Diagnostic Prediction Using Discomfort Drawings with IBTM

Cheng Zhang, Hedvig Kjellström {CHENGZ, HEDVIG}@KTH.SE
Robotics, Perception and Learning (RPL), KTH Royal Institute of Technology
RPL/CAS, CSC, KTH, 100 44 Stockholm, Sweden

Carl Henrik Ek CARLHENRIK.EK@BRISTOL.AC.UK
University of Bristol, UK

Bo C. Bertilson BO.BERTILSON@KI.SE
Department of Neurobiology, Care Sciences and Society, Karolinska Institutet
Alfred Nobels all 12, 14183 Huddinge, Stockholm, Sweden

Abstract

In this paper, we explore the possibility to apply machine learning to make diagnostic predictions using discomfort drawings. A discomfort drawing is an intuitive way for patients to express discomfort and pain related symptoms. These drawings have proven to be an effective method to collect patient data and make diagnostic decisions in real-life practice. A dataset from real-world patient cases is collected for which medical experts provide diagnostic labels. Next, we use a factorized multimodal topic model, Inter-Battery Topic Model (IBTM), to train a system that can make diagnostic predictions given an unseen discomfort drawing. The number of output diagnostic labels is determined by using mean-shift clustering on the discomfort drawing. Experimental results show reasonable predictions of diagnostic labels given an unseen discomfort drawing. Additionally, we generate synthetic discomfort drawings with IBTM given a diagnostic label, which results in typical cases of symptoms. The positive result indicates a significant potential of machine learning to be used for parts of the pain diagnostic process and to be a decision support system for physicians and other health care personnel.

1. Introduction

A discomfort drawing is a drawing on the image of a body where a patient may shade all areas of discomfort in preparation for a medical appointment. The drawing has been shown to be able to make diagnostic predictions - especially to discern neuropathic from nociceptive and psychiatric diseases [Bertilson et al. (2007)]. The use of drawings (pain drawing) to collect data from patients was first reported by Palmer in 1949 [Palmer (1949)] and has been studied in clinical settings showing high diagnostic predictive value especially in spine related pain by [Ohnmeiss et al. (1999); Vucetic et al. (1995); Albeck (1996); Tanaka et al. (2006)]. The pain drawing, where different signs mark different kind of pain, is still in use at many clinics. As a more recent method, the discomfort drawing (a revised pain drawing) instructs the patient to shade all areas of discomfort. This method may have some possible benefits compared to pain drawings due to the fact that many different symptoms may arise from disfunction of the same body organ and /or nerve [Bertilson et al. (2003, 2007, 2010)]. Hence, we focus on the use of discomfort drawings.

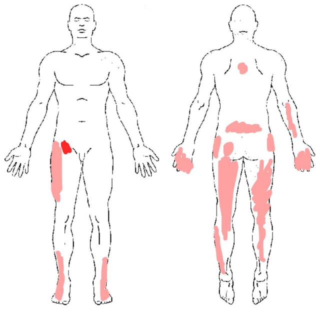
. This research has been supported by the Swedish Research Council (VR), Stiftelsen Promobilia and Vinnova.

To find high-quality diagnostic prediction methods is a goal of health care as well as the machine learning community. For example, machine learning for electrocardiogram (EKG) diagnostic prediction [Kukar et al. (1999)] has been used for years as a decision support system for health care personal. However, the most common and most costly medical problem is unspecific pain and discomfort [Upshur et al. (2010)] to which machine learning has not been applied yet. In this paper, we focus on applying machine learning for diagnosing pain-related problems using discomfort drawings.

Topic models [Blei et al. (2003)], a type of generative models, have been successfully applied in different domains, such as information retrieval and computer vision [Wang et al. (2009); Newman et al. (2006); Hospedales et al. (2011); Zhang et al. (2013)]. With efficient inference algorithms [Hoffman et al. (2010); Ranganath et al. (2013)], these models can handle both small and big datasets, in complete data and in incomplete scenarios. Additionally, they are highly interpretable and can be used to generate missing data. In our application of using discomfort drawings for diagnostic prediction, the data consist of multiple modalities (drawings and labels). Hence, a multi-modal topic model [Blei and Jordan (2003); Wang et al. (2009); Zhang et al. (2016)] is needed. Traditional multi-modal topic model [Blei and Jordan (2003); Zhang et al. (2013)] represent all the information contained in the data, hence these models are not robust to noise. A recent advancement in multi-modal topic models shows that Inter-Battery Topic Model (IBTM) [Zhang et al. (2016)] is robust to noise in the data by explaining away irrelevant parts of the information. Therefore, in this paper IBTM is adapted to predict diagnostic labels given a discomfort drawing. IBTM was originally proposed for representation learning and applyed for classification tasks. In this paper, we adapt the framework for diagnostic label prediction and use mean-shift clustering [Comaniciu and Meer (2002)] to determine the number of diagnostic predictions that the system needs to make.

The main contribution of this paper lies in the modification and use of IBTM for diagnostic prediction with discomfort drawings. This is a novel application of a principled framework. For this purpose, a dataset was collected from real-world clinical cases with medical expert labels. The experiments show that the adapted IBTM makes reasonable diagnostic predictions. Additionally, the model also contributes to the interpretability of the data for humans and may further provide insight into the diagnostic procedure. Our approach shows that the use of machine learning in the assessment of discomfort drawings is a promising direction.

2. Problem Statement



Symptom diagnoses: Interscapular discomfort; R arm discomfort; B hands discomfort; Lumbago; B crest of the ilium discomfort; L side thigh discomfort; B back thigh discomfort; B calf discomfort; B achilles tendinitis; B shin discomfort; R inguinal discomfort;

Pattern diagnoses B L5 Radiculopathy; B S1 Radiculopathy; B C7 Radiculopathy;

Pathophysiological diagnoses DLI L4-L5; DLI S1-S2; DLI C6-C7

Table 1: Discomfort drawings (left) and diagnoses by medical expert (right). R stands for right-side, L stands for left-side and B stands for bilateral. DLI refers to discoligament injury.

At some clinics which treat pain-related problems, a patient is asked to shade all areas of discomfort on a drawing of a body. The intensity of shade should indicate the level of discomfort. The patient is typically also asked to specify what type of discomfort they experience and furthermore to describe the discomfort-level over time. During a patient interview additional information regarding symptoms, prior treatment and experiences may be added to provide the health care personnel with sufficient information to make a diagnostic prediction that can guide the treatment.

In this paper we focus on diagnostic prediction solely based on areas of discomfort which is the key information. Table 1 shows an example of discomfort drawings and their diagnoses. On a standard body contour the discomfort regions are marked in red. The right column shows the diagnostic label provided by medical experts which are roughly ordered by symptom diagnoses, possible pattern diagnoses and possible pathophysiological diagnoses. The dataset was collected in a Swedish clinic based on real-world patient cases, hence the diagnostic labels are originally given in Swedish. These labels were translated into english by the authors to ease the readability of the paper. The later part of the labels focuses on the underlying pathophysiology of the discomfort.

Our task is to build a system that makes high quality diagnostic predictions given a discomfort drawing. This could be extended into a decision support system, which could increase the effectiveness and precision of the care for a large group of less favored patients [Upshur et al. (2010)].

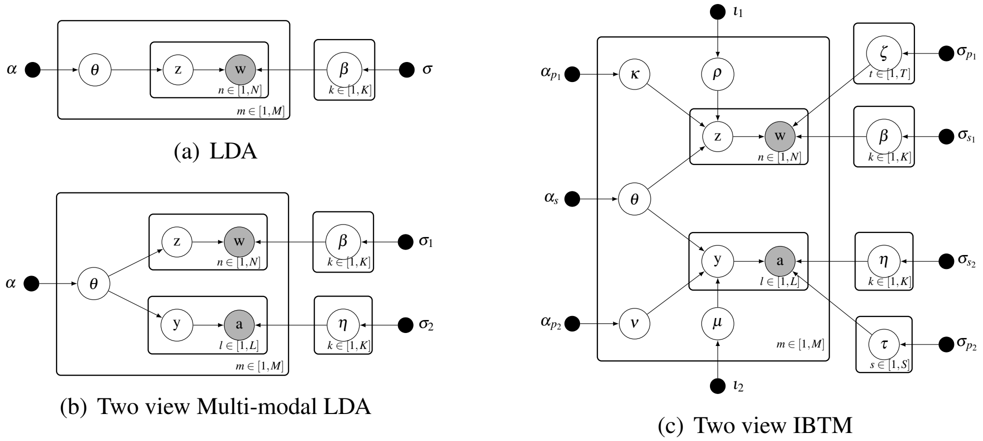
3. Model

For this application, we adapt IBTM, which is a generative model. One advantage of generative models is that they achieve good performance even on small data sets. As it is expensive to collect data in the health care system and there is a big variance in the frequency of different types of diseases, this is highly important. Secondly, generative models have the advantage of being able to handle missing data. In this preliminary work, we are only dealing with two modalities, discomfort drawings and diagnostic labels. Even in this simplistic setting, the diagnostic labels are not complete. In health care systems, there exists a variety of examinations and tests that are only partially used for different patients. Hence, a system that can handle missing data is desired in such application. Finally, a probabilistic interpretation of the symptoms and diagnostic decisions is desirable. IBTM is a factorized multi-modal topic model which enjoys all the properties of generative models and is robust to noise in the data.

3.1 Inter-Battery Topic Model

LDA Topic models, a group of generative models based on Latent Dirichlet Allocation (LDA) [Blei et al. (2003)], have been successfully applied to many scenarios, mainly focusing on discrete data. The graphical representation of LDA is shown in Figure 1(a). LDA assumes that each word in a document is generated by sampling from a per document topic distribution $\theta \sim Dir(\alpha)$ and per topic word distribution $\beta \sim Dir(\sigma)$. The document here stands for an information piece, such as a visual document (picture or video) or any collection of text. The topics are latent representations which can be topics in text documents or symptom groups in medical documents.

Multi-Modal Topic Model LDA is designed for data with a single modality. In our application, we want to jointly model discomfort drawings and the diagnostic labels. Additionally, we want to learn a system that is able to give high quality predictions of diagnostic labels given an unseen discomfort drawing. Hence, a multi-modal topic model is needed. There exist a number of approaches to extend LDA to capture multi-modal data in a joint fashion [Blei and Jordan (2003);



(a) LDA

(b) Two view Multi-modal LDA

(c) Two view IBTM

Figure 1: Graphical representations of three topic models.The nodes with grey shadows indicate an observation while all other nodes are latent variables that need to be learned.

Wang et al. (2009); Wang and Mori (2011); Hospedales et al. (2011); Zhang et al. (2013)], among which some were designed for special applications and based on more assumptions. Multi-modal LDA (MMLDA) [Blei and Jordan (2003); Zhang et al. (2013)] is the most natural multi-modal extension of LDA. Figure 1(b) shows the graphic representation of MMLDA in a two modality case, where w and a represent the observations for each modality and θ is the joint latent representation.

IBTM As shown in Figure 1(b), MMLDA forces the two modalities to completely share a latent space θ. However, real-life data is noisy and incomplete in general and might have shared and disjunct latent sources. In our application, we also need to deal with exchangeable clinical terms and possible missing labels. IBTM is proposed to make MMLDA more robust with respect to complex real-life data. Compared to MMLDA, a private topic space for each modality (κ and ν) is introduced to explain away irrelevant information. By this, the shared topic space can provide qualitatively better latent representations of the structure of the data. In IBTM, $\rho \sim Beta(\iota_1)$ and $\mu \sim Beta(\iota_2)$ are portions of the information that can be shared between the two modalities, where ι_1 and ι_2 are two dimensional pairs of beta distribution hyper-parameters.

In our task, the first modality is the discomfort drawing, where a bag-of-words representation of the discomfort areas is used as the observation w. The second modality is the diagnostic labels y which are only available in the training phase. Each document m contains a discomfort drawing and its corresponding diagnostic labels. Both modalities share the same per document topic distribution θ which can be interpreted as the combination of symptoms that generate a drawing and its diagnostic labels. For each modality, the private topic distributions κ and ν are used to encode the information that cannot be simultaneously explained by both modalities which are noises per se in general. The β is the per shared topic distribution for the drawing locations and η is the per shared topic distribution for the diagnostic labels. These encode the essential information that will be used for prediction. Similarly, the ζ is the per private topic distribution for the drawing location and τ is the per private topic distribution for the diagnostic labels. These encode irrelevant information that needs to be explained away. Each topic is a latent variable, which indicates the problem of the patient. For example, in the case shown in Table 1, one topic in θ may be an injury between the 4th and 5th lumber vertebrae, which generates the discomfort drawing (w) in the upper tie and lower back and generates the diagnostic labels (y) L4 Radiculopathy and DLI L4-L5.

To learn all latent parameters in IBTM, mean field variational inference is used in this work, because variational inference is efficient and can easily be adapted to online settings [Hoffman et al. (2010); Ranganath et al. (2013); Wang et al. (2011)]. In real health applications, online learning is desirable. A standard batch update is used for the experiments due to the small amount of the data. An online version of the IBTM for diagnosis prediction is derived based on the batch vision in Zhang et al. (2016) and implemented for long-term usage for this application.

3.2 Diagnostic Prediction using IBTM

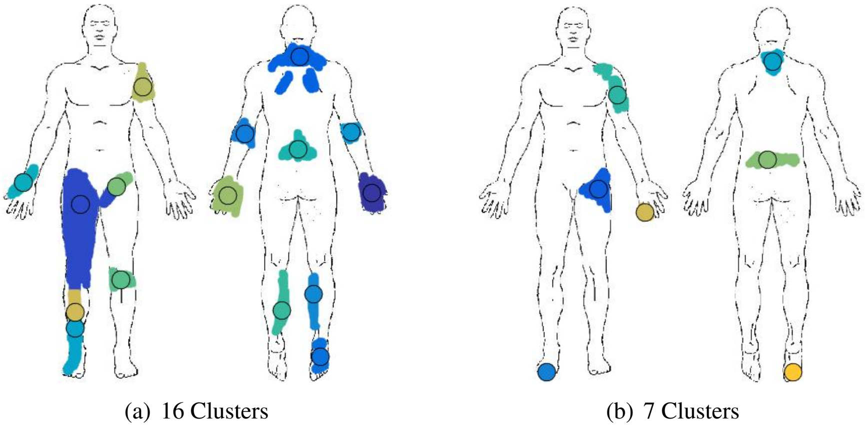
Diagnostic Prediction In the training phase, all latent variables will be learned. In the testing phase, given an unseen observation w without diagnostic labels y, we will estimate the per document distributions θ and κ using the learned per topic word distribution i.e. the global parameters β and ζ. Given the estimated θ for the new document, possible y can be easily generated with the help of the learned parameter η. Hence, we can predict diagnostic labels given a new discomfort drawing. Using IBTM, only the shared topic distribution θ is used for diagnostic prediction, which is similar to MMLDA. However, the latent representation is of higher quality due to the private topics κ that explain away irrelevant information.

Using IBTM, we can generate all possible diagnostic labels y for a drawing with different probabilities. However, it is difficult to decide how many diagnostic labels are actually needed since there exists no universal probability threshold. One patient may have a broken toe for which one or two labels are needed. Another patient may have several discoligament injuries causing discomfort in multiple areas, where more than 50 labels may be needed. To determine how many diagnostic labels are required, we would like to know how many discomfort regions are contained in the test pain drawing. Intuitively, the number of diagnostic labels are positively correlated with the number of discomfort regions. Thus, we use the mean shift clustering algorithm [Comaniciu and Meer (2002)] to cluster the noisy, irregular drawing locations into coherent groups. Mean shift clustering is non-parametric, hence it let data determine the number of clusters. Figure 2 shows examples of the output of mean shift clustering on test discomfort drawings. In this paper, we use twice the number of clusters as the number of prediction labels since the labels are a mixture of symptom diagnostic labels and pattern/pathophysiological diagnostic labels.

Diagnostic Interpretation Besides automatic diagnostic prediction, it is useful to investigate which features the models learn to make these diagnostic predictions. Based on big datasets, the model may give us insights into diagnostic procedures. Hence, in this work, we also generate synthetic discomfort drawings given a diagnostic label. In an ideal case, the model will provide knowledge about the typical discomfort drawing of each diagnostic label. This is done in a similar way as diagnostic prediction. Instead of a test drawing, we give the model a test diagnostic label without a drawing. In the following, the model is used to generate possible drawings. In the experiments section, we will evaluate the diagnostic interpretation of these generated synthetic discomfort drawings.

4. Experiments

Dataset A dataset of 174 real-world patient discomfort drawings was collected from clinical records with diagnostic labels from medical experts. The clinic in question is specialized on di-



(a) 16 Clusters

(b) 7 Clusters

Figure 2: Examples of the clustering output with mean shift. The shades were drawn by patients while the circles indicate the mean location of an identified cluster of the discomfort region.

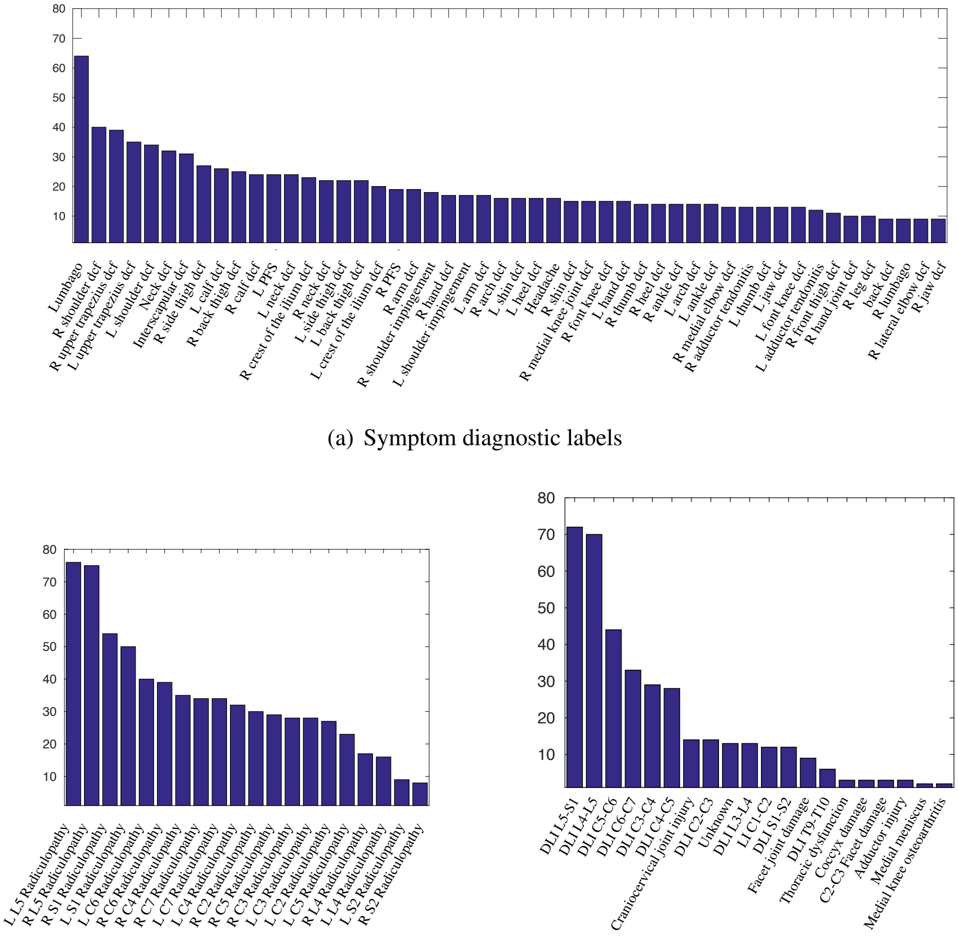
agnosing unspecific pain and discomfort and presented patient cases often have neuropathic pain syndromes. Since bilateral diagnostic labels indicates the problem shows in both sides, we preprocess the data breaking all bilateral labels into left side and right side labels. Taken the example in Figure 1, the preprocessed labels are:

Interscapular discomfort; R arm discomfort; L hand discomfort; R hand discomfort; Lumbago; L crest of the ilium discomfort; R crest of the ilium discomfort; L side thigh discomfort; L back thigh discomfort; R back thigh discomfort; L calf discomfort; R calf discomfort; L achilles tendinitis; R achilles tendinitis; L shin discomfort; R shin discomfort; R inguinal discomfort; L L5 Radiculopathy; R L5 Radiculopathy; L S1 Radiculopathy; R S1 Radiculopathy; L C7 Radiculopathy; R C7 Radiculopathy; DLI L4-L5; DLI S1-S2; DLI C6-C7.

Symptoms such as Interscapular discomfort are kept using only the name without the indication of left or right since it lies in the middle of the body. However, when the discomfort is occurring on one side of the body, medical experts also indicate this information. Moreover, it is common that the same symptoms can be termed differently in different systems. After consulting medical personal, we treat the diagnostic labels listed in Table 2 as exchangeable, which means that they will be treated as the same label. Whether these medical terms are equivalent is a point of discussion, but this is not in the range of this work. We believe that the equivalence of the labels listed in Table 2 is assured. After preprocessing the diagnostic labels, the number of labels per patient ranged from 2 to 50. Figure 4 shows a histogram of the top symptom, pattern and pathophysiological diagnostic labels. About 30% of these diagnostic labels appear only once in the dataset.

Exchangeable labels		Exchangeable labels	
Medial elbow dcf	Golfer's elbow	Lateral elbow dcf	Tennis elbow
Nerve strain effect	Myelopathy	Medial knee arthrosis	Gonarthrosis
Medial meniscus	Medial gonarthrosis	Jew dcf	Bruxism
Back thigh dcf	Hamstrings dcf	Hand joint dcf	Carpal Tunnel Syndrome
Heel dcf	Calcaneodynia	Upper abdominal dcf	Gastritis
Side thigh dcf	Piriformis tendonitis	Crest of the ilium dcf	Trochanter
Throat dcf	Globus hystericus	Coxarthrosis	Hip joint arthritis

Table 2: List of exchangeable labels. "dcf" stands for discomfort.



(a) Symptom diagnostic labels

(b) Pattern diagnostic labels

(c) Pathophysiological diagnostic labels

Figure 3: Histogram of different types of diagnostic labels appeared in the dataset. The x-axis shows different diagnostic labels and the y-axis shows the number of occurrences of each label.

4.1 Diagnostic Prediction Evaluation

We randomly split the dataset into two halves. One half is used for training the model where both discomfort drawings and diagnostic labels are used while the other half is used for testing (i.e. only the discomfort drawings are available). We cluster all painted point locations on the drawing using K-means clustering with 256 clusters. Subsequently, each discomfort drawing is represented with help of a bag-of-location words. In this work, we only use those discomfort area which have been confirmed by medical experts to be the most relevant. The diagnostic labels are used as the second modality. To balance the number of words in both modalities [Tang et al. (2014); Zhang and Kjellstrom (2014)], the diagnostic labels are scaled up by 10 in the experiment so that the number of words in both modalities is in the same order of magnitude.

We use the average F-measure on the predicted diagnostic terms to evaluate the prediction performance. The number of predicted diagnostic terms is determined by mean shift clustering for each test drawing. Additionally, we set a minimum number of 5 labels and a maximum number of

50 labels. The F-measure is defined as:

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{1}$$

The dataset is randomly split 10 times for evaluation and the performance is reported in Table 3 with mean and standard deviation for these 10 groups of experiments with different numbers of shared topics[1]. The number of private topics is set to $T = 5, S = 5$ in all the experiments. The hyper-parameters are set to $\alpha_* = 0.8$, $\sigma_* = 0.6$ and $\iota_* = (1,1)$.

	$K = 5$	$K = 10$	$K = 20$	$K = 30$	$K = 50$
F-measure	$34.31 \pm 1.35\%$	$36.7 \pm 1.37\%$	$38.32 \pm 1.1\%$	$38.56 \pm 1.23\%$	$38.81 \pm 1.08\%$

Table 3: Prediction performance

Table 4 shows typical examples of test results. We found that reasonable diagnostic labels can be suggested using IBTM. The first example in Table 4 shows a case of high prediction accuracy with a small number of predicted labels.

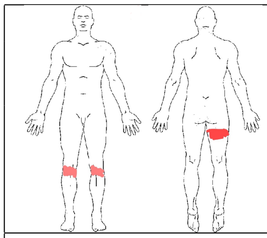
The second and third example show typical predictions for which the F-Measure is in the same level as the mean F-measure. The second example produced a large number of diagnostic labels. 50 Prediction labels are generated and 50 ground truth labels are given. Approximately half of the predictions match the ground truth. However, the mismatched labels are reasonable as well. For example, IBTM predicted Headache, L neck discomfort R neck discomfort, Neck discomfort; while the ground truth gives L back headache and R back headache and only a general Neck discomfort. This is caused by different levels of specificity. These could in fact be considered as correct labels under a more systematic labelling level. The same applies to the prediction of toe joint discomfort while the ground truth only includes big toe discomfort. In the drawing, the big toe and toe joint region are overlapping. Additionally, interscapular discomfort is predicted but not named in the ground truth, although it is marked in the drawing. Hence, when many labels are required, missing labels are easily encountered in both medical expert judgement and machine learning systems.

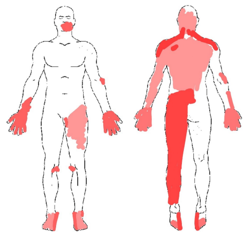
The third example shows a case where the number of predictions exceeds significantly the ground truth labels. This is due to the difficulty to determine the number of required predictions, where clustering methods can only aid us but can be inaccurate. Mismatched predictions are the result. The predicted label L upper trapezius discomfort and R upper trapezius discomfort are apparent in the drawing but not included in the ground truth. Additionally, nerves that next to each other are hard to differ due to overlapping symptoms.

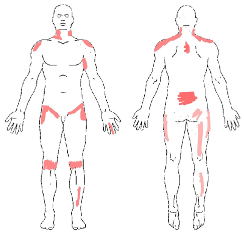
The last example demonstrates highly inaccurate predictions, as the patient has a single local problem. This is a rare case in this dataset. IBTM still suffers from imbalanced data. However, the predicted label PFS is describing knee problems which seems to be a reasonable prediction.

In the end, manually judging the predicted labels, 80% of the labels are in fact reasonable. The measurement in Table 3 is a rather rough measure without considering more fitting metrics. With a systematic evaluation standard, for example, considering the predicted label that has a correspondence on the drawing (upper trapezius discomfort in the third example) as a correct prediction; and considering the labels within a coherent group as correct predictions (back headache is one type of headache), the F-Measure can be easily recomputed around 70%. Hence, we believe that with more data and more systematic diagnostic labels, machine learning algorithms can achieve high quality diagnostic predictions. We identify this as an important direction of future work.

[1] For each experiment setting, 10 random seeds were considered and the best result is used.

	**6 Prd:** R back thigh dcf; L PFS (Patellofemoral pain syndrome); R PFS;
	L L5 Rdc; R L5 Rdc; DLI L4-L5;
	6 GT: R back thigh dcf; L PFS; R PFS;
	L L5 Rdc; R L5 Rdc; DLI L4-L5;

	**50 Prd:** Headache; L neck dcf; R neck dcf; Neck dcf; L upper trapezius dcf; R upper trapezius dcf; L shoulder dcf; L hand dcf; R hand dcf; Interscapular dcf; Lumbago; Lateral abdominal dcf; L groin dcf; L side thigh dcf; R side thigh dcf; L calf dcf; L back thigh dcf; L crest of the ilium dcf; R crest of the ilium dcf; R foot arch dcf; L toe joint dcf; R toe joint dcf; L medial elbow dcf; L ankle dcf; R ankle dcf; L foot arch dcf; L PFS; L dorsal knee dcf; R medial knee dcf;
	L C2 Rdc; R C2 Rdc; L C3 Rdc; R C3 Rdc; L C4 Rdc; R C4 Rdc; R C6 Rdc; L C6 Rdc; L C7 Rdc; R C7 Rdc; L L5 Rdc; R L5 Rdc; L S1 Rdc; R S1 Rdc; DLI C2-C3; DLI C3-C4; DLI C5-C6; DLI C6-C7; DLI L4-L5; DLI L5-S1; OB;
	50 GT: L back headache; R back headache; Neck dcf; L jaw dcf; L upper trapezius dcf; R upper trapezius dcf; L arm dcf; R arm dcf; L lateral elbow dcf; R lateral elbow dcf; L hand joint dcf; R hand joint dcf; L hand dcf; R hand dcf; L thumb dcf; R thumb dcf; L finger dcf; R finger dcf; Lumbago; L groin dcf; L back thigh dcf; L calf dcf; L medial knee dcf; L ankle dcf; R ankle dcf; R medial knee dcf; R big toe dcf; L big toe dcf;
	L C2 Rdc; R C2 Rdc; L C3 Rdc; R C3 Rdc; L C4 Rdc; R C4 Rdc; L C5 Rdc; R C5 Rdc; L C6 Rdc; R C6 Rdc; L C7 Rdc; R C7 Rdc; L L4 Rdc; L L5 Rdc; R L5 Rdc; L S1 Rdc; R S1 Rdc; Craniocervical joint injury; DLI C4-C5; DLI L3-L4; DLI L4-L5; DLI L5-S1;

	**36 Prd:** L neck dcf; L shoulder impingement; R shoulder impingement; L shoulder dcf;R shoulder dcf; L upper trapezius dcf; R upper trapezius dcf; Lumbago;L crest of the ilium dcf; R crest of the ilium dcf; L adductor tendonitis; R back thigh dcf; L PFS; R PFS; R calf dcf; L back thigh dcf; R anterior knee dcf;Coccydynia; L anterior knee dcf; R medial knee dcf;
	L C4 Rdc; R C4 Rdc; L C6 Rdc; L C7 Rdc; L L5 Rdc; R L5 Rdc; R S1 Rdc; L S1 Rdc; L S2 Rdc; R S2 Rdc; DLI C3-C4; DLI C5-C6; DLI C6-C7; DLI L4-L5;DLI L5-S1; DLI S1-S2;
	27 GT: L neck dcf; R neck dcf; L shoulder impingement; R shoulder impingement; L shoulder dcf; R shoulder dcf;; Interscapular dcf; L PFS; R PFS; Lumbago; L crest of the ilium dcf; R crest of the ilium dcf; L adductor tendonitis; R adductor tendonitis; R sciatica; L shin discomfort; R side thigh dcf;
	L C5 Rdc; R C5 Rdc; L C7 Rdc; L L5 Rdc; R L5 Rdc; R S1 Rdc; DLI C5-C6; DLI C6-C7;DLI L4-L5; DLI L5-S1;

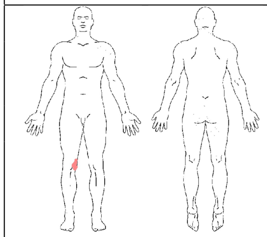
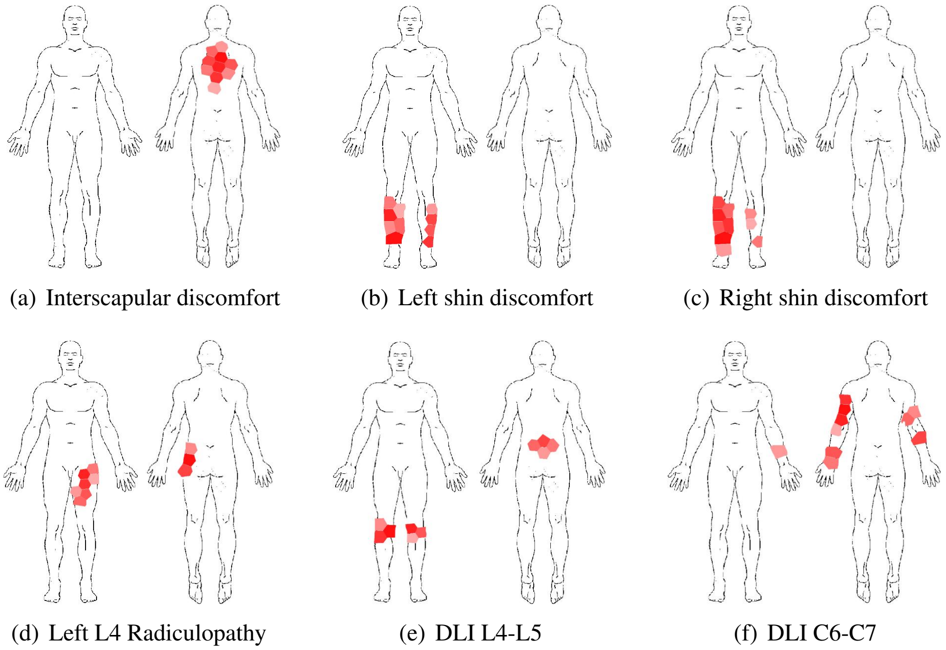
	**5 Prd:** L PFS; R PFS;
	R L5 Rdc; DLI L4-L5; L L5 Rdc;
	2 GT: R Medial knee joint dcf;
	R Medial meniscus;

Table 4: Prediction examples: The left column shows the input discomfort drawing. The right column shows the predicted diagnostic labels using IBTM after **Prd:** and the ground truth diagnostic labels given by medical experts after **GT:**. The number of diagnostic labels is indicated in front of **Prd:** and **GT:**.Correctly predicted labels are marked in blue, while the wrong ones are marked in red. All labels are given in the order of symptom, pattern and pathophysiology. Rdc stands for Radiculopathy and bcf stands for discomfort in the table.

4.2 Diagnostic Interpretation Evaluation

In this section, we investigate the structure that the model learned as described in the second part of Section 3.2. In this evaluation, we train IBTM with all available data. In the predication phase, we provide one diagnostic label as the label modality and generate a discomfort drawing. Figure 4 shows examples of generated drawings with top 10 location words plotted with decreasing intensity for less probable areas. The first row shows examples given a symptom label and the second row shows examples given pattern and pathophysiology label. In the first row, we find that IBTM can generate typical drawings for each symptom, however, it does not always differ between left and right side correctly. This is caused by the large amount of bilateral symptom labels in the data. In the second row, Figure 4 (d) shows a very typical case of left side L4 nerve radiculopathy and (e) and (f) show typical drawings of DLI L4-L5 and DLI C6-C7. This means that the model is able to learn diagnostic patterns automatically. With a large amount of data, this information can potentially be used to help humans to differentiate between different factors in the diagnostic process.



(a) Interscapular discomfort

(b) Left shin discomfort

(c) Right shin discomfort

(d) Left L4 Radiculopathy

(e) DLI L4-L5

(f) DLI C6-C7

Figure 4: Generated discomfort drawings given a diagnostic label

5. Discussion

In this paper, we used IBTM for automated assessment of discomfort drawings. A dataset containing real-world discomfort drawings and corresponding diagnostic labels was collected. Reasonable diagnostic predictions were found in the experiments. This preliminary work on this application area shows a promising research direction. We will continue to enlarge and refine the dataset and improve the model. At the same time, we will investigate how to present machine learning results to real-life health care personnel. We believe that applying machine learning for diagnostic prediction on discomfort drawings may have a significant impact on the health care system. It may lead to decision support systems that can help health care personnel to increase effectiveness and precision in diagnosis and treatment of patients.

References

M. J. Albeck. A critical assessment of clinical diagnosis of disc herniation in patients with monoradicular sciatica. *Acta neurochirurgica*, 138(1):40–44, 1996.

B. C. Bertilson, M. Grunnesjö, and L. E. Strender. Reliability of clinical tests in the assessment of patients with neck/shoulder problems?impact of history. *Spine*, 28(19):2222–2231, 2003.

B. C. Bertilson, M. Grunnesjöand S.E. Johansson, and L.E. Strende. Pain drawing in the assessment of neurogenic pain and dysfunction in the neck/shoulder region: Inter-examiner reliability and concordance with clinical examination. *Pain medicine*, 8(2):134–146, 2007.

B. C. Bertilson, E. Brosjö, H. Billing, and L. E. Strender. Assessment of nerve involvement in the lumbar spine: agreement between magnetic resonance imaging, physical examination and pain drawing findings. *BMC musculoskeletal disorders*, 11(1):1, 2010.

D. M. Blei and M. I. Jordan. Modeling annotated data. In *International Conference on Research and Development in Information Retrieval*. ACM, 2003.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.

M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, 2010.

T. M. Hospedales, S. G. Gong, and T. Xiang. Learning tags from unsegemented videos of multiple human actions. In *International Conference on Data Mining*, 2011.

M. Kukar, I. Kononenko, C. Grošelj, C. Kralj, and J. Fettich. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial intelligence in medicine*, 16(1):25–50, 1999.

D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2006.

D. Ohnmeiss, H. Vanharanta, and J. Ekholm. Relation between pain location and disc pathology: a study of pain drawings and ct/discography. *The Clinical journal of pain*, 15(3):210–217, 1999.

H. Palmer. Pain charts; a description of a technique whereby functional pain may be diagnosed from organic pain. *The New Zealand medical journal*, 48(264):187–213, 1949.

R. Ranganath, C. Wang, D. Blei, and E. Xing. An adaptive learning rate for Stochastic Variational Inference. In *International Conference on Machine Learning*, 2013.

Y. Tanaka, S. Kokubun, T. Sato, and H. Ozawa. Cervical roots as origin of pain in the neck or scapular regions. *Spine*, 31(17):E568–E573, 2006.

J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. In *International Conference on Machine Learning*, 2014.

C. C. Upshur, G. Bacigalupe, and R. Luckmann. They don't want anything to do with you: Patient views of primary care management of chronic pain. *Pain Medicine*, 11(12):1791–1798, 2010.

N. Vucetic, H Määttänen, and O Svensson. Pain and pathology in lumbar disc hernia. *Clinical orthopaedics and related research*, 320:65–72, 1995.

C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

C. Wang, J. Paisley, and D. M. Blei. Online variational inference for the hierarchical dirichlet process. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Y. Wang and G. Mori. Max-margin Latent Dirichlet Allocation for Image Classification and Annotations. In *British Machine Vision Conference*, 2011.

C. Zhang and H. Kjellstrom. How to Supervise Topic Models. In *European Conference on Computer Vision-workshop on Graphical Models in Computer Vision*, 2014.

C. Zhang, D. Song, and H. Kjellstrom. Contextual Modeling with Labeled Multi-LDA. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013.

C. Zhang, H. Kjellström, and C. H. Ek. Inter-battery topic representation learning. In *European Conference on Computer Vision*, 2016.