Providing Insight into Machine Learning Through Visualization

Anirudh Kaushik
College of Computer and Information Science
Northeastern University
Boston, MA
kaushik.an@ccs.neu.edu

Futai Lin
College of Engineering
Northeastern University
Boston, MA
lin.fu@husky.neu.edu

Bingyu Wang
College of Computer and Information Science
Northeastern University
Boston, MA
rainicy@ccs.neu.edu

***Abstract*—Enterprise investments in machine learning are growing faster than ever. In medical diagnosis, traditional machine learning algorithms, such as Bayesian classifier and decision trees, are widely applied on patients' disease predictions. Many other state-of-the-art machine learning methods, like neural networks and conditional Bernoulli mixtures(CBM), outperform these traditional algorithms, but they are rarely used in production due to lacking of good explanation of trained models. In this paper, we propose a visualization solution to present a more sophisticated model, CBM. In addition, we provide the quantified analysis of approximation between CBM and decision tree. Furthermore, we recommend a pipeline system for hospital industry, which starts with raw patients data and provides the insight into CBM predictions for doctors.**

I. Introduction

In machine learning (ML), applying the latest and best algorithms from research to a production environment requires a lot of work relating the two areas. For example, in the medical domain ML has been used for determining which disease or condition explains a person's symptoms, which is a Multi-label classification[12] problem, wherein symptoms are features as input and diseases are multiple labels as output. Many traditional ML algorithms, such as Naive Bayes[6], [5], Boosting[3], and Decision Tree[10], have been used in clinical diagnostic systems. The main reason these less-sophisticated methods are preferred, despite their weaker performance, is that doctors can better understand these models. Doctors, as the primary communication channel with patients, require an understanding of how the algorithm makes a decision before they can apply its diagnosis to the patient. Algorithms like Deep Learning[9] and Conditional Bernoulli Mixtures (CBM)[4], can achieve much higher performances, but are not production-quantified, because the clarity and mechanisms of the algorithm are lost to the users (doctors).

CBM was specifically designed for medical data, and was designed through a partnership with Massachusetts General Hospital. Unfortunately, despite the model's accuracy over the medical data set, doctors felt uneasy about using the model for diagnosis because of its lack of clarity and their lack of understanding of the model. Doctors felt they would be unable to explain to patients why the model made a specific diagnosis, which is an essential part of a doctor's job. This tradeoff was not one doctors were willing to take.

Our main goal is to develop a visualization solution for CBM to reduce the gap between the CBM model and humans' understanding of the model, especially for non-technical people in hospitals. To help accomplish this task, we established three design goals for our project, and the effectiveness of our visualization will be measured based on its effectiveness to accomplish these goals:

1) Exposing the underlying mechanisms of CBM
2) Providing insightful information about the patient
3) Providing novel insights about the diseases

Furthermore, we are proposing an integrated pipeline framework for the hospital industry to effectively utilize this visualization solution. This pipeline would create an end-to-end solution for hospitals to provide raw patient data and receive a detailed dashboard visualizing the mechanisms of the machine learning model and its analysis of individual patients.

The paper first discusses the mechanisms of the CBM model and the nature of the data we are working with. We then describe our contributions, which include describing the mechanisms of the CBM in medical diagnosis, approximating the CBM model to a decision tree, and visualizing the output of the CBM in an integrated dashboard. We also describe a data pipeline solution for hospitals to utilize in order to maximize the effectiveness of this tool with their data.

II. Partners

This project is a cooperation with Prof. Javed Aslam[1], and Prof. Virgil Pavlu[2], who are from the College of Computer and Information Science in Northeastern University. Their research is focusing on solving multi-label classification problems and applying their results to the healthcare industry. Cheng Li[3] and Yuyu Xu[4], are the PhD candidates, who are also working on the same problems. They have been collaborating with Massachusetts General Hospital[5] to apply multi-label methods to patients diagnosis. We have been working with them to develop a visualization that they can potentially apply to production for medical staff in MGH in the future.

[1] https://www.ccis.northeastern.edu/people/jay-javed-aslam/
[2] https://www.ccis.northeastern.edu/people/virgil-pavlu/
[3] https://www.ccis.northeastern.edu/people/cheng-li/
[4] https://scholar.google.com/citations?user=CwHkgE4AAAAJ&hl=en
[5] http://www.massgeneral.org/

III. Related Works

Many visualizations exist that attempt to provide visual insight into complex machine learning algorithms. All of them aim to achieve the same goal as we have set out to reach: to reduce the gap between the model itself and human's understanding of the model. One of the most well-known publications dedicated to this goal is Distill[1], which focuses on publications aimed to clearly explain machine learning research and its results. Distill is not only focusing on explaining the ML algorithm, but also working on explanations of training procedures. They have got excellent job done on interpreting state-of-the-art ML methods, such as Neural Networks, Momentum and Sequential modeling etc. Another famous visual introduction to ML was designed by R2D3[11], described by themselves as "an experiment in expressing statistical thinking with interactive design". In their visualization, they give a detailed description of how machine learning and statistical methods can identify patters in data. Their visualization showcases a visually impressive assortment of graphics that guide the reader through an explanation of how a decision tree algorithm works to differentiate homes in New York City and San Francisco. Visualization techniques, such a histogram, scatterplot matrix and bar chart, have been also applied. It concludes with a dynamic decision tree to classify the homes in their data set. This design, shown in Figure 1, is extremely effective in allowing a reader to recognize the patterns in the data set. Our final decision tree design in our visualization dashboard is based heavily on the R2D3 decision tree, as their design was extremely effective in showcasing the same information about the data as we wished to show. We owe the entire design of our decision tree graphic to the team at R2D3. Utilizing this design allowed us to give better visual insights into CBM, making the overall visualization more informative.

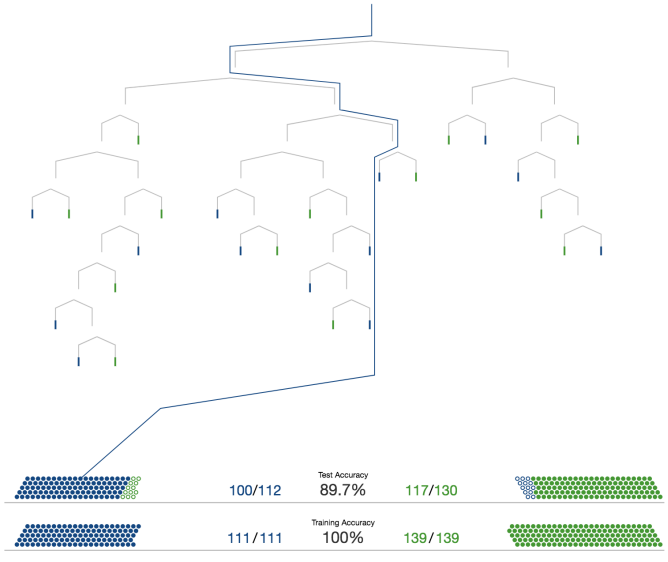


Fig. 1. Decision Tree in R2D3

IV. CBM and Dataset

In this section, we introduce the CBM model and the medical dataset. Since our potential clients are doctors and they are more familiar with patients' data, we focus more on the CBM explanation and less on the details of the dataset itself.

A. Conditional Bernoulli Mixtures

Conditional Bernoulli Mixtures (CBM) is a mixture model in machine learning, proposed in previous work[10]. CBM relies on a mixture to approximate the conditional joint probability over all labels, which is defined as:

$$p(\underbrace{\boldsymbol{y}}_{diseases} \mid \underbrace{\boldsymbol{x}}_{patient}) = \underbrace{\sum_{k=1}^{K} \pi(z=k|x;\alpha)}_{\substack{\text{\#K experts and} \\ \text{his/her own credibility}}} \underbrace{\prod_{l=1}^{L} b(y_l|x;\beta_l^k)}_{\substack{\text{disease predicted} \\ \text{by each expert}}} \quad (1)$$

Applied in the scenario of a hospital, over medical data, the CBM's task could be taking a patient's symptoms as input($\boldsymbol{x}$), and predicting the probability of this patient with diseases($\boldsymbol{y}$), which is shown on the left part of equation (1). The right side of equation (1) shows how the CBM calculates this probability. The best way to interpret this model is using an analogy involving a mixture of experts. We provide a visualization to explain the mechanism of CBM to doctors through this analogy. The mechanism of CBM in medical domain is shown in Figure 2.

The analogy can be described as thus: Suppose a patient (he) comes to a hospital to figure out if he has caught the flu. First, he describes the symptoms to a nurse, such as fever and sore throat. The nurse will assign two experts (doctors), according to his symptoms, to give a diagnosis of the patient. Suppose one is infectious specialist (**I**) and another is otolaryngologist, or an ENT, (**O**). Since the inquiring disease is flu, these two experts have different confidences on their ability to diagnose the flu. For instance, **I** has a $C(\boldsymbol{I}) = 90\%$ confidence to diagnose the flu, but **O** only has a $C(\boldsymbol{O}) = 10\%$ confidence. These confidences can be also interpreted as credibility of specialists, which was learned from the experts' past history of diagnosing the flu correctly. This process is demonstrated in the first step of Figure 2.

These two experts will see this patient individually and ask him bunch of different questions. This determines the patient's symptoms, which may include the old ones (fever and sore throat) and new ones (e.x. coughing, headache, trouble breathing). Based on the updated symptoms, the experts are able to give the probability of this patient having flu. For example, **I** has $D(\boldsymbol{I}) = 30\%$ confidence that the patient has the flu, and **O** has $D(\boldsymbol{O}) = 80\%$ confidence of a flu diagnosis. This is shown in the second step in Figure 2.

In the last step of Figure 2, the CBM combines the two experts' diagnosis and give the final confidence of this patient having the flu, which is 0.35. This is basically a marginal probability calculation, as shown in Equation (2). Since the final confidence of flu is less than 0.5, the threshold the CBM sets for labeling a patient, our patient can be more sure that he does not have the flu.

$$p(flu) = C(\boldsymbol{I}) * D(\boldsymbol{I}) + C(\boldsymbol{O}) * D(\boldsymbol{O}) = 0.35 \quad (2)$$

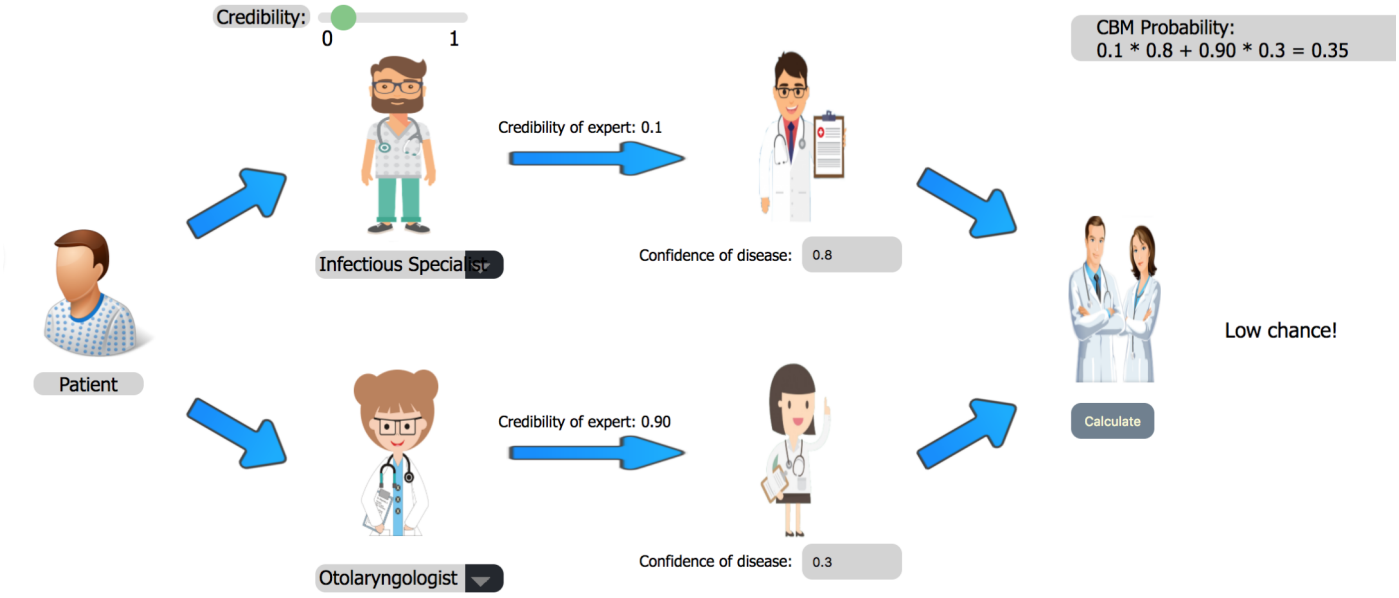


Fig. 2. The mechanisms of CBM in medical diagnosis

For the demo purpose, we only show two experts in this example. In reality however, a production CBM model could have dozens of experts who may belong to the same/different specialists-types, such as cardiologist, allergist, internist and so on. The CBM model also makes more sense in real-world diagnosis. When we are curious about a disease we may have, the more advice diagnosed by experts, the more confidence we may have about this disease. We created an online demo[6] that explains these mechanisms of CBM to educate people who are not familiar with the details of the model.

B. Dataset

All our visualization demos online are based on a publicly available data set. This data, called Medical[8], consists of clinical text reports of individual patients labeled with their diseases. This dataset is based on the data made available during the Computational Medicine Centers 2007 Medical Natural Language Processing Challenge 10. It consists of 978 clinical free text reports labeled with one or more diseases, out of 45 possible diseases. Due to confidentiality, all the diseases are encoded into disease codes, which is a combination of alphabets and digits.

Additionally, we also have data from the output of the trained CBM model. This consists of the probability each "expert" has given on the diagnosis of a patient, as well as the symptoms each expert considered when making its diagnosis.

In summary, The data we have can be split into two parts:

1) Raw patients' clinical text report and the final diagnosis of the patient (as a disease code)
2) Data trained from CBM.

We will elaborate on these two components of our data and how we utilize them in our visualization below:

[6]https://github.com/TaitaiLin/CBMGraph

Raw Data. The patients' clinical text report consists mainly of each patient's symptoms and diagnosis. This data was collected by doctors and nurses in hospitals in full English sentence, a form that can be easily understood by specialists. Specialists are able to indicate a patient's diagnosis using one or more symptoms, because they were educated with professional knowledge. Essentially, a specialist has built a clear pattern/logic to identify a patient's diagnosis using symptoms. One of the ways that they got trained is using clinical experiments or patients' clinical text records. However, training a specialist via professional education costs a lot of time and money.

CBM Data. Instead of educating a real doctor, we want to build a diagnosis checking AI, which can provide the confidences of diseases by given symptoms. As discussed previously, CBM is able to learn the pattern through three steps:

1) a nurse roughly checks symptoms and refers bunch of experts with different credibility;
2) each expert checks symptoms and gives his/her own confidence of the diagnosis;
3) combining all experts' opinions to output the final confidence of the diagnosis.

The mechanism of CBM now is clear to us. Unfortunately, it is still not clear which symptoms are checked by the nurse and which symptoms are used for each expert to build the confidence within each step. Nevertheless, all these questions can be directly answered through the CBM model, because all these connections between symptoms and confidence of a diagnosis are already captured by the model. Unfortunately, there is a gap in understanding between AI language and human language, as this data from the model is not an easily-consumable format of information. This problem is where we derive our second design goal: reducing this gap in understand-

ing.

Before reducing the gap, let us first take a look at what information we have in the CBM data.

TABLE I. SCORING SYSTEM EXAMPLE FOR CBM

scores for symptoms	Credibility of experts		Confidence of flu	
	I	**O**	**I**	**O**
Fever	10	5	8	-3
Sore throat	4	9	3	7

CBM has built numerical connections between each symptom and each diagnosis after empirical learning. It is more like a scoring system, where it takes a symptom and outputs scores for credibility of experts and confidence of diagnosis separately. Given the same example in Figure 2, suppose we have two symptoms (**Fever** and **Sore throat**) and two experts (infectious specialist(**I**) and otolaryngologist(**O**)). CBM builds a scoring matrix shown in TABLE I.

The scores in the table have two parts:

1) credibility of expert
2) confidence of disease.

1) is a positive number between 0 and 1, and a larger number represents the veracity of the expert. 2) can be either positive or negative. The larger number is, the higher the likelihood of this symptom contributing to the disease we are diagnosing, and the lower the number, the lower the likelihood. In TABLE I, the **Fever** symptom gives more credibility (10) to the infectious specialist, but less credibility (5) to the otolaryngologist on flu diagnosis. On the other hand, **Sore throat** gives a higher trust to the otolaryngologist (9) over the infectious specialist (4). This scoring system in CBM also indicates what connections exist between symptoms and the confidence of diagnosis from different experts. For instance, **Fever** gives higher confidence to **I**(8) than **O**(-3) to believe patient is infected by flu.

Now we see the CBM data in a scoring system perspective. This still does not provide obvious answers to our questions, such as how good is this scoring system (CBM performance), which symptom is the most related to flu, and whether a new patient has a flu or not. In the next section, we build up from this tabular view of the CBM to derive the visualization of the model itself.

V. VISUALIZATION OF CBM

Ultimately, we decided on three visualizations to best accomplish our goals of visualizing the CBM model. We used Python to process and prepare our data, and the D3.js and React libraries to complete the interactive visualization. We will now discuss the specifics of each part of the design.

A. Interpreting CBM Data

As discussed earlier, CBM has been interpreted in three ways: Equation (1), Figure 2 and TABLE I. However, none of them can be used to figure out the patterns between symptoms and diagnosis. The closest solution would be the last table structure, which already gives us insights into how each symptom plays its role in a diagnosis (along with the credibility of experts and confidence of disease).

To simplify our problem, let us compare this scoring system with a classroom grading system. Let us consider each symptom as a student: **Fever** and Sore throat as two students; treat credibility of **I** and **O** as homework and project percentage contributions to final grades respectively; and consider the confidence of **I** and **O** as the grades who these students have received for their homework and project.

Let us understand the disadvantages of table structure by comparing with the grading system. This will allow us to build a visualization that overcomes these disadvantages:

1) Lack of overall contributions from an individual symptom to the disease, which means there is no single numerical value connecting a symptom to the disease. In a grading system, we do not directly know each student's final grade.
2) No comparison between symptoms in terms of contributing to the disease. For example in TABLE I, there is no way to figure out **Fever** is more/less important than **Sore throat** for predicting flu. In the grading system analogy, you cannot tell which student is doing better than the other.
3) No measurement of the CBM's scoring system to indicate whether it is good or bad, e.g. what is the accuracy of predicting flu? Equivalently, what's the average grades over all students in a class?

We start with solving the first issue from the list, which measures the importance between a symptom and disease, by TABLE I. Inspired by the mechanism of CBM, we see that the final decision step uses the weighted average scores from different experts, which is a method of computing the arithmetic mean of a set of numbers in which some elements of the set carry more importance (weight) than others[7]. It is also the same mechanism in a grading system, when you grade a stuent's final grade based on grades from homework/project and their respective contribution percentages. Now it is straightforward to see the final score for each symptom: the weighted average scores of symptoms **Fever** and **Sore throat** are calculated in the last column in TABLE II. After the weighted average, we can clearly see the final contribution of each symptom in predicting the disease. Now that we have addressed the first issue, the second issue also becomes easy to address: the higher weighted score a symptom has, the more important it is. Under this situation, **Sore throat**(75) plays a more important role in predicting flu than **Fever**(65).

TABLE II. WEIGHTED SCORE SYSTEM FOR CBM

scores for symptoms	Credibility		Confidence		Weighted score
	I	**O**	**I**	**O**	
Fever	10	5	8	-3	$10 \times 8 + 5 \times -3 = 65$
Sore throat	4	9	3	7	$4 \times 3 + 9 \times 7 = 75$

Now we address the last issue: what iss the average grade of students in a class? If we do the same thing in CBM scoring system, take the average scores of two symptoms, which is $70 = (65 + 75)/2$. Unfortunately, 70 in the CBM scoring system is meaningless. In reality, we only care about the accuracy of the system, such as how accurate it is when a disease has been predicted comparing with the ground truth.

[7]Mathwords: http://www.mathwords.com/w/weighted_average.htm

If a patient has flu and the scoring system prediction is positive (meaning it predicted flu), then it is a successful classification. Otherwise, it is unsuccessful. The proportion of total successful classifications is the model's accuracy[13]. However, the current scoring system in TABLE II is not able to classify any patients, so we must do something else to determine the accuracy of the model.

Let us first assume that we have three patients: patient-1 (**P1**) only has **Fever** symptom without the flu, patient-2 (**P2**) just has **Sore throat** symptom with the flu and patient-3 (**P3**) has both symptoms with the flu. According to different symptoms of these three patients (using two elements vector to indicate symptoms, such as $(1,0),(0,1),(1,1)$), their weighted scores in CBM are shown in TABLE III along with their ground truth of flu.

TABLE III. WEIGHTED SCORES FOR PATIENTS WITH DIFFERENT SYMPTOMS

scores for symptoms	Weighted score	P1 (1,0)	P2 (0,1)	P3 (1,1)
Fever	65	65	0	65
Sore throat	75	0	75	75
Ground Truth(flu)	N/A	0	1	1

With all patients' information now converted to the form as shown in TABLE III, we are able to build another classifier on top of CBM scoring system, wherein:

- each patient is regarded as an instance
- each symptom is considered as a feature
- each weighted score is considered the feature value
- ground truth is used as a binary label.

Now this problem is a normal binary classification problem to test if a patient has certain disease or not[14]. Well-known binary classifiers include logistic regression[2], [7], naive Bayesian[7], [6], and decision tree[10] have all been proven successful at accomplishing this task. Among all these classifiers, the decision tree stands out for a few reasons. Firstly, its simple structure means its ability to be explained to non-technical persons is quite straightforward. Additionally, existing work by R2D3[11] has created a visually impressive representation of a decision tree. This design allows us to showcase the final decision tree classifier in an elegant way, helping non-technical staff understand the patterns in the data and better explain the model's mechanisms.

B. Decision Tree

The decision tree view visualizes the approximation we have made from a CBM to a decision tree, as described in the previous section. The visual design of this tree comes directly from R2D3. We used R2D3's design because they did a very effective job designing a decision tree that conveys all the relevant information we were looking to showcase, namely the features at the split point, the training process, and the paths of each data point (the patients). The training and testing process can be started for the tree with buttons at the bottom. This shows all the data points falling down the tree through their respective paths. Once all the points have fallen down the tree, the training and testing accuracies for the tree can be seen. Later in this paper we will show an analysis of these accuracies. Once the training and testing has completed, hovering over each point will illuminate the path that point took down the tree.

Motion is a major component of the information this tree conveys. By showing the dots falling down the tree, this visualization shows doctors how a decision tree model is built and trained. It also clearly visualizes how the model makes a decision about each patient. Each leaf is colored with the appropriate color for the label that it assigns, so that a doctor can see how the data points are labeled as they pass down the tree towards a specific leaf.

We envision doctors using this view to understand what features the CBM ultimately used to make the decision it did. For example, if a node passes through forks in the tree with features cough, pneumonia, chest, and renal, the doctor can understand that the CBM for this disease believes "cough" and "penumonia" are the most relevant in determining whether a patient is diseased or not. Not only does this information give doctors insight about the mechanisms of CBM, but it also provides a clear way to communicate with patients how the machine learning model made their diagnosis.

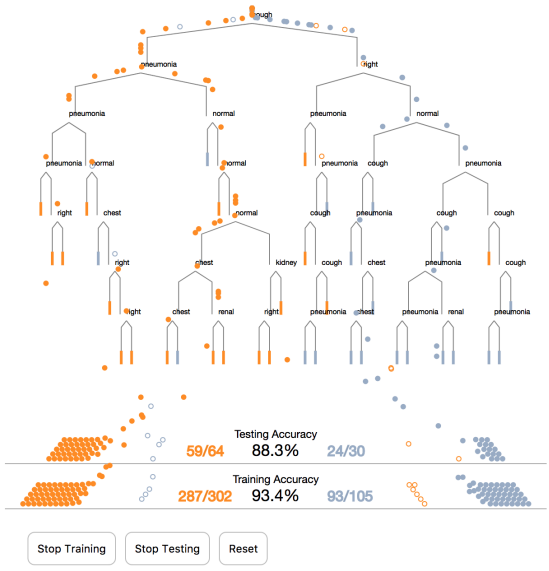


Fig. 3. Decision Tree for Disease Class-4-753_0

C. Bar Chart

The bar chart visualizes raw output from the CBM model. It incorporates the credibility of the experts, confidence of labeling a patient as diseased, and the feature score into once cohesive graphic. Each bar chart visualizes the CBM labeling of one patient, which allows a user to identify patient-specific factors influencing their final classification (diagnosis). As the CBM model is a multi-label classification model, the bar chart shows classification data for all eight diseases in the data set, shown on the x-axis. The disease of focus is shown in shades of green, while the rest are in shades of blue. The figure below shows an example of our design. We plot the three distinct experts who give confidence of eight different diseases. In the bar chart, the first three bars represent experts c1, c2,

and c3. The last bar represents the marginal probability of the patient having this disease, based on the decisions made by the three experts. As explained previously, marginal is the weighted mean across all three experts. Hovering over each bar shows the features score used by the expert to determine a probability. The numbers provided are relevance each expert placed on the specified feature. This information all ultimately facilitates our understanding of the significance of symptoms for each particular patient's diagnosis.

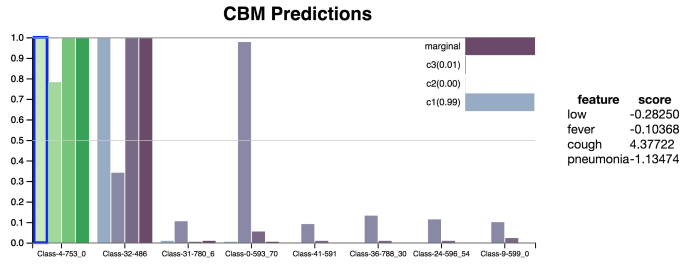


Fig. 4. The Mechanism of CBM in Medical Diagnosis

D. Node-Link Graph

The node-link visualization provides us with a comprehensive overview of the entire patient sample. This graph utilizes a force-directed layout to position all the nodes on the canvas. The set of nodes in this visualization includes all the patients analyzed by the CBM model (colored based on their classification) and the set of features relevant to the particular disease (colored red). An edge connects two nodes only if one node is a patient and one node is a feature. The weight of the edge is equal to the weight of the feature for that patient, as explained in section V-A.

The usefulness of this visualization is in an analysis of the disease as a whole. Because nodes are colored based on their diagnosis, it is easy to see clusters of related diseased nodes or related non-diseased nodes. Because the force-directed layout pulls nodes with heavier edges closer together, nodes are closer to the features they have high values for. Recall that the higher the feature value, the more impact that feature had on the diagnosis of the patient. This means that clusters of nodes are closer to the features that impacted their diagnosis. It also means that features are closer to nodes that they had relevance to. Ultimately, through an analysis of the visualization, one can determine which features are likely to have the most impact on the diagnosis of a disease across a large sample of patients.

For example, if the feature nodes **Penumonia**, **Cough**, and **chest** are the features closest to the clusters of diseased nodes, a doctor viewing this graphic can have an understanding that this disease is likely related to those symptoms. Additionally, if the feature nodes **Renal** and **Ultrasound** are only in proximity with non-diseased nodes, this also gives an indication that those symptoms were relevant in determining that the disease is not present.

We foresee this visualization being effective at deriving novel information about new diseases. Suppose an epidemic breaks out and doctors are not sure about what researchers should focus on in order to develop a cure for the epidemic. While doctors are aware of the symptoms victims are experiencing, it is possible that many of them are simply side-effects of other conditions. Valuable information for a researcher attempting to develop a cure would be which symptoms are extremely important in diagnosing the disease. As doctors take notes on the symptoms of patients, this information is fed into the CBM model and visualized using our visualization tools. The node-link visualization here would show to doctors clearly which symptoms had the most impact on the epidemic. They could then relay this information to researchers, who then know which symptoms they should focus on in order to have a higher chance of detecting something important.

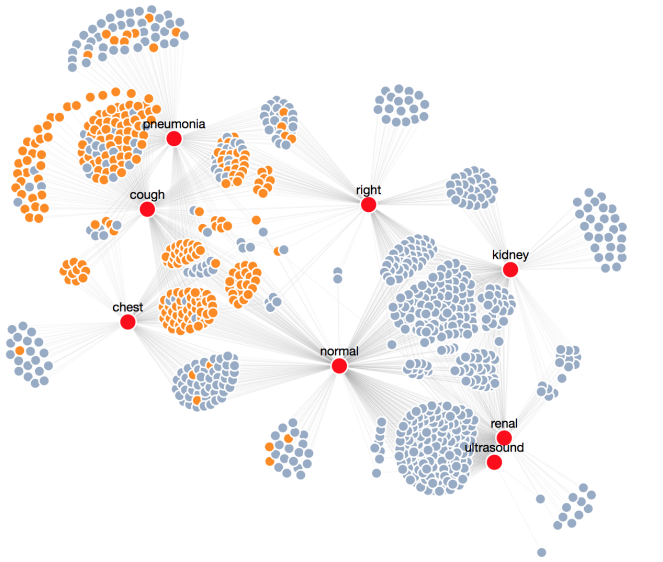


Fig. 5. Node-Link representation of patient data

E. Interactions

Interactions from the user are what make the separate elements of this visualization connected. Hovering over a node in the decision tree highlights the path that node followed through the decision tree. This allows a user to quickly see some patterns in the paths certain nodes have taken. Hovering over a bar in the bar chart provides the user with the scoring information the CBM produced for that expert, disease, and patient.

The most valuable interaction is a user clicking a node (on either the decision tree or the node-link graph). Clicking a node highlights that node on both graph visualizations. It highlights the path that node took down the decision tree, and it highlights the features it is connected to in the node-link graph. It also redraws the bar chart with the CBM output for that node. Clicking a node therefore provides you with a variety of patient-specific data.

VI. Pipeline Framework Solution

CBM algorithm has been demonstrated to outperform most of the state-of-the-art machine learning algorithms in term of accuracy[4]. Besides this, it also shows very promising performance on patients' data owned by MGH. Currently, CBM is implemented in a public machine learning package,

called Pyramid[8], which has been used by MGH for a while. MGH uses the current Pyramid package to clean/index raw patients' data, and train CBM model to be used in production, shown as the components outside the dotted line in Figure 6.

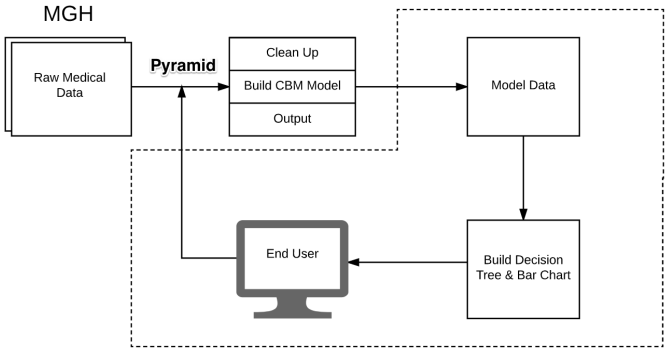


Fig. 6. The Pipeline Framework for Hospital Industry

Unfortunately, the current problem for MGH, is that doctors are struggling in understanding CBM model, which prevents CBM being applied fully in production to diagnose patients. Our current visualization project can be summarized as the inside of the dotted box in Figure 6, which takes over the CBM model, converts CBM to DT and creates the coordinate views from different perspectives, and provides insightful information about patients and diseases to doctors/users. Furthermore, we want to re-train our CBM model based on user feedback. Via our proposed visualization tool, the specialists and doctors can figure out some unrelated symptoms being used or some important symptoms being ignored from the trained model. According to their feedback/constrains, we could re-train our CBM model in order to get higher performances.

Our final goal is to integrate our visualization tools and Pyramid packages, to provide not only machine learning solution but also insights in ML. Therefore, in the future works, we want to improve our visual designs according to our users' feedback and integrate it into Pyramid package, for being used in industrial productions.

This pipeline utilizes the following flow to build the visualization dashboard:

1) Input raw patient data
2) Train CBM model
3) Convert CBM into decision tree
4) Provide insights about each patient and disease to doctors through visualization.
5) Collect feedback from doctors and improve the trained CBM model.

This flow is summarized in Figure 6.

VII. Analysis of Visualization

In this section, we do some analysis on our visualization designs mainly through two parts: quantifying the accuracy changes from CBM model to decision tree in R2D3 and feedback from our clients and potential users.

[8]https://github.com/cheng-li/pyramid

A. Quantify The Accuracy

Since the proposed decision tree (DT) is converted from CBM, the paramount concern from a user perspective is whether we lose any accuracy via the conversion and, if so, how much loss of accuracy?

To answer this question, we have to build the comparison between the CBM model and our DT. The only challenge is that CBM reports the example-based accuracy[12], while the DT only has individual disease accuracy. Individual disease accuracy is already introduced in binary classification accuracy (see section V-A), while example-based accuracy is an exact match ratio—that is, the prediction is correct only when it classifies every single disease correctly for that patient. What we can do then is decompose the example-based accuracy into individual accuracy for each disease. We randomly selected 8 diseases out of 45 from Medical dataset and reported the individual accuracy of each diseases. The comparisons between CBM model and retrained DT model is shown in Figure 7. Figure 7 indicates that the accuracy on training data varies from disease to disease, both for CBM and DT. CBM training accuracy on diseases **Class-4-753_0**, **Class-32-486** and **Class-9-599_0** outperform DT training accuracy, while on the rest of diseases, DT has better performance. On testing accuracy report, CBM testing accuracy outperform DT across the board. The largest gap happens on disease **Class-32-486**, which has almost 5% drops from CBM to DT. Despite the loss of accuracy between CBM and DT, the loss is not significant enough to say that DT is no longer a good representation of the CBM model.

Since substantial testing accuracy loss exists for some diseases when converting CBM to DT, some future work to improve this has proposed by our partners:

- Highlight the patients for who the DT model predicted differently from CBM, both in decision tree visual and node-link graph.
- Find the reasons (i.e. different symptoms used in between CBM and DT) why the two models made opposite predictions.

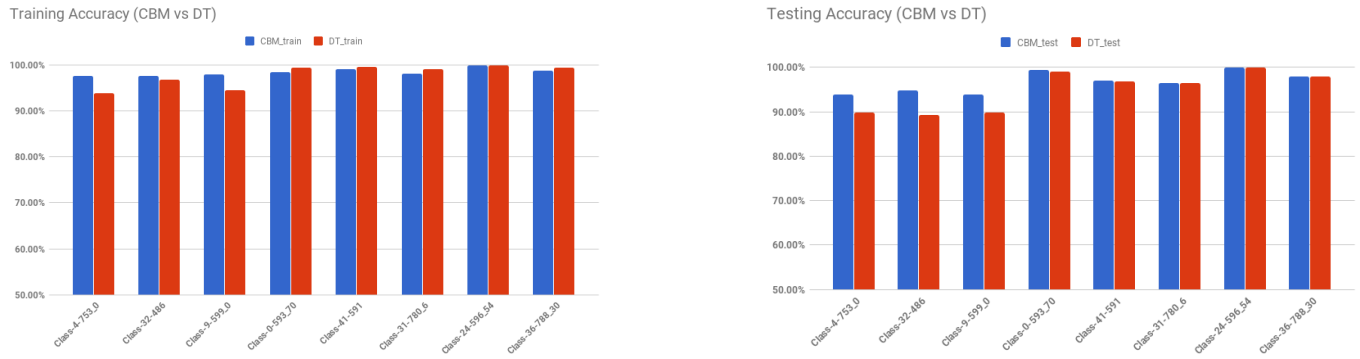
B. Feedback From Users

To test the efficacy our visualization, we have four experts, who are familiar with either CBM model or medical diagnosis. To verify if our visualization provide the insightful information, we have following questions:

1) **Mechanism:** Would you be able to expose what symptoms CBM considered when diagnosing a patient via the design in Figure 2?
2) **Insights into patient:** Would you be able to identify the significant symptoms in predicting each patient?
3) **Insights into disease:** Would you be able to seek the important symptoms in predicting the disease?

We collect all the answers and feedback from the user study and list them as below:

Mechanism: all of the users are able to explain CBM mechanism through the Figure 2 designs. They like the interactions, which include 1) selecting different specialists; 2)



(a) Training Accuracy

(b) Testing Accuracy

Fig. 7. Training and Testing Accuracy Comparison between CBM and DT on 8 Diseases

distributing the different credibility for each specialist; 3) assigning confidence for each specialist; and 4) calculating the final weighted average confidence in the end. Some suggestions provided to us in order to improve the user experience on this visual element is

- Instead of fixing two specialists, users should be able to choose the number of experts.
- "Confidence of disease" input value box can be changed to slider bar with value between 0 to 1.
- Put captions to help user to understand the mechanisms.

Insights into patient: For each disease, the most significant symptom can be identified via decision tree design, which locates on the top layer. The prediction details in CBM for each patient can be explained by clicking on this patient's dot in decision tree. Furthermore, in the bar chart, the scores of symptoms are listed in the side table, which points out the importance of each symptoms. Through the network view, the highlight links help them to locate the related symptoms, which is in red node.

Insights into disease: Our users also agree with that: both decision tree and networks are able to provide the insightful information for a disease:

- the more interesting and influential symptoms for a disease are located on the upper levels of the DT visualization
- it is clear to see the central symptoms related to this disease in the node-link view, and they are the red nodes connecting the most orange nodes

However, our potential clients gave us some drawbacks/suggestions in DT, BC and node-link designs:

- No way to identify the specific patient. Searching function for a patient can be helpful.
- Each patient has his/her own clinical notes. Those notes are not visible in our current design. Adding it as a coordinated view into the dashboard would allow doctors to read the original text note. Further functions, like highlight the words in the note, can be also introduced.
- In the bar chart, add a sorting functions in the table to the side of the chart, either by features name or scores.
- Users are not able to switch from disease to disease. And currently the design is only for one disease.

VIII. Future Work

Potential next steps with this visualization primarily focus on performance improvements and incorporating user feedback. One major issue our partners provided after using the tool for a bit was that it was limited to visualizing one disease. The system we have built can produce data for any number of diseases, and it is possible to view these diseases separately. When attempting to place all 8 diseases together in one dashboard, we found that performance could easily crash a browser window. This meant that our demo had to be limited to only one disease. One of the first steps we will take in the future is to improve the performance of the decision tree visualization so that multiple trees can be visualized in the same browser window. If this is accomplished, a user would then be able to click on a different disease in the bar chart and see the DT and node-link graph for that disease too.

Conducting a larger user study with actual hospital staff, who are our potential future clients for this tool, would also provide more detailed analysis of the effecacy of our tool. The limitations of our user study were primarily that the users in the study were machine learning experts. This prevented us from accurately determining if the tool helped explain the mechanisms of CBM to non-technical users.

An additional group to analyze in a future user study is patients. As this tool is designed to help doctors understand the mechanisms of a patient's diagnosis, it is possible that doctors will be showing this visualization to patients too. We must therefore also understand the nature of information conveyed to patients, especially when it is information about their own diagnosis.

Incorporating the explanation of the CBM mechanisms, as in 2, into the demo itself is something that needs to be done. As of now, they exist as two separate entities.

Lastly, building infrasturcture to connect this visualization dashboard to the raw output of the CBM model from the Pyramid package would allow hospitals like MGH to simply upload their data and get detailed visualizations for diseases of their choice. This component is represented by the arrow entering the area inside the dotted line in Figure 6. This would complete an end-to-end solution for a hospital to get detailed and accurate analysis of their patients and specific diseases in a format non-experts can understand and appreciate.

IX. Conclusion

In this paper, we present a collection of integrated visualizations that provides doctors and other hospital staff with the power to leverage advanced machine learning models without having the knowledge and experience of machine learning engineers. The visualization incorporates a decision tree visualization (based on the R2D3 decision tree), a bar chart, and a node-link graph with a force-directed layout. The visualization we propose aims to provide users with insight into the mechanisms of the machine learning model for explanatory purposes, either to their peers or to patients seeking an explanation of their diagnosis. It also potentially provides novel information about a particular disease, specifically information about which symptoms have the most influence in diagnosis. We believe that this information is all useful to doctors and can be provided through CBM and this visualization.

Additionally, we present a pipeline system that can directly incorporate the CBM model and our proposed visualization to provide hospitals a means to convert raw patient data directly into our powerful visualization. It is imperative that hospitals utilize the advancements researchers are making in machine learning and artificial intelligence, but it is also essential that doctors can still communicate these ideas to patients, as this doctor-patient communication is an essential part of the healthcare system. Fully implementing this system would provide all of the benefits of machine learning while still maintaining that valuable communication channel.

Acknowledgment

The authors would like to thank Prof. Cody Dunne[9], Michail Schwab[10] and Danielle Nguyen, for supervision, technical and writing supports, especially for the suggestion of providing quantified analysis between CBM and DT by Prof. Cody Dunne. The authors also like to thank Prof. Javed Aslam, Virgil Pavel, Cheng Li and Yuyu Xu, for partnership, supervision and user feedback. Additionally, the authors would like to thank the existing works from people like the R2D3 team and Mike Bostock of D3.js for design inspiration on this project.

References

[1] Shan Carter and Chris Olah. Distill. https://distill.pub/about/.

[2] David G Kleinbaum and Mitchel Klein. Analysis of matched data using logistic regression. In *Logistic regression*, pages 389–428. Springer, 2010.

[3] Aurélie Lemmens and Christophe Croux. Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, 43(2):276–286, 2006.

[4] Cheng Li, Bingyu Wang, Virgil Pavlu, and Javed Aslam. Conditional bernoulli mixtures for multi-label classification. In *International Conference on Machine Learning*, pages 2482–2491, 2016.

[5] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Madison, WI, 1998.

[6] Kevin P Murphy. Naive bayes classifiers. *University of British Columbia*, 2006.

[7] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[8] John P Pestian, Christopher Brew, Paweł Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen, and Włodzisław Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics, 2007.

[9] Jesse Read and Fernando Perez-Cruz. Deep learning for multi-label classification. *arXiv preprint arXiv:1502.05988*, 2014.

[10] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[11] Tony Chu Stephanie Yee. R2D3. http://www.r2d3.us/visual-intro-to-machine-learning-part-1/.

[12] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3), 2006.

[13] WIKIPEIDA. Accuracy and precision. https://en.wikipedia.org/wiki/Accuracy_and_precision#In_binary_classification.

[14] WIKIPEIDA. Binary classification. https://en.wikipedia.org/wiki/Binary_classification.

[9] https://www.ccis.northeastern.edu/people/cody-dunne/

[10] https://www.ccis.northeastern.edu/people/michail-schwab/