Identifying Rare Diseases from Behavioural Data: A Machine Learning Approach

Haley MacLeod, Shuo Yang, Kim Oakes, Kay Connelly, Sriraam Natarajan
{hemacleo, shuoyang, kimoakes, connelly, natarasr}@indiana.edu
School of Informatics & Computing
Indiana University
Bloomington, Indiana, USA

***Abstract*—Rare diseases are hard to identify and diagnose. Our goal is to use self-reported behavioural data to distinguish people with rare diseases from people with more common chronic illnesses. To this effect, we adapt a state of the art machine learning algorithm to make this classification. We find that using this method, and an appropriate set of questions, we can accurately identify people with rare diseases.**

I. Introduction

Researchers have dedicated themselves to designing and understanding technologies to support individuals with chronic illnesses in managing their health (e.g. [1]–[4]). These interventions allow people to change their behaviour (e.g. [5]), learn about their disease (e.g. [6]), get support from similar others (e.g. [7]), or track information about themselves (e.g. [8]). However, as far as we are aware, there are not many technologies that focus on rare diseases. These diseases are not generally as well understood in the medical literature and do not have the same constrained set of symptoms to support through design.

There are, however, *experiences* in common between people with rare diseases, and we argue that these experiences are distinct from the experiences of people with common chronic illnesses [9]. There are a wealth of opportunities to support these experiences through the design of appropriate technologies.

Rare diseases impact approximately 10% of the world's population [10]. Despite impacting a substantial number of people, rare diseases are hard to diagnose; receiving a diagnosis can take over five years in the UK and over seven years in the US. Patients often receive 2–3 misdiagnoses before converging on a lasting diagnosis [10]. Physicians are often taught to focus on the most likely diagnosis (*"When you hear hoofbeats, think horses not zebras"*) and it can take visits to many different physicians and specialists to actually identify and diagnose a rare disease [11], [12].

This extended diagnosis process can be extremely frustrating for an individual experiencing an undiagnosed disease. It is crucial to be able to identify and diagnose these conditions in a timely manner. The drawn out diagnosis process can have serious implications for the individuals with rare diseases (health management, finances, work, personal stress, and many other aspects of life). This can cause a huge strain on relationships [9].

As an example, we can consider an individual with an undiagnosed case of amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease. This is a debilitating neurodegenerative disease, but it has a gradual onset which makes it difficult to diagnose early. ALS has a mean survival time of 3–5 years from after the onset of the disease, but this can be prolonged with treatment [13]. Getting diagnosed quickly is imperative—we are willing to screen many individuals who *may* have ALS, knowing that many of them will not have it, to ensure that we do catch the positive cases. This allows those people to receive treatment early and manage the condition as best as is possible.

In this work, we created a survey where the questions were inspired by the results of our previous work [9]. We distributed this survey to people with common chronic illnesses as well as people with rare diseases to help better understand the differences between these two groups and especially to identify the rare examples. Specifically, we ask the following question: Can advanced machine learning techniques, when combined effectively with data about the experiences of people with different health conditions, effectively predict the occurrence of rare diseases? The challenge here is that, given the rarity of the diseases, our survey responses are skewed towards people with common chronic illnesses (i.e. we have more responses from people with common illnesses than we do from people with rare diseases). This problem is called *class imbalance* in machine learning [14]. If we consider people with a common chronic illness as one class and people with a rare disease as another, there is a huge skew towards people with common diseases (by the definition). Simply adopting standard learning techniques can achieve high accuracy by predicting all the examples as the majority class (common chronic illnesses).

Consequently, there has been research in the machine learning community on specific algorithms that can handle the imbalance in a principled manner. Recently, we developed an algorithm [15] that can learn from imbalanced relational data by explicitly trading-off between false positives and false negatives. While the algorithm was designed for learning relational dependency networks, we adapt the algorithm to learn a probabilistic model with standard feature vectors.

The specific contributions of this work are:

1) the adaptation and application of soft-margin functional gradient boosting [15] to a new, real world problem

space (i.e. identifying people with rare diseases),
2) a demonstration of the potential to learn about health from self-reported behavioural data (as an alternative to clinical/medical data), and
3) a discussion of the differences between rare disease and common chronic illness populations to bolster and extend previous qualitative work on this topic [9].

In this paper, we begin by providing background information on rare diseases as well as approaches to learning about health from online behaviour. We discuss the design and distribution of our survey, and the responses we received. We introduce our adapted approach to soft-margin functional gradient boosting, and present the results of experiments on this approach. We conclude by discussing the resulting model and its interpretation of the differences between common chronic illness populations and people with rare diseases, as well as a brief discussion of possible future work in this area.

II. Background

A. Rare Diseases

Rare diseases are conditions that, by definition, impact an extremely small number of people. In the US, rare diseases are those that impact less than 200,000 people (or 0.06% of the population) [10]. In Europe, they are defined as affecting no more than 5 out of every 10,000 people (0.05% of people). While each individual disease is rare, it is estimated that 10% of people world wide have one of the approximately 7,000 rare diseases [10].

Although rare diseases have not been widely studied in consumer health informatics communities, some previous work has begun to examine the needs of people with rare diseases, and how they are similar to or different from the needs of people with common chronic illnesses (i.e. diseases like diabetes, asthma, or arthritis that impact large numbers of people). In our previous work [9], we found that people with rare diseases face a unique set of challenges because they have diseases that few have even heard of, let alone understand. Although some have strong support systems, many have friends and family members that do not understand well enough to really be helpful. Additionally, many of these family members are in denial about the prognosis of the disease, so people with rare diseases rely a lot more heavily on online communities for social and emotional support. These communities also serve as a place to exchange information, because health care providers are also often unfamiliar with rare diseases and patients quickly become researchers and experts in their own conditions. Our past work provided a discussion of the needs of people with rare diseases from a human computer interaction (HCI) perspective, and suggested that these needs should be considered differently than what already exists in the literature about people with common chronic illnesses. We build on this work by:

1) collecting data on a larger scale to validate these qualitative findings, and
2) collecting data from both rare *and* common chronic illness populations, and modelling the differences with the goal of better understanding the ways in which rare disease populations are unique.

B. Learning about Health From Online Behaviour

Much of what we know about health and medicine is the result of rigorous clinical and medical research. Recent research however has seen value in using social media data to identify large scale public health patterns such as influenza spread [16]–[18], detecting depression [19], or Ebola outbreaks [20]. Additional research has been able to identify drug-drug interactions or adverse drug events from Twitter [21], [22] or Instagram [23]. These studies all rely primarily on people's posts about clinical information (i.e. drug names, disease/condition names, symptoms). In this work, we are interested in the value of *behavioural* data in identifying people with rare diseases. That is, we are not looking for social media posts about the diseases symptoms or treatments, but instead explore how people use the Internet and social media to access health information and support.

Some researchers have taken a similar behavioural approach to identifying people with a specific health condition. Saeb et al. [24] found that GPS and usage data from cellphones were strongly related to the severity of depressive symptoms. These types of approaches allow us to learn a great deal about conditions that are otherwise under-reported (e.g. because of stigma). Identifying people with rare diseases presents an additional challenge because we are not targeting one single rare disease, but instead targeting a large class of diseases that are united by their rarity but may have wildly different symptoms. This approach allows us to identify people with diseases about which very little clinical information is known, and ultimately provide them with social or informational support, without needing to know specifically what disease they have.

III. Survey

We created and distributed a survey to better understand the differences between the behaviour of people with rare diseases and people with common chronic illnesses. This study was approved by the Institutional Review Board (IRB) at Indiana University.

A. Survey Design & Distribution

We selected topics for the survey questions that our previous work (as well as existing literature on common chronic illnesses) had demonstrated to be areas where people with rare diseases and people with common chronic illnesses were different—these primarily had to do with technology use, information seeking, and perception of health care professionals. We prepared the survey questions based on a similar Pew Internet & American Life Project survey [25] but tailored the questions to our specific research questions, removing questions that seemed irrelevant and modifying questions to better incorporate prior knowledge. The survey contained 35 questions, divided into 4 themes (Table I).

Theme	Questions	Answers
Demographic Information	Age	Number entry
	Gender	Male, Female, Other
	Country of Residence	Text entry
	Marital Status	{Married, Living with a Partner, Divorced, Separated, Widowed, Single, Other}
	Employment	{Full time, Part time, Retired, Student, Disabled, Not Employed for Pay}
	Education	{Less than grade 8, some high school, completed high school, technical/trade/vocational school, some college/university, completed college/university, some post-graduate education, completed post-graduate education}
Disease Information	Disease name	Text entry
	How many years has it been since you first started experiencing symptoms?	Number entry
	How many years has it been since you were diagnosed?	Number entry
	How severely do your symptoms impact your life?	5 point scale from No impact to Extreme impact
Technology Use	How often do you use the internet?	{Several times a day, About once a day, Several times a week, Every few weeks, Less often, Never}
	Do you own any of the following technologies? (Check all that apply)	{Desktop computer, Laptop computer, Cell phone, e-Reader, MP3 Player, Game console, Tablet}
	On your cell phone, do you have any applications that help you track or manage your health?	{Yes, No, I dont have a cellphone}
	Do you ever use your cell phone to look up health or medical information?	{Yes, No, I dont have a cellphone}
	Have you ever looked online for information about any of the following? (Check all that apply)	{Information about a specific disease or medical problem, Information about a certain medical treatment or procedure, Information about doctors or other health professionals, Information about hospitals or other medical facilities, Information related to health insurance, Information about environmental health hazards, Information about drug safety or recalls, Information about managing chronic pain, Information about medical test results, Information about memory loss, Information about any other health issue}
	Have you ever done any of the following? (Check all that apply)	{Signed up to receive email updates or alerts about health or medical issues, Read someone elses commentary or experience about health or medical issues on an online group, website, or blog, Watched an online video about health or medical issues, Gone online to find others who might have health concerns similar to yours, Tracked your weight, diet or exercise routine online, Tracked other health indicators or symptoms online}
	Have you posted comments, questions or information about your health or medical issues on any of the following? (Check all that apply)	{In an online discussion specific to your condition, In a health related online discussion not specific to your condition, On a blog, On Facebook, Twitter or another social networking site, On YouTube or other video sharing site, On a website of any kind}
	Have you ever used a social networking site to do any of the following? (Check all that apply)	{Get health information, Start or join a health-related group, Follow your friends personal health experiences or health updates, Raise money or draw attention to a health related issue or cause, Remember or memorialize others who suffered from a certain health condition}

	Have you been helped by following medical advice or health information found on the internet?	7 point scale from "Yes, major help" to "No serious harm"
	Have you ever done any of the following online? (Check all that apply)	{Consulted online rankings or reviews of doctors or other providers, Consulted online rankings or reviews of hospitals or other medical facilities, Consulted online reviews of particular drugs or medical treatments, Posted a review online of a doctor, Posted a review online of a hospital, Posted your experience with a particular drug or medical treatment online}
Health care professionals	Do you have a personal or family doctor that you rely on if you need medical care?	{Yes, No}
	How many specialists did you see in the last two years?	Number entry
	Think about the doctor or healthcare professional that you get most of your medical care from. • How helpful is this person in giving you an accurate medical diagnosis? • How helpful is this person in providing emotional support? • How helpful is this person in providing the medical or health information that you need? • How helpful is this person in finding effective treatment strategies for you? • How helpful is this person in coordinating your overall health care?	5 point scale from "Very helpful" to "Very unhelpful"
	Overall, who is most helpful... • ...when you need an accurate medical diagnosis? • ...when you need emotional support in dealing with a health issue? • ...when you need practical advice for coping with day to day health situations? • ...when you need information about alternative treatments? • ...when you need information about prescription drugs? • ...when you need a quick remedy for an everyday health issue? • ...when you need a recommendation for a doctor or specialist? • ...when you need a recommendation for a hospital or other medical facility?	{Health professionals, Friends and family, Fellow patients, Online sources}

TABLE I
SURVEY QUESTIONS

We tested this survey internally to identify potential problems and estimate the time it would take to complete. We then distributed the survey to 26 different Facebook groups for people with specific chronic illnesses. We also distributed our survey on Reddit, targeting 8 subreddits of specific chronic illnesses. We did not recruit more widely on social media because we did not want to create additional privacy risks by encouraging people to indicate to a social network that they had a chronic disease if they had not already done so (by their participation in a disease specific group). We received 341 responses to our survey overall.

B. Survey Responses

Of the 341 responses, we omitted 13 incomplete responses (i.e. the respondent had skipped one or more questions). This left us with 328 responses in our data.

Disease names were entered as text by survey respondents (in many cases respondents had multiple commorbidities and listed these in the text box). One side effect of this approach is that it excluded anyone undiagnosed; everyone who completed the survey knew the name of the condition(s) they had. We discuss the potential for extending this work to undiagnosed populations in Section VII below.

We discretized the disease names as one or more rare diseases (30.34%), or one or more common chronic illnesses (60.66%). For the 6 respondents having at least one common chronic illness *and* at least one rare disease, we created duplicate entries, labelling one entry as common and one entry as rare. We did not have enough of these examples to consider them as a third category, but believed they may have characteristics of either population. We define "rare disease" using the NIH Genetic and Rare Disorder Information Center's database[1].

Respondents ranged in age from 18 to 71 ($\bar{x} = 31.71, s = 12.71$). We discretized these into 5 year bins. The number of years the respondent had been experiencing symptoms and the number of years since the respondent had been diagnosed were similarly discretized into bins.

The data set consisted of responses from 39.00% male respondents and 59.53% female respondents (1.46% of respondents did not identify with either gender).

We received responses from 22 different countries, primarily the US (58.46%), Canada (12.46%), UK (10.39%), and Australia (8.01%). Other countries included Czech Republic, Denmark, France, Germany, Hungary, Israel, Latvia, Malaysia, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Russia, Sri Lanka, and Sweden. We defined "rare disease" using the NIH's database, given that the majority of survey respondents were from individuals living in the US but we recognize that this is a limitation of this work; diseases that are rare in the US may not be rare in other countries (and vice versa). Given that we are trying to capture the *experience* of having a disease that is rare and not the clinical or medical information about the disease, an ideal data set would take into account whether or not the disease was rare in the country where the survey respondent lived. However, rare disease databases do not exist in many of the countries represented in our data.

Many of the questions in the survey asked the respondents to "check all that apply". Each of these checkboxes was treated as a binary feature (checked/unchecked). This provided us with a total of 70 features.

[1]https://rarediseases.info.nih.gov/

IV. Approach

In this study, our main interest is to identify people with rare diseases. By definition, the number of people with rare diseases is much smaller than the number of people with common diseases. In machine learning, this is referred to as *class imbalance*. The nature of class imbalance problems is that the cost associated with misclassifying the rare class (in our case, rare diseases) is higher than misclassifying the common class (common chronic illnesses). The cost matrix does not contain uniform cost values but rather has a higher cost for the instances that are rare diseases but classified as common chronic illnesses than for those that are common chronic illness but are classified as rare. If we assume the class of rare disease as the positive class, then we prefer *recall* (how many relevant items are selected) rather than precision (how many selected items are relevant).

In this section, we show how to incorporate such domain knowledge into the cost matrix when using machine learning. Particularly, expanding on our prior work on relational models [15], we introduce a penalty term into the objective function of a learning algorithm that allows for the trade-off between the precision and recall to be tuned during the learning process.

The learning algorithm we employ is called *functional-gradient boosting* (FGB) [15], [26]–[30]. We use y to denote the class variable (in our case rare/common disease), $\mathbf{X}$ to be the set of features, x^j to be the j^{th} feature and the suffix i to denote the i^{th} example. Thus $\hat{y}_i$ is the true label for the i^{th} example and $\mathbf{X}_i$ is the set of all features for the i^{th} example. In our case, each survey respondent is an example, their disease type (rare/common) is the class, and their answers to all the survey questions is the feature values.

The goal of many probabilistic models is to learn the distribution $P(y_i|\mathbf{X}_i)$ for all examples i. Given a training set, most methods optimize the loglikelihood over the training set, where loglikelihood is given as $LL = \sum_i log(P(y_i|\mathbf{X}_i))$. Gradient-descent is usually performed on the loglikelihood to find the best set of parameters that model the training data.

FGB methods employ a similar approach with two key differences. First, they represent the conditional distributions using a sigmoid function,

$$P(y_i|\mathbf{X}_i) = \frac{e^{\psi(y_i=\hat{y}_i|\mathbf{X}_i)}}{\sum_{y'_i} e^{\psi(y_i=(y'_1|\mathbf{X}_i)}} \quad (1)$$

where the denominator is the sum over all the possible label values of the example. Given that our classification task is

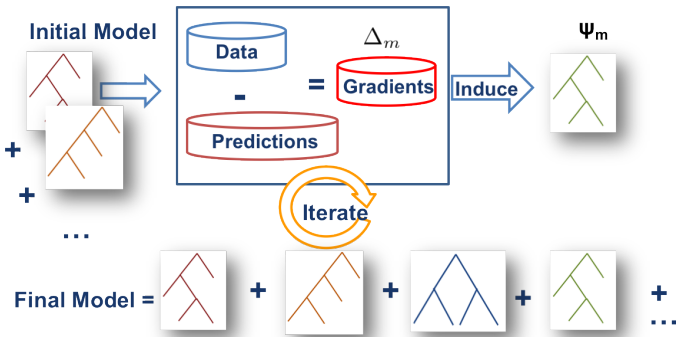

Fig. 1. Functional Gradient Boosting Process

binary, this sum is performed over the common and rare disease probabilities. Now the loglikelihood can be written as,

$$\log LL = \sum_i [\psi(y_i = \hat{y}_i; \mathbf{X}_i) - \log \sum_{y'_i} \exp\{\psi(y'_i; \mathbf{X}_i)\}] \quad (2)$$

The second key difference between an FGB approach and a standard gradient-descent approach is the process of obtaining the gradients. While standard methods differentiate LL with respect to $P(y|\mathbf{X})$ (the parameters of the distribution), FGB methods differentiate LL with respect to ψ (the function that models this conditional distribution). Friedman [26] took this derivative with respect to each training example i. This is an approximation of the true gradient but led to excellent results in many real-tasks [15], [26]–[30].

The gradient of equation 2 with respect to the potentials (ψ_i) for each example i is:

$$\Delta(y_i) = I(\hat{y}_i = Rare) - P(y_i = Rare; \mathbf{X}_i), \quad (3)$$

where I is an indicator function which returns 1 for rare diseases and 0 for common chronic illnesses.

This gradient depends on how well the current model fits the true label of the example, and then is assigned to each example to generate a regression dataset. If it fits well, the gradient would approach 0 and if it does not, the predicted probability of the example would be far from the true label and hence make the boosting algorithm attach a high weight to that example. Then in the next iteration, we learn a regression function to fit the current regression dataset and add this to the model to improve the probabilistic predictions. As these iterations go on, the probability of all positive examples is pushed towards 1, and all the negative examples towards 0. This process is shown in Figure 1.

While this method is successful, it still treats the misclassified negative examples and positive ones equally. In this study, it is important that we classify rare diseases more correctly; the goal is to achieve a higher recall for rare diseases. This can be achieved by modifying the objective function to incorporate this cost difference. Specifically, we add a term to the objective function so the positive and negative examples can be penalized differently.

Following the work of Gimpel and Smith [31] and Yang et al. [15], we introduce a cost function into the objective function:

$$c(\hat{y}_i, y) = \alpha\, I(\hat{y}_i = 1 \wedge y = 0) + \beta I(\hat{y}_i = 0 \wedge y = 1),$$

where $\hat{y}_i$ is the true label of the i^{th} instance and y is the predicted label. $c(\hat{y}_i, y) = 0$ when the example has been correctly classified. $c(\hat{y}_i, y) = \alpha$ when it is a rare disease example but classified as common, while $c(\hat{y}_i, y) = \beta$ when it is a common chronic illness example but classified as rare. The key difference from Yang et al. [15] is that their method assumed a relational representation, while we use the answers to the survey questions as features. They were interested in learning joint models, while we employ the algorithm in the context of learning a single probabilistic function. That is, we are estimating the conditional probability $P(disease = rare|responses)$. Given this difference, our new objective function is now:

$$\log J = \sum_i [\psi(y_i; \mathbf{X}_i) - \log \sum_{y'_i} \exp\{\psi(y'_i; \mathbf{X}_i) + c(\hat{y}_i, y'_i)\}] \quad (4)$$

Note the difference to the original FGB function in Equation 2. We now include a cost function that essentially allows for different misclassification costs for different labels. In our cases, missing rare diseases is costlier than missing common diseases. The gradient of the objective function with respect to $\psi(y_i = Rare; \mathbf{X}_i)$ can be shown as:

$$\frac{\partial \log J}{\partial \psi(y_i = Rare; \mathbf{X}_i)} = I(y_i = Rare; \mathbf{X}_i) - \frac{P(y = Rare; \mathbf{X}_i) e^{c(y_i, y = Rare)}}{\sum_{y'_i} [P(y'_i; \mathbf{X}_i) e^{c(y_i, y'_i)}]} \quad (5)$$

The gradients of the objective function can be rewritten compactly as:

$$\Delta = I(\hat{y}_i = 1) - \lambda P(y_i = 1; \mathbf{X}_i). \quad (6)$$

Where:

$$\lambda = \frac{e^{c(\hat{y}_i, y=1)}}{\sum_{y'} [P(y'; \mathbf{X}_i)\ e^{c(\hat{y}_i, y')}]}.$$

For Rare disease examples, we have:

$$\lambda = \frac{1}{P(y' = 1; \mathbf{X}_i) + P(y' = 0; \mathbf{X}_i) \cdot e^{\alpha}}.$$

As $\alpha \to \infty$, which amounts to putting a large positive cost on the false negatives, $\lambda \to 0$ and the gradients ignore the predicted probability as the gradient is pushed closer to 1 ($\Delta \to 1$), indicating a harsher penalty on misclassified positive examples. On the other hand, when $\beta \to -\infty$, the gradients are pushed closer to 0 ($\Delta \to 0$), indicating more tolerance on misclassified negative. By setting the parameters $\alpha > 0$ and $\beta < 0$, the different costs of false positive and false negative examples can be incorporated into the learning process, hence the trade-off between precision and recall can be controlled.

function SOFTRFGB($Data$)
  for $1 \leq m \leq M$ **do**
    $S :=$GENSOFTMEGS($Data; F_{m-1}$)
    $\Delta_m :=$FITRELREGRESSTREE(S)
    $F_m := F_{m-1} + \Delta_m$
  end for
end function
function GENSOFTMEGS($Data, F$)
  $S := \emptyset$
  for $1 \leq i \leq N$ **do**
    $p_i = P(y_i = 1|\mathbf{x}_i) = sigmoid(F(y_i; \mathbf{x}_i))$
    if $\hat{y}_i = 1$ **then**
      $\lambda = 1/(p_i + (1 - p_i) \cdot e^{\alpha})$
    else
      $\lambda = 1/(p_i + (1 - p_i) \cdot e^{-\beta})$
    end if
    $\Delta(y_i; \mathbf{x}_i) := I(y_i = 1) - \lambda P(y_i = 1|\mathbf{x}_i)$
    $S := S \cup [(y_i), \Delta(y_i; ; \mathbf{x}_i))]$
  end for
return S ▷ Return regression examples
end function
function FITRELREGRESSIONTREE(S)
  Tree := createTree($P(X)$)
  Beam := {root(Tree)}
  while $numLeaves(Tree) \leq L$ **do**
    Node := popBack(Beam) ▷ Node w/ worst score
    C := createChildren(Node) ▷ Create children
    BN := popFront(Sort(C, S)) ▷ Node w/ best score
    addNode(Tree, Node, BN)
      ▷ Replace Node with BN
    insert(Beam, BN.left, BN.left.score)
    insert(Beam, BN.right, BN.right.score)
  end while
return Tree
end function

Fig. 2. Soft-RFGB: RFGB with Soft-Margin

We show our approach in Figure 2. We iterate through M steps and in each iteration, we generate examples based on the soft-margin gradients. We learn a relational regression tree to fit the examples using FITRELREGRESSIONTREE which is added to the current model. We limit our trees to have maximum L leaves and greedily pick the best node to expand.

For generating the regression examples (GENSOFTMEGS function), we iterate through all the examples (N in the algorithm). For each example, we calculate the probability of the example being true (p_i) based on the current model. We then calculate the gradients based on a simplification of Equation 6. The example and its gradient is added to the set of regression examples, S.

V. EXPERIMENTS

Our experiments focus on answering two key questions:

Q1: How effective is a class imbalance approach in detecting rare diseases using self-reported behavioural data?

Q2: How effective is our method in handling class imbalance as part of the classification process (as opposed to changing the class distribution)?

To answer **Q1**, we compare four standard, widely used supervised classification methods (Naïve Bayes, Logistic Regression, 5-Nearest Neighbours, and Decision Trees) that are not specifically designed to handle class imbalance against four class imbalance methods (random undersampling, random oversampling, Synthetic Minority Oversampling TEchnique (SMOTE) [32] and our cost-sensitive boosted probabilistic classifier).

To answer **Q2**, we compare our cost-sensitive boosted probabilistic classifier against two random sampling methods, as these methods have been shown to be effective in imbalanced datasets [33], [34]. We additionally compare our classifier against SMOTE [32], a state of the art oversampling method that generates synthetic examples along the line segments joining the minority examples with their 5 nearest neighbours.

In all experiments, we performed 10-fold cross validation, and the test sets were consistent between experiments. The results of these experiments (Table II) are presented as averages of the results from the 10 test sets. Standard evaluation metrics include accuracy, Area Under ROC curve (AUC-ROC) and F_1 (the harmonic mean of precision and recall). We include these metrics, which measure accuracy with a balanced weight between positive and negative examples, but we focus primarily on evaluation metrics that assign higher weights to higher recall regions. Specifically, we report on two additional metrics: F_3, and F_5, where this F-measure is:

$$F_\beta = (1 + \beta^2) \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

where β is the importance given to recall over precision (i.e. a higher β indicates more emphasis on recall and a smaller β indicates more emphasis on precision). We use F_3 and F_5 to increase the importance of recall over precision. We also report the confusion matrix for each experiment.

The class imbalance methods generally outperform the standard methods at identifying rare examples (TP, FN, F_3, and F_5). Thus, **Q1** can be answered affirmatively in that class imbalance methods are more effective at identifying people with rare diseases. There is a slight trade off when using these methods in our ability to identify common examples (see FP, TN, Accuracy). We see this trade off as acceptable; we would rather falsely identify someone as having a rare disease who actually has a common chronic illness than miss someone who actually has a rare disease. If we can identify someone who *may* have a rare disease from their online behaviour, we can direct them towards appropriate medical and clinical resources to find out for sure (or at least with greater certainty).

Additionally, we can see that our approach comfortably outperforms other class imbalance methods in many respects. It performs substantially better on every class imbalance evaluation metric (TP, FN, F_3, F_5) without losing too much performance on the standard evaluation metrics. Sampling methods alter the class distribution in the training set, which

		Standard Classification Methods				*Class Imbalance Classification Methods*			
		Naïve Bayes	**Logistic Regression**	**5-Nearest Neighbours**	**Decision Trees (J48)**	**Random Oversampling**	**Random Undersampling**	**SMOTE**	**Soft-FGB**
Standard Evaluation Metrics	**FPR***	0.185	0.263	0.088	0.188	0.226	0.292	0.214	0.758
	TNR	0.815	0.737	0.912	0.812	0.774	0.708	0.786	0.242
	AUC-ROC	0.749	0.615	0.684	0.612	0.742	0.759	0.685	0.717
	Accuracy	0.713	0.641	0.724	0.690	0.687	0.688	0.690	0.467
	F	0.514	0.401	0.360	0.437	0.491	0.566	0.478	0.523
Class Imbalance Evaluation Metrics	**TPR**	0.507	0.413	0.288	0.408	0.511	0.657	0.472	0.972
	FNR*	0.492	0.587	0.712	0.592	0.489	0.343	0.528	0.028
	F3	0.507	0.400	0.299	0.412	0.504	0.632	0.471	0.825
	F5	0.507	0.402	0.293	0.410	0.508	0.647	0.472	0.909

TABLE II
EXPERIMENTAL RESULTS FOR PREDICTING RARE DISEASES.
GREEN INDICATES BEST PERFORMANCE.
ASTERISK INDICATES THE METRIC SHOULD BE MINIMIZED (ALL OTHER METRICS SHOULD BE MAXIMIZED).

makes the training set not as reflective of the true test set. It is therefore unsurprising that a method that handles class imbalance in a principled manner would outperform a sampling method. Thus, **Q2** can also be answered affirmatively.

VI. DISCUSSION

Our findings from this study illustrate it is possible to identify people with rare diseases based on self-reported behavioural data using a soft-FGB approach. This suggests that, as we reported in our qualitative interview study [9], people with rare disease have unique challenges that are distinctly different from people with common chronic illnesses and this presents design opportunities not yet addressed by existing interventions and human computer interaction research. When we examine the trees produced by the soft-FGB, we see that the features used to identify people with rare diseases are well-aligned with these previous qualitative findings.

A sample of one of the learned trees is provided in Figure 3. If we examine this tree node by node, we can understand the behaviours that distinguish people with common chronic illnesses from people with rare diseases. Our previous work indicated that people with rare disease take on a much more active role in managing and seeking information about their health [9]. Given the rarity of the conditions, people with rare diseases do not often have in-person access to others with the same conditions, so they turn to online communities to connect with others [9], [10], [35]–[37]. It makes sense that they would be more engaged in activities like watching videos by people with similar health circumstances. Additionally, people with rare diseases are known to contribute to these online communities by posting their own data/test results and talking extensively about their experiences [9], [38]–[41] (including posting their own videos online to share with others).

People with common chronic illnesses do not have the same need to engage in this online support and information seeking behaviour, because there is more information available through medical professionals in the first place and more support available locally. It seems reasonable that many people with common chronic illnesses would have never joined a health group on a social network, never posted a review of a hospital, never memorialized or remembered someone suffering from a health condition, never used the Internet to seek information about medical test results, etc. (as is shown in Figure 3). It is also unsurprising that people with common chronic illnesses have only 1–2 specialists (where it is known that people with rare diseases may have many [10]).

Of particular interest is the few areas where people with common chronic illnesses *do* use technology or the Internet to manage their health; people with common chronic illnesses have used a smartphone app to track health information. We suspect this is because many of the mobile health applications available commercially are designed to target the symptoms, causes, and management of specific chronic illnesses. These conditions are well studied in the medical literature and have a specific and known set of symptoms and treatments, so technologies are customized to specific illnesses. The equivalent applications for rare diseases simply do not exist. We believe that if they did exist, people with rare diseases would be equally likely to use them.

VII. FUTURE WORK

In our study, survey respondents provided a name of a disease or health condition. This meant they were already diagnosed and knew what condition they had. Additionally, all of the data represented self-reported accounts of behaviour (which may be different than actual behavioural data that could be gathered in real-time). We see value in extending this work to use social network data (as in [19], [24]) in addition to self-reported behavioural data. A social platform or tool capable of doing this type of classification or identification in real time could identify people who are not yet diagnosed.

An additional factor to consider would be identifying people with rare diseases amongst healthy populations. Our current study focused on the distinction between common chronic illness and rare disease populations, but it may be interesting to identify those groups from broader social media data. We can imagine that even people who do not have a chronic illness diagnosis may still engage in health information seeking online, if they are interested in general wellness for themselves, or are seeking information for a close friend or family member.

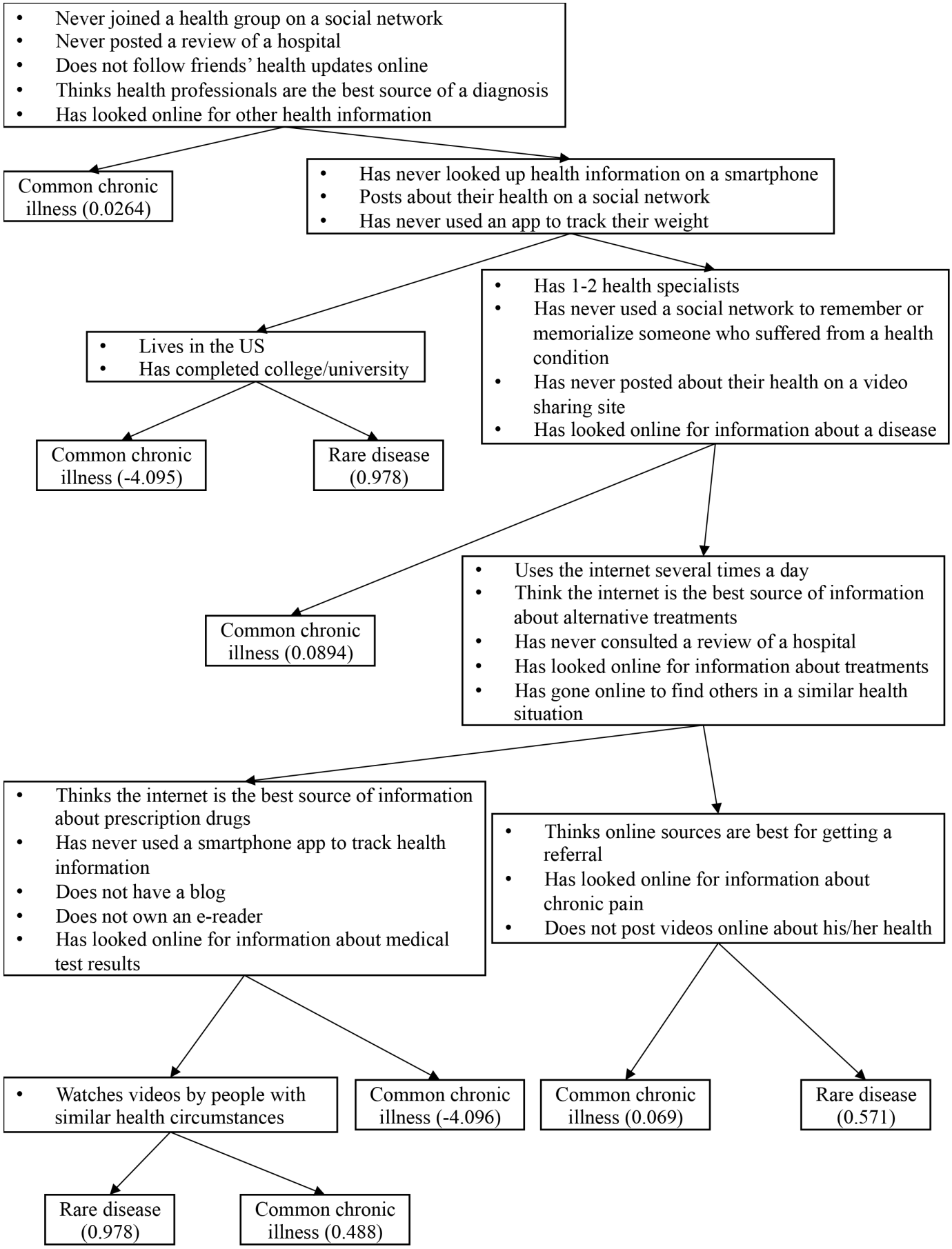

Fig. 3. Sample tree learned in soft-FGB

ACKNOWLEDGMENTS

SY and SN gratefully acknowledge National Science Foundation grant no. IIS-1343940. HM and KC gratefully acknowledge CLEAR Health Information's support of the RARE project.

REFERENCES

[1] L. Mamykina, E. Mynatt, P. Davidson, and D. Greenblatt, "MAHI: investigation of social scaffolding for reflective thinking in diabetes management," in *CHI '08*, 2008, pp. 477–486.

[2] K. Siek, K. Connelly, and Y. Rogers, "Pride and prejudice: learning how chronically ill people think about food," in *CHI '06*. ACM, 2006, pp. 947–950.

[3] L. Liu, J. Huh, T. Neogi, K. Inkpen, and W. Pratt, "Health vlogger-viewer interaction in chronic illness management," in *CHI '13*, 2013, pp. 49–58.

[4] T. Yun and R. Arriaga, "A text message a day keeps the pulmonologist away," in *CHI '13*, 2013, pp. 1769–1778.

[5] S. Consolvo, P. Klasnja, D. W. McDonald, D. Avrahami, J. Froehlich, L. LeGrand, R. Libby, K. Mosher, and J. A. Landay, "Flowers or a robot army?: encouraging awareness activity with personal, mobile displays," in *UbiComp '08*. ACM, 2008, pp. 54–63.

[6] M. D. Choudhury, M. R. Morris, and R. W. White, "Seeking and sharing health information online: comparing search engines and social media," in *CHI '14*. ACM, 2014, pp. 1365–1376.

[7] T. Ammari and S. Schoenebeck, "Networked empowerment on facebook groups for parents of children with special needs," in *CHI '15*. ACM, 2805–2814, pp. 2805–2814.

[8] H. MacLeod, A. Tang, and S. Carpendale, "Personal informatics in chronic illness management," in *GI '13*, 2013, pp. 149–156.

[9] H. MacLeod, K. Oakes, D. Geisler, K. Connelly, and K. Siek, "Rare world: Towards technology for rare diseases," in *CHI '15*. ACM, 2015, pp. 1145–1154.

[10] Shire Human Genetic Therapies, "Rare disease impact report: Insights from patients and the medical community," Shire Human Genetic Therapies, Tech. Rep., 2013.

[11] K. Hunter, ""Don't think zebras": Uncertainty, interpretation, and the place of paradox in clinical education," *Theoretical Medicine*, vol. 17, no. 3, pp. 225–241, 1996.

[12] W. R. Phillips, "Zebras on the commons: Rare conditions in family practice," *Journal of the American Board of Family Medicine*, vol. 17, no. 4, pp. 283–286, 2004.

[13] L. C. Wijesekera and P. N. Leigh, "Amyotrophic lateral sclerosis," *Orphanet journal of rare diseases*, vol. 4, no. 1, p. 3, 2009.

[14] N. Chawla, "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*, 2010, pp. 875–886.

[15] S. Yang, T. Khot, K. Kersting, G. Kunapuli, K. Hauser, and S. Natarajan, "Learning from imbalanced data in relational domains: A soft margin approach," in *ICDM '14*. IEEE, 2014.

[16] H. Kautz, "Data mining social media for public health applications," in *IJCAI '13*. AAAI Press, 2013.

[17] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic," *PLoS ONE*, vol. 6, no. 5, 2011.

[18] A. Sadilek, H. Kautz, and V. Silenzio, "Modeling spread of disease from social interactions," in *ICWSM '12*. AAAI Press, 2012.

[19] M. D. Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *WebSci '13*. ACM, 2013, pp. 47–56.

[20] E. Seltzer, N. Jean, E. Kramer-Golinkoff, D. Asch, and R. Merchant, "The content of social media's shared images about ebola: a retrospective study," *Public Health*, vol. 129, no. 9, pp. 1273–1277, 2015.

[21] A. A. Hamed, X. Wu, R. Erickson, and T. Fandy, "Twitter kh networks in action: Advancing biomedical literature for drug search," *Journal of biomedical informatics*, pp. 157–168, 2015.

[22] A. Sarkera, R. Ginna, A. Nikfarjama, K. O'Connora, K. Smithc, S. Jayaramanb, T. Upadhayab, and G. Gonzaleza, "Utilizing social media data for pharmacovigilance: A review," *Journal of Biomedical Informatics*, vol. 54, pp. 202–212, 2015.

[23] R. B. Correia, L. Li, and L. M. Rocha, "Monitoring potential drug interactions and reactions via network analysis of instagram user timelines," in *Pacific Symposium on Biocomputing*, 2016.

[24] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr, "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study," *Journal of Medical Internet Research*, vol. 17, no. 7, 2015.

[25] Pew Research Center. September 2010 – health tracking. [Online]. Available: http://www.pewinternet.org/datasets/september-2010-health-tracking/

[26] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, 2001.

[27] T. G. Dietterich, A. Ashenfelter, and Y. Bulatov, "Training conditional random fields via gradient tree boosting," in *ICML '04*. ACM, 2004, pp. 28–36.

[28] K. Kersting and K. Driessens, "Non–parametric policy gradients: A unified treatment of propositional and relational domains," in *ICML '08*, 2008, pp. 456–463.

[29] S. Natarajan, T. Khot, K. Kersting, B. Gutmann, and J. Shavlik, "Gradient-based boosting for statistical relational learning: The Relational Dependency Network case," *Machine Learning*, vol. 86, no. 1, pp. 25–56, 2012.

[30] T. Khot, S. Natarajan, K. Kersting, and J. Shavlik, "Learning markov logic networks via functional gradient boosting," in *ICDM '11*, 2011, pp. 320–329.

[31] K. Gimpel and N. A. Smith, "Softmax-margin CRFs: training log-linear models with cost functions," in *HLT '10*. Association for Computational Linguistics, pp. 733–736.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[33] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *IC-AI '00*, 2000.

[34] C. X. Ling and C. Li, "Data mining for direct marketing problems and solutions," in *KDD '98*. AAAI Press.

[35] S. Aymé, A. Kole, and S. Groft, "Empowerment of patients: lessons from the rare diseases community," *The Lancet*, vol. 371, no. 9629, pp. 14–20, 2008.

[36] N. S. Coulson, H. Buchanan, and A. Aubeeluck, "Social support in cyberspace: A content analysis of communication within a Huntington's disease online support group," *Patient Education and Counseling*, vol. 68, no. 2, pp. 173–178, 2007.

[37] T. Gundersen, "'one wants to know what a chromosome is': the internet as a coping resource when adjusting to life parenting a child with a rare genetic disorder," *Sociology of Health & Illness*, vol. 33, no. 1, 2010.

[38] J. H. Frost and M. P. Massagli, "Social uses of personal health information within PatientsLikeMe, an online patient community: What can happen when patients have access to one another's data," *Journal of Medical Internet Research*, vol. 10, no. 3, 2008.

[39] G. R. Polich, "Rare disease patient groups as clinical researchers," *Drug Discovery Today*, vol. 17, no. 3–4, pp. 167–172, 2012.

[40] H. MacLeod, B. Jelen, A. Prabhakar, L. Oehlberg, K. Siek, and K. Connelly, "Asynchronous remote communities (arc) for researching distributed populations," in *PervasiveHealth '16*, 2016.

[41] P. Wicks, M. Massagli, J. Frost, C. Brownstein, S. Okun, T. Vaughan, R. Bradley, and J. Heywood, "Sharing health data for better outcomes on PatientsLikeMe," *Journal of medical Internet research*, vol. 12, no. 2, 2010.