Interpret Your Care: Predicting the Evolution of Symptoms for Cancer Patients

Rupali Bhati [1], Jennifer Jones [2], Audrey Durand [1,3]

[1] Université Laval
[2] Princess Margaret Cancer Centre, University Health Network
[3] CIFAR AI Chair

Abstract

Cancer treatment is an arduous process for patients and causes many side-effects during and post-treatment. The treatment can affect almost all body systems and result in pain, fatigue, sleep disturbances, cognitive impairments, etc. These conditions are often under-diagnosed or under-treated. In this paper, we use patient data to predict the evolution of their symptoms such that treatment-related impairments can be prevented or effects meaningfully ameliorated. The focus of this study is on predicting the pain and tiredness level of a patient post their diagnosis. We implement an interpretable decision tree based model called LightGBM on real-world patient data consisting of 20163 patients. There exists a class imbalance problem in the dataset which we resolve using the oversampling technique of SMOTE. Our empirical results show that the value of the previous level of a symptom is a key indicator for prediction and the weighted average deviation in prediction of pain level is 3.52 and of tiredness level is 2.27.

Introduction

Nearly half of Canadians are expected to receive a diagnosis of cancer in their lifetime (Brenner et al. 2020). Unlike other chronic diseases where disability is commonly caused by the disease process, disability associated with cancer is often caused more by treatment than the disease itself (Short PF 2008). Individual cancer patients may undergo different treatments, for different amounts of time and may even suffer from multiple diagnoses. Such experiences render the patients with distinct symptoms along their treatment, for example pain, fatigue, sleep disturbances, or cognitive impairments. There is a need for tailored models of rehabilitation that account for the fact that "one size is not likely to fit all" needs of different cancer survivors (Chan et al. 2021). To this aim, in many cancer centres, patient symptom data is collected via surveys which are filled by the patient when they come for their cancer clinic visits. However, asking patients to fill in surveys about their health conditions too often can be very burdening. Therefore, there is a need to systematically identify the adverse effects of cancer such that they can be treated efficiently.

Our aim is to leverage previously collected patient reported data to predict the evolution of a patient's symptoms. Predicting when a patient encounters high levels of symptom burden may help clinicians take precautionary measures to avoid such situations or be able to effectively manage the patient's conditions. For a predictive model to be used within such an application, it should be interpretable by clinicians (Zhu et al. 2014). The clinician not only needs to see the predicted value of the symptom but also needs to understand the criteria based on which the prediction is made.

In this paper we, therefore, propose an interpretable pipeline for predicting the level of two different symptoms, i.e. pain and tiredness levels, using LightGBM (Ke et al. 2017), a decision tree (DT) based model which is widely used for its efficiency and interpretability. DT based methods have a wide application in healthcare (Podgorelec, Kokol, and et al. 2002; Rufo et al. 2021) and cancer-related applications, for example to predict breast cancer survivability (Khan et al. 2008). Although other, possibly more complex, predictive models have displayed good performances at predicting future symptoms of oncology patients (Papachristou N 2018), there has been resistance to adopting these machine learning based methods in clinical practice due to a perception that these are "black-box" techniques and incompatible with decision-making based on evidence and clinical experience (Zhu et al. 2014). Therefore, the proposed pipeline rather relies on an interpretable approach based on DTs, combined with a training procedure based on oversampling (Chawla et al. 2002) to compensate for the important data imbalance observed across the different values of symptom level. By evaluating this approach against a naive predictor based solely on prior distributions of symptoms, we highlight the necessity of using evaluation metrics that compensate for data imbalance in order to evaluate fairness in model performances. Finally, although a feature importance analysis of our models reveals that the previous symptom value plays an important role in predicting the current symptom value, our results show a clear benefit in leveraging clinical variables in the predictions and, in some cases, the correlation between symptoms.

Problem Statement

Post diagnosis, a patient visits the doctor or clinician either when they are going through a treatment or require some assistance. These visits can take place at irregular intervals. During such visits, the patient is asked to fill a survey doc-

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

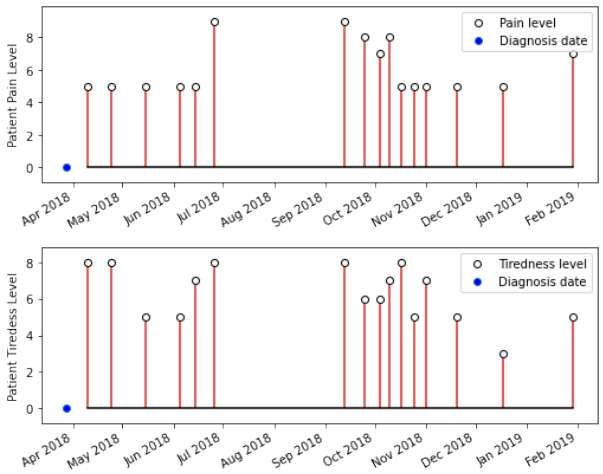

Figure 1: Example of symptom evolution for one patient

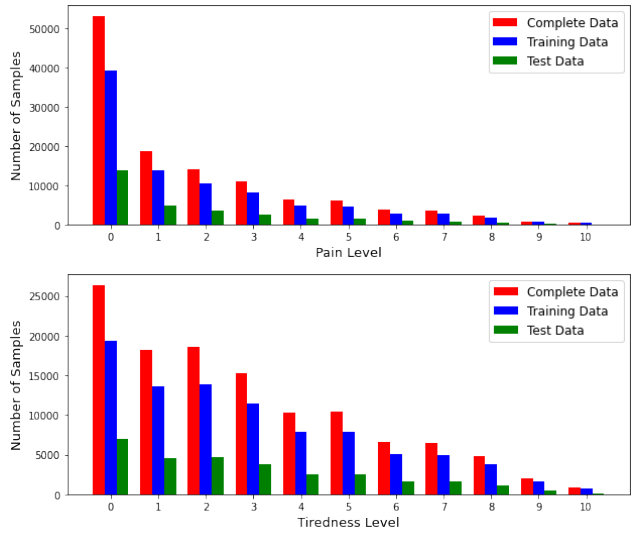

Figure 2: Distribution of number of samples over classes for pain and tiredness levels

umenting their symptoms. Pain and tiredness are two symptoms which acutely affect the overall quality of life of a patient undergoing cancer treatment. Both symptoms are documented by the patient as a discrete value between 0 and 10, where 0 is the lowest level of pain/tiredness a patient can feel and 10 is the highest based on the Edmonton Symptom Assessment Scale (ESAS) (Watanabe, Nekolaichuk, and Beaumont 2012). To demonstrate how the symptoms of a patient can evolve over a period of time, Figure 1 shows an example of the evolution of a patient's symptoms along a timeline.

We have access to a dataset of real cases collected for years 2013-2019 on 20163 patients. On average, there are 5.95 observations per patient. Four cancer types are included in this analysis, namely, breast, colorectal, lymphoma and head and neck. The dataset includes patient information, i.e., sex, age, cancer type, and surgery data, in addition to the survey data, i.e., information about the patient's symptoms, including ESAS values and Patient Reported Functional Status (PRFS) measures (Popovic et al. 2018) which monitor overall nutritional status in patients. For the purpose of this study, we focus on predicting some important symptoms using clinical and demographic information of the patient and their previous symptom values which are summarised in Table 1 along with the values each feature can take.

To avoid a situation where the future observations of a patient are in the training set while the older observations are in the test set, we split the dataset based on a date. The surveys before the date '2017-10-04' are in the training set and the ones after this date are in the test set. This leads to a 75%-25% split between the training and test set, respectively. The distribution of number of samples in each class can been seen in Figure 2 for each symptom. We observe that there are class imbalance problems, where $\sim 44\%$ and $\sim 22\%$ of the outputs belong to one class (level 0) for pain and tiredness, respectively. It is also noted that the number of samples in a class are inversely proportional to the symptom level. This occurs because high levels of pain and tiredness levels are relatively rarer events. This bias in the data can lead to unfairness in the model's predictions and therefore inequity in a patient's treatment and may exacerbate health care disparities (Rajkomar et al. 2018).

Variable	**Possible values**
Sex	male, female
Age	18 to 93
Cancer type	breast, head and neck, lymphoma, colorectal
Number of days between the diagnosis and the survey date	5 to 23091 ($\sim$ 63 years)
Number of days between the previous and the current survey date	1 to 2002 ($\sim$ 5 years)
Previous pain level	{0,1,2,3,4,5,6,7,8,9,10}
Previous tiredness level	{0,1,2,3,4,5,6,7,8,9,10}

Table 1: Variables available for prediction

Methodology

In this section, we propose a pipeline to predict the symptom levels of a patient while trying to maintain equal performance as much as possible over the different classes. As the values to predict are discrete (they belong to the finite, discrete set of symptom levels), we formulate this problem as a classification problem. We propose to evaluate candidate strategies using an interpretable metric that accounts for class imbalance, ensuring that the performance over all groups of patients (based on their symptoms) is equally represented in the global evaluation score.

Predictive model

Clinical providers and other decision makers in healthcare note interpretability of model predictions as a priority for implementation and utilization (Ahmad, Teredesai, and Eckert 2018). To this end, we use DTs for prediction as each decision point can be visualised individually by clinicians.

Among several DT algorithms, we select LightGBM for several reasons. Firstly, LightGBM is a gradient boosting DT algorithm. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models which is more beneficial as compared to other DT algorithms that either train the different models separately before aggregating them or do not use multiple models at all. Secondly, instead of building trees that have a uniform depth across all branches, LightGBM expands the tree leaf-wise (up to a maximum tree depth given as a parameter), resulting in much lighter trees and making it suitable for high-speed and low memory consumption.

For predicting each symptom (pain and tiredness), we consider two variants of LightGBM models. The first variant relies on clinical variables (first five rows of Table 1) and previous level of the predicted symptom (row 6 or 7 of Table 1) in the prediction, while the second variant uses all clinical variables in addition to previous levels of both symptoms (all rows of Table 1). We refer to the first variants as LP1 and LT1, and to the second variants as LP2 and LT2, for pain and tiredness respectively.

Fighting data imbalance Inspired by previous works (Yang Zhao and Tsui 2018), we rely on Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla et al. 2002) to address class imbalance by generating synthetic samples. At training time, SMOTE over-samples the minority classes so that the trained model sees the same number of data points of these classes as for the majority class. A minority class is over-sampled by taking each of its samples and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours. For our experiments, we use $k = 5$.

Training To determine the maximum tree depth parameter for LightGBM, we perform a 5-fold cross-validation on the training set (90046 samples from 16674 patients) for tree depths values between 1 and 25 (as the performance saturates after 25). The maximum tree depth value minimizing the WMAE metric (described below using Equations 1, 2, 3) averaged over the 5 folds is then used for retraining a LightGBM model using on the whole training dataset.

Baselines

We compare the performance of the LightGBM models with two baselines models. We refer to these baselines as NP and PV for naive prior and previous value, respectively.

Naive prior This baseline is a naive strategy that always predicts the dominant (most frequent) symptom level (i.e., 0 for both pain and tiredness).

Previous value This baseline is slightly smarter than NP. It predicts the current symptom level to be the same as the previous symptom level, which could have been recorded at any time in the past because surveys are collected at irregular intervals. Therefore, the previous value may or may not be a good indicator of the current value of the symptom.

Metrics

In case of imbalanced classes, there can often be misleading evaluation metrics where mistakes on a minority class can be hidden by good performance on a dominant class. This can lead to unfairness in the treatment of patients, for example, due to the high frequency of low pain level in the data, we might underestimate the level of pain a patient might be in. To this end, we use weighted metrics such that the model can focus on achieving good performance on individual classes as well as overall. Let $\mathcal{C}$ denote the set of classes. We measure the performance of a given strategy on each class $c \in \mathcal{C}$ using the mean absolute error (MAE) :

$$\text{MAE}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} |y_i - \hat{y}_i| \,, \tag{1}$$

where y_i is the true value of the ith sample in class c, $\hat{y}_i$ is the value predicted by the strategy for the ith sample in class c, and N_c is the number of samples belonging to class c. The MAE corresponds to the absolute deviation in the symptom level making it interpretable for clinicians to infer and analyse. Therefore, a good model would have a low MAE so that for the correct predictions, it would be zero and for the incorrect predictions, the model would predict a symptom level closer to the true symptom level. We combine these values using the weighted-MAE (WMAE), which essentially gives value to the MAE of each class inversely proportional to the number of samples in that class. Therefore, mistakes on a class for which there is little data account for more in the WMAE. To calculate the WMAE, we first calculate the weight of each class c as

$$w_c = \frac{\max_{c' \in \mathcal{C}} N_{c'}}{N_c}. \tag{2}$$

The weight will therefore be of value 1 for the class with the largest amount of data, and would be 10 times larger for a class with 10 times less data. We can then compute the global WMAE as

$$\text{WMAE} = \frac{\sum_{c \in \mathcal{C}} w_c \text{MAE}_c}{\sum_{c \in \mathcal{C}} w_c}. \tag{3}$$

Results

For each symptom, we evaluate the performance of the four considered strategies: the two baselines (NP and PV) and the two variants of LightGBM models (LP1/LT1 and LP2/LT2). We consider both the performance per class (MAE_c) and the global performance (WMAE).

Table 2 shows the results obtained for the task of pain level prediction. Recall that the dominant class for pain level is class 0, therefore, the NP model predicts 0 everywhere. If we did not account for class imbalance, the average of the MAE_c's would be 55/11 = 5, which is lower than the WMAE. Even worse, the global MAE would be 1.65. This shows how low MAE values for high frequency classes can compensate for the high MAE values of low frequency classes, ultimately leading to an unfair metric to evaluate the performance of a model. In that case, the model would

	NP	PV	LP1	LP2
MAE_0	0	0.52	0.8	0.8
MAE_1	1	0.91	1.34	1.35
MAE_2	2	1.3	1.89	1.88
MAE_3	3	1.5	2.11	2.09
MAE_4	4	1.82	2.3	2.26
MAE_5	5	2.04	2.45	2.46
MAE_6	6	2.39	2.54	2.55
MAE_7	7	2.76	2.67	2.72
MAE_8	8	2.89	2.65	2.62
MAE_9	9	2.98	2.59	2.64
MAE_{10}	10	5.03	4.41	4.53
WMAE	8.72	3.88	**3.52**	3.57

Table 2: Prediction performance per pain level (MAE_c) and global performance (WMAE) for each strategy

	NP	PV	LT1	LT2
MAE_0	0	0.7	0.84	0.83
MAE_1	1	0.99	1.21	1.21
MAE_2	2	1.22	1.6	1.59
MAE_3	3	1.41	1.86	1.86
MAE_4	4	1.55	2.02	2.01
MAE_5	5	1.71	2.23	2.22
MAE_6	6	1.91	2.29	2.27
MAE_7	7	2.06	2.17	2.18
MAE_8	8	2.2	2.17	2.15
MAE_9	9	2.4	2	2.03
MAE_{10}	10	3.25	2.63	2.52
WMAE	8.43	2.65	2.33	**2.27**

Table 3: Prediction performance per tiredness level (MAE_c) and global performance (WMAE) for each strategy

predict high frequency classes more often and still be able to maintain a good average MAE. In application, the model would predict high levels of pain less often which would be detrimental to the patients who might need immediate care. The LP1 model performs the best overall suggesting that using the previous tiredness level is not useful to predict the pain level. Although the discrepancy between class performance is slightly attenuated with the LP1 model compared with other strategies, we should still note that, its performance (MAE_c) is still inversely proportional to the class density. This indicates that the LP1 model is still unfair towards individuals with stronger pain and further work would probably be needed to make it clinically acceptable.

Table 3 shows the results obtained for the 4 models used to predict the tiredness level. The dominant class here is once again class 0. Unsurprisingly, we observe lower WMAE when predicting tiredness than pain since the former suffers less class imbalance. Moreover, it can be observed that LT2 performs better than LT1 indicated by the lower WMAE. This suggests that adding the previous pain level is a good indicator to predict the tiredness level.

It is interesting to note that the advantage of LightGBM-based models is observed for higher symptom levels, i.e., MAE_c for $c > 6$ is higher with PV than with LP1/LP2 and MAE_c for $c > 7$ is higher with PV than with LT1/LT2. This suggests that patients with higher symptom levels would benefit the most from more advanced models.

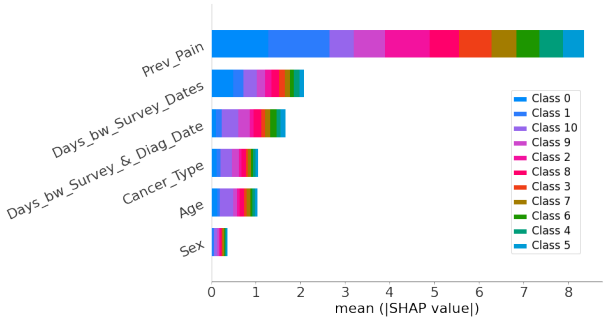

Figure 3: Feature Importance obtained from LightGBM model to predict pain (LP1)

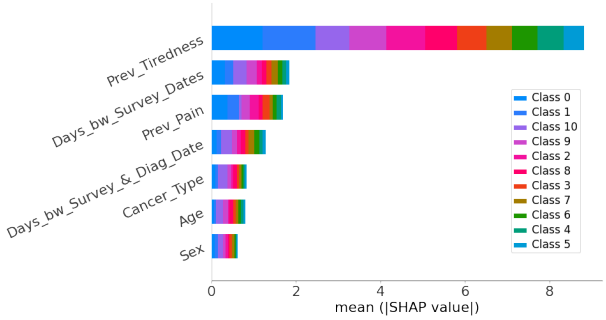

Figure 4: Feature Importance obtained from LightGBM model to predict tiredness using previous pain level (LT2)

Figures 3 and 4 show the importance of each feature in terms of their SHAP (Lundberg and Lee 2017) values for the LP1 and LT2 models, respectively. The SHAP value of a feature indicates the impact of the feature on the magnitude of the output, helping to understand the contribution of each feature to each class. For both symptoms, we observe that previous symptom levels strongly contribute to the prediction for current symptom levels.

Conclusion

In this study, we predict the pain and tiredness level of a patient using their clinical and demographic data and previous observations of the symptom levels. The results demonstrate the need for weighted metrics and the importance of previous symptom levels. In the future, we want to formulate this problem as a a reinforcement learning problem and model the context-reward relationship using the DTs as discussed in (Elmachtoub et al. 2018). We also plan to predict more qualitative measures like the indicator of quality of life.

Acknowledgements

We thank Mathieu Godbout for his feedback on the paper and the CanRehab team for their generous support.

References

Ahmad, M. A.; Teredesai, A.; and Eckert, C. 2018. Interpretable Machine Learning in Healthcare. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 447–447.

Brenner, D. R.; Weir, H. K.; Demers, A. A.; Ellison, L. F.; Louzado, C.; Shaw, A.; Turner, D.; Woods, R. R.; and Smith, L. M. 2020. Projected estimates of cancer in Canada in 2020. *CMAJ*, 192(9): E199–E205.

Chan, R.; Nekhlyudov, L.; Duijts, S.; and et al. 2021. Future research in cancer survivorship. *J Cancer Surviv (2021) https://doi.org/10.1007/s11764-021-01102-x*.

Chawla, N.; Bowyer, K.; Hall, L.; and Kegelmeyer, W. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.*

Elmachtoub, A. N.; McNellis, R.; Oh, S.; and Petrik, M. 2018. A Practical Method for Solving Contextual Bandit Problems Using Decision Trees. arXiv:1706.04687.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Khan, M. U.; Choi, J. P.; Shin, H.; and Kim, M. 2008. Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5148–5151.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777.

Papachristou N, B. P. C. B. H. X. M. R., Puschmann D. 2018. Learning from data to predict future symptoms of oncology patients. *arXiv preprint arXiv:1804.09028*.

Podgorelec, V.; Kokol, P.; and et al., S. B. 2002. Decision Trees: An Overview and Their Use in Medicine. *In: Journal of Medical Systems 26, 445–463 (2002). https://doi.org/10.1023/A:1016409317640*.

Popovic, G.; Harhara, T.; Pope, A.; Al-Awamer, A.; Banerjee, S.; Bryson, J.; Mak, E.; Lau, J.; Hannon, B.; Swami, N.; et al. 2018. Patient-reported functional status in outpatients with advanced cancer: correlation with physician-reported scores and survival. *Journal of pain and symptom management*, 55(6): 1500–1508.

Rajkomar, A.; Hardt, M.; Howell, M. D.; Corrado, G.; and Chin, M. H. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of internal medicine, 169(12), 866–872. https://doi.org/10.7326/M18-1990*.

Rufo, D. D.; Debelee, T. G.; Ibenthal, A.; and Negera, W. G. 2021. Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). *Diagnostics*, 11(9).

Short PF, B. R., Vasey JJ. 2008. Work disability associated with cancer survivorship and other chronic conditions. *Psycho-oncology, 17(1), 91–97.*

Watanabe, S. M.; Nekolaichuk, C. L.; and Beaumont, C. 2012. The Edmonton Symptom Assessment System, a proposed tool for distress screening in cancer patients: development and refinement. *Psycho-Oncology*, 21(9): 977–985.

Yang Zhao, Z. S.-Y. W.; and Tsui, K. L. 2018. A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection. *In: Journal of Healthcare Engineering https://doi.org/10.1155/2018/6275435*.

Zhu, M.; Cheng, L.; Armstrong, J. J.; Poss, J. W.; Hirdes, J. P.; and Stolee, P. 2014. Using machine learning to plan rehabilitation for home care clients: Beyond "black-box" predictions. In *Machine learning in healthcare informatics*, 181–207. Springer.