Developing Parameters for a Technology to Predict Patient Satisfaction in Naturalistic Clinical Encounters

Tianyi Tan(✉), Enid Montague, Jacob Furst, and Daniela Raicu

College of Computing and Digital Media, DePaul University, Chicago, IL 60604, USA
ttan6@mail.depaul.edu,
{emontagl, jfurst, draicu}@cdm.depaul.edu

Abstract. Patient-centered communication is crucial in the clinical encounter. Previous studies on patient satisfaction have focused on nonverbal cues and demographics of the patient separately; the integrated influence of both aspects is yet to be explored. This study aims to build a model to learn the quantitative relationship among nonverbal behaviors such as mutual gaze and social touch, demographics of the patients such as age, education and income, and patient perceptions of clinicians. Using 110 videotaped clinical encounters of patients from a study of assessing placebo, Echinacea, and doctor-patient interaction in the acute upper respiratory infection and a decision tree machine learning approach, duration per mutual gaze, percentage of mutual gaze, age, and social touch were identified as the top four important features in predicting how much patients liked their clinicians. Patients of older age, with higher percentage of mutual gaze, longer social touch duration and moderate duration per mutual gaze tended to report greater rating on likeness towards their clinicians. Findings from this study will be used to inform the design of a real-time automatic feedback system for physicians. By using the decision tree machine learning approach, the findings help determine the parameters required for the design of a real-time monitoring and feedback system of the quality of care and doctor-patient interaction in natural environments.

Keywords: Healthcare IT & Automation · Quality and safety in healthcare · Machine learning · Patient satisfaction · Decision tree · Automatic feedback system

1 Introduction

Effective patient-centered communication is integral to the patient-provider relationship and has been identified as a dimension of physician competency. The quality of clinician communication is associated with patient outcomes, such as understanding recommendations for treatment, adherence to therapy, and health outcomes [1]. Patient satisfaction, the key identifier of the quality of the communication [2], is defined as patients' reactions to salient aspects of the clinical experience including cognitive evaluations and emotional reactions [3]. However, accurate measurement of the quality

© Springer Nature Switzerland AG 2020
V. G. Duffy (Ed.): HCII 2020, LNCS 12198, pp. 473–490, 2020.
https://doi.org/10.1007/978-3-030-49904-4_35

of the patient-provider relationship can be challenging. Standardized questionnaires are commonly used as a quantitative method to assess patient satisfaction while unobtrusive observation, video recording, and shadowing are common approaches of qualitative methods [2]. The survey results might be affected by false memory and recall, internal (e.g., emotions) and external factors (e.g. measurement effects). Qualitative data may be difficult to collect, summarize and interpret [3]. The timing of measurement might also affect the ratings due to recall inaccuracies [4]. Mixed methods that incorporate both methods with real-time feedbacks may provide more accurate evaluations. Previous studies have examined correlations between patient satisfaction and eye contact, touch with specific social meaning such as handshaking (social touch) [5], demographics such as gender [6], age, and literacy [7]. Few studies focus on the integration effects of both aspects on patient satisfaction. Effective guidelines and reliable evaluation of physician-patient interactions are needed for practical innovations such as a dynamic feedback system, which can help physicians emphasize positive interactions and build better relationships for longer periods of time.

The purpose of this study was to determine how to develop parameters for a system to provide real-time feedback about the quality of patient-physician interactions in naturalistic settings. In this study, behavioral data from videotaped clinical visits and self-reported surveys were analyzed using a decision tree machine learning approach. The findings can inform the future design of clinical settings and computational health tools focused on patient care. They also provide guidance for personnel recommendation as well as procedures and the care system.

1.1 Assessment of Communication

The effectiveness of communication can be accessed by patient satisfaction which is dependent on good communication skills demonstrated by care providers [4, 5]. Patient satisfaction can be the key identifier of the communication and reliable judgment of the quality of clinical experience [2]. The empathy which provides supportive interpersonal communication is an essential aspect to the patient-clinician relationship and has been linked to greater patient satisfaction [10]. The effectiveness of empathy is related to patient satisfaction, adherence, anxious and stressful emotions, patient enablement, diagnostics and clinical outcomes [11]. Empathy has been studied in healthcare services [12, 13] and linked with satisfaction and nonverbal behavior to health encounter outcomes [5]. There is a general lack of research on the role of empathy regarding clinical outcomes in primary care [12]. This study will further explore the relationship between empathy and patient satisfaction by predicting the level of satisfaction based on nonverbal interaction.

1.2 Nonverbal Interaction in Clinician-Patient Communication

Both verbal and non-verbal communication have been studied by significant research on clinician-patient communication. Most tools available to analyze physician-patient interactions are found based on verbal cues such as the process analysis system, the verbal response mode, or the Roter Interaction Analysis System (RIAS) but the role of nonverbal interaction has comparatively less focused in the literature [14]. Nonverbal

behavior, however, plays an important role in physician-patient communication and interpersonal judgment mainly depends on nonverbal cues [14]. For example, research found that distancing behaviors of physical therapists such as the absence of smiling and lack of eye contact were associated with a decrease in physical and cognitive functioning of the patients [15]. Another research has shown that patient satisfaction was related to physician expressiveness: less time for medical chart reading, more gazing, more forward lean, more nodding, more gestures and closer interpersonal distance [14, 16]. Montague et al. [5] revealed that there was a positive correlation between the length of the visit and eye contact and the patient's assessment of clinician empathy. In their study [5], apart from eye gaze, touch was found to be also important for patient satisfaction. Social touch such as handshake and hug (defined as a touch with specific social meaning as opposed to task touch defined as a touch with clinical purpose) were also linked to the perception of clinician empathy. In our proposed study, we will not only quantify the importance of different nonverbal cues but also learn how to make predictions of patient satisfactions based on interaction data annotated from videos.

1.3 Patient Demographics and Communication

Interpreting the meaning of specific nonverbal cues can be also affected by patient demographics. A study shows that many factors affect whether and how a specific nonverbal interaction is associated with patient satisfaction [14] such as gender [6], age, race, literacy, and optimism [7]. For example, older, non-White, optimistic, and literacy deficient patients had greater satisfaction among the low-income populations [7]. Although there is much research about the effects of demographics on patient satisfaction, less research has studied both demographic and nonverbal effects. This study will incorporate the demographics of the patients and explore the interaction effects between the demographics and nonverbal cues.

1.4 Machine Learning Techniques Related to Patient Satisfaction and Healthcare Industry

Several machine learning techniques have been used to understand patient satisfaction in the context of the health industry. Li et al. [17] identified clinical risk factors such as self-evaluation of health, education level, race, treatment and new medication prescription associated with patient satisfaction with the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm, which analyzed binary variables and identified risk factors for various aspects of a hospital through correlations. Galatas et al. [18] applied forward selection and Naïve Bayes to predict patient satisfaction.

2 Methods

The study was motivated by the findings of Montague et al. [5, 19] and served to give direction for the development of the design of an interactive feedback system based on user needs and guidelines for information as well as communication technologies in

clinical practice. The study aims to identify the important factors and quantify the relationships among nonverbal interactions, demographics of the patient and patient satisfaction using decision tree algorithm by answering the following questions:

1. Which factors in the demographic features and nonverbal interaction features have more deterministic contributions to patient satisfaction?
2. What are the quantitative relationships between important features and patient satisfaction?

Our methodology is illustrated in Fig. 1.

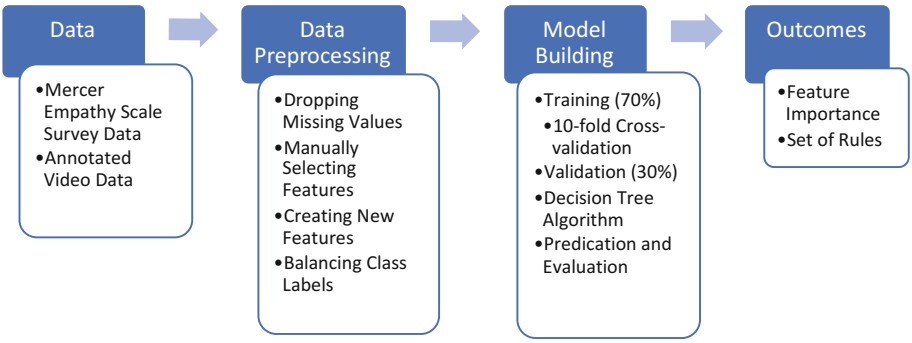

Fig. 1. Methodology diagram

2.1 Data Sources

The data set was a subset of the data collected for a study of assessing placebo, Echinacea, and doctor-patient interaction in the acute upper respiratory infection (common cold). It contained the annotated data from videotaped clinical encounters of patients with the common cold. The methodology of the study design was published in the research of Barrett et al. [20] previously. The clinical encounters took place in two different locations in Dane County, Wisconsin between April 2004 and February 2006. The protocols were approved by the University of Wisconsin School of Medicine and Public Health and clinical review boards. Patient rights and privacy were protected by following the Health Insurance Portability and Accountability Act (HIPPA) strictly.

There were 719 patients and 6 clinicians involved in this dataset. Participants were randomly assigned to three groups of different interaction mode: standard interaction, enhanced interaction and no clinical encounter [19, 20]. Data from 110 of the videotaped encounters were included in the study. The videos were of high quality and reliability of nonverbal interaction evaluation.

The detailed procedure of the annotation was published in the research of Montague et al. [5]. The non-verbal behaviors was classified using a coding scheme developed by Montague et al. [21]. The start and the stop time for each behavior were coded with Noldus Observer XT 9.0. The duration of each behavior over the course of the encounter was recorded. A coding procedure was developed by researchers [5] to ensure the reliability of the coders. The coders coded the behaviors of patients and

clinicians separately with video reduced to half-normal speed. Each coder had training videos to practice and to be evaluated on. During the final coding, the same video was assigned to all the coders each week to check the agreements by reliability tests using Cohen's Kappa coefficient. The average reliability among high-quality videos was 0.76 which was considered excellent reliability based on the study by Bakeman [5, 22].

The survey data was obtained by the survey instruments completed by participants immediately after the consultation. Questionnaires measured the perception of the patients on the clinician empathy using the Consultation and Relational Empathy (CARE) Measure (Table 1) which was a patient-rated measure of the clinician's the communication skills, the reliability of which has been validated [5, 10].

Table 1. Survey questions in Mercer Empathy Scale (The CARE Measure)

"How was the clinician at…?" with Options (Poor; Fair; Good; Very Good; Excellent)
1.Making you feel at ease…… *(being friendly and warm towards you, treating you with respect; not cold or abrupt)*
2. Letting you tell your "story" …… *(giving you time to fully describe your illness in your own words; not interrupting or diverting you)*
3. Really listening…… *(paying close attention to what you were saying; not looking at the notes or computer as you were talking)*
4. Being interested in you as a whole person…… *(asking/knowing relevant details about your life, your situation; not treating you as "just a number")*
5. Fully understanding your concerns…… *(communicating that he/she had accurately understood your concerns; not overlooking or dismissing anything)*
6. Showing care and compassion…… *(seeming genuinely concerned, connecting with you on a human level; not being indifferent or "detached")*
7. Being positive…… *(having a positive approach and a positive attitude; being honest but not negative about your problems)*
8. Explaining things clearly…… *(fully answering your questions, explaining clearly, giving you adequate information; not being vague)*
9. Helping you to take control…… *(exploring with you what you can do to improve your health yourself; encouraging rather than "lecturing" you)*
10. Making a plan of action with you…… *(discussing the options, involving you in decisions as much as you want to be involved; not ignoring your views)*
Satisfaction with Options (Very Little; Not Very Much; Somewhat; Quite A Lot; Very Much)
11. How much did you like this doctor?
12. How connected did you feel to him/her?

The questionnaire had two sections. The first section was the CARE Measure with ten questions, each with options from 1 to 5 (Poor; Fair; Good; Very Good; Excellent). The second section measured patient satisfaction containing the likeness and connectedness towards the clinician options from 1 to 5 (Very Little; Not Very Much; Somewhat; Quite A Lot; Very Much). In this study, Likeness ("How much did you like this doctor?") was chosen as the class label and an indicator of patient satisfaction, which has strong correlations with all variables in the CARE survey.

2.2 Data Preprocessing

The original dataset contained 110 records and 98 features. There were 41 features annotated from videos including the duration, frequency and proportion of the duration of the visit of the non-verbal interactions. There were 13 features containing information of patients including age, gender, race, education, income and smoking history. 15 Features extracted or calculated from the Mercer Empathy Scale survey indicated the patients' perception of empathy. 29 features obtained from other surveys aimed to access placebo. 59 features were manually selected from the data set as variables of interests due to their high relevance to the research questions and fewer missing records for each feature.

The records with 59 features containing 11 missing values in education, 8 missing value in household income and 1 missing values in features extracted from survey. After dropping all the records with missing values, the cleaned dataset had 93 records with 59 features. The features included all the video data with total time, frequency and percentage, the demographic data with age, gender, education and household income, patient satisfaction data with mercer scores and all the sections in the Mercer Empathy Scale Survey. From the non-verbal interaction features, 3 features that were directly related to the research goal were manually selected for simplicity and popularity in literature: Total duration of social touch (*Social touch (total time)*), percentage of visit in mutual gaze (*% of visit in mutual gaze*), percentage of visit in gazing chart together (*% gaze chart together*). There were 2 new calculated interaction features added to test if the length of each interaction might relate to patient satisfaction: Time per Mutual Gaze calculated by dividing the total time of mutual gaze by the frequency and Time per Social Touch. Both features were not analyzed by previous literature which provided a new angle of understanding the non-verbal interaction in the clinical encounters. To incorporate demographics of the patients, Age, Gender(*1 = Male; 2 = Female),* Education (*1: High School or High School Grad/GED; 2: Some college/tech school; 3: College Grad (bachelor's)*), Household Income (*1 =< $15 K; 2 = $15–25 K; 3 = $25–50 K; 4 = $50–75 K; 5 = $75–100 K; 6 = Over $100 K*) were selected as relevant features mentioned in the literatures from the data set. The original levels of Education were re-categorized for balanced distributions for different levels. There were in total 9 features related to nonverbal interactions and patient demographics analyzed by the model.

The possible values to the Likeness class variable ("How much did you like this doctor?") had a scale from 1 to 5 indicating different level of likeness to the clinician shown in Fig. 2.

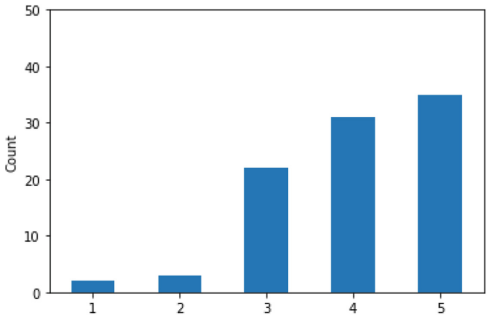

Fig. 2. Bar chart of likeness before discretization

The class distribution of Likeness was very imbalanced for level 1 ("Very little") and level 2 ("Not very much"). An empirical study has shown that classifier tended to have worse performance on the minority class and balanced class distribution provided better results with fixed amount of the data in the training set [23]. Due to the fact that the size of the data set was limited, the first three levels were binned together as a group, which provided a more balanced class distribution.

2.3 Decision Tree Algorithm

To build the classification model, decision trees approach was used considering its simplicity of result understanding, interpreting and validating. It also has no assumptions on the distribution of the data. More importantly, it helps identify the most significant attributes with the highest differential influence for prediction. The important features in the model provide insights for clinicians and inspirations for the future design of interactive technology between clinicians and patients. A decision tree has a flowchart structure, where each non-terminal node serves as a test for an attribute, each branch denotes test results, and each terminal node represents a class label [24]. Decision trees are constructed in a top-down recursive manner. It starts with the topmost node which is the root node and ends with the terminal nodes which hold the class prediction [25].

There are two commonly used impurity measures used to optimize the model performance: Information Gain and Gini Index.

ID3 decision tree algorithm uses information gain. Assume node N represents the instances of partition D. The attribute with the least entropy is chosen as the splitting attribute for node N. The information needed to classify an instance in D is calculated by the formula (1) *where* p_i = nonzero probability that tuple in D in class C_i, m = m classes:

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{1}$$

Suppose the tuples are tested to be partitioned by attribute A with v distinct values $\{a_1, a_2, \ldots, a_v\}$. D is split into v partitions, $\{D_1, D_2, \ldots, D_v\}$, where D_j contains tuples

in D with the value of a_j of A attribute. $Info(D)$ denotes average amount of information needed to identify tuple class label in D, $\frac{|D_j|}{|D|}$ denotes the weight of jth partition, $Info_A(D)$ denotes expected information required to classify a tuple from D based on the partitioning by A.

$$Info_A(D) = \sum_{i=1}^{v} \frac{|D_j|}{|D|} \times Info(D_i) \tag{2}$$

Information gain is then calculated by obtaining the difference between the information required for partition D and the new requirement obtained after partitioning on certain attribute [24].

$$Gain(A) = Info(D) - Info_A(D) \tag{3}$$

The Gini index is used in CART. It measures the impurity of partition D $wherep_i$ = nonzero probability that tuple in D in class C_i, m = m classes:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{4}$$

With the notation previously described, the Gini index determines the best binary split on attribute A with v distinct values. All the possible subsets $(2^v - 2)$ can be formed and considered using values of A. For example, if for one binary split, D is partitioned into D_1 and D_2, the Gini index of D:

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{5}$$

The decrease in impurity by a binary split:

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \tag{6}$$

The splitting attribute which maximizes the reduction in impurity will be selected [24].

Pre-pruning approach prunes the tree by halting the growing process early. The growth of the tree can be halted by choosing a maximum depth and setting values of minimum number of nodes to split and minimum number of nodes in the leaf after splitting. With all the parameters, the tree will not further split at a given node [24].

The feature importance is calculated as the normalized total reduction of the node impurity brought by that feature, known as Gini importance. Assuming only two child nodes, the importance of node j is calculated by formula (7) where ni_j = the importance of node j, w_j = weighted number of samples reaching node j, C_j = the impurity value of node j, $w_{left(j)}$ = weighted number of samples of child node from left split on node j, $w_{right(j)}$ = weighted number of samples of child node from right split on node j,

$C_{left(j)}$ = the impurity value of child node from left split on node j, $C_{right(j)}$ = the impurity value of child node from right split on node j:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{7}$$

The importance for each feature on decision tree is then calculated as the formula where fi_i = the importance of feature I, ni_j = the importance of node j:

$$fi_i = \frac{\sum_{j:\, node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in K(all\ nodes)} ni_j} \tag{8}$$

2.4 Parameter Tuning and Model Validation

The classifier model built in the experiment and trials was evaluated based on the accuracy performance of the validation set. To obtain reliable results, a stratified sampling validation was used with 70% training and 30% validation for all experiments. The validation set was selected using stratified sampling regarding the distribution of the class variable. The best parameters were determined by grid search techniques. The validation set would be held out for final evaluation. The training set was split into 10 smaller sets. The model was trained using 9 of the folds as training data. The resulting model is validated on the remaining data. The performance metrics reported by 10-fold cross-validation were the average of the accuracy computed on each fold. The best parameters chosen will then build the model tested by the validation set. The accuracy of the training and the validation set will be compared to avoid overfitting. The hyperparameters of the Decision Tree model were tuned by trying all the combination of the parameters in a given range. Parameter tuning using grid search provided the best performance by searching for the best combinations of parameters. The score function was leveraged to determine model performance. Accuracy was used for measuring classification performance.

3 Results

3.1 Sample Characteristics

There were 93 records in the dataset. The total visit time of the patients was 189.14 s on average and ranged from 26.29 s to 642.45 s. The duration of the encounter was rather short. During such a short period of time, the physicians might not be able to provide sufficient nonverbal interactions. As shown in Table 2, both distributions of the total time of social touch and time per social touch in the dataset were right-skewed. % of visit in mutual gaze was on average around 25% of the visit length and approximately 20% of the visit time were spent for gazing chart together. The duration of a mutual gaze was 3.32 s but can range from 0.54 s to 21.64 s. As shown in Table 3, patients were from different age groups with the minimum age of 14.21 to maximum age of 71.76. There were more female patients (60) than male patients (33) in the dataset. 64 of the 93 patients graduated from college. A majority of the patient had household income larger than 25 K.

Table 2. Descriptive statistics for numeric data

Feature	Social touch (total time) (s)	Time per social touch (s)	% of mutual gaze	% gaze chart together	Time per mutual gaze (s)	Age
mean	1.12	0.57	25.34	20.03	3.32	35.63
std	1.65	0.81	17.27	16.53	3.13	14.65
min	0	0.00	1.59	0.00	0.54	14.21
25%	0	0.00	10.61	1.62	1.38	23.27
50%	0	0.00	25.07	21.84	2.40	33.60
75%	1.74	1.09	40.15	30.85	3.92	46.66
max	7.55	5.34	73.04	61.08	21.64	71.76

Table 3. Descriptive statistics for nominal/ordinal data

Feature	Gender		Education			Household income						Likeness		
Levels	**M**	**F**	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**
Counts	33	60	17	22	64	6	10	21	19	22	15	27	31	35

As the correlation between the features in the CARE questionnaires were all moderate to strong and significant at 0.01 level (2-tailed test) evaluated by Spearman correlation for ordinal data shown in Table 4, the study chose Likeness ("How much did you like this doctor?") as the class label as an indicator of patient satisfaction which has strong correlation (0.854, $p <= 0.01$) with Connectedness ("How connected did you feel to this doctor?") and also moderate to strong correlation (0.60–0.86) with other 10 features measuring patient satisfaction.

Table 4. Correlation matrix for Mercer empathy scale variables (each feature indicated by question numbers in the survey; results of Likeness (11) is in bold type; F. means features)

F.	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	0.71	0.79	0.76	0.67	0.87	0.73	0.73	0.69	0.70	**0.78**	0.65
2	0.71	1.00	0.81	0.66	0.72	0.74	0.67	0.71	0.66	0.58	**0.60**	0.53
3	0.79	0.81	1.00	0.78	0.73	0.82	0.75	0.75	0.71	0.72	**0.71**	0.64
4	0.76	0.66	0.78	1.00	0.78	0.86	0.81	0.71	0.81	0.77	**0.86**	0.79
5	0.67	0.72	0.73	0.78	1.00	0.84	0.76	0.79	0.73	0.75	**0.73**	0.69
6	0.87	0.74	0.82	0.86	0.84	1.00	0.86	0.85	0.83	0.78	**0.85**	0.72
7	0.73	0.67	0.75	0.81	0.76	0.86	1.00	0.84	0.79	0.73	**0.76**	0.70
8	0.73	0.71	0.75	0.71	0.79	0.85	0.84	1.00	0.73	0.75	**0.71**	0.61
9	0.69	0.66	0.71	0.81	0.73	0.83	0.79	0.73	1.00	0.81	**0.76**	0.69
10	0.70	0.58	0.72	0.77	0.75	0.78	0.73	0.75	0.81	1.00	**0.76**	0.72
11	**0.78**	**0.60**	**0.71**	**0.86**	**0.73**	**0.85**	**0.76**	**0.71**	**0.76**	**0.76**	**1.00**	**0.85**
12	0.65	0.53	0.64	0.79	0.69	0.72	0.70	0.61	0.69	0.72	**0.85**	1.00

3.2 Decision Tree Model Building

The dataset was divided into training and validation subsets using stratified sampling. 30 random repeated independent trials were conducted to obtain the mean accuracy and confidence interval for decision tree models with different hyperparameters. The hyperparameters consisted of different splitting criteria (Gini index or Entropy), minimum number of samples needed in a parent node (ranging from 4 to 20), minimum number of samples needed in a child node (ranging from 2 to 10), and maximum depth (ranging from 3 to 9). The model with validation accuracy closest to the average was chosen to provide stable results. The model with chosen parameters was tested and evaluated in training and validation set with 30 trials. After building the models, the importance of each feature was recorded and compared. The top four features were then chosen based on the rank of feature importance to build simpler models with high performance.

As shown in Table 5, the average accuracy of the classifier with the corresponding width of 95% confidence interval, generated by 30 trials, was $59.74\% \pm 1.78\%$. The mean accuracy gained from validation set was $52.74\% \pm 3.24\%$.

Table 5. Summary table for decision tree results for with different features and hyperparameters (mean accuracy and 95% confidence interval constructed by 30 trials)

Mean accuracy	All features	Four most important features
Training (10-Fold Cross-validation)	$59.74\% \pm 1.78\%$	$61.29\% \pm 2.27\%$
Validation	$52.74\% \pm 3.24\%$	$54.29\% \pm 4.69\%$
Final model	Splitting criteria: Entropy; Maximum tree depth:3; Minimum number of samples needed in a child node: 5; Minimum number of samples needed in a parent node:13	Splitting criteria: Gini Index; Maximum tree depth: 4; Minimum number of samples needed in a child node: 3; Minimum number of samples needed in a parent node:13
Training	$53.03\% \pm 2.28\%$	$52.30\% \pm 2.12\%$
Validation	$52.02\% \pm 3.01\%$	$51.67\% \pm 2.77\%$

The performance of the final model with all features and chosen parameters shown in the row of final model and the second column in Table 5 was $53.03\% \pm 2.28\%$ in training and $52.02\% \pm 3.01\%$ in validation set. The difference between the mean accuracies of training and validation was 1.01% which did not indicate an overfitting problem.

The average feature importance was calculated among the 30 trials from decision trees with the chosen parameters. There were three features had the average importance much higher than other features: Time per Mutual Gaze, % of Visit in Mutual Gaze and Age. Social Touch (Total Time) ranked the fourth most importance features (Fig. 3).

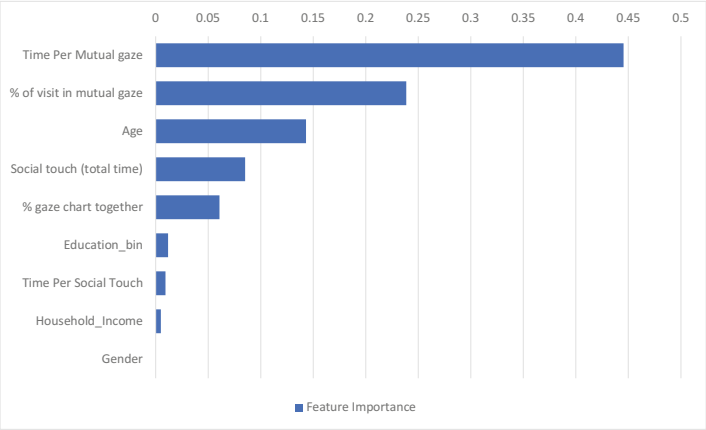

Fig. 3. Bar chart of the average feature importance

The rank was learned by the models and was consistent across random training and validation sets. To have a more efficient prediction, the top four features were selected to construct simpler trees with a similar process.

To build a simpler model with only 4 features, the dataset was divided into training and testing subsets using stratified sampling and 30 random repeated independent trials were conducted to obtain the mean and confidence interval for the performance metrics. The average accuracy of the classifiers with the corresponding width of 95% confidence intervals across the 30 trials was $61.29\% \pm 2.27\%$. The accuracy gained from the validation set was $54.29\% \pm 4.69\%$. The parameters with its validation accuracy closest to the mean testing accuracy was chosen for final model shown in the third column and the row of final model. The mean accuracy generated by the cross-validation among all different sampling of given optimal parameters was $52.30\% \pm 2.12\%$. The accuracy gained from validation set was $51.67\% \pm 2.77\%$. The difference between the accuracies of training and validation was 0.63% which did not indicate the overfitting problem. The results above have shown that the performance of four features was comparable to the performance of all features. The model with only four features helped learn simple and interpretable rules.

3.3 Rule Extraction

While the above analysis provided feature importance, further analysis also was conducted with the intent to understand how these features were combined to produce classification rules. To accomplish the rule extraction, a sample tree was chosen with validation accuracy close to the mean validation accuracy of the 30 trials, without overfitting issues and contained all four features. The training accuracy for this sample tree was 58.75% and the validation accuracy was 57.14%.

Table 6 concluded other important performance metrics for the model on the validation set. Precision indicated the percentage of correct predictions of the level of likeness among all positive predictions. Recall indicated the percentage of correct

predictions of the level of likeness among all actual positive cases. F1-score was calculated and indicated the weighted average of Precision and Recall, which took both false positives and false negatives into account. The model performed better in classifying the lower and upper level of the class variable. The model performed better for the class label 'Excellent' with highest precision, recall, and F1-score, which might due to the fact that there were more cases for the 'Excellent' label.

Table 6. Other performance metrics for sample tree

Likeness metrics	Precision	Recall	F1-score
1: Very Little to Somewhat	0.40	0.25	0.31
2: Quite A Lot	0.50	0.67	0.57
3: Very Much	0.73	0.73	0.73

Rules extracted by the model shown in Fig. 4 from the training set were summarized in Table 7. All class labels had rules with the class probability larger than 80%. Both class label "Quite A Lot" and "Very Much" had rules which correctly predicted all the cases at the terminal nodes.

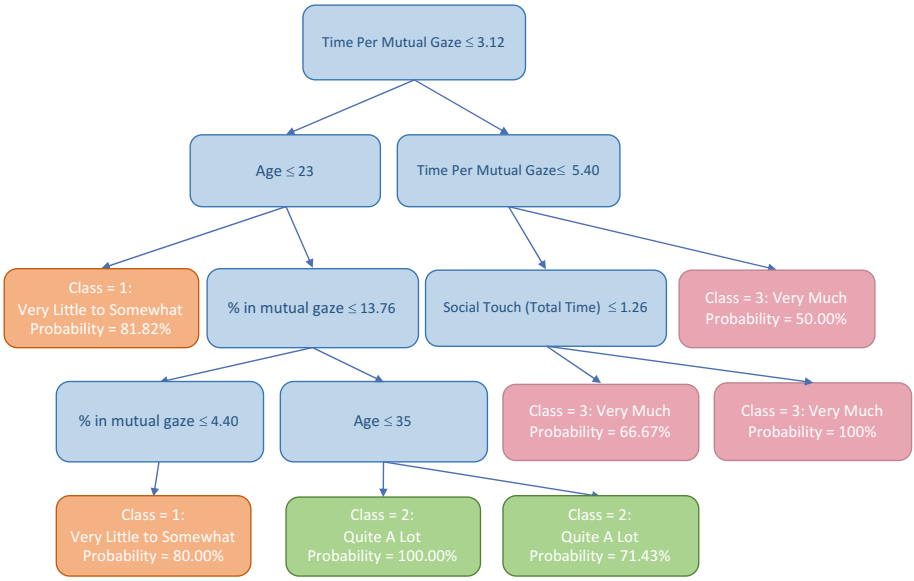

Fig. 4. Sample decision tree graph (different colors for different class levels, left arrow: yes, right arrow: no) (Color figure online)

Table 7. Rules extracted by pruned sample tree

Rules	Class: "How much did you like this doctor?"	Class probability
Time Per Mutual Gaze <= 3.12 Age <= 23	**1: Very Little to Somewhat**	81.82%
Time Per Mutual Gaze <= 3.12 Age > 23% of Mutual Gaze <= 4.40	**1: Very Little to Somewhat**	80.00%
Time Per Mutual Gaze <= 3.12 23 < Age <= 35% of Mutual Gaze > 13.76	**2: Quite A Lot**	100%
Time Per Mutual Gaze <= 3.12 Age > 35% of Mutual Gaze > 13.76	**2: Quite A Lot**	71.43%
3.12 < Time Per Mutual Gaze <= 5.40 Social Touch (Total Time) > 1.26	**3: Very Much**	100%
3.12 < Time Per Mutual Gaze <= 5.40 Social Touch (Total Time) <= 1.26	**3: Very Much**	66.67%
Time Per Mutual Gaze >= 5.40	**3: Very Much**	50%

4 Discussion

The results showed that the duration for each mutual gaze, percentage of eye contact, age and social touch were important features associated with patient satisfaction. It was consistent with previous research of Montague [5] which identified the percentage of eye contact and social touch as important indicators. Many studies had shown that eye contact had a robust effect on communication and satisfaction. However, considerably less attention had been devoted to investigating the relationship between the length of each mutual gaze and patient satisfaction in medical encounters. The importance of the length of each mutual gaze was analyzed by few studies in the medical encounters but was considered as a crucial factor in other fields. A study [26] suggested that longer durations of eye contact with fewer eye shifts were more likely to have a perception of higher intelligence.

The performance of the model with selected features was comparable with models with all features. The rules extracted by the sample tree provided some insights on the quantitate relationship for those non-verbal cues and demographic features with patient perceptions. The extracted rules reveal that patients of older age (>23), with a higher percentage of mutual gaze (>13.76%), longer social touch duration (>1.26 s) and moderate duration per mutual gaze (larger than 3.12 s but smaller than 5.40 s) tended to report greater rating on likeness towards their clinicians. To suggest simple and reliable rules, the rule for each class label providing the highest purity was analyzed in detail. When time per mutual gaze was smaller than 3.12 and age was smaller than 23,

the patient tended to have a low rating for their likeness towards the clinician. When time per mutual gaze was smaller than 3.12, age was larger than 23 and smaller than 35 and percentage of mutual gaze was larger than 13.76 the patient would rate "Quite a Lot". When time per mutual gaze was larger than 3.12 but smaller than 5.40 with time for social touch larger than 1.26, the patient would rate "Very much" as the answer to the likeness. Thus, patients of older age tended to report greater likeness which was consistent with the findings of research on demographic influence on patient-provider communication and satisfaction [7]. A higher percentage of mutual gaze would lead to higher satisfaction towards clinician which was demonstrated in the study of Montague which identified positive beta for eye contact on empathy scores (beta = 0.43, $R^2 = 0.18$) [5]. Longer social touch duration tended to result in higher likeness ratings. It confirmed the results of a study on touch in primary care consultations that patients were sensitive to the nonverbal communications and social touch improved the interactions, especially in situations of severe distress [27]. The rules also revealed that insufficient duration per mutual gaze resulted in low rating in the likeness and moderate duration led to a higher rating. The result of a study on preferred mutual gaze duration was consistent with this rule that longer gazes were preferred to frequent and short eye contact, but gazing for too long or overly short glance can be discomforting [28].

The survey data was self-reported and the video data to extract the non-verbal interaction was annotated by watch videos. Both might provide noisy and inconsistent data. Also, video-recording might influence the behavior of patients and clinicians. The external validity of the research might also be limited due to the fact that the sample was collected from a certain community associated with a certain symptom (common cold), which might result different conclusions for different locations and contexts. There were many exogenous effects might affect patient satisfaction. The study mainly focused on the demographics of the patients and did not have the features for clinicians. However, the gender of the physician can also affect the interpretation of the nonverbal communication of the patients [29].

5 Conclusion

This study identified four most important features in the dataset for patient perceptions of their clinicians: the duration for each mutual gaze, percentage of eye contact, age and social touch. Decision trees approach provided simple and interpretable results. Patients of older age, with a higher percentage of mutual gaze, longer social touch duration and moderate duration per mutual gaze were more likely to report greater likeness towards their clinicians. Although the decision tree algorithm provides great performance with only four most relevant characteristics in the physician-patient interaction, further study with more features and analysis of more data can help increase the performance and support decision making in health settings with more accurate results learning from the entire process.

With the quantitative rules extracted by a sample decision tree model, this study provided insights for the future design of the clinical environment and health information technology. These findings suggest that sufficient training of non-verbal communication skills for physicians can help improve patient satisfaction and outcome.

To help busy primary care clinicians in a highly interruptive and time-pressured environment [30], the future design of the clinical environment and health information technology should facilitate and remind clinicians of satisfactory non-verbal interactions. An automatic feedback system can be built to provide feedback on clinical visits based on real-time analysis of the important features identified by the model. The feedback system customized for primary care visits and with defined functionality of improving the non-verbal communication will provide high effectivity and efficiency [31]. Dynamic feedbacks will not only serve as a reminder but also reinforce important factors in decision making [32].

Acknowledgments. This research was supported by NSF Division of Information & Intelligent Systems Award - "CHS: Small: Extracting affect and interaction information from primary care visits to support patient-provider interactions" (Grant No: 1816010).

References

1. King, A., Hoppe, R.B.: "Best practice" for patient-centered communication: a narrative review. J. Grad. Med. Educ. **5**, 385–393 (2013). https://doi.org/10.4300/JGME-D-13-00072.1
2. Al-Abri, R., Al-Balushi, A.: Patient satisfaction survey as a tool towards quality improvement. Oman Med. J. **29**, 3–7 (2014). https://doi.org/10.5001/omj.2014.02
3. Cleary, P.D., McNeil, B.J.: Patient satisfaction as an indicator of quality care. Inquiry **25**, 25–36 (1988)
4. LaVela, S.L., Gallan, A.S.: Evaluation and measurement of patient experience. Patient Exp. J. **1**(1), 28–36 (2014)
5. Montague, E., Chen, P., Xu, J., Chewning, B., Barrett, B.: Nonverbal interpersonal interactions in clinical encounters and patient perceptions of empathy. J. Participat. Med. **5**, e33 (2013)
6. Hall, J.A., Irish, J.T., Roter, D.L., Ehrlich, C.M., Miller, L.H.: Gender in medical encounters: an analysis of physician and patient communication in a primary care setting. Health Psychol. **13**, 384–392 (1994)
7. Jensen, J.D., King, A.J., Guntzviller, L.M., Davis, L.A.: Patient-provider communication and low-income adults: age, race, literacy, and optimism predict communication satisfaction. Patient Educ. Couns. **79**, 30–35 (2010). https://doi.org/10.1016/j.pec.2009.09.041
8. Berman, A.C., Chutka, D.S.: Assessing effective physician-patient communication skills: "Are you listening to me, doc?". Korean J. Med. Educ. **28**, 243–249 (2016). https://doi.org/10.3946/kjme.2016.21
9. Collins, L.G., Schrimmer, A., Diamond, J., Burke, J.: Evaluating verbal and non-verbal communication skills, in an ethnogeriatric OSCE. Patient Educ. Couns. **83**, 158–162 (2011). https://doi.org/10.1016/j.pec.2010.05.012
10. Bikker, A.P., Fitzpatrick, B., Murphy, D., Forster, L., Mercer, S.W.: Assessing the Consultation and Relational Empathy (CARE) Measure in sexual health nurses' consultations. BMC Nurs. **16**, 71 (2017). https://doi.org/10.1186/s12912-017-0265-8
11. Derksen, F., Bensing, J., Lagro-Janssen, A.: Effectiveness of empathy in general practice: a systematic review. Br. J. Gen. Pract. **63**, e76–e84 (2013). https://doi.org/10.3399/bjgp13X660814
12. Mercer, S.W., Reynolds, W.J.: Empathy and quality of care. Br. J. Gen. Pract. **52**(Suppl), S9–12 (2002)

13. Jolliffe, D., Farrington, D.P.: Development and validation of the Basic Empathy Scale. J Adolesc. **29**, 589–611 (2006). https://doi.org/10.1016/j.adolescence.2005.08.010
14. Mast, M.S.: On the importance of nonverbal communication in the physician–patient interaction. Patient Educ. Couns. **67**, 315–318 (2007). https://doi.org/10.1016/j.pec.2007.03.005
15. Ambady, N., Koo, J., Rosenthal, R., Winograd, C.H.: Physical therapists' nonverbal communication predicts geriatric patients' health outcomes. Psychol. Aging **17**, 443–452 (2002). https://doi.org/10.1037/0882-7974.17.3.443
16. Hall, J.A., Harrigan, J.A., Rosenthal, R.: Nonverbal behavior in clinician—patient interaction. Appl. Prev. Psychol. **4**, 21–37 (1995). https://doi.org/10.1016/S0962-1849(05)80049-6
17. Li, L., Lee, N.J., Glicksberg, B.S., Radbill, B.D., Dudley, J.T.: Data-driven identification of risk factors of patient satisfaction at a large urban academic medical center. PLoS ONE **11**, e0156076 (2016). https://doi.org/10.1371/journal.pone.0156076
18. Galatas, G., Zikos, D., Makedon, F.: Application of data mining techniques to determine patient satisfaction. In: Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments - PETRA 2013, pp. 1–4. ACM Press, Rhodes (2013)
19. Montague, E.: An intervention study of clinician-patient nonverbal interactions and patient perceptions of visits. J. Healthc. Commun. **3** (2018). https://doi.org/10.4172/2472-1654.100120
20. Barrett, B., et al.: Rationale and methods for a trial assessing placebo, echinacea, and doctor-patient interaction in the common cold. Explore (NY). **3**, 561–572 (2007). https://doi.org/10.1016/j.explore.2007.08.001
21. Montague, E., Xu, J., Chen, P.-Y., Asan, O., Barrett, B.P., Chewning, B.: Modeling eye gaze patterns in clinician-patient interaction with lag sequential analysis. Hum. Factors **53**, 502–516 (2011). https://doi.org/10.1177/0018720811405986
22. Bakeman, R.: Behavioral observation and coding. In: Handbook of Research Methods in Social and Personality Psychology, pp. 138–159. Cambridge University Press, New York (2000)
23. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inf. Sci. **250**, 113–141 (2013). https://doi.org/10.1016/j.ins.2013.07.007
24. Han, J., Kamber, M., Pei, J.: 9 - Classification: advanced methods. In: Han, J., Kamber, M., Pei, J. (eds.) Data Mining (Third Edition), pp. 393–442. Morgan Kaufmann, Boston (2012)
25. Patel, B.N., Prajapati, S.G., Lakhtaria, K.I.: Efficient Classification of Data Using Decision Tree. Presented at the (2012)
26. Wheeler, R.W., Baron, J.C., Michell, S., Ginsburg, H.J.: Eye contact and the perception of intelligence. Bull. Psychon. Soc. **13**, 101–102 (1979). https://doi.org/10.3758/BF03335025
27. Cocksedge, S., George, B., Renwick, S., Chew-Graham, C.A.: Touch in primary care consultations: qualitative investigation of doctors' and patients' perceptions. Br. J. Gen. Pract. **63**, e283–e290 (2013). https://doi.org/10.3399/bjgp13X665251
28. Nicola, B., Charlotte, H., Antoine, C., Alan, J., Isabelle, M.: Pupil dilation as an index of preferred mutual gaze duration. Roy. Soc. Open Sci. **3**, 160086 (2016). https://doi.org/10.1098/rsos.160086
29. Mast, M.S., Hall, J.A., Köckner, C., Choi, E.: Physician gender affects how physician nonverbal behavior is related to patient satisfaction. Med. Care **46**, 1212–1218 (2008). https://doi.org/10.1097/MLR.0b013e31817e1877

30. Kim, M.S., Clarke, M.A., Belden, J.L., Hinton, E.: Usability challenges and barriers in EHR training of primary care resident physicians. In: Duffy, V.G. (ed.) DHM 2014. LNCS, vol. 8529, pp. 385–391. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07725-3_39
31. Bundschuh, B.B., et al.: Quality of human-computer interaction - results of a national usability survey of hospital-IT in Germany. BMC Med. Inf. Decis. Mak. **11**, 69 (2011). https://doi.org/10.1186/1472-6947-11-69
32. Hartwig, M., Windel, A.: Safety and health at work through persuasive assistance systems. In: Duffy, V.G. (ed.) DHM 2013. LNCS, vol. 8026, pp. 40–49. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39182-8_5