy, 2015; Bjorck, Gomes, Selman & Weinberger, 2018) are employed for models' stability, and balancing the classes by using the Bayes Imbalance Impact Index (Lu *et al.*, 2019) to deal with the imbalanced classes. Then, these hyperparameters were carefully chosen to achieve optimal performance and prevent overfitting.

The data was divided into 80% training and 10% validation and 10% testing. To assess the performance of our method, metrics including accuracy, precision, recall, and F1 score were used (Goutte & Gaussier, 2005). These metrics are defined as follows.

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

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

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

$$\text{F1-Score} = \frac{2^{\star}\text{Precision}^{\star}\text{Recall}}{\text{Precision}\quad \text{Recall}}$$

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.

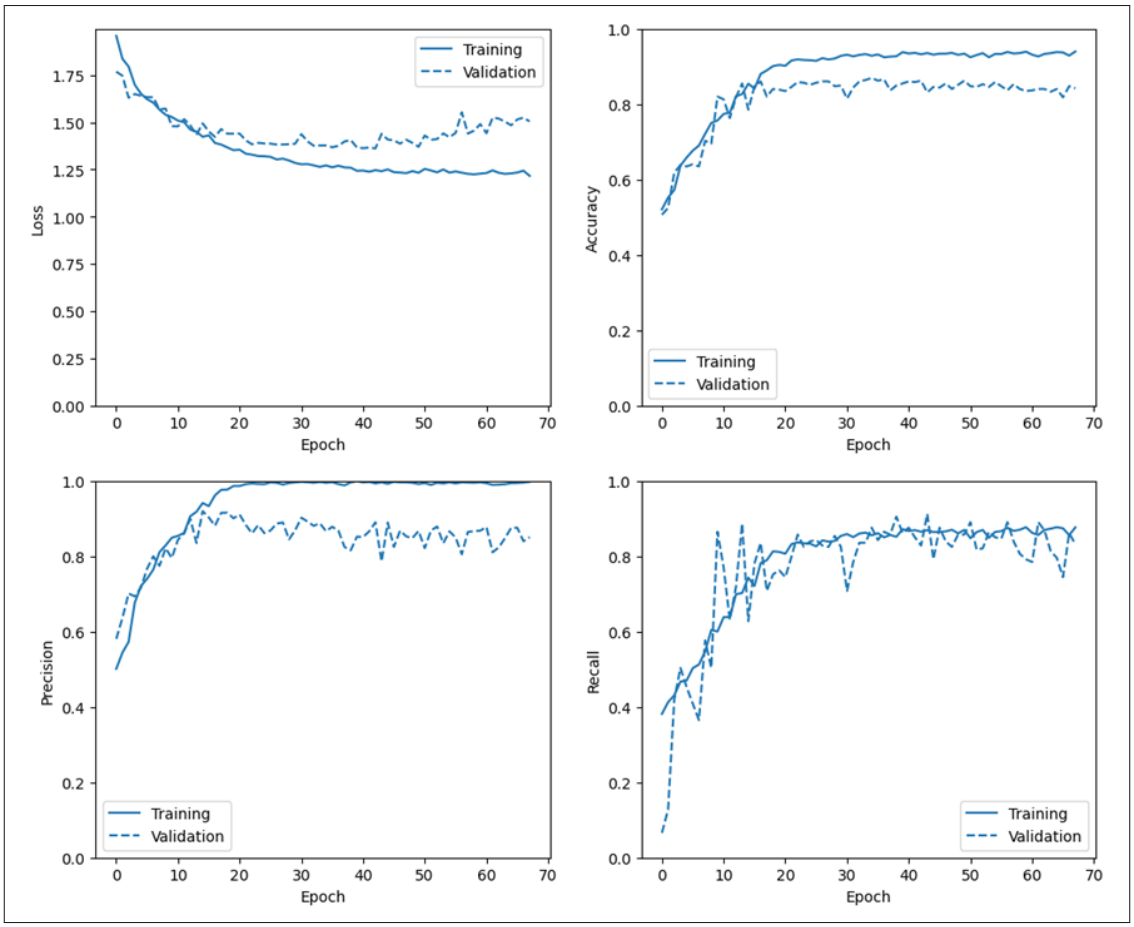

Figure 4.4 Training and validation performance results from Switch Transformer model

4.5 Results and Discussion

During training and validation shown in Fig. 4.4, the Switch Transformer model showed a gradual decrease in loss with increasing epochs. The loss started to converge after around 20 epochs and reached its minimum at the 30th epoch. Applying the early stopping at this point helped prevent the model's overfitting. The accuracy and precision of the model showed a smooth convergence to their optimal values for both the training and validation phases. However, the recall values for the two phases were observed to be quite fluctuating. The model's overall performance was good, with high accuracy, precision, and recall. The model's ability to reach its optimal values with smooth convergence and with the help of early stopping indicates the model's effectiveness in the given task.

The results presented in the table 4.3 indicate that careful hyperparameter tuning can result in better performance of Transformer models over pre-trained BERT-based models for the given task. The table compares the performance of six classifiers with metrics such as accuracy, precision, recall, F1, and AUC. The classifiers include DistillBERT, CamemBERT, FlauBERT, FrALBERT, Transformer, and Switch Transformer. The results show that the best-performing classifier in accuracy, precision, recall, F1, and AUC is Switch Transformer, with an accuracy score of 0.87, precision of 0.87, recall of 0.85, F1 score of 0.86, and AUC of 0.92. The Transformer model has the second-best performance with an accuracy score of 0.85. DistillBERT, CamemBERT, and FrALBERT perform comparably well, with accuracy scores ranging from 0.80 to 0.83. The Switch Transformer and Transformer models achieved the best accuracy, precision, recall, F1 score, and AUC. These models demonstrated faster training and evaluation times than others, making them the most suitable options for the given task. However, it is essential to note that FlauBERT achieved the best precision, recall, F1 score, and AUC among all models, although it required longer training and evaluation times. Compared to other methods (excluding fine-tuning), mixture-of-experts (MoEs) is more efficient regarding the computational resources required (Artetxe *et al.*, 2021). The study suggests that Switch Transformer and Transformer models are the most suitable for the given task, given their high performance and faster training and evaluation times. Overall, these findings suggest that careful selection of Transformer-based models and hyperparameter tuning can significantly improve the performance of small clinical narrative classification.

Table 4.3 A comparison performance of different classifiers

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

Fig. 4.5 compares the confusion matrices obtained from six models. Each confusion matrix presents the number of true positives (TP), false positives (FP), false negatives (FN), and true

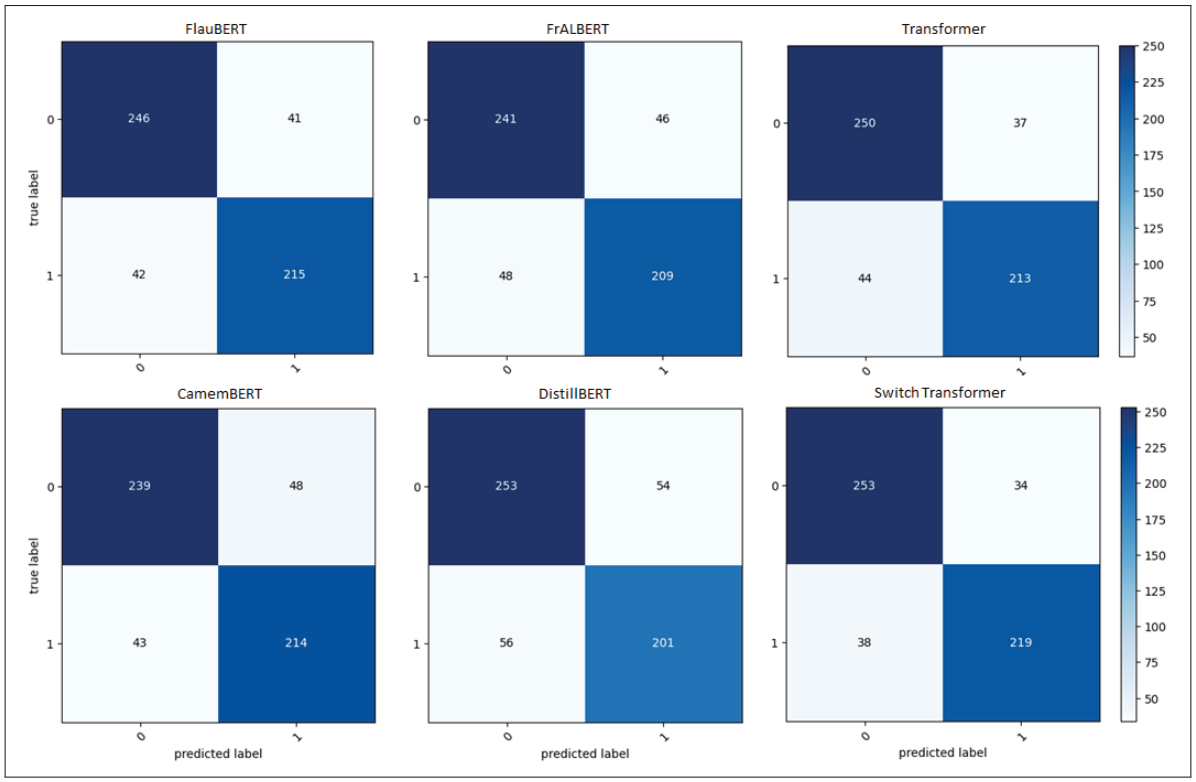

Figure 4.5 Confusion matrix comparison for all classifiers

negatives (TN) for binary classification tasks. This study labels the classes '0' for 'Negative' and '1' for 'Positive.' The Switch Transformer model obtained the highest number of TP and TN, with 253 and 219, respectively. It misclassified 34 instances as false positives and 38 instances as false negatives. DistillBERT, on the other hand, obtained 253 TP and 201 TN, with 54 instances misclassified as false positives and 56 instances as false negatives. FlauBERT and FrALBERT models had similar results with 246 TP and 215 TN and 241 TP and 209 TN, respectively. Both models misclassified around 15% of instances. CamemBERT model obtained 239 TP and 214 TN, with 48 and 43 instances misclassified as false positives and false negatives, respectively. Finally, the Transformer model obtained 250 TP and 213 TN, with 37 and 44 instances misclassified as false positives and false negatives, respectively. In summary, the Switch Transformer model achieved the highest number of correct classifications and the lowest number of misclassifications, followed closely by the DistilBERT and Transformer models. The FlauBERT and FrALBERT models performed similarly, with slightly higher misclassifications.

However, the CamemBERT model had the lowest number of correct classifications and a relatively high number of misclassifications. These results can guide the selection of models for future classification tasks. Particularly, it suggests that simpler models (in terms of the number of parameters) may perform better for non-English and limited clinical narrative datasets.

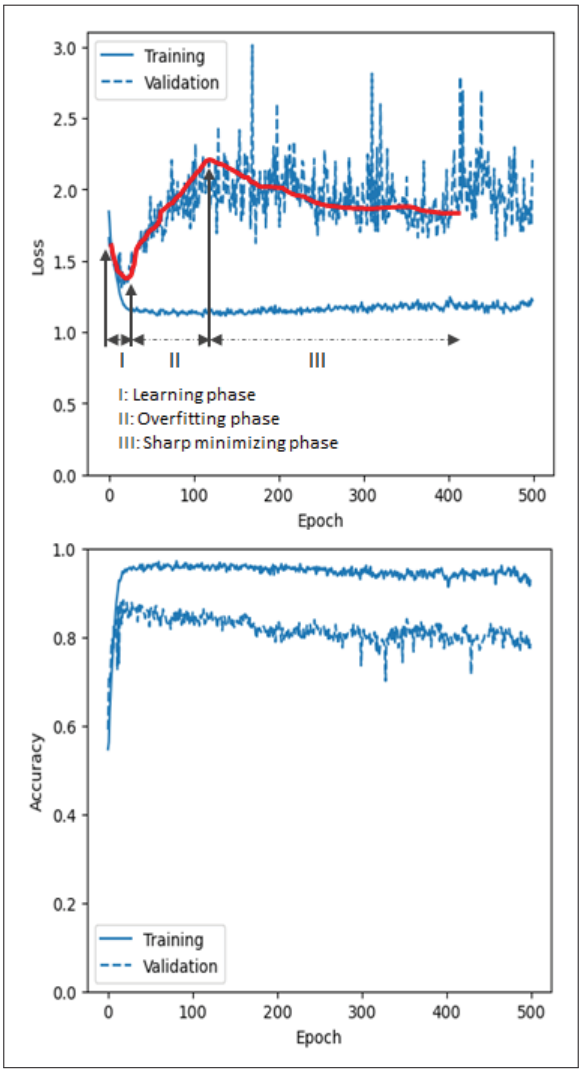

Figure 4.6 Generalization gap and sharp minima during training Switch Transformer without early stopping

Although the Switch Transformer outperforms several other models, including DistillBERT, CamemBERT, FlauBERT, FrALBERT, and the conventional Transformer model, its performance falls short when compared to two of our previous studies (Le *et al.*, 2022, 2023c) that extensively

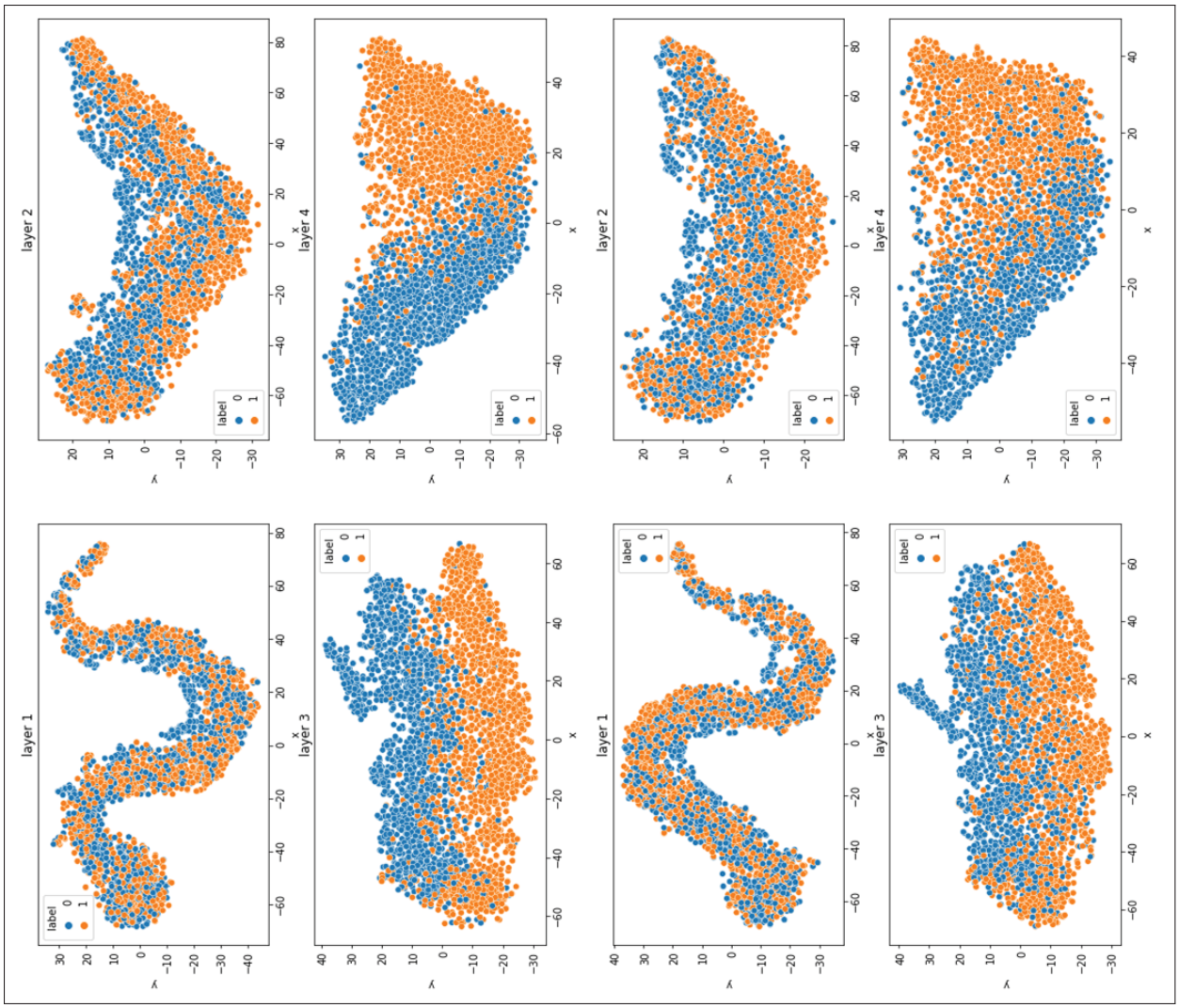

Figure 4.7 Hidden embedding visualization during training (4 left figures) and validation (4 right figures) for the Switch Transformer at the 30th epoch

analyzed a conceptual framework for detecting a patient's health condition from contextual input to output. The proposed framework in those studies utilized a combination of TF-IDF (term frequency-inverse document frequency) and MLP-NN (multilayer perceptron neural network), achieving an overall classification performance of 89% accuracy, 88% recall, and 89% precision. Moreover, sparsity reduction significantly affected classifier performance in downstream tasks, and a generative AE (autoencoder) learning algorithm effectively leveraged sparsity reduction to help the MLP-NN classifier achieve 92% accuracy, 91% recall, 91% precision, and 91%

F1-score. These findings suggest that the simpler frameworks are effective for this specific context and highlight the limitations of the Switch Transformer model.

While the Switch Transformer model has demonstrated promising results in clinical text classification, there is still room for further improvement of its performance. One possible area of investigation is the training methodology, as suggested by previous research (Hoffer, Hubara & Soudry, 2017; Nakkiran *et al.*, 2021). 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. 4.6. 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 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 (Keskar, Mudigere, Nocedal, Smelyanskiy & Tang, 2017).

Furthermore, we propose a novel perspective on this behavior and find a better illustration, viewing them through hidden embedding visualization for each layer during training and validation to explain their behavior. To illustrate this perspective, we present detailed visualizations of the Switch Transformer embedding for each layer (from 1 to 4) in Figure 4.7. We utilize t-SNE, a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data into lower-dimensional data (2 dimensions in our case). By analyzing the hidden embedding from the model, we successfully observe the difference between the training and validation processes. The four top figures illustrate that after the 30th epoch, the model successfully separates the two classes (1: positive, 0: negative) in each hidden layer. Remarkably, the last

hidden layer (4th layer) achieves perfect classification accuracy of 98% on the training set. However, this level of performance does not carry over to the validation set at the same epoch. The four bottom figures demonstrate that the two classes overlap, and the model cannot learn a clear boundary between them, resulting in only 87% validation accuracy. Therefore, we observe a generalization gap between the training and validation for a large model with small data.

4.6 Misclassification Interpretability

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

Technically, Integrated Gradients (IG) (Sundararajan *et al.*, 2017) are a powerful interpretability technique for explaining the predictions of deep learning models, including the Transformers model used in clinical text classification. IG provides a way to attribute importance to the input features of a model, allowing clinicians and researchers to gain insight into how the model is making its predictions. Then, we compared this information with the information from the classifier based on the IG methods. This helped us identify misclassification sources and improve our classifiers' accuracy in detecting cardiac failure.

The results in Fig. 4.8 demonstrate the Transformer model's ability to calculate attribution scores to predict output based on input features. The sign of the attribution score indicates the direction of the feature's influence on the output: a positive score means that the feature positively influences the output, while a negative score indicates a negative influence. However, the model did not perform well on the task at hand. The correct labeling of the data requires

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 géméllaire 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 4.8 The highlighted misclassification cases from the Switch Transformer model

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 (Bear Don't Walk IV *et al.*, 2021). Second, the tasks in the clinical domain often have a low signal-to-noise ratio, where the presence of a few essential keywords may suffice to determine a specific label. In contrast, Transformer's training process involves learning intricate and nuanced relations between all words in the pretraining corpus, which may not be relevant for the classification task and may shift attention away from the critical keywords (Gao *et al.*, 2021).

4.7 Conclusion

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

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

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

These findings suggest that careful selection of Transformer-based models and hyperparameter tuning can significantly improve the performance of clinical narrative classification tasks. Especially the CDSS at CHUSJ is currently under development. By combining this NLP algorithm to detect the absence of heart failure with the two other algorithms already developed on hypoxemia detection (Sauthier *et al.*, 2021) and chest, X-ray analysis (Zaglam *et al.*, 2014; Yahyatabar *et al.*, 2020), the next step of our study is to implement the resulting CDSS (integration

of the three algorithms) within the cyberinfrastructure of the pediatric intensive care unit (PICU) at Sainte-Justine Hospital to diagnose ARDS early. We will then verify the ability of the CDSS to detect ARDS prospectively once the integration with the PICU e-Medical infrastructure is completed.

4.8 Future Works

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

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

CONCLUSION AND RECOMMENDATIONS

With the rise of Artificial Intelligence, Quebec has announced its ambition to revolutionize three sectors by developing advanced Artificial Intelligence technologies. These three sectors are health, finance, and the intelligent city reported in "Strategy for the Development of Quebec's Artificial Intelligence Ecosystem", a mandate from Quebec Économie, Science et Innovation in May 2018. In its 2018-2019 budget, the Quebec government has tagged specific priority sectors in health, in which more than $5.4B in additional investment is planned. The goal is to engage health informatics, data analytics in health databases, and machine learning techniques. Therefore, the successful development of the proposed predictive model to identify heart failure, based on natural language processing by using clinical notes and relevant prescribed parameters, yields multiple uses in health from Quebec's Artificial Intelligence strategy.

First, it prevents the continuous loss of scientific information from significant data points ($4x10^8$) on more than 1300 patients, uniquely stored in the database at CHU Sainte-Justine. Moreover, the data is continuously collected; it is a good starting point to learn from it effectively. Hence, it will be innovative for casual data tracking for later use in clinical settings. This work highlights the value of heterogeneous datasets combined with data analytics in improving proactive health prevention and intervention programs and accelerating precision care.

Second, educational use of the drawn result allows students and researchers to learn how cardiac disease evolves through historical clinical narratives. Standard criteria and other variables that may directly or indirectly impact cardiac disease do not explain the relation. The expected outcomes can help overcome diagnosis issues and improve disease prevention. In particular, wide-scale uptake and implementation that impact decision-making will lead to changes in the health system's care delivery and patient empowerment.

Third, it promotes the optimization and consistency of care for children with pediatric acute respiratory distress syndrome by detecting a cardiac failure as recommended by experts in

pediatric acute lung injury. Then, patients are predicted for their early-stage health condition and have an appropriate treatment regimen. The syndrome of acute lung injury and cardiac failure are very similar; it is reported by LUNG SAFE Investigators ESICM Trials Group (2016) that 40% of cases of acute respiratory distress syndrome were not recognized at any time during a patient's stay in the intensive care unit.

Furthermore, the theoretical and practical issues plaguing the applications of machine learning and statistical learning representation techniques will be effectively contributed to biomedical information processing at Laboratoire de traitement de l'information en santé (LATIS), École de technologie supérieure, and other academic researchers. The contributions include scientific attention to clinical data interpretation, hyper-parameter optimization, training acceleration, and improving machine learning algorithm prediction accuracy.

Finally, this research is part of a research program that aims to develop a clinical decision-support system in real-time for the management of critically ill patients at CHU Sainte-Justine. Therefore, the developed algorithm can bring the success of a clinical decision support system implementation. The system helps intensive care clinicians to make informed decisions concerning the evolution of cardiac disease. Also, the data validation process of variables in acute respiratory distress syndrome analysis will be effectively elaborated. Consequently, it would enable clinicians to anticipate potentially related events in a child's health that may require special attention. This research will help the healthcare system in Montreal to be recognized as an international leader, strengthening the position of École de technologie supérieure and CHU Sainte-Justine Research Center internationally.

Technically, this study presents a framework for detecting cardiac failure in children at CHUSJ using natural language processing (NLP) techniques. We employ both learning representation and machine learning algorithms to process French clinical text and identify a patient's health condition from the contextual input to the contextual output. Our framework combines TF-IDF

and MLP-NN, and we demonstrate that feature selection from the learning representation vector space can further improve performance. Our case study also shows that encoding decimal points as a string "DOT" helps retain the information from numerical values in clinical notes. Our proposed framework achieves an overall classification performance with 89% accuracy, 88% recall, and 89% precision.

Furthermore, we show that using an autoencoder (AE) in training can effectively compress the feature space of TF-IDF. Compared to other approaches such as PCA, NCA, LAE, and stacked AE, the AE with a nonlinear activation function achieves the best reconstruction capacity at 86% compared to the original data. The AE can learn the best representation of the training data due to its lossless compression capacity, making it an effective mechanism for interpretability and transparency in our CDSS system.

The second step involves using an MLP-NN to predict the health status based on the compressed feature space. We show that the sparsity reduction for the feature space strongly affects the classifier performance in the downstream task, and the AE learning algorithm effectively leverages the sparsity reduction. Our efficient ensemble model achieves 92% accuracy, 91% recall, 91% precision, and 91% f1-score, outperforming all alternative approaches.

We also compare the performance of six classifiers on a binary classification task and demonstrate that careful hyperparameter tuning can significantly improve the performance of Transformer models over pre-trained BERT-based models. The Switch Transformer model achieves the highest performance in Accuracy, Precision, Recall, F1, and AUC, with an accuracy score of 0.87, precision of 0.87, recall of 0.85, F1 score of 0.86, and AUC of 0.92. The results also indicate the need for further research to improve Transformer models' performance in clinical settings.

Finally, we present the confusion matrices obtained from six models, with the Switch Transformer model obtaining the highest number of correct classifications and the lowest number of misclassifications. Our study demonstrates the effectiveness of our proposed framework and provides valuable insights for developing NLP techniques in clinical settings.

Future research should carefully consider the potential effects of numerical values alongside unstructured notes. One promising approach is to investigate an algorithm that can automatically extract and represent numerical values from clinical notes using a semantic neural network to determine the boundaries and extract the numerical values from the text. Furthermore, the algorithm's effectiveness can be evaluated using generative learning (Dua *et al.*, 2019). In addition, the weak supervision approach should be explored as it has proven effective in maximizing unlabeled data at scale and can be applied to expand the dataset (Fries *et al.*, 2019).

To improve the generalizability of the findings, future work should consider expanding the dataset to include multiclass classification tasks and examining the impact of fine-tuning on Transformer-based models' performance. Potential solutions to improve the performance of the study include model optimization, data augmentation, domain-specific pre-training, and explainable AI. For example, Transformer-based models can be optimized using distillation or pruning methods to reduce computational requirements while maintaining accuracy. Data augmentation techniques such as synthetic data generation or unsupervised learning can increase the amount of labeled data available for training Transformer-based models. Pre-trained Transformer-based models on clinical text data can be employed to improve their understanding of domain-specific language and performance on clinical text classification. Lastly, researchers can develop techniques to make Transformer-based models more interpretable using attention visualization or sensitivity analysis to understand which parts of the input text the model focuses on during prediction.

In conclusion, there are several potential future research directions as follows:

Addressing the challenges inherent in our research domain, the initial focus lies in enhancing domain-specific training, given the constraints of limited data availability. As established in our study (Le *et al.*, 2022), harnessing numerical attributes effectively leads to an improved classifier for the ultimate end-task classification. However, the reliance on manually designed methods remains a drawback in current approaches. To surmount this limitation, future endeavors should be directed towards autonomous acquiring representations from these numerical attributes. Simultaneously, considering the overarching goal of facilitating ADRS diagnosis, our research has concentrated on singular modalities: chest X-ray infiltrations, laboratory vital sign time series data for oxygenation level estimation, and cardiac failure absence identification. To amplify the impact of our work, future investigations should pivot towards embracing a multimodal learning paradigm, as depicted in Figure 5.1. This innovative framework holds the potential to extract insights from diverse data modalities, thereby enriching the real-time diagnosis of ADRS in a more comprehensive and effective manner.

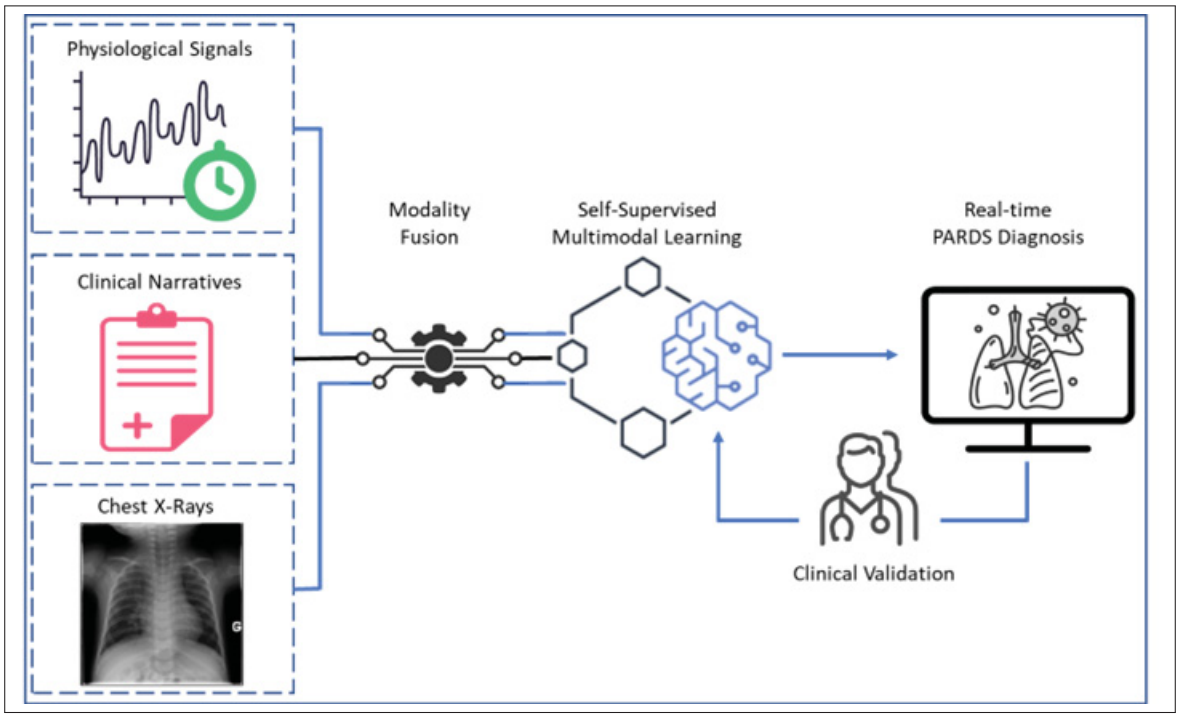

Figure 5.1 A proposed self-supervised multimodal learning to combine tri-modality for real-time PARDS diagnosis

Additionally, our utilization of autoencoders (Le *et al.*, 2023c) has proven to be instrumental in effectively mitigating the sparsity issue inherent in TF-IDF's learning representation feature space. However, the scalability of these techniques may be eclipsed by the potential of Transformers with expanded data availability. Nonetheless, the efficacy of training Transformer models with modest datasets poses a significant challenge, evident in their susceptibility to a generalization gap and convergence challenges when applied to smaller data contexts (Le, Jouvet & Noumeir, 2023a; Macabiau, Le, Albert, Jouvet & Noumeir, 2023). As such, finding strategies to adapt Transformer models within constrained datasets becomes crucial. And model optimization that can improve and adapt the Transformer models for a small dataset is necessary. A potential avenue involves the integration of the Gated Residual Network (GRN) as an intermediate component within Transformer-based classifiers (Le, Macabiau, Jouvet & Noumeir, 2023b). The GRN, characterized by its Gated Linear Units (GLUs), addresses the intricacies of uncertain relationships between inputs and targets, providing nonlinear processing only when essential. By capitalizing on GLUs, the GRN harmonizes information emphasis or suppression based on task-specific requirements. This innovation significantly improves the Transformer convergence, resulting in smoother loss curves during training and validation. These advancements culminate in substantial gains in performance, showcasing the remarkable performance of the GRN-Transformer. Navigating this complexity holds the key to unlocking the full potential of Transformers in these scenarios.

Finally, our experiments confirmed certain limitations of the Transformer architecture in generalizing for small clinical text classification tasks. This limitation often arises due to the model's reliance on specific keywords or phrases contributing to the classifier's final output. Nevertheless, one can mitigate this shortcoming by integrating human feedback and leveraging reasoning capabilities derived from large language models (LLMs), particularly those built upon the advanced architecture of the Transformer.

The recent emergence of large language models has reignited discussions about the potential of these models to replicate human cognitive capacities, especially when trained with extensive data. A focal point of these discussions is the LLMs' ability to reason about unfamiliar challenges on a zero-shot basis without any prior direct training. Such an ability is reminiscent of human cognitive reasoning, often driven by analogy.

A study (Webb, Holyoak & Lu, 2023) revealed that GPT-3, in particular, demonstrated a commendable aptitude for abstract pattern recognition, often equating or even exceeding human abilities in various contexts. Preliminary evaluations of GPT-4 have hinted at even superior performance metrics. Our results suggest that LLMs, including GPT-3, can intuitively solve a wide array of problems by drawing analogies, even when confronted with them on a zero-shot basis.

Another study (Singhal *et al.*, 2023) substantiates that as the scale of these models grows and instruction prompts are fine-tuned, there's a noticeable improvement in knowledge recall and reasoning. This suggests the promising potential of LLMs in the medical domain. However, our human evaluations also spotlight the existing limitations of the current LLMs. This underscores the significance of developing robust evaluation frameworks and refining methods to craft safe and efficient LLMs suitable for clinical applications.

In short, as we navigate the intricacies of domain-specific training in the face of limited data, the apparent path lies in the fusion of automated representation acquisition and embracing a machine learning paradigm. Our research has shown the power of conventional neural networks (MLP-NN), yet as the landscape evolves, the potential of Transformer models, incredibly when fine-tuned for smaller datasets, cannot be overlooked. The recent strides in large language models, exemplified by GPT-3, offer a glimpse into the future of clinical text classification. Their remarkable pattern recognition abilities, even in zero-shot scenarios, with human feedback and advanced architectures, hold immense promise. However, with every advancement, we must

address existing limitations and continuously refine our methods, ensuring we harness the best technology for holistic, accurate, and real-time diagnoses in medical applications.

BIBLIOGRAPHY

Afzal, N., Sohn, S., Abram, S., Scott, C. G., Chaudhry, R., Liu, H., Kullo, I. J. & Arruda-Olson, A. M. (2017). Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. *Journal of vascular surgery*, 65(6), 1753–1761.

Agarwal, A., Baechle, C., Behara, R. & Zhu, X. (2017). A natural language processing framework for assessing hospital readmissions for patients with COPD. *IEEE journal of biomedical and health informatics*, 22(2), 588–596.

Alammar, J. (2018). The illustrated transformer. *The Illustrated Transformer–Jay Alammar–Visualizing Machine Learning One Concept at a Time*, 27, 1-2.

Alimova, I., Tutubalina, E. & Nikolenko, S. I. (2021). Cross-Domain Limitations of Neural Models on Biomedical Relation Classification. *IEEE Access*, 10, 1432–1439.

Alomrani, M. A. (2021). A Critical Review of Information Bottleneck Theory and its Applications to Deep Learning. *arXiv:2105.04405*, 1-2.

AlShuweihi, M., Salloum, S. A. & Shaalan, K. (2021). Biomedical corpora and natural language processing on clinical text in languages other than English: a systematic review. *Recent Advances in Intelligent Systems and Smart Applications*, 491–509.

Althari, G. & Alsulmi, M. (2022). Exploring transformer-based learning for negation detection in biomedical texts. *IEEE Access*, 10, 83813–83825.

Anowar, F., Sadaoui, S. & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, 40, 100378.

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4, 40–79.

Artetxe, M., Bhosale, S., Goyal, N., Mihaylov, T., Ott, M., Shleifer, S., Lin, X. V., Du, J., Iyer, S., Pasunuru, R. et al. (2021). Efficient large scale language modeling with mixtures of experts. *arXiv preprint arXiv:2112.10684*, 1-2.

Atrio, À. R. & Popescu-Belis, A. (2021). Small Batch Sizes Improve Training of Low-Resource Neural MT. *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pp. 18–24.

Bartlett, P. L. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2), 525–536.

Bear Don't Walk IV, O. J., Sun, T., Perotte, A. & Elhadad, N. (2021). Clinically relevant pretraining is all you need. *J Am Med Inform Assoc*, 28(9), 1970–1976.

Beaulieu-Jones, B., Finlayson, S. G., Chivers, C., Chen, I., McDermott, M., Kandola, J., Dalca, A. V., Beam, A., Fiterau, M. & Naumann, T. (2019). Trends and focus of machine learning applications for health research. *JAMA network open*, 2(10), e1914051–e1914051.

Bellani, G., Laffey, J. G., Pham, T., Fan, E., Brochard, L., Esteban, A., Gattinoni, L., van Haren, F., Larsson, A., McAuley, D. F. et al. (2016). Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *Jama*, 315(8), 788–800.

Bengio, Y., Courville, A. & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798–1828.

Berner, E. S. (2007). *Clinical decision support systems*. Springer.

Bhattamishra, S., Patel, A. & Goyal, N. (2020). On the Computational Power of Transformers and Its Implications in Sequence Modeling. *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 455–475.

Bjorck, N., Gomes, C. P., Selman, B. & Weinberger, K. Q. (2018). Understanding batch normalization. *Advances in Neural Information Processing Systems*, 31, 1-2.

Blanco, A., Pérez, A. & Casillas, A. (2021). Exploiting ICD Hierarchy for Classification of EHRs in Spanish Through Multi-Task Transformers. *IEEE J. Biomed. Health Inform.*, 26(3), 1374–1383.

Cai, T., Zhang, L., Yang, N., Kumamaru, K. K., Rybicki, F. J., Cai, T. & Liao, K. P. (2019). EXTraction of EMR numerical data: an efficient and generalizable tool to EXTEND clinical research. *BMC medical informatics and decision making*, 19(1), 226.

Cattan, O., Servan, C. & Rosset, S. (2021). On the Usability of Transformers-based Models for a French Question-Answering Task. *International Conference on Recent Advances in Natural Language Processing, 2021*, pp. 244–255.

Chen, Q., Yao, L. & Yang, J. (2016). Short text classification based on LDA topic model. *IEEE International Conference on Audio, Language and Image Processing*.

Chollet, F. (2015). keras.

Clauwaert, J. & Waegeman, W. (2020). Novel transformer networks for improved sequence labeling in genomics. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 19(1), 97–106.

Curto, S., Carvalho, J. P., Salgado, C., Vieira, S. M. & Sousa, J. M. (2016). Predicting ICU readmissions based on bedside medical text notes. *2016 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, pp. 2144–a.

Deléger, L. & Grouin, C. (2012). Detecting negation of medical problems in French clinical notes. *Proceedings of the 2nd ACM sighit international health informatics symposium*, pp. 697–702.

Demner-Fushman, D., Chapman, W. W. & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5), 760–772.

Demuth, H. B. (2014). *Neural network design*. Martin Hagan.

Deng, Z., Cai, Y., Chen, L., Gong, Z., Bao, Q., Yao, X., Fang, D., Yang, W., Zhang, S. & Ma, L. (2022). Rformer: Transformer-based generative adversarial network for real fundus image restoration on a new clinical benchmark. *IEEE J. Biomed. Health Inform.*, 26(9), 4645–4655.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S. & Gardner, M. (2019). DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378.

Dubois, S., Romano, N., Kale, D. C., Shah, N. & Jung, K. (2017). Learning effective representations from clinical notes. *stat*, 1050, 15.

Dynomant, E., Lelong, R., Dahamna, B., Massonnaud, C., Kerdelhué, G., Grosjean, J., Canu, S., Darmoni, S. J. et al. (2019). Word embedding for the French natural language in health care: comparative study. *JMIR medical informatics*, 7(3), e12310.

Dzisevič, R. & Šešok, D. (2019). Text classification using different feature extraction approaches. *IEEE Open Conference of Electrical, Electronic and Information Sciences*.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V. et al. (2021). Beyond English-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1), 4839–4886.

Fan, Y. & Zhang, R. (2018). Using natural language processing methods to classify use status of dietary supplements in clinical notes. *BMC medical informatics and decision making*, 18, 15–22.

Fedus, W., Zoph, B. & Shazeer, N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res*, 23, 1–40.

Fodeh, S. J., Li, T., Jarad, H. & Safdar, B. (2019). Classification of patients with coronary microvascular dysfunction. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 17, 1-2.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3, 1-2.

Fries, J. A., Varma, P., Chen, V. S., Xiao, K., Tejeda, H., Saha, P., Dunnmon, J., Chubb, H., Maskatia, S., Fiterau, M. et al. (2019). Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nature Communications*, 10(1), 1–10.

Gao, S., Alawad, M., Young, M. T., Gounley, J., Schaefferkoetter, N., Yoon, H. J., Wu, X.-C., Durbin, E. B., Doherty, J., Stroup, A. et al. (2021). Limitations of Transformers on Clinical Text Classification. *IEEE journal of biomedical and health informatics*, 1-2.

Gárate-Escamila, A. K., El Hassani, A. H. & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, 19, 100330.

Garg, S. & Liang, Y. (2020). Functional Regularization for Representation Learning: A Unified Theoretical Perspective. *Advances in Neural Information Processing Systems*, 33, 1-2.

Gehring, J., Miao, Y., Metze, F. & Waibel, A. (2013). Extracting deep bottleneck features using stacked auto-encoders. *IEEE international conference on acoustics, speech and signal processing*.

Geiger, B. C. (2021). On Information Plane Analyses of Neural Network Classifiers–A Review. *IEEE Trans. Neural Netw. Learn. Syst.*, 1-2.

Geiger, B. C. & Kubin, G. (2020). Information Bottleneck: Theory and Applications in Deep Learning. Multidisciplinary Digital Publishing Institute.

Ghosh, J. & Shuvo, S. B. (2019). Improving Classification Model's Performance Using Linear Discriminant Analysis on Linear Data. *IEEE International Conference on Computing, Communication and Networking Technologies*.

Gillioz, A., Casas, J., Mugellini, E. & Abou Khaled, O. (2020). Overview of the Transformer-based Models for NLP Tasks. *15th Conference on Computer Science and Information Systems (FedCSIS)*, pp. 179–183.

Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256.

Gold, R., Larson, A. E., Sperl-Hillen, J. M., Boston, D., Sheppler, C. R., Heintzman, J., McMullen, C., Middendorf, M., Appana, D., Thirumalai, V. et al. (2022). Effect of Clinical Decision Support at Community Health Centers on the Risk of Cardiovascular Disease: A Cluster Randomized Clinical Trial. *JAMA Network Open*, 5, 1-2.

Goldberger, J., Hinton, G. E., Roweis, S. & Salakhutdinov, R. R. (2005). Neighbourhood components analysis. *NeurIPS*, pp. 513–520.

Google. (2019). Machine Learning Guides Text Classification. Retrieved on 2019-10-21 from: https://developers.google.com/machine-learning/guides/text-classification/step-2-5.

Gotmare, A., Keskar, N. S., Xiong, C. & Socher, R. (2019). A Closer Look at Deep Learning Heuristics: Learning rate restarts, Warmup and Distillation. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Goutte, C. & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *European conference on information retrieval*, pp. 345–359.

Group, P. A. L. I. C. C. et al. (2015). Pediatric acute respiratory distress syndrome: consensus recommendations from the Pediatric Acute Lung Injury Consensus Conference. *Pediatric critical care medicine: a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies*, 428.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1), 1–23.

Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Havrlant, L. & Kreinovich, V. (2017). A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, 46(1), 27–36.

Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786), 504–507.

Hoffer, E., Hubara, I. & Soudry, D. (2017). Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in Neural Information Processing Systems*, 30, 1-2.

Huang, N., Nie, F., Ni, P., Luo, F. & Wang, J. (2020). Sacall: a neural network basecaller for oxford nanopore sequencing data based on self-attention mechanism. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 19(1), 614–623.

Huang, Y.-J., Lin, Y.-T., Liu, C.-C., Lee, L.-E., Hung, S.-H., Lo, J.-K. & Fu, L.-C. (2022). Assessing schizophrenia patients through linguistic and acoustic features using deep learning techniques. *IEEE IEEE Trans. Neural Syst. Rehabilitation Eng.*, 30, 947–956.

Huddar, V., Desiraju, B. K., Rajan, V., Bhattacharya, S., Roy, S. & Reddy, C. K. (2016). Predicting complications in critical care using heterogeneous clinical data. *IEEE Access*, 4, 7988–8001.

Hunter, D., Yu, H., Pukish III, M. S., Kolbusz, J. & Wilamowski, B. M. (2012). Selection of proper neural network sizes and architectures—A comparative study. *IEEE Transactions on Industrial Informatics*, 8(2), 228–240.

Hurley, N. & Rickard, S. (2009). Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10), 4723–4741.

Ilias, L. & Askounis, D. (2022). Explainable identification of dementia from transcripts using transformer networks. *IEEE J. Biomed. Health Inform.*, 26(8), 4153–4164.

Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*, pp. 448–456.

Jain, D. & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, 19(3), 179–189.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European conference on machine learning*, pp. 137–142.

Johnson, A. E., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A. & Clifford, G. D. (2016). Machine learning and decision support in critical care. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, 104(2), 444.

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S. & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5(1), 7–16.

Karparthy, A. (2020). Neural Networks 1. Retrieved on 2020-04-01 from: https://cs231n.github.io/neural-networks-1.

Kenton, J. D. M.-W. C. & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. & Tang, P. T. P. (2017). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Kessler, J. (2017). Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. *Proceedings of ACL 2017, System Demonstrations*, pp. 85–90.

Kim, H. K., Park, Y., Park, Y., Choi, E., Kim, S., You, H. & Bae, Y. S. (2023). Identifying alcohol-related information from unstructured bilingual clinical notes with multilingual transformers. *IEEE Access*, 1-2.

Kim, S.-W. & Gil, J.-M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*, 9(1), 30.

Kingma, D. P. & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kjell, O. N., Sikström, S., Kjell, K. & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports*, 12(1), 3918.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, 14(2), 1137–1145.

Kolyvakis, P., Kalousis, A., Smith, B. & Kiritsis, D. (2018). Biomedical ontology alignment: an approach based on representation learning. *J. Biomed. Semantics*, 9, 1-2.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37, 233–243.

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M. & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of Biomedical Informatics*, 73, 14–29.

Kumar, V., Recupero, D. R., Riboni, D. & Helaoui, R. (2020). Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification from Clinical Notes. *IEEE Access*, 1-2.

Laghmati, S., Cherradi, B., Tmiri, A., Daanouni, O. & Hamida, S. (2020). Classification of Patients with Breast Cancer using Neighbourhood Component Analysis and Supervised Machine Learning Techniques. *IEEE International Conference on Advanced Communication Technologies and Networking*.

Lai, S., Liu, K., He, S. & Zhao, J. (2016). How to Generate a Good Word Embedding. *IEEE Intelligent Systems*, (6), 5–14.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kocisky, T., Ruder, S. et al. (2021). Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34, 29348–29363.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L. & Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2479–2490.

Le, T.-D., Noumeir, R., Rambaud, J., Sans, G. & Jouvet, P. (2021). Machine Learning Based on Natural Language Processing to Detect Cardiac Failure in Clinical Narratives. *36e Congres de la recherche au CHU Sainte-Justine*.

Le, T.-D., Noumeir, R., Rambaud, J., Sans, G. & Jouvet, P. (2022). Detecting of a Patient's Condition From Clinical Narratives Using Natural Language Representation. *IEEE Open Journal of Engineering in Medicine and Biology*, 3, 142–149.

Le, T.-D., Jouvet, P. & Noumeir, R. (2023a). A Small-Scale Switch Transformer and NLP-based Model for Clinical Narratives Classification. *arXiv preprint arXiv:2303.12892*.

Le, T.-D., Macabiau, C., Jouvet, P. & Noumeir, R. (2023b). GRN-Transformer: Enhancing Motion Artifact Detection in PICU Photoplethysmogram Signals. *arXiv preprint arXiv:2308.03722*.

Le, T.-D., Noumeir, R., Rambaud, J., Sans, G. & Jouvet, P. (2023c). Adaptation of Autoencoder for Sparsity Reduction From Clinical Notes Representation Learning. *IEEE Journal of Translational Engineering in Health and Medicine*, 1-2.

Lee, S. & Jo, J. (2021). Information Flows of Diverse Autoencoders. *Entropy*, 23(7), 862.

Levy, O., Goldberg, Y. & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3, 211–225.

Leyli-Abadi, M., Labiod, L. & Nadif, M. (2017). Denoising autoencoder as an effective dimensionality reduction and clustering of text data. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Li, R., Li, Q., Wang, H., Li, S., Zhao, J., Yan, Q. & Wang, L. (2022a). DDPTransformer: Dual-Domain With Parallel Transformer Network for Sparse View CT Image Reconstruction. *IEEE Trans Comput Imaging*, 8, 1101–1116.

Li, Y., Yao, L., Mao, C., Srivastava, A., Jiang, X. & Luo, Y. (2018). Early prediction of acute kidney injury in critical care setting using clinical notes. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 683–686.

Li, Y., Mamouei, M., Salimi-Khorshidi, G., Rao, S., Hassaine, A., Canoy, D., Lukasiewicz, T. & Rahimi, K. (2022b). Hi-BEHRT: Hierarchical Transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE J. Biomed. Health Inform.*, 1-2.

Li, Y. & Yuan, Y. (2017). Convergence Analysis of Two-layer Neural Networks with ReLU Activation. *Advances in Neural Information Processing Systems*, 30, 597–607.

Lin, T., Wang, Y., Liu, X. & Qiu, X. (2022). A survey of transformers. *AI Open*, 1-2.

Liu, F., Pradhan, R., Druhl, E., Freund, E., Liu, W., Sauer, B. C., Cunningham, F., Gordon, A. J., Peters, C. B. & Yu, H. (2019a). Learning to detect and understand drug discontinuation events from clinical narratives. *Journal of the American Medical Informatics Association*, 26(10), 943–951.

Liu, X., Chen, Y., Bae, J., Li, H., Johnston, J. & Sanger, T. (2019b). Predicting Heart Failure Readmission from Clinical Notes Using Deep Learning. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2642–2648.

López-García, G., Jerez, J. M., Ribelles, N., Alba, E. & Veredas, F. J. (2021). Transformers for clinical coding in Spanish. *IEEE Access*, 9, 72387–72397.

López-García, G., Jerez, J. M., Ribelles, N., Alba, E. & Veredas, F. J. (2023). Explainable clinical coding with in-domain adapted transformers. *Journal of Biomedical Informatics*, 1-2.

Loshchilov, I. & Hutter, F. (2019). Decoupled Weight Decay Regularization. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Lu, P., Wang, C., Hagenah, J., Ghiasi, S., Zhu, T., Thwaites, L., Clifton, D. A. et al. (2022). Improving Classification of Tetanus Severity for Patients in Low-Middle Income Countries Wearing ECG Sensors by Using a CNN-Transformer Network. *IEEE Trans. Biomed. Eng.*, 1-2.

Lu, Y., Cheung, Y.-M. & Tang, Y. Y. (2019). Bayes imbalance impact index: A measure of class imbalanced data set for classification problem. *IEEE transactions on neural networks and learning systems*, 31(9), 3525–3539.

Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5(1), 1–16.

Macabiau, C., Le, T.-D., Albert, K., Jouvet, P. & Noumeir, R. (2023). Label Propagation Techniques for Artifact Detection in Imbalanced Classes using Photoplethysmogram Signals. *arXiv preprint arXiv:2308.08480*.

Maimon, O. Z. & Rokach, L. (2014). *Data mining with decision trees: theory and applications*. World scientific.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de La Clergerie, É. V., Seddah, D. & Sagot, B. (2020). CamemBERT: a Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203–7219.

Martinez, A. M. & Kak, A. C. (2001). Pca versus lda. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(2), 228–233.

Matthay, M. A., Zemans, R. L., Zimmerman, G. A., Arabi, Y. M., Beitler, J. R., Mercat, A., Herridge, M., Randolph, A. G. & Calfee, C. S. (2019). Acute respiratory distress syndrome. *Nature reviews Disease primers*, 5(1), 18.

Matton, M.-P., Toledano, B., Litalien, C., Vallee, D., Brunet, F. & Jouvet, P. (2016). Databases and Computerized Systems in PICU: Electronic Medical Record in Pediatric Intensive Care: Implementation Process Assessment. *Journal of pediatric intensive care*, 5(3), 129.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), 115–133.

Meng, Y., Speier, W., Ong, M. K. & Arnold, C. W. (2021). Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE J. Biomed. Health Inform.*, 25(8), 3121–3129.

Mienye, I. D., Sun, Y. & Wang, Z. (2020). Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. *Informatics in Medicine Unlocked*, 18, 100307.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pp. 3111–3119.

Mondal, A. K., Bhattacharjee, A., Singla, P. & Prathosh, A. (2021). xViTCOS: explainable vision transformer based COVID-19 screening using radiography. *IEEE J. Transl. Eng. Health Med.*, 10, 1–10.

Mugisha, C. & Paik, I. (2022). Comparison of Neural Language Modeling Pipelines for Outcome Prediction From Unstructured Medical Text Notes. *IEEE Access*, 10, 16489–16498.

Musen, M. A., Middleton, B. & Greenes, R. A. (2021). Clinical decision-support systems. In *Biomedical informatics* (pp. 795–840). Springer.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B. & Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12), 124003.

Névéol, A., Dalianis, H., Velupillai, S., Savova, G. & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1), 1–13.

Ng, A. & Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14(2002), 841.

Noumeir, R. (2003). DICOM structured report document type definition. *IEEE Transactions on information technology in biomedicine*, 7(4), 318–328.

Olsen, C. R., Mentz, R. J., Anstrom, K. J., Page, D. & Patel, P. A. (2020). Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. *American Heart Journal*, 1-2.

Olsen, C., Meyer, P. E. & Bontempi, G. (2008). On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *Journal on Bioinformatics and Systems Biology*, 2009, 1-2.

Otter, D. W., Medina, J. R. & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 1-2.

Paleyes, A., Urma, R.-G. & Lawrence, N. D. (2020). Challenges in deploying machine learning: a survey of case studies. *arXiv preprint arXiv:2011.09926*, 1-2.

Pasini, A. (2015). Artificial neural networks for small dataset analysis. *Journal of thoracic disease*, 7(5), 953.

Pasupa, K. & Sunhem, W. (2016). A comparison between shallow and deep architecture classifiers on small dataset. *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 1–6.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011a). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011b). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Perlich, C., Provost, F. & Simonoff, J. (2003). Tree induction vs. logistic regression: A learning-curve analysis. 1-2.

Pham, T., Tran, T., Phung, D. & Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of biomedical informatics*, 69, 218–229.

Phan, H., Mikkelsen, K., Chén, O. Y., Koch, P., Mertins, A. & De Vos, M. (2022). Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Trans. Biomed. Eng.*, 69(8), 2456–2467.

Pluim, J. P., Maintz, J. A. & Viergever, M. A. (2003). Mutual-information-based registration of medical images: a survey. *IEEE Trans. Med. Imaging*, 22(8), 1-2.

Popel, M. & Bojar, O. (2018). Training tips for the transformer model. *arXiv preprint arXiv:1804.00247*, 1-2.

Quiroz, J. C., Laranjo, L., Kocaballi, A. B., Berkovsky, S., Rezazadegan, D. & Coiera, E. (2019). Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digital Medicine*, 1-2.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1), 5485–5551.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C. & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 12116–12128.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M. et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1), 18.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G. & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access*, 8, 54776–54788.

Rizwan, M., Mushtaq, M. F., Akram, U., Mehmood, A., Ashraf, I. & Sahelices, B. (2022). Depression Classification From Tweets Using Small Deep Transfer Learning Language Models. *IEEE Access*, 10, 129176–129189.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*, 1-2.

Roitero, K., Portelli, B., Popescu, M. H. & Della Mea, V. (2021). DiLBERT: Cheap embeddings for disease related medical NLP. *IEEE Access*, 9, 159714–159723.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V., McCoy, T. & Perlis, R. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry*, 6(10), e921–e921.

Sahlgren, M. & Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 975–980.

Salton, G. & Yang, C.-S. (1973). *On the specification of term values in automatic indexing*.

Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 1-2.

Santos, F., Macedo, H., Bispo, T. & Zanchettin, C. (2020). Morphological Skip-Gram: Using morphological knowledge to improve word representation. *arXiv preprint arXiv:2007.10055*, 1-2.

Sauthier, M., Tuli, G., Jouvet, P. A., Brownstein, J. S. & Randolph, A. G. (2021). Estimated Pao2: A Continuous and Noninvasive Method to Estimate Pao2 and Oxygenation Index. *Critical care explorations*, 3(10), 1-2.

Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., Osmani, V. et al. (2019). Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Medical Informatics*, 7(2), e12239.

Shi, X., Hu, Y., Zhang, Y., Li, W., Hao, Y., Alelaiwi, A., Rahman, S. M. M. & Hossain, M. S. (2016). Multiple disease risk assessment with uniform model based on medical clinical notes. *IEEE Access*, 4, 7074–7083.

Shi, Y., Lei, M., Ma, R. & Niu, L. (2019). Learning robust auto-encoders with regularizer for linearity and sparsity. *IEEE Access*, 7, 17195–17206.

Shwartz-Ziv, R. & Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 1-2.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S. et al. (2023). Large language models encode clinical knowledge. *Nature*, 1–9.

Soguero-Ruiz, C., Hindberg, K., Rojo-Alvarez, J. L., Skrøvseth, S. O., Godtliebsen, F., Mortensen, K., Revhaug, A., Lindsetmo, R.-O., Augestad, K. M. & Jenssen, R. (2014). Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records. *IEEE journal of biomedical and health informatics*, 20(5), 1404–1415.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929–1958.

Steinmeyer, C. & Wiese, L. (2020). Sampling methods and feature selection for mortality prediction with neural networks. *Journal of Biomedical Informatics*, 111, 103580.

Sundararajan, M., Taly, A. & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, pp. 3319–3328.

Suresh, H., Hunt, N., Johnson, A., Celi, L. A., Szolovits, P. & Ghassemi, M. (2017). Clinical intervention prediction and understanding with deep neural networks. *Machine Learning for Healthcare Conference*, pp. 322–337.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N. & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1), 1–10.

Tapia, N. I. & Estévez, P. A. (2020). On the information plane of autoencoders. *IEEE International Joint Conference on Neural Networks*.

Tharwat, A., Gaber, T., Ibrahim, A. & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI communications*, 30(2), 169–190.

Tishby, N., Pereira, F. C. & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*, 1-2.

Tolles, J. & Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama*, 316(5), 533–534.

Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. *Machine Learning for Healthcare Conference*, pp. 359–380.

Tripathy, J. K., Sethuraman, S. C., Cruz, M. V., Namburu, A., Mangalraj, P., Vijayakumar, V. et al. (2021). Comprehensive analysis of embeddings and pre-training in NLP. *Computer Science Review*, 42, 100433.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010.

Viola, P. & Wells III, W. M. (1997). Alignment by maximization of mutual information. *International journal of computer vision*, 24, 137–154.

Wallace, E., Wang, Y., Li, S., Singh, S. & Gardner, M. (2019). Do NLP Models Know Numbers? Probing Numeracy in Embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5310–5318.

Wang, Y., Yao, H. & Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing*, 184, 232–242.

Wang, Y., Zhou, Z., Jin, S., Liu, D. & Lu, M. (2017). Comparisons and selections of features and classifiers for short text classification. *Iop conference series: Materials science and engineering*, 261(1), 012018.

Ware, L. B. & Matthay, M. A. (2000). The acute respiratory distress syndrome. *New England Journal of Medicine*, 342(18), 1334–1349.

Webb, T., Holyoak, K. J. & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 1–16.

Weng, W.-H., Wagholikar, K. B., McCray, A. T., Szolovits, P. & Chueh, H. C. (2017). Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC medical informatics and decision making*, 17(1), 1–13.

Xiong, Y. & Lu, Y. (2020). Deep feature extraction from the vocal vectors using sparse autoencoders for Parkinson's classification. *IEEE Access*, 8, 27821–27830.

Xue, F., Shi, Z., Wei, F., Lou, Y., Liu, Y. & You, Y. (2022). Go wider instead of deeper. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8), 8779–8787.

Xue, J.-H. & Titterington, D. M. (2008). Comment on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Neural processing letters*, 28(3), 169–187.

Yahyatabar, M., Jouvet, P. & Cheriet, F. (2020). Dense-Unet: a light model for lung fields segmentation in Chest X-Ray images. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1242–1245.

Yahyatabar, M., Jouvet, P., Fily, D., Rambaud, J., Levy, M., Khemani, R. G. & Cheriet, F. (2023). A web-based platform for the automatic stratification of ARDS severity. *Diagnostics*, 13(5), 933.

Yang, X., Bian, J., Hogan, W. R. & Wu, Y. (2020). Clinical concept extraction using transformers. *J Am Med Inform Assoc*, 27(12), 1935–1942.

Yang, Z., Huang, Y., Jiang, Y., Sun, Y., Zhang, Y.-J. & Luo, P. (2018). Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Scientific reports*, 8(1), 6329.

Young, T., Hazarika, D., Poria, S. & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3), 55–75.

Yu, S. & Principe, J. C. (2019). Understanding autoencoders with information theoretic concepts. *Neural Networks*, 117, 104–123.

Yu, T. & Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689*, 1-2.

Zafar, M. B., Donini, M., Slack, D., Archambeau, C., Das, S. & Kenthapadi, K. (2021). On the Lack of Robust Interpretability of Neural Text Classifiers. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3730–3740.

Zaglam, N., Jouvet, P., Flechelles, O., Emeriaud, G. & Cheriet, F. (2014). Computer-aided diagnosis system for the acute respiratory distress syndrome from chest radiographs. *Computers in biology and medicine*, 52, 41–48.

Zeng, X., Linwood, S. L. & Liu, C. (2022). Pretrained transformer framework on pediatric claims data for population-specific tasks. *Scientific Reports*, 12(1), 3651.

Zhang, R., Ma, S., Shanahan, L., Munroe, J., Horn, S. & Speedie, S. (2017). Automatic methods to extract New York heart association classification from clinical notes. *2017 ieee international conference on bioinformatics and biomedicine (bibm)*, pp. 1296–1299.

Zhang, Y., Jin, R. & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4), 43–52.

Zhou, B., Yang, G., Shi, Z. & Ma, S. (2022). Natural language processing for smart healthcare. *IEEE Rev Biomed Eng*, 1-2.

Zhou, C., Jia, Y. & Motani, M. (2018). Optimizing autoencoders for learning deep representations from health data. *IEEE J. Biomed. Health Inform.*, 1-2.