# Towards multimodal modeling of physicians' diagnostic confidence and self-awareness using medical narratives

**Joseph Bullard**[†] **Cecilia Ovesdotter Alm**[‡]
**Qi Yu**[†] **Pengcheng Shi**[†] **Anne Haake**[†]
[†]College of Computing and Information Sciences
[‡]College of Liberal Arts
Rochester Institute of Technology
jtb4478@cs.rit.edu
coagla|qi.yu|spcast|arhics@rit.edu

## Abstract

Misdiagnosis is a problem in the medical field, often related to physicians' cognitive errors. Overconfidence is considered a major cause of such errors. Intelligent diagnostic support systems could benefit from understanding how aware physicians are of their performance when they estimate their confidence in a diagnosis (i.e. a physician's *diagnostic self-awareness*). Shedding light on the cognitive processes related to such awareness could also help improve medical education. We use a multimodal dataset of medical narratives to computationally model diagnostic confidence and self-awareness based on physicians' linguistic and eye movement behaviors. Dermatologists viewed images of cutaneous conditions, providing a description, diagnosis, and certainty level for each image case, while their speech and eye movements were recorded. We define both a generalized and a personalized approach to binning confidence levels, used in classification experiments. We also introduce truly multimodal features, which focus on combining linguistic and eye movement data into multimodal attributes. Results indicate that combinations of multiple modalities can outperform their constituent modalities in isolation for these problems.

## 1 Introduction

Misdiagnosis in the medical field is estimated to be as high as 10%-15% (Berner and Graber, 2008; Croskerry, 2009). Such errors can result in incorrect or delayed treatment, causing patients to experience additional suffering. Graber et al. (2002) describe three types of diagnostic errors: *no-fault* errors, resulting from atypical disease presentation or limitations of medical knowledge; *system* errors, resulting from problems with the health care system; and *cognitive* errors, resulting from biases or faulty interpretation on the part of a physician. Cognitive errors in particular have potential for substantial reduction through education and training aimed at developing clinicians' metacognitive skills. Understanding the cognitive processes of physicians during diagnosis is also of critical importance for building human-centered diagnostic support systems, which could help detect and flag problematic diagnostic self-awareness cases. Examples of cognitive errors include settling on a final diagnosis too early, without ever considering the correct diagnosis (Berner and Graber, 2008), or confirmation bias, in which only evidence to confirm a diagnostic hypothesis is considered (Croskerry, 2003). Overconfidence is generally thought to be a major cause of such errors (Berner and Graber, 2008; Croskerry, 2008). For example, an overconfident physician may not question her original thoughts or explore alternative diagnoses until later in the treatment process. In general, overconfidence may be a systemic problem, reinforced by patients' preferences for confident doctors, and by a professional environment that favors decisive actions (Katz, 1984). Similarly, underconfidence can erode patients' trust in their providers. In this study, we view the interplay between confidence[1] and correctness as a two-dimensional problem (see Figure 1). Ideally, physicians would have high confidence when correct and low confidence when incorrect, indicated by the upper-left and lower-right quadrants in Figure 1.

[1] For consistency, this paper uses the term *confidence*, treated as interchangeable with *certainty* and similar synonymous expressions which may have been used by clinicians in the medical narratives, such as *sure*, *certain*, *confident*, etc.

*Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1718–1727, Dublin, Ireland, August 23-29 2014.

| | **Correct** | **Incorrect** |
|---|---|---|
| **Confident** | Appropriate Confidence | Overconfidence |
| **Not confident** | Underconfidence | Appropriate Confidence |

Figure 1: Two-dimensional view of the confidence and correctness relationship as it relates to *diagnostic self-awareness*. A similar conceptual model is presented by Pon-Barry and Shieber (2011). Ideally, physicians should have high confidence when they are correct and low confidence when incorrect.

**Contribution** Diagnostic self-awareness is an important phenomenon with implications for clinical training and practice, yet has received little focus from a computational perspective. We report on computational modeling for predicting the confidence and correctness interplay in diagnosis using features of physicians' speech, eye movements, and combinations thereof, as dermatologists performed medical image inspection tasks while narrating their diagnostic thought process. In dermatology, visual expertise and clinical knowledge are both important. A motivation behind our multimodal approach is that medical image inspection relies on both the physician's visual perceptual expertise and conceptual knowledge base, each of which can be regarded as expressed by eye movement behavior and linguistic behavior, respectively. We aim to apply this decision modeling to intelligent diagnostic support and clinical tutoring systems. Here we solve a foundational problem by successfully modeling the complex relationship between physicians' confidence in and correctness of their diagnoses. We also make contributions in multimodal and linguistic feature analysis: carefully assessing feature modalities that represent physicians' behaviors, and introducing a novel multimodal feature type that focuses on fusing eye movement and verbal data.

## 2 Previous Work

Although there are many causes of diagnostic errors (Graber et al., 2005), those resulting from cognitive errors may be the most challenging to reduce (Croskerry, 2003; Graber et al., 2002), while their reduction provides high impact. Examples of such errors include flawed perception, biased heuristics, and settling on a final diagnosis too early (Graber et al., 2002), all of which can be caused by overconfidence (Berner and Graber, 2008; Croskerry, 2008). Underconfidence may also be a problem if it prevents a physician from pursuing a correct diagnosis (Friedman et al., 2005).

There is evidence for links between speech and confidence in terms of prosodic features, such as pitch and loudness (Scherer et al., 1973; Pon-Barry and Shieber, 2011; Kimble and Seidel, 1991), as well as other characteristics of spoken language, such as speech disfluencies (Womack et al., 2012) and hedges (Smith and Clark, 1993). Prosodic features have been identified and successfully used in intelligent tutoring systems (Liscombe et al., 2005), where a student's confidence (or lack thereof) can play a key role in effective system response. In medical diagnosis, prosodic and lexical features have been useful indicators of physicians' confidence and diagnostic correctness, individually (Womack et al., 2013; McCoy et al., 2012). Other potentially useful information may be evident in speech as well. In a study by Womack et al. (2012) on a similar dataset, the authors found a relationship between speech characteristics and physician experience: attending (experienced) physicians used more filled pauses and spoke more than resident (in-training) physicians. Additionally, verbal features may expose differences in diagnostic reasoning that may be useful predictors of confidence. Rogers (1996) analyzed a dataset of spoken chest X-ray examinations by radiologists, remarking that reasoning styles influence physicians' expectations, and confirmations or contradictions of those expectations can affect their self-reported confidence levels.

Most relevant literature focuses on linguistic features. Language, as the primary form of human expression, is certainly critical. However, analyzing meaning may require going beyond linguistic inference, depending on the context or application. Previous studies have successfully incorporated multiple expressive modalities when examining linguistic and cognitive processes, such as facial expressions for video sentiment analysis (Pérez-Rosas et al., 2013) and pointing gestures for referring actions (Gatt and Paggio, 2013). In such studies, the additional modalities were carefully chosen based on the nature of the performed tasks. Here, we deal with experts (dermatologists) inspecting images (skin conditions) for diagnostic purposes, a task that heavily involves their use of visual perceptual expertise, in addition to conceptual domain knowledge. For this reason, we incorporate features of their eye movements in our study. There is evidence for ties between perceptual expertise and eye movements during image inspection tasks (Li et al., 2012b), and we explore if such ties may also relate to a physician's confidence and diagnostic self-awareness.

Integrating different expressive modalities is challenging. Previous work involving multimodality has predominantly treated each in isolation. We further address this challenge by identifying and exploring truly *multimodal* features that focus on combining verbal and eye movement data into complex multimodal attributes, as it seems reasonable that the two modalities together could be more informative if linked, and that such complex features represent a natural interactive extension of multimodal semantics. Evidence for ties between speech and eye movements specifically was found by Li et al. (2012a), in which sequences of fixations and saccadic eye movements were identified to predominantly align with particular conceptual units of thought (e.g. primary lesion type) expressed verbally in medical narratives.

## 3 Data Description and Analysis

This study takes advantage of a dataset previously reported on by Womack et al. (2013), which is briefly described here for clarity, as Womack et al.'s work ignored the eye movement data. A group of 29 dermatologists (11 attending physicians, 18 residents) were each shown a series of 30 images of dermatological conditions in random order and asked to narrate their diagnosis of each condition. They were asked to provide a description of the case, a list of differential diagnoses to consider, a final diagnosis, and their certainty of their final diagnosis, as a percentage. The physicians' verbal descriptions were recorded as audio and later manually transcribed in detail, including pauses, disfluencies, and other speech phenomena.[2] During this process, the physicians' eye movements were also tracked. Each image was displayed on a 22" LCD monitor (1650x1050 pixels) with an attached 250Hz SensoMotoric Instruments RED remote eye-tracker while IViewX software was recording the eye movements.

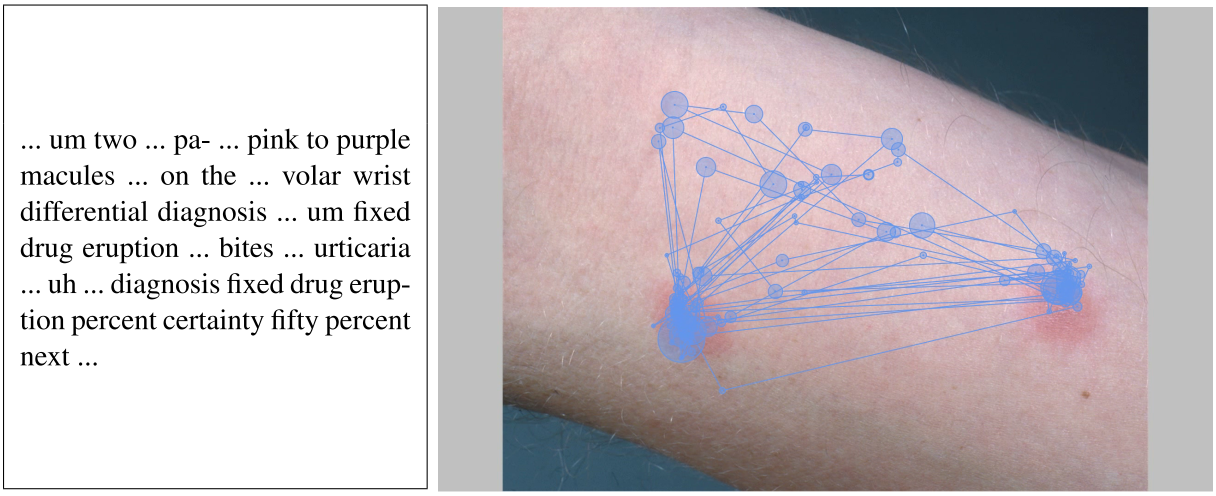
In this study, the time-aligned pair of verbal description and eye movements for one physician viewing one image is henceforth called a *narrative*. Figure 2a shows an example of a verbal description for one narrative and Figure 2b shows a visualization of the corresponding eye movements. The correct diagnoses for all images were known for the experiment and each narrative was assigned a binary label of *correct* or *incorrect*.[3] For the purposes of this multimodal study, 238 of the 870 narratives were excluded due to technical issues that had occurred with the eye tracking or audio capture equipment, or because the physicians had provided no confidence values for their diagnoses. The remaining 632 narratives were used for the analysis and experimentation reported on in this paper.

### 3.1 Case Studies towards Understanding Physicians' Confidence and Correctness

The physicians tended to evaluate their confidence towards the upper end of the spectrum, with a median of 70% confident over all narratives. But diagnostic confidence may be affected by many factors, including professional experience, case difficulty, and personality. We examine both individual images and physicians at the extremes of confidence to gain insight into the relationship between confidence and correctness in the dataset. Table 1 summarizes information for the three image cases that received the

[2] Some transcription imperfections may occur.

[3] A limited number of narratives in the dataset were labeled *half* correct if one of two final diagnoses given was correct, and *partially* correct if the final diagnosis was too broad. Here, we consider *half* to be correct, because in such cases the correct diagnosis was still identified, but *partial* to be incorrect, because the correct diagnosis was technically not identified.



... um two ... pa- ... pink to purple macules ... on the ... volar wrist differential diagnosis ... um fixed drug eruption ... bites ... urticaria ... uh ... diagnosis fixed drug eruption percent certainty fifty percent next ...

(a) Sample verbal description. Ellipses ("...") show pauses.

(b) Sample eye movement visualization. Circles represent *fixations*, where the center is the point of fixation and the radius is proportional to the time fixating at that point. Lines represent *saccades* (movements) between fixation points.

Figure 2: Sample verbal description and eye movements for one narrative. The final diagnosis is correct and the physician was 50% confident.

| Confidence | Conf. | % Correct | Rank |
|---|---|---|---|
| **Highest** | 100 | 100 | 2 |
| | 90 | 100 | 5 |
| | 90 | 100 | 1 |
| **Lowest** | 50 | 24 | 25 |
| | 50 | 35 | 29 |
| | 45 | 0 | 20 |

Table 1: Images receiving highest and lowest median confidence values. Difficulty ranking provided by a dermatology expert with 1 reflecting the easiest image and 30 the most difficult.

| Confidence | Conf. | % Correct | Exp. |
|---|---|---|---|
| **Highest** | 90 | 53 | R |
| | 85 | 50 | A |
| | 85 | 41 | R |
| **Lowest** | 38 | 39 | A |
| | 30 | 48 | R |
| | 15 | 37 | R |

Table 2: Most and least confident physicians by median confidence values given over all images. The last column shows experience level: experienced attending (A) or resident (R) physician.

highest median confidence values and the three that received the lowest. A domain expert (dermatologist and clinical educator) who was not a subject in the experiment gave each image a unique difficulty ranking from 1 to 30, where the image ranked number 1 was considered the easiest to support a correct diagnosis, and 30 the most difficult. As expected, the highest confidence images were among the easiest, and vice versa. Accordingly, the higher confidence images were correctly diagnosed by every physician, while those receiving the lowest confidence were correctly diagnosed much less often. The negative correlation between image difficulty and median physician confidence was significant using Spearman's rank correlation ($r_s = -0.544$, $p < 0.005$). In other words, higher levels of case difficulty were associated with lower levels of physician confidence. In contrast, examination of the most and least confident physicians yields less intuitive results. The physicians with the highest and lowest median confidence values are shown in the top and bottom halves of Table 2, respectively. Notably, each of the two groups contained both resident dermatologists-in-training and attending physicians with careers spanning multiple decades. Also, the most confident physicians were only correct roughly half of the time, and the least confident physicians' correctness appears quite similar. While this may reflect the sample size, the observation is interesting nonetheless. Clearly, this points to how complicated diagnostic self-awareness is, and how potentially useful it would be to computationally infer a physician's self-awareness for diagnostic cases based on their behaviors.

### 3.2 Confidence Binning

Nearly all confidence values given were multiples of five, or simply numbers close to 100, such as 99%.[4] This makes discretization preferable to using real-numbered values for confidence. Additionally, the analyses in Section 3.1 revealed patterns of over- or underconfidence in individual physicians. What this indicates is that "high" and "low" confidence involve different numerical values in the minds of different physicians. This subjectivity could be problematic in doctor-patient interactions and it adds complexity for predictive modeling involving confidence. To explore the impact, we devise two alternative binary binning schemes: *generalized* bins, based on the performance of all physicians in the dataset, and *personalized* bins, based on each individual physician's performance in the training data only. In terms of application, consider a diagnostic support system which could establish a history for each physician who uses it. Such a system could implement a generalized binning scheme and predictive model for new users, and later, after learning from repeated exposure to a given physician, switch to a model based on that physician's individual performance. In addition, binning choice may be influenced by context: in a clinical tutoring system, it may be preferable to compare learners to experienced physicians as a target population. For the *generalized* binning scheme, a confidence value greater than or equal to the median over all physicians is considered *high*, while a value below is considered *low*. This results in a slight imbalance towards *high* confidence (56% of narratives).[5] We construct the *personalized* binning scheme similarly, but using a given physician's own median confidence in the training data as the dividing line. In this case, *high* confidence accounts for 58% of the narratives, similar to that of the generalized bins. Calling a physician's median confidence *high* lets us better distinguish the problem cases: cases of underconfidence should be strictly less than their "typical" confidence, while cases of overconfidence should be at or above typical. The binning scheme used does not affect the correctness value for each narrative, but it does change the distribution of high and low confidence, with the *generalized* scheme favoring over- and underconfidence, and the *personalized* scheme favoring appropriate confidence. Arguably, the latter is a better reflection of the expected: over- and underconfidence as the minority classes.

## 4 Approach and Methodology

There are many ways to approach the problem of predicting physicians' diagnostic self-awareness. Here we formulate two classification problems, each tested under both binning schemes, yielding a total of four classification models. We also outline the performance evaluation experiments for the models.

### 4.1 Classification Problems

We define two classification problems based on the chart in Figure 1 (above). First, we define *Confidence Only*, which ignores correctness (the horizontal dimension of Figure 1) and predicts only confidence as a binary *high* or *low*. Intuitively, low confidence might be considered a warning sign for a diagnosis, alerting a physician to seek additional insight or information.[6] This first problem was used as a stepping stone to explore and better understand confidence, before incorporating correctness. Next, we define *Confidence & Correctness*, which relates confidence with the correctness of the diagnosis (considering all four quadrants in Figure 1, individually) to better address the more problematic, but interesting, cases. Distinguishing these four classes could be of use to intelligent tutoring or clinical support systems, which could respond differently to over- or underconfident users. In general, the full separation of these classes could ultimately allow for deeper analysis of physician self-awareness.

### 4.2 Model Evaluation

Before any development took place, the 632 narratives were randomly divided into three subsets: 442 (70%) for training (*dev-train*), 95 (15%) for testing during development and tuning (*dev-test*), and 95

[4] There were only a few exceptions: one physician gave three values of 3%, another gave a 33% and a 66% (rounded down from "two-thirds"), and a third gave a 33%. The latter three cases could also seem intuitive depending on how many conditions were listed in the differential diagnosis. For example, 66% might indicate that one disease seemed twice as likely as a another.

[5] Other simple binning schemes dividing up the 0-100% range were explored, but this binary version allowed for a more systematic approach to both generalized and personalized binning, without sacrificing performance.

[6] Normally, a physician would likely administer tests after the differential diagnosis, before reaching a final diagnosis.

(15%) for final summative evaluation after all development was completed (*heldout-test*). All three subsets have similar class distributions. Each of the four classification models were evaluated in two ways: (1) by training the model on the union of the *dev-train* and *dev-test* sets and testing on the *heldout-test* set, and (2) by running 50 randomized iterations of 10-fold cross-validation on the entire collection of 632 narratives. The first evaluation experiment addresses the problem of overfitting by excluding the *heldout-test* set from all development, while the second addresses the problem of sampling bias in the initial set divisions. The results are described in Section 5.2.

## 5 Models and Results

Here we describe the development and performance of each of the four computational models outlined in Section 4. We report on logistic regression, which had the best performance in all metrics for all experiments, after dimensionality reduction (see Section 5.1). The feature selection and modeling was implemented in Python with the `scikit-learn` machine learning library (Pedregosa et al., 2011).

### 5.1 Feature Extraction and Selection

A total of 60 features were examined (see Table 3). The features represented three modalities, motivated by the task the physicians performed and knowledge about dimensions of clinical expertise in this domain: *verbal*, composed of lexical, prosodic, and structural features of the narratives; *eye movement*, consisting of features of fixations and saccadic eye movements; and truly *multimodal* features, consisting of overlapping or simultaneously occurring features from the other two modalities, to reflect integrated multimodal semantics. Continuing with the theme of personalization, we also created a fourth category of *personal* features, with demographics of the physician and statistics about their confidence and correctness in the training data, in order to model their “past” performance. The latter simulates how a system could learn from experience with a particular physician.

As discussed in Section 2, verbal features of confidence have been studied before, and many of the verbal features used here are inspired by previous work. Some verbal features are based on word choice, such as *amplifiers* (e.g. *definitely*, *sure*) and *modals* (e.g. *could*, *might*),[7] while other have to do with silences (or pauses) or prosody. The eye movement and multimodal features are mostly concerned with fixations, as it seems intuitive that fixation may be associated with thoughtfulness about a particular area of the image, which may in turn reflect a physician’s confidence.

Initial feature selection was performed on the development data (*dev-train* and *dev-test*) using `scikit-learn`’s random forest ensemble classifier. This allowed for human-friendly inspection of useful features. Random forests (Breiman, 2001) are an ensemble method in which numerous decision trees are constructed, each trained on a randomized subset of the development data, which allows for the utility of features to be evaluated on many sub-distributions of the data. The importance of a feature can then be approximated as the sum of the error reduction at each node that splits on that feature, weighted by the population size at that node. This reflects the fact that features used near the root of the tree often handle a larger number of individuals. The importance values for all features will sum to 1. We consider any feature that appeared in the top 20 of the ranked features for any model to be important, and all such types of features are marked in bold in Table 3. Interestingly, the useful features for all classification models were almost the same, with a few transpositions in the ordering. The exception was *past confidence*, which was useful under generalized, but disappeared under personalized, as expected, since the personalized scheme effectively normalizes each physician’s confidence values.

Interpreting the results for the verbal features, *silence duration* (statistics about the durations of all silences) and the *duration of narrative* were most useful. Intuitively, this may relate to thoughtfulness or contemplation. Additionally, *words per second*, or speech rate, was also useful, again perhaps relating to more careful or thorough inspection/diagnosis. As discussed earlier, ties between speech and confidence have been well-studied, while eye movements are underreported. It seems intuitive that eye movement

[7] Such word-choice features were mostly based on lexical lists, and some overlap may occur. The *cutaneous conditions* feature contained multiword expressions. These could be improved by using resources such as UMLS (`http://www.nlm.nih.gov/research/umls/`) or WordNet (`http://wordnet.princeton.edu/`).

| Verbal (29) | | Multimodal (14) |
|---|---|---|
| **Duration of narrative** | Pronouns 1st ($n$, %) | **% of initial silence time fixating** |
| Number of silences | Pronouns 3rd ($n$, %) | **% of total silent time fixating** |
| **Silence duration** ($\underline{\Sigma}$, $\underline{\mu}$, $\underline{\sigma}$) | Modals ($n$, %) | **% of total fixation time silent** |
| Duration of initial silence | Amplifier words ($n$, %) | Words per second during fixation |
| Number of filled pauses | Speculative words ($n$, %) | **Pitch during fixations** ($\underline{\mu}$, $range$) |
| Word type-token ratio | Negations ($n$, %) | **Intensity during fixations** ($\underline{\mu}$, $range$) |
| **Words per second** | **Pitch** ($\underline{m}$, $M$, $\mu$) | Pitch of filled pauses ($m$, $M$, $\mu$) |
| Cutaneous conditions ($n$, %) | Intensity ($m$, $M$, $\mu$) | Intensity of filled pauses ($m$, $M$, $\mu$) |

| Eye movement (11) | | Personal (6) | |
|---|---|---|---|
| **Fixation duration** ($\underline{\Sigma}$, $\mu$, $\sigma$) | **Number of fixations** | Attending vs. Resident | Past correctness |
| **Saccade duration** ($\underline{\Sigma}$, $\underline{\mu}$, $\underline{\sigma}$) | **% image area fixated** | Years of experience | |
| **Saccade amplitude** ($\underline{\Sigma}$, $\underline{\mu}$, $\underline{\sigma}$) | | **Past confidence** ($\underline{m}$, $M$, $\mu$) | |

Table 3: Features examined for classification (60 total), grouped by modality. Symbols in parentheses indicate statistics over all occurrences of a feature in a narrative: raw count ($n$), raw count divided by the total number of words (%), sum ($\Sigma$), mean ($\mu$), standard deviation ($\sigma$), min ($m$), max ($M$), range ($range$). Useful features are boldfaced. If a feature has multiple statistics, the useful ones are underlined.

features may be more related to correctness. For example, the most useful eye movement feature was *% image area fixated*, computed using a grid overlaid onto the image. If more of the image was fixated upon, then it may have contained more areas of interest, or more visual evidence may have been sought, which may also be related to case difficulty. Similarly, features of *saccade amplitude* (the angle of a saccadic eye movement) may reflect physicians feeling a need to explore additional visual evidence by switching focus between distant areas in an image. It is not surprising that the useful individual features from verbal and eye movement modalities were also useful when combined as multimodal features. In particular, simultaneous silence and fixation were the most useful, which again might indicate contemplation and analytical cognitive processing. This suggests that expression of confidence and diagnostic self-awareness is at least partially a multimodal phenomenon.

Although the random forest method could be used for dimensionality reduction, we instead use Principle Component Analysis (PCA) in evaluation below, as it gave better performance gains in development. The purpose of the random forest method was to examine which verbal, eye movement, and multimodal features were most informative for classification, as we are interested in understanding how these modalities relate to confidence and correctness. The latent features resulting from PCA are linear combinations of the features, and thus would not allow for such inspection. The number of PCA components was optimized for classification accuracy in cross-validation for each of the four classification models. Each problem had a different number of principal components, indicating that both the binning scheme and the classification problem type affected which features were identified as more collectively discriminative by PCA.

### 5.2 Results and Evaluation

**Heldout narratives** We addressed the problem of overfitting by withholding 15% ($n = 95$) of the narratives as an unseen final evaluation set. All predictive models performed well above their respective majority class baselines (see Table 4). The Confidence Only models were able to reach higher accuracy, precision, and recall than the joint Confidence & Correctness models. The exception is the accuracy relative to baseline for personalized Confidence Only, which may be due to its higher baseline. As mentioned in Section 3.2, the generalized binning scheme is biased towards over- and underconfidence, and the personalized towards appropriate confidence. The per-class metrics (not shown here) reflect this fact, with overconfidence having higher precision and recall under generalized binning than under personalized. Additionally, under the personalized scheme underconfidence is particularly underrepresented and thus more difficult to predict.

| Binning | Problem | N | Majority Class | % BL | % Acc. | P | R |
|---|---|---|---|---|---|---|---|
| Generalized | Conf. Only | 2 | High Confidence | 53 | 76 (+23) | 0.76 | 0.76 |
| | Conf. & Corr. | 4 | Overconfidence | 37 | 53 (+16) | 0.42 | 0.42 |
| Personalized | Conf. Only | 2 | High Confidence | 65 | 77 (+12) | 0.75 | 0.73 |
| | Conf. & Corr. | 4 | Appropriate High | 37 | 53 (+16) | 0.38 | 0.42 |

Table 4: Performance metrics for the *heldout-test* set under each binning scheme with logistic regression and PCA. All four models performed well above the majority class baselines (% BL) of their respective problems (each with $N$ many class labels). Precision (P) and recall (R) are each macro-averaged.

**Random cross-validation** A potential drawback of the initial development strategy used here is that the initial random splits may bias classification models. To address this problem, after the heldout testing, 50 randomized iterations of 10-fold cross-validation were performed on the total collection of narratives, the results of which are in Table 5. The personalized binning scheme was designed to mimic a system that could adapt to a physician's performance history, and thus the statistics used for personalized confidence binning were recomputed on the training data within each individual cross-validation fold. It is therefore not possible to establish a baseline for the personalized confidence binning outside of a given fold. Instead, we take the mean of the percent accuracy *above baseline* from each test fold ($\frac{1}{k}\sum_{i=1}^{k}(accuracy_i - baseline_i)$). All models performed well above their respective baselines, which is in line with observations from heldout testing.

| **Binning** | Generalized | | Personalized | |
|---|---|---|---|---|
| **Problem** | C.O. | C&C | C.O. | C&C |
| **Acc. above baseline** | +14 | +9 | +13 | +12 |
| **Precision** | 0.70 | 0.25 | 0.69 | 0.32 |
| **Recall** | 0.70 | 0.38 | 0.57 | 0.37 |

Table 5: Performance metrics for logistic regression with 50 randomized iterations of cross-validation using all narratives for Confidence Only (C.O.) and Confidence & Correctness (C&C). We average the accuracy above baseline from each individual fold. Precision and recall are each macro-averaged for each problem.

| **Feature modality** | Generalized | | Personalized | |
|---|---|---|---|---|
| | C.O. | C&C | C.O. | C&C |
| V | +13 | +9 | +12 | +11 |
| E | +7 | +6 | +11 | +10 |
| MM | +7 | +4 | +6 | +5 |
| V+E | +13 | +9 | +13 | +11 |
| V+MM | +14 | +8 | +11 | +11 |
| E+MM | +10 | +6 | +13 | +11 |
| V+E+MM | +14 | +9 | +13 | +12 |

Table 6: Modality study with cross-validation for Verbal (V), Eye movement (E), and Multimodal (MM) features, measured in accuracy above respective baselines, averaged over all folds. Most modality combinations equaled or slightly improved on constituent modalities in isolation.

**Modality study** We also performed a study within the cross-validation testing to investigate the impact of different feature modality combinations on classification (see Table 6). Importantly, the verbal modality alone was more powerful than the eye movement or multimodal features, but most combinations of modalities resulted in slightly higher or equal accuracy compared to their isolated constituent modalities. This suggests that, as we projected, considering multiple modalities of a physician's behavior can help reveal their confidence and self-awareness, but also that verbal features are the most informative, likely since verbal expression is the primary means to tap into physicians' rich and tacit conceptual understanding of a diagnostic case. The multimodal features, which focused on combining verbal and eye movement data, did not improve performance over baselines as much as the simple combination of the individual verbal and eye movement features. One reason for this could be that a person's speech and eye movements are not perfectly temporally aligned (Vaidyanathan et al., 2012), and this asynchronous relationship may affect the meaningfulness of our multimodal feature measurements. Additionally, these eye movement features may be at a much finer spatial or temporal scale than the verbal features.

## 6 Conclusions

This study examined a dataset of medical narratives consisting of verbal descriptions, eye movements, and self-reported confidence values, and used it to model physicians' confidence in diagnosis, as well as their diagnostic self-awareness. The Confidence Only problem involves the expression of confidence based on clinicians' belief, but it is important to understand the relationship to clinicians' actual diagnostic performance. This distinction is key because, while predicting confidence alone is a stepping stone, self-awareness is the ability to additionally align one's confidence with unknown correctness, which involves human intuitive and analytical reasoning (another topic of interest to the medical field, see Hochberg et al. (2014)). Case studies of the most and least confident physicians revealed a complex relationship between confidence and correctness, and highlighted the need for exploring clinical self-awareness. We also defined a personalized binning scheme for physician confidence levels, taking into account a physician's past confidence when drawing the line between high and low confidence, and compared this to a generalized binning scheme based on performance of all physicians. In tandem, these approaches to confidence binning could be used by an intelligent diagnostic support system.

We incorporated previously unused eye movement information from this dataset, and introduced truly multimodal features which directly combined physicians' verbal and eye movement behaviors. While physicians' eye movement and multimodal features were not individually as powerful as verbal features, combinations of the three groups mostly produced classification improvements that were slightly better than, or at least as good as, their constituent feature groups in isolation. The best performance for the majority of models was achieved by considering features from all three modalities. This suggests that eye movements help convey confidence and diagnostic self-awareness. The multimodal features did not help as much, which we believe is explained by the more flexible temporal relationship between speech and eye movements in the human mind. We leave the multimodal alignment challenge to future work. Some pitch features implemented without speaker-dependent analysis were useful for classification, but future work may benefit from pitch feature representations that adapt to demographic variation. Another area for future work beyond the scope of this study includes examining alternative ways of combining confidence and correctness classes, such as merging the diagonals of Figure 1 into a binary classification of appropriate vs. inappropriate (i.e. the union of over- and underconfidence). Such alternatives may present additional challenges for classification, but could also provide benefits for simpler clinical support applications that may not be concerned with differentiating all four classes.

## Acknowledgements

This work was supported by a seed award, and its dissemination partially by a Kodak Endowed Chair award, both from the Golisano College of Computing and Information Sciences at RIT. The original data collection was supported by NIH grant 1 R21 LM01003901A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors also thank Rui Li, and appreciate the helpful comments from reviewers.

## References

Eta S. Berner and Mark L. Graber. 2008. Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, 121(5A):S2–S23.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Pat Croskerry. 2003. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, 78(8):775–780, August.

Pat Croskerry. 2008. Overconfidence in clinical decision making. *The American Journal of Medicine*, 121(5A):S24–S29.

Pat Croskerry. 2009. A universal model of diagnostic reasoning. *Academic Medicine*, 84(8):1022–1028, August.

Charles P. Friedman, Guido G. Gatti, Timothy M. Franz, Gwendolyn C. Murphy, Frederic M. Wolf, Paul S. Heckerling, Paul L. Fine, Thomas M. Miller, and Arthur S. Elstein. 2005. Do physicians know when their diagnoses are correct? *Journal of General Internal medicine*, 20:334–339, April.

Albert Gatt and Patrizia Paggio. 2013. What and where: An empirical investigation of pointing gestures and descriptions in multimodal referring actions. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 82–91, Sofia, Bulgaria, August 8-9.

Mark Graber, Ruthanna Gordon, and Nancy Franklin. 2002. Reducing diagnostic errors in medicine: What's the goal? *Academic Medicine*, 77(10):981–992, October.

Mark L. Graber, Nancy Franklin, and Ruthanna Gordon. 2005. Diagnostic error in internal medicine. *Archives of Internal Medicine*, 165:1493–1499, July 11.

Limor Hochberg, Cecilia Ovesdotter Alm, Esa M. Rantanen, Caroline M. DeLong, and Anne Haake. 2014. Decision style in a clinical reasoning corpus. BioNLP 2014.

Jay Katz. 1984. Why doctors don't disclose uncertainty. *Hastings Center Report*, 14:35–44.

Charles E. Kimble and Steven D. Seidel. 1991. Vocal signs of confidence. *Journal of Nonverbal Behavior*, 15:99–105.

Rui Li, Jeff Pelz, Pengcheng Shi, Cecilia Ovesdotter Alm, and Anne Haake. 2012a. Learning eye movement patterns for characterization of perceptual expertise. In *ETRA 2012 Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 393–396, Santa Barbara, CA, March 28-30.

Rui Li, Jeff Pelz, Pengcheng Shi, and Anne Haake. 2012b. Learning image-derived eye movement patterns for characterization of perceptual expertise. In *Proceedings of CogSci 2012*, pages 1900–1905.

Jackson Liscombe, Julia Hirschberg, and Jennifer J. Venditti. 2005. Detecting certainness in spoken tutorial dialogues. In *Proceedings of Interspeech 2005*, pages 1837–1840, Lisbon, Portugal.

Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff B. Pelz, Pengcheng Shi, and Anne Haake. 2012. Linking uncertainty in physicians' narratives to diagnostic correctness. In *Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM-2012)*, pages 19–27, Jeju, Republic of Korea, 13 July.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Phillippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 973–982, Sofia, Bulgaria, August 4-9.

Heather Pon-Barry and Stuart M. Shieber. 2011. Recognizing uncertainty in speech. *EURASIP Journal on Advances in Signal Processing*, 2011(251753).

Erika Rogers. 1996. A study of visual reasoning in medical diagnosis. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pages 213–218, La Jolla, California, 12-15 July.

Klaus R. Scherer, Harvey London, and Jared J. Wolf. 1973. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, 7:31–44, June.

Vicki L. Smith and Herbert H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language*, 32(1):25–38.

Preethi Vaidyanathan, Jeff Pelz, Wilson McCoy, Cara Calvelli, Cecilia Ovesdotter Alm, Pengcheng Shi, and Anne Haake. 2012. Visualinguistic approach to medical image understanding. In *Proceedings of the AMIA 2012 Annual Symposium*, Chicago, Illinois, November.

Kathryn Womack, Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff B. Pelz, Pengcheng Shi, and Anne Haake. 2012. Disfluencies as extra-propositional indicators of cognitive processing. In *Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM-2012)*, pages 1–9, Jeju, Republic of Korea, 13 July.

Kathryn Womack, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff B. Pelz, Pengcheng Shi, and Anne Haake. 2013. Markers of confidence and correctness in spoken medical narratives. In *Proceedings of Interspeech 2013*, pages 2549–2553, Lyon, France, August 25-29.