Identifying Patients with Pain in Emergency Departments using Conventional Machine Learning and Deep Learning

Thanh Vu[1], **Anthony Nguyen**[1], **Nathan J Brown**[2,3], **James Hughes**[2,4]
[1] Australian e-Health Research Centre, CSIRO, Brisbane, Australia
[2] Emergency & Trauma Centre, Royal Brisbane & Women's Hospital, Brisbane, Australia
[3] Faculty of Medicine, University of Queensland, Brisbane, Australia
[4] School of Nursing, Queensland University of Technology, Brisbane, Australia
{thanh.vu, anthony.nguyen}@csiro.au
{nathan.brown3, james.hughes}@health.qld.gov.au

Abstract

Pain is the main symptom that patients present with to the emergency department (ED). Pain management, however, is often poorly done aspect of emergency care and patients with painful conditions can endure long waits before their pain is assessed or treated. To improve pain management quality, identifying whether or not an ED patient presents with pain is an important task and allows for further investigation of the quality of care provided. In this paper, machine learning was utilised to handle the task of automatically detecting patients who present at EDs with pain from retrospective data. Experimental results on a manually annotated dataset show that our proposed machine learning models achieve high performances, in which the highest accuracy and macro-averaged F_1 are 91.00% and 90.96%, respectively.

1 Introduction

There are over 8 million presentations to Australian public hospital emergency departments (EDs) each year (AIHW, 2018). Pain is the most common symptom for patients seeking care in EDs (Karwowski-Soulié et al., 2006; Hatherley et al., 2016; Todd, 2017; Varndell et al., 2018). In particular, a study of 726 presentations (Karwowski-Soulié et al., 2006) showed that 78% of patients presented to EDs with pain. Despite the large number of ED patients with pain, pain is often poorly assessed and treated within the ED (Hatherley et al., 2016; Varndell et al., 2018). This leads to increases in waiting times for patients until pain is assessed and pharmacological analgesia is offered (Hatherley et al., 2016; Varndell et al., 2018).

In this paper, we propose the task of identifying patients presenting to EDs with pain, with the view of improving care quality and the management of pain. In particular, the identification of patients presenting to EDs with pain allows for the scoping of groups who may be receiving poor pain care (Pletcher et al., 2008; Hwang et al., 2014; Todd, 2017). The ease of identification of patients with pain also allows for the evaluation of targeted interventions to improve care using large datasets.

However, manually handling the identification task at a large scale, such as tens, hundreds of thousands of ED patients is challenging as it requires expensive human effort to determine whether a patient presented to the ED with pain or not. To handle this problem, we propose to use machine learning including both conventional feature-based and deep learning models (Scholkopf and Smola, 2001; Liaw et al., 2002; Elman, 1990; LeCun et al., 1998; Chung et al., 2014) to automatically learn the hidden patterns to solve the task.

Machine learning research in healthcare has shown success in handling many other predictive tasks, such as cancer staging from pathology reports (McCowan et al., 2007), disease or diagnosis coding from health records (Koopman et al., 2015; Mullenbach et al., 2018), predicting in-hospital mortality, unplanned readmissions (Rajkomar et al., 2018), atrial fibrillation risk (Nguyen et al., 2019), and opioid overdose risk (Che et al., 2017; Lo-Ciganic et al., 2019). To this end, we construct a dataset of ED patients from an Australian hospital in order to evaluate our proposed machine learning models, in which each patient is assigned with either a "Pain" or "No-Pain" label if the patient is with or without the presence of pain, respectively. Our main contributions are as follows:

- We formally introduce the task of identifying whether or not a patient presented at an ED with pain.

- We propose conventional machine learning as well as deep learning models to handle the task.
- We perform extensive experiments on the task-specific annotated dataset to show the effectiveness of the models.

The remainder of the paper is structured as follows. In Section 2, we present related work on improving the pain management and care quality, as well as the application of machine learning in healthcare predictive tasks. Section 3 describes the pain identification task as well as how we constructed the annotated dataset. In Section 4, we first describe our machine learning models. We then report the experimental results of the models on the annotated dataset. Section 5 concludes the paper.

2 Related Work

An overview of research to help improve the pain management and care quality at emergency departments, as well as the application of machine learning in the healthcare domain, will be presented.

2.1 Pain-related Studies

Pain is the most common symptom of patients presenting at EDs (Cordell et al., 2002; Karwowski-Soulié et al., 2006). In particular, Cordell et al. (2002) and Karwowski-Soulié et al. (2006) revealed that pain accounts for up to 70% and 78% of ED visits, respectively. Much research attention has focused on the need for improving the pain management and care quality (Doherty et al., 2013; Georgiou et al., 2015; Hatherley et al., 2016; Todd, 2017; Varndell et al., 2018). Historically, the detection, assessment and management of pain are often neglected (Georgiou et al., 2015; Varndell et al., 2018). This results in patients being forced to wait extra time before getting assessed and/or treated (Doherty et al., 2013), which leads to negative outcomes for the patient and the healthcare system.

The use of opioids is a popular approach for pain treatment (Todd, 2017). However, recent studies of prescription opioid misuse and abuse revealed that in Australia, it has been increasing to levels of harm (Häuser et al., 2017). This leads to a more urgent need of improving the pain management and care quality in Australia, which motivates this study.

2.2 Machine Learning in Healthcare

With the availability of Electronic Medical Record (EMR) systems, electronic health records (EHRs) of patients, collected during the patient clinical encounters, have been increasingly available. This creates a great opportunity for using machine learning to improve care management and quality (McCowan et al., 2007; Koopman et al., 2015; Rajkomar et al., 2018; Mullenbach et al., 2018; Che et al., 2017). In particular, McCowan et al. (2007) utilised Support Vector Machine (SVM) (Scholkopf and Smola, 2001) to automatically infer and classify cancer stages from patient pathology reports. (Koopman et al., 2015) applied SVM to classify cancer-related International Classification of Diseases (ICD) codes from free-text death certificates.

Especially, with the huge amount of EHR data, deep learning has shown to obtain state-of-the-art (SOTA) results in many predictive tasks (Che et al., 2017; Rajkomar et al., 2018; Mullenbach et al., 2018). In particular, deep learning models based on recurrent neural network (RNN) (Elman, 1990) and convolutional neural network (CNN) (LeCun et al., 1998) have achieved SOTA performances in a number of tasks, such as large-scale ICD coding from hospital free-text discharge summaries (Mullenbach et al., 2018) and prediction of opioid overdose risk (Che et al., 2017). Che et al. (2017) presented an RNN-based model for classifying categories of opioid users and achieved robust results on a large-scale dataset of over a hundred thousand opioid users. Mullenbach et al. (2018) proposed a CNN-based model to tackle the ICD coding task and showed that the model yielded SOTA performances on the MIMIC III dataset (Johnson et al., 2016). Rajkomar et al. (2018) demonstrated that deep learning models achieved high accuracy for predictive tasks, such as in-hospital mortality, unplanned readmission and prolonged length of stay. In this study, we take the advantages of both conventional machine learning and deep learning models to handle a new task of pain-related identification which is described in Section 3.

3 Task Description and Dataset

In this section, the formulation of the task and the dataset used for the task evaluation is presented.

3.1 Task Description

Given a patient who presents at an ED, the aim is to identify whether or not that patient is with the presence of pain on admission. This task can be formulated as a two-class (binary) classification problem, in which the ED data of the patient was used to predict the pain class (i.e., either *"Pain"* to represent patients *with* pain or *"No-Pain"* to represent patients *without* pain).

Unstructured free-text ED data fields, namely "presenting problem" and "nurse assessment", were used for the classification task. These were the two free text fields that ED nurses fill out on a patient's arrival to the ED. Table 1 illustrates examples of patients with and without pain. Note that short-hand notations, abbreviations and typographical errors are common in the patient ED data, which presents additional challenges to the task.

Table 1: Examples of the ED data associated with patients with/without pain. The class of each patient was manually annotated.

Patient ED Data	Class
Presenting problem: 4/7 cough, tight chest, myalgia// recent dx t2dm; *Nurse assessment*: a=patent b= spontaneous, rr 19, reports it hurts to breath and having difficulty breathing c= strong reg radial pulse, tachycardic 120 very dry mucous membranes not maintaining oral intake d= gcs 15;	Pain
Presenting problem: mdma yesterday anxious insomnia subjective tongue swelling tachycardic; *Nurse assessment*: hr 120bpm dry tongue doe not appear swollen;	No-Pain

It is worth noting that the focus was on identifying patients with pain at admission. Other potentially useful ED information, such as ICD-10 diagnosis codes, would not be available until the patient is discharged.

3.2 Dataset

A dataset of patients presenting at EDs was constructed by randomly extracting 2,000 ED adult patients from an Australian hospital with the arrival date from August to October 2018[1]. The dataset was annotated by an experienced medical student under the supervision of a senior medical nurse, in which a patient was assigned a "Pain" label if the patient presented with pain, and "No-Pain" otherwise. In particular, the annotator was provided with a list of pain related keywords (Hughes et al., 2019) to look for when reviewing the triage nursing assessment. In addition to these keywords, the annotator also reviewed the documentation for a pain score, such as "xx/10", "severe pain". When the annotator believed that the triage nursing assessment indicated pain but was outside of the definition, discussion was held between the student and the senior medical nurse about whether this indicated the patient arrived in pain. After annotating the data, the annotated dataset contained a total of 938 and 1,062 instances of "Pain" and "No-Pain" labels, respectively.

[1]Research ethics was obtained from the Metro North Hospital and Health Service Human Research Ethics Committee.

Table 2: Basic dataset statistics

Dataset	#Patients	#Pain	#No-Pain
Training	1,200	574	626
Development	400	171	229
Test	400	193	207

The annotated dataset was split into training, development and test sets containing 60%, 20%, 20% instances of the annotated dataset, respectively. Table 2 shows the dataset statistics.

4 Methods

This section details the machine learning models used for the pain classification task. Specifically, Support Vector Machine (SVM) and Random Forest (RF) were used as our conventional feature-based models, and Recurrent Neural Network (RNN) and Convolutional Neural Networks (CNN) were used as our deep learning models.

4.1 Conventional feature-based models

SVM (Scholkopf and Smola, 2001) and RF (Liaw et al., 2002) were used as our conventional feature-based models to handle the classification task. Figure 1 shows the general architecture of the conventional models. Here, both the models used the same set of lexical and semantic features according to (Yang et al., 2016; Vu et al., 2018) as follows:

Lexical features: Lexical features included n-grams at both word and character levels (i.e. sequences of n words or characters). n-grams at the character level were used to handle out-of-vocabulary (OOV) words. For each type of n-gram, only the top k most frequent n-grams from the training set were kept. The value of

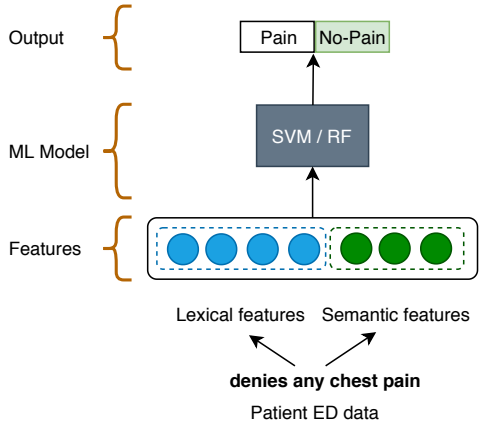

Figure 1: Conventional machine learning models. The model output is "No-Pain" for the input of "denies any chest pain".

each n-gram feature was calculated using the term frequency-inverse document frequency (tf-idf) weighting scheme.

Semantic features: Two approaches were applied to semantically represent the patient. Firstly, the average of pre-trained embeddings of words in the patient ED data was used as the representation of that patient. Secondly, latent semantic indexing (LSI) (Papadimitriou et al., 2000) was used to capture the underlying semantics of the dataset.

Implementation details: For the experimental dataset, tokens that contained no alphabetic or numeric characters were removed (for example, removing "//" but keeping "3/7"). All the remaining tokens were lowercased. The "presenting problem" and "nurse assessment" fields of a patient were concatenated to form the single text for the patient. Regarding lexical features, we set k, the top most frequent n-grams to 2,000 for both word and character levels. For the semantic features, fastText (Bojanowski et al., 2016) was applied to train a subword embedding model on a large-scale dataset of 8 million hospital clinical notes. We found that the best experimental results on the development set were achieved with the pre-trained embedding size of 200. Moreover, we set the LSI output size to 100.

The Scikit-learn implementations for both the SVM and RF models (Pedregosa et al., 2011) were used. For each model, a grid-search on hyper-parameters was performed to find the best performing model on the development set. Specifically, for each hyper-parameter setting, we trained the machine learning model using the training set and then evaluate the trained model on the development set. The best-trained model was selected using the macro-averaged F_1 scores of "Pain" and "No-Pain" labels on the development set. After that, the best model was used for the evaluation on the test set.

For SVM, the "linear" kernel performed better than the "polynomial", "radial basis function (rbf)" and "sigmoid" kernels. The grid search was performed over *loss_function* $\in$ {"squared_hinge", "hinge"}; $C \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0, 3.0, 5.0\}$; *max_iteration* $\in \{100, 200, 500, 1000, 2000\}$. We also set the *penalty_norm* parameter to l_2. The highest average-macro F_1 score was archived with *loss_function* = "hinge", $C = 0.3$, and *max_iteration* = 1000.

For RF, the grid search was performed over *max_depth* $\in [1 - 10]$ and *number_of_trees* $\in \{10, 50, 100, 200, 300, 500, 1000\}$. A *max_depth* = 6 and *number_of_trees* = 500 produced the highest F_1 score on the development set.

4.2 Deep learning models

RNN (Elman, 1990) and CNN (LeCun et al., 1998) were used as deep learning models as they have proved to work well in the tasks of modelling sequence data (Mikolov et al., 2010) and text classification (Kim, 2014; Yang et al., 2016). Figure 2 shows the architecture of the deep learning models. For both models, the base/embedding layer was obtained by concatenating the pre-trained word embeddings (from a large-scale hospital clinical note dataset) with the character-level embeddings of that word (Kim et al., 2016). The pre-trained word embeddings were fixed, while character-level word embeddings were simultaneously trained with the other model parameters. The character-level word embeddings help handle the OOV words. The representations of words in the ED data of a patient were concatenated to form a sequence of word representation vectors.

RNN model: The sequence of word vectors were fed into an RNN encoder to learn the representation of the patient. As RNN has struggled with long-term dependencies, we applied a prominent variant of RNN, Gated recurrent unit (GRU) (Chung et al., 2014), which can handle the problem.[2] The hidden state vector of the last word in

[2]GRU performed better than long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) as well as their bi-directional counterparts for this task in our experiments.

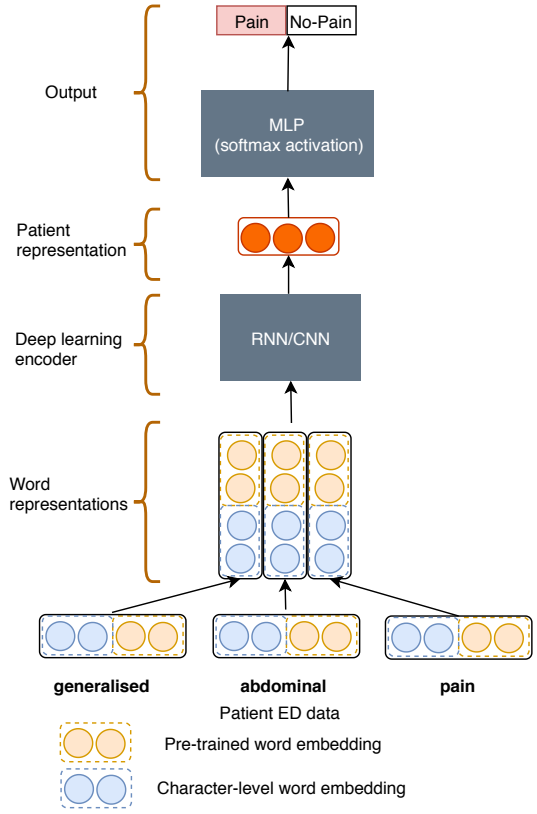

Figure 2: Deep learning models. The model output is "Pain" for the input of "generalised abdominal pain".

the patient data produced by GRU was used as the patient representation. Finally, the representation vector was fed into a one-hidden layer multilayer perceptron (MLP) with softmax output for classification which returns the class probabilities for the patient.

CNN model: The sequence of word vectors was fed into a CNN encoder which performed convolution operations and max-pooling to produce the representation vector of the patient. Similar to the RNN model, the representation vector was fed into a one-hidden layer MLP with softmax output to produce the class prediction.

Implementation details: The same pre-trained embeddings model and data preprocessing detailed in Section 4.1 was used. The deep learning models were implemented using Pytorch (Paszke et al., 2017). We trained each model using the Adam optimiser (Kingma and Ba, 2014) with the default learning rate of 0.001 and a fixed random seed. The batch size and the number of training epochs were respectively set to 32 and 50. The CNN-based model proposed by (Kim et al., 2016) was used to learn the character-level embeddings for each word, in which the window size and the number of filters were set to 2 and 50, respectively. For both models, we applied a dropout mechanism to both the word representations (before the encoder) and the patient representation (before MLP) with the probability of 0.5.

For RNN, a grid search was performed over *hidden_size* $\in \{100, 200, 300, 400, 500\}$ and the number of layers, *n_layers* $\in \{1, 2, 3, 4, 5\}$. The best performance was achieved with *hidden_size* $= 400$ and *n_layers* $= 5$. For CNN, we performed a grid search of the number of filters, *n_filters* $\in \{100, 200, 300, 400, 500\}$ and the window sizes, *kernel_sizes* $\in \{2, 3, 4, 5, (2, 3), (4, 5), (3, 4, 5)\}$. The setting of *n_filters* $= 400$ and *kernel_sizes* $= 3$ produced best performances.

4.3 Evaluation

Baseline: The proposed machine learning models were compared with a rule-based baseline, *RULE*. Specifically, the baseline used inclusion and exclusion criteria predefined by a senior ED nurse as detailed in a recent publication (Hughes et al., 2019), for example, terms containing "pain', "discomfort", "stab", "burn", "ache". In the baseline, Context/Negex (Chapman et al., 2011) was also used to handle negation in the patient ED data, for example, "denies neck pain".

Metrics: The metrics used to evaluate the models included *accuracy*, *precision*, *recall* and F_1 which are standard evaluation metrics for classification tasks. For all metrics, higher values represent better performances.

4.4 Results

Overall performance: Table 3 shows the main experimental results of the models on the test set. The rule-based model, RULE, achieves an accuracy of 84.75% indicating that the model performed well for this task. The drawback of this model is that it required expert knowledge for the construction of the inclusion and exclusion criteria. The results also show that all the proposed machine learning models achieved higher performances than RULE. Specifically, the machine learning models achieve absolute improvements of at least 3.3% over the RULE baseline with respect to the macro-averaged F_1 scores. This indicates that the proposed machine learning models can handle the task well even without expert knowledge.

Noteworthy, without feature engineering, the deep learning models (i.e., CNN and RNN) per-

Table 3: Experimental results (%) on the test set. $*$ indicates that the performance difference between the machine learning model and the RULE baseline is significant at the significance level α of 0.1 using the Approximate Randomisation test (Chinchor, 1992; Dror et al., 2018), with N= 5,000.

Model	Accuracy	Macro-averaged			Pain			No-Pain		
		Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
RULE	84.75	84.87	84.63	84.69	86.26	81.35	83.73	83.49	87.92	85.65
SVM	88.00	88.13*	87.90*	87.96*	89.62	84.97	87.23*	86.64	90.82	88.68*
RF	88.00	88.31*	87.85*	87.93*	90.96*	83.42	87.03*	85.65	92.27*	88.84*
RNN	**91.00***	**91.21***	**90.88***	**90.96***	**93.37***	87.56*	**90.37***	89.04*	**94.20***	**91.55***
CNN	88.25	88.25*	88.30*	88.25*	86.50	**89.64***	88.04*	**90.00***	86.96	88.45

formed competitively or better than conventional models (i.e., SVM and RF). RNN produced the highest performance with an accuracy and macro-averaged F_1 score of 91.00% and 90.96%, which were 6.25% and 6.27% absolute higher than RULE, respectively. With regards to the "Pain" label, RNN achieved the highest precision of 93.37% and F_1 score of 90.37%. Meanwhile, CNN achieved the highest recall of 89.64%. In terms of the "No-Pain" label, CNN obtained the highest precision of 90.00%, while RNN achieved the highest recall of 94.20% and F_1 of 91.55%.

Table 3 also presents the statistical significance between the difference in performances between the machine learning models and the RULE baseline, using the Approximate Randomisation test (Chinchor, 1992; Dror et al., 2018), with N = 5,000 and the significance level α of 0.1. The Approximate Randomisation test is a popular non-parametric statistical significance test for NLP tasks (Chinchor, 1992; Dror et al., 2018). We found that the difference in performances between RNN and RULE were statistically significant across all classes as well as metrics. The deep learning models were also consistently better in terms of absolute values when compared to conventional models.

Ablation study: In the previous sections, deep learning models were shown to work better than conventional machine learning models. In this section, we evaluate the effectiveness of the pre-trained and character-level word embeddings on the performance of the deep learning models. Specifically, we perform an ablation study on the development set, in which we evaluate the best performing model, RNN, with different ablation settings as follows: w/o *char embeds*, w/o *word embeds* and w/o *both* denoting that the model was trained without character-level word embeddings, without pre-trained word embeddings[3] and neither the embeddings, respectively. We also presented the results of RNN when using only the "presenting problem", namely only *PP* and using only the nurse assessment, namely only *NA* ED data fields.[4]

As can be seen from Table 4, without character-level word embeddings slightly decreases the RNN performance. Without pre-trained word embedding significantly degrades the RNN performances by about 4%. The largest decline in the performance of about 4.8% was observed when both embedding types were not used. We further see that without using presenting problem data (i.e. "only NA") results in a significant decrease of more than 23% in the performance. This is perhaps caused by the fact that there were 276 out of 2,000 ($\sim$ 14%) patients who did not have any nursing assessment data. Another reason may be due to the presenting problem field being more informative in terms of containing more pain-related information than in the nurse assessment field. Further investigation revealed that 38% of the presenting problems in the development set contained pain-related keywords (detailed in the baseline, RULE) compared to 25.50% of the non-empty nurse assessment data. We also found that without using nurse assessment (i.e. "only PP") degrades the performance by about 3.8%. This indicates that the concatenation of the two free-text fields was important to the task.

4.5 Application

The immediate application of the research is to provide machine learning assistance to process and analyse very large datasets for the purposes of research or clinical audit. Apart from that, it can also be a potential real-time clinical application,

[3] In this case, the word embeddings were initialised randomly and then fine-tuned with the training of other model parameters.

[4] For the ED data field ablation study, we used the RNN with both pre-trained word and character-level embeddings.

Table 4: Ablation study performance (%) on the development set.

Model	Accuracy	F_1
RNN	90.00	89.80
w/o *char embeds*	$89.75_{-0.25}$	$89.61_{-0.19}$
w/o *word embeds*	$86.00_{-4.00}$	$85.50_{-4.30}$
w/o *both*	$85.25_{-4.75}$	$84.99_{-4.81}$
only *PP*	$86.25_{-3.75}$	$85.88_{-3.92}$
only *NA*	$67.42_{-22.58}$	$66.32_{-23.48}$

such as a smart support assistant to help improve the quality of triage related to presentations that involve or are likely to involve pain. Specifically, in the scenario of a patient who presents to triage, the triage nurse asks about the problem/symptoms and records in electronic notes. If pain or a condition likely to be associated with pain is recorded then the triage nurse should also ask the patient about the level of pain and record a pain score. The smart support assistant will monitor the electronic notes in real-time and if a pain score is not recorded in the notes when it should be, then it will provide a suggestion of adding the information to the triage nurse. Pain is a common symptom that the ED sees everyday but still does not do a good job at assessing. These applications are able to be expanded to different hospital departments and units, such as Intensive Care Units where assessing pain may also be challenging (Suominen et al., 2009).

4.6 Limitations and Future Work

As in the previous section, the best accuracy on the development set was 90%, achieved using RNN. This meant that there were 10% $\sim$ 40 instances, namely the "error" set, where RNN produced incorrect labels. The "error" set was reviewed by a senior ED nurse to determine the underlying reasons for the system discrepancies.

On review, we found cases where the ED nurse had difficultly in classifying pain. In these more difficult cases, half of the "error" cases could have been classified differently.[5] This shows that even with a medical background, there exist the more difficult cases where there may be uncertainty in the labels. In future, we plan to handle the uncertainty problem by involving multiple annotators and an adjudicator.

[5]This indicates that our proposed machine learning approaches could have achieved higher performances if the dataset labels relating to "error" cases were corrected.

Another limitation of the proposed deep learning model was that although it produced the highest performance, it was still difficult to understand and locate the evidence it used for prediction, which is an important aspect of text analytics in the healthcare domain. In future work, we aim to integrate neural attention mechanisms to our deep learning models to make it interpretable (Bahdanau et al., 2014; Luong et al., 2015; Vaswani et al., 2017).

5 Conclusions

In this paper, we presented the task of identifying patients who presented to EDs with pain. Both conventional feature-based machine learning and deep learning models were proposed to handle the task. Experimental results on a 2,000 ED patient annotated dataset showed that our machine learning models performed well on this task with the highest accuracy and macro-averaged F_1 score of 91.00% and 90.96%, respectively.

It was shown that the machine learning models achieved higher results than a rule-based baseline. Moreover, deep learning models performed competitively or better than conventional models. The ablation study indicated that pre-trained word embeddings and character-level word embeddings played an important role leading to the success of the deep learning models. These learnings are beneficial for similar research on other clinical tasks but also sets a solid foundation for further improving performances on the "pain" models as well as improve the clinical utility of the model through explainability, with the aim to scale the "pain" study to other hospitals and regions.

References

AIHW. 2018. Emergency department care 2017-2018: Australian hospital statistics. *Australian Institute of Health and Welfare*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Brian E Chapman, Sean Lee, Hyunseok Peter Kang, and Wendy W Chapman. 2011. Document-level classification of ct pulmonary angiography reports

based on an extension of the context algorithm. *Journal of biomedical informatics*, 44(5):728–737.

Zhengping Che, Jennifer St Sauver, Hongfang Liu, and Yan Liu. 2017. Deep learning solutions for classifying patients on opioid use. In *AMIA Annual Symposium Proceedings*, volume 2017, page 525. American Medical Informatics Association.

Nancy Chinchor. 1992. The statistical significance of the muc-4 results. In *Proceedings of the 4th conference on Message understanding*, pages 30–50. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

William H Cordell, Kelly K Keene, Beverly K Giles, James B Jones, James H Jones, and Edward J Brizendine. 2002. The high prevalence of pain in emergency medical care. *The American journal of emergency medicine*, 20(3):165–169.

Steven Doherty, Jonathan Knott, Scott Bennetts, Mitra Jazayeri, and Sue Huckson. 2013. National project seeking to improve pain management in the emergency department setting: Findings from the nhmrcnics n ational p ain m anagement i nitiative. *Emergency Medicine Australasia*, 25(2):120–126.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhikers guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Evanthia Georgiou, Maria Hadjibalassi, Ekaterini Lambrinou, Panayiota Andreou, and Elizabeth DE Papathanassoglou. 2015. The impact of pain assessment on critically ill patients outcomes: a systematic review. *BioMed research international*, 2015.

Claire Hatherley, Natasha Jennings, and Rachel Cross. 2016. Time to analgesia and pain score documentation best practice standards for the emergency department–a literature review. *Australasian Emergency Nursing Journal*, 19(1):26–36.

Winfried Häuser, Stephan Schug, and Andrea D Furlan. 2017. The opioid epidemic and national guidelines for opioid therapy for chronic noncancer pain: a perspective from different continents. *Pain reports*, 2(3).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

James A. Hughes, Nathan J. Brown, Jacqui Chiu, Brandon Allwood, and Kevin Chu. 2019. The relationship between time to analgesic administration and emergency department length of stay: A retrospective review. *Journal of Advanced Nursing*, 0(0):1–8.

Ula Hwang, Laura K Belland, Daniel A Handel, Kabir Yadav, Kennon Heard, Laura Rivera-Reyes, Amanda Eisenberg, Matthew J Noble, Sudha Mekala, Morgan Valley, et al. 2014. Is all pain is treated equally? a multicenter evaluation of acute pain care by age. *Pain®*, 155(12):2568–2574.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Fabienne Karwowski-Soulié, Stéphanie Lessenot-Tcherny, Agathe Lamarche-Vadel, Sébastien Bineau, Christine Ginsburg, Olivier Meyniard, Brigitte Mendoza, Pascale Fodella, Gwenaelle Vidal-Trecan, and Fabrice Brunet. 2006. Pain in an emergency department: an audit. *European Journal of Emergency Medicine*, 13(4):218–224.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Bevan Koopman, Guido Zuccon, Anthony Nguyen, Anton Bergheim, and Narelle Grayson. 2015. Automatic icd-10 classification of cancers from free-text death certificates. *International journal of medical informatics*, 84(11):956–965.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

Wei-Hsuan Lo-Ciganic, James L Huang, Hao H Zhang, Jeremy C Weiss, Yonghui Wu, C Kent Kwoh, Julie M Donohue, Gerald Cochran, Adam J Gordon, Daniel C Malone, et al. 2019. Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. *JAMA network open*, 2(3):e190968–e190968.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Iain A McCowan, Darren C Moore, Anthony N Nguyen, Rayleen V Bowman, Belinda E Clarke, Edwina E Duhig, and Mary-Jane Fry. 2007. Collection of cancer stage data by classifying free-text medical reports. *Journal of the American Medical Informatics Association*, 14(6):736–745.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Dat Quoc Nguyen, Karin Verspoor, and Blanca Gallego Luxan. 2019. Risk prediction using electronic health records of patients with atrial fibrillation. In *Proceedings of the Advances in Data Science conference abstracts*.

Christos H Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. 2000. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mark J Pletcher, Stefan G Kertesz, Michael A Kohn, and Ralph Gonzales. 2008. Trends in opioid prescribing by race/ethnicity for patients seeking care in us emergency departments. *Jama*, 299(1):70–78.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18.

Bernhard Scholkopf and Alexander J Smola. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Hanna Suominen, Heljä Lundgrén-Laine, Sanna Salanterä, and Tapio Salakoski. 2009. Evaluating pain in intensive care. In *Nursing Informatics*, pages 192–196.

Knox H Todd. 2017. A review of current and emerging approaches to pain management in the emergency department. *Pain and therapy*, 6(2):193–202.

Wayne Varndell, Margaret Fry, and Doug Elliott. 2018. Quality and impact of nurse-initiated analgesia in the emergency department: A systematic review. *International emergency nursing*, 40:46–53.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thanh Vu, Dat Quoc Nguyen, Xuan-Son Vu, Dai Quoc Nguyen, Michael Catt, and Michael Trenell. 2018. Nihrio at semeval-2018 task 3: A simple and accurate neural network model for irony detection in twitter. *arXiv preprint arXiv:1804.00520*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.