# Visualization for Classification in Deep Neural Networks

Junghoon Chae* Shang Gao† Arvind Ramanthan‡ Chad Steed§ Georgia D. Tourassi¶
Oak Ridge National Laboratory

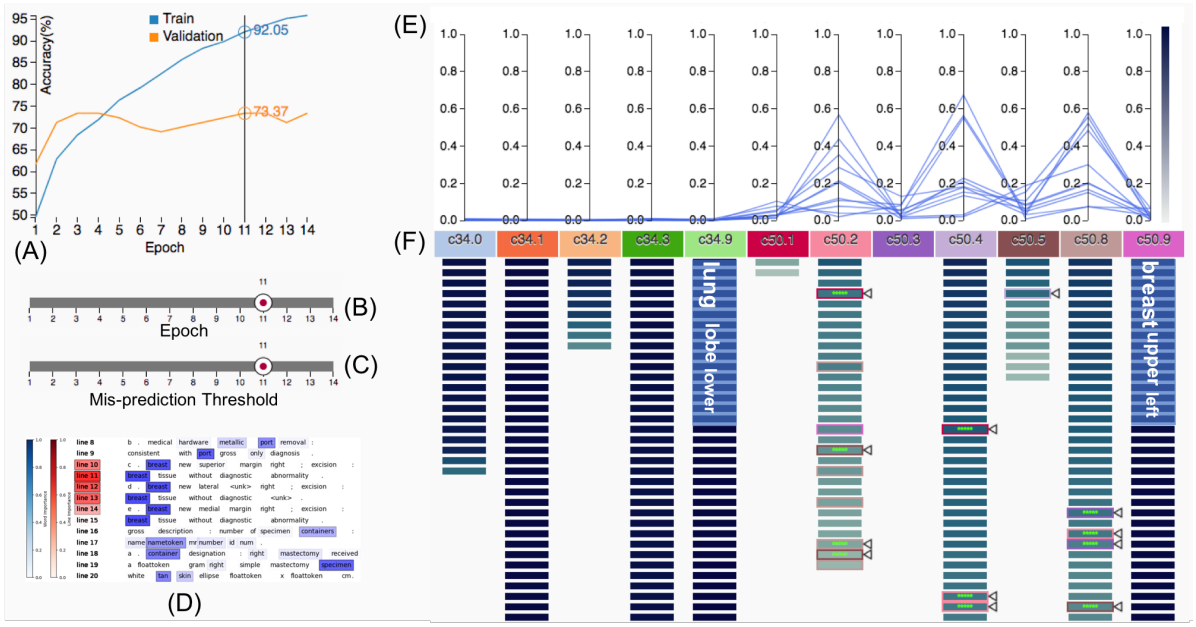


Figure 1: A visual analytics tool to understand classification results and suggest potential directions during the development of a Deep Neural Networks model.

## Abstract

Recently, the techniques based on Deep Neural Networks (DNNs) have achieved a great performance in classification tasks in a wide range of applications, such as image recognition and natural language processing. However, DNN developers face a lot of trial and error during the development process and spend their efforts in developing their network model through analyzing and understanding the classification results. As such, tools are needed that help the developers not only understand the results, but also suggest the ways to improve their model. In this paper, we propose a visual analytics tool for visualizing the classification results during the iterative development pipeline of a DNN model. Our tool enables exploring the classification results from any type of neural network models, identifying misclassified samples, examining the predicted score distributions of samples, and showing how the outcomes progressively change during the training process.

**Index Terms:** Visualization, classification, deep neural networks

## 1 Introduction

Classification is one of the major tasks in a wide range of data analysis tasks. Recently, the techniques based on Deep Neural Networks (DNNs) have achieved a great performance in classification tasks for a wide range of applications, such as image [9], video [7], text [8], and natural language [3]. However, DNN developers face a lot of trial and error to develop a satisfying DNN model. During the trial and error process, they spend their efforts in developing their network model through analyzing and understanding intermediate experimental results. As such, tools are needed that help them not only understand the outcomes, but also suggest the ways to improve their model.

In this paper, we propose a visual analytics tool to understand classification results and suggest potential directions during the iterative development pipeline of a DNN model. Our visual analytics tool allows users to explore the classification results from any type of neural network models, to identify misclassified samples, to examine the predicted score distributions of samples, and to show how the outcomes progressively change during the training process. Eventually, our visualization helps the users understand, diagnose, and improve DNN models [10]. Also, our visual design can be applied to a wide range of data classification tasks in deep learning. In this paper, we focus on the task of classifying clinical pathology reports by DNNs. This work is in progress now. So, we focus on introducing the current system and showing preliminary results.

*e-mail: chaej@ornl.gov
†e-mail:gaos@ornl.gov
‡e-mail:ramanathana@ornl.gov
§e-mail:csteed@ornl.gov
¶e-mail:tourassig@ornl.gov

## 2 Related Work

The technologies based on DNNs have been used widely in data science, since DNNs have shown great promise in image recognition [9]. Also, a number of researchers have proposed new models beyond the state-of-the-art techniques. In this section we introduce some previous studies related to this work. We discuss how visualizations have played in development of DNN models and the differences between our work and the previous work.

Although DNNs make great achievements, understanding of what computations performed at intermediate layers of DNNs is still limited. Some previous studies have developed tools that help understand the processes in DNNs through visualization techniques. They have proved such tools can help DNN developers improve their network models [10, 20, 24]. Here, we discuss the roles of visualization with regard to the following three topics:

**Understanding of the Underlying Processes in Networks:** The most important purpose of visualizations for DNNs is to understand the underlying processes in networks. The visualizations can be classified into two categories according to which aspects of the network are visualized: **Inner Process** and **Learned Features**.

- **Inner Process:** Visualizations in this category directly represent the interactions between neurons in a network. These visualizations enable intuitive understanding of the inner processes in the networks. Tzeng and Ma [20] employed a directed acyclic graph (DAG) to show the interactions between neurons in the three different layers: input, hidden, and output. The visualization, however, has a visual clutter issue when handling a large number of neurons. Liu et al. [10] also used a similar type of visualization, but they mitigate the issue. They clustered layers and neurons in networks, selected representative layers and neurons, and illustrated those by a hybrid visualization. Harley [5] developed an interactive and intuitive visualization system for DNNs. Given an input image drawn by an user, the visualization system shows not only the features learned by networks, but also the behavior (i.e., activation and interaction) of neurons and layers.

- **Learned Features:** This type of visualizations focuses on visualizing the features learned by networks rather than the behavior in networks. Since Convolutional Neural Networks (CNNs), a specific type of DNNs, have been widely used for image classification, many visualization approaches have been proposed to understand how images are classified by CNNs [4, 13, 24, 25]. The visualizations display synthetic images produced by gradient-based techniques including deconvolution, code inversion, activation maximization—these techniques are based on high activations (learned features) of neurons of a network. The images (although they do not look natural) would provide insight into the network model and can suggest potential directions to improve the model. In addition to CNNs, Recurrent Neural Networks (RNNs) are getting attention since RNNs have achieved a great performance for sequential data, such as streaming, speech, and document. Strobelt [19] proposed a visual analytics system to understand a RNN-based model, where the system helps users explore hidden states in RNNs and find similar patterns in data.

The previous studies have made great achievements and shown the importance of visualization in development of DNNs. However, they still have limitations. For the clustering approach, when handling a large network, finding an appropriate abstraction level of networks would be a challenge. Also, the DAG-based visualizations have a limitation in supporting other types of networks beside CNN-based networks. For visualizing learned features, when the number of samples and classes is huge, it would be not easy to gain insights into the networks by analyzing the learned features of every single

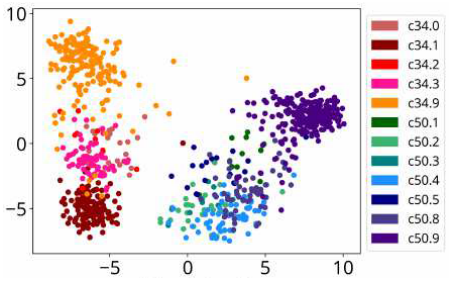


Figure 2: 2D-embedding of cancer pathology reports using PCA. The colors of the points denote their classes.

samples. Although our approach is close to the second category, we focus on visualizing and understanding classification results of any types of DNNs.

**Visualization for Classification Result:** Visualization enables a better understanding of the classification results of neural networks. Especially, when handling a large size of samples, suitable visualization techniques are required. Most techniques have used a 2D-embedding type of visualization (Figure 2) to represent the classification results. An output layer of neural networks usually produce predicted scores for each sample across all classes. They project the prediction scores on a 2D space by dimensionality reduction techniques, such as t-SNE, PCA, etc [6, 11, 14, 16]. This type of visualizations has a visual clutter issue caused by the large number of data points. If the number of classes is large, the issue will become even worse. Also, the visualizations are not able to reveal the assigned predicted score distribution over the classes. Our visual representation solves these issues. Our stripe visualization has a better scalability, resolves the visual clutter issue, and effectively shows the probability distribution by integrating a parallel coordinated visualization into our tool. Details are described in Section 5. Ren et al. [17] proposed a visualization for performance analysis in machine learning. Their major visual metaphor and the way they use that metaphor is similar to our approach. In this paper, our visualization tool focuses on the understanding of classification results rather than performance analysis. Besides, we visualize how the outcome changes, which samples are continuously misclassified, and what features are learned by the neural networks as the training progress. This provides opportunities to improve the model.

**Direct Manipulation:** Few previous studies proposed approaches that support direct manipulation of neural networks and a real-time visualization of the networks [1, 18]. Both Playground [18] and ReVACNN [1] provide similar features. They support real-time visualization of how neural networks are trained and real-time model steering—selecting filters, adding/removing neurons/layers, and adjusting weights. The tools help users gain an intuition about deep neural networks. Playground, however, is designed for an educational purpose rather than real-world applications.

## 3 Design Requirements

Our hypothesis is that the understanding of classification results can be effective to improve a neural network model. In the real-world, the application domain for classification can be very specific, for example, medical records, CT scan images, scientific images. In most cases, however, neural network developers would be not experts in a specific application domain. If the model developers have no background knowledge about the data they deal with and even they cannot expect any classification outcomes, starting with the understanding of data and classification outputs using our visualization tool should help them improve their network model for a specific classification task. Through this preceding step, they can more clearly understand the inner working and the learned features

of their neural networks than without the step.

We discussed visual design requirements for understanding classification with deep learning experts many times. We realized that they need not only basic visualizations for classification, but also specific requirements for their DNN models. The basic visual design needs to support the following requirements: *exploration of the classifications*, *a summary of each class*, and *a detail view of each sample*. In addition to these basic requirements, we identified the following specific visualization requirements for supporting the aspects of DNNs.

- **R1: Showing training result changes as a training progressed.** Visualization needs to allows experimenters to monitor the accuracy and outcome of a network model during its training process. This allows to find a best performance point or stop the network from over-fitting [22]. Also, the experts need to identify the misclassified or ambiguously classified samples.

- **R2: Examining classification probability distributions.** Neural networks usually produce score distributions over all classes. The class with the highest score is selected as the predicted class. It is important to visualize the score distribution. For example, they need to see the second and third highest probability of classification of a misclassified or ambiguous data set to find possible solutions and refine their network model.

- **R3: Revealing the features of data learned by neural networks.** The experts want to view the learned features of each training sample. They examine the features to discover the issues of their network model. For example, they can recognize noise in training data that highly affects the networks and then handle the noise to refine their model.

## 4 Background

In this work, we focus on classification of clinical pathology reports. Clinical pathology reports contains highly valuable information. However, many experts have manually classified the huge volume of pathology reports and extracted information from the reports. Also, the reports are usually unstructured and have highly varied formats because they are generated from hundreds of different medical facilities and providers.

Recently, Deep Learning (DL) based approaches in Natural Language Processing (NLP) have been applied for analysis of pathology reports and health records [12, 15, 23]. Yoon et al. [23] proposed a multi-task learning model to extract important keywords from cancer pathology reports. Moitto et al. [12] presented clinical predictive model using a unsupervised DL to derive patient representations that improve clinical prediction and decision system. Recurrent Neural Networks (RNNs) have achieved a great performance in NLP tasks. Yang et al. [21] developed a Hierarchical Attention Network (HAN) which is based on RNNs. They designed the networks to capture a hierarchical structure of documents (words form sentences, sentences form a document). HAN contains two levels of attention mechanisms to extract important words and sentences in a document. In this work we use a set of model snapshot data resulted from the hierarchical attention layers and the final softmax output layer in HAN.

## 5 Visualization

Our visualization tool as shown in Figure 1 satisfies not only the basic visual design requirements, but also the motivated specific requirements mentioned in Section 3. The tool consists of the following components:

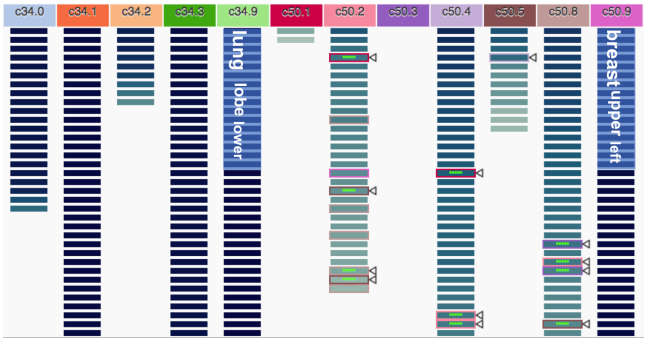


Figure 3: Classification View: Samples (small narrow boxes) are visualized according to their predicted classes. The box colors represent their predicted scores. Outlined boxes are incorrectly predicted samples. Small triangles denote the samples whose the misclassified number is more than mis-prediction threshold value.

- *Training Accuracy Graph:* This standard line graph (Figure 1 (A)) shows the accuracy changes of train and validation sets according to each epoch (R1).

- *Classification View:* The classification view (Figure 1 (F)) visualizes each categorized samples according to the predicted classes and the calculated predicted scores for a selected epoch (R1).

- *Predicted Probability Distribution View:* This parallel coordinate visualization (Figure 1 (E)) reveals the predicted score distributions over classes of selected samples by users (R2).

- *Detail View:* Our detail view (Figure 1 (D)) shows sample's features learned by neural networks, for example important sentences and words (R3).

The following subsections describe these components in detail.

### 5.1 Classification View

Our classification view is to explore classification results and to identify misclassified samples. For each sample, neural networks usually produce predicted scores across all classes. The class with the highest score is selected as the predicted class. The columns in the view (Figure 3) represent the classes and each column head has its class name and an unique color. Under the heads, each narrow box corresponds a sample (a pathology report in this paper) and the boxes are lined along their predicted class columns and ordered by their scores. For each class, the stacked boxes as a whole is a bar that is a suitable visual variable for comparing quantity (the number of samples for each class) [2]. A sample is selected by users, we add green markers inside the box. Once users change the epoch using the slider control (Figure 1 (B)) and then the classification view dynamically updates the result corresponding to the current selected epoch.

The solid boxes for each column represent samples correctly predicted while the outlined boxes represent samples incorrectly predicted. The fill color of each box indicates the predicted score assigned to the predicted class. The score is varied from 0 to 1 and the sum of the scores is 1. In other words, the boxes with the dark blue color are predicted with high confidence while the boxes with the light blue color are predicted with low confidence. This allows to inspect the samples that are ambiguously classified—its predicted scores are evenly distributed across all classes. In addition, the color gradient of stacked boxes as a whole shows that how confidently the model classifies the samples for the corresponding class. It possibly suggests the directions to increase the accuracy of the model. The

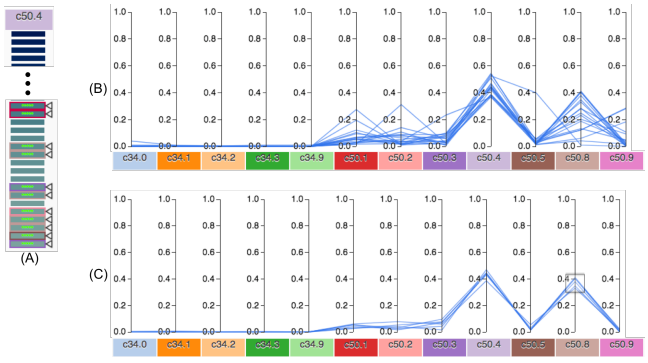


Figure 4: Selected samples in the class, c50.4 (A). The predicted score distributions of the selected samples (B). Filtering the samples with its score of the class, c50.8 over 0.3 (C). The model confuses the two classes: c50.4 and c50.8

outline color of a misclassified sample indicates the class which the sample is labeled as. The misclassified samples usually have relatively low confidence as shown in Figure 3. Also, we provide a different type of visualization for the classification view as an alternative option as shown in Figure 6. This type can handle more number of samples and is useful when we use a small size of display.

If the number of incorrect prediction for a sample is greater than the value of the mis-prediction threshold slider (Figure 1 (C)) as epochs progress, we indicate the samples with the small triangles. This allows to see that what samples are continuously mis-predicted during the training process. Users can examine the samples using the detail view to find reasons for misclassification.

Also, we extract a set of words considered to be important by HAN from samples for each class and allow to display those over the column by user selection (This function is not completely implemented yet). The font size of each word indicates its importance. For example, we display a set of three keywords for each of the two selected columns: *c34.9* and *c50.9* as shown in Figure 3. We can see what are the major keywords of each class even though we have no background knowledge on the reports and classes.

### 5.2 Predicted Probability Distribution View

Our visual analytics system utilizes a standard parallel coordinate visualization to enable examining of the predicted score distributions over classes for the selected samples. As we mentioned previously, to increase the classification accuracy, it is important to know what samples are incorrectly predicted and how their predicted scores are distributed. We show an example case in Figure 4. We select the samples that are predicted as a class (c50.4) with low confidence (Figure 4 (A)). The numbers of incorrect prediction of the selected samples are greater than the mis-prediction threshold value set as 10. We can see their predicted score distributions as shown in Figure 4 (B). We filter the samples whose scores assigned to another class (c50.8) are over 0.3 as shown in Figure 4 (C). We can see that the model confuses the two classes, *c50.4* and *c50.8*. This can provide an opportunity to investigate why the model cannot clearly distinguish between the two different classes.

### 5.3 Detail View

Finally, our detail view visualizes the features of each sample learned by the neural networks. An example is shown in Figure 5. We utilize the Hierarchical Attention Network (HAN) described in Section 4 to extract important words and sentences from a huge volume of pathology reports and to classify the reports. For each report, we visualize the words and the lines (sentences) with their importance in the detail view. The dark red and dark blue colors denote highly important lines and words, respectively. On the other hand, the light blue and red colors denote little important lines and words. In the example view in Figure 5, we can see the highlighted important words and lines which usually include the important words. In this paper the detail view supports only a document type of data, but it is possible to visualize the learned features of other types of data. For example, in image classification, we can highlight the learned features on an image using other feature detection techniques [26]. We leave this as a future work.

## 6 Conclusion

In conclusion, we described our visual analytics tool for understanding of classification results in Deep Neural Networks (DNNs). Our visual components of the tool allows DNN experts to track the accuracy of their model, explore the classification results, examine the predicted score distributions of samples, and see the learned features of samples. Eventually, our visualization helps the experts understand and diagnose the DNN model and suggest potential directions to improve the model. This tool still has limitations in understanding inner process of DNNs and handling and visualizing a huge volume of samples. For future work, we will investigate the ways to understand inner transactions of neurons and layers in DNNs and improve the current visual design to cover bigger size of data.

## Acknowledgments

T his work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE- AC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725. This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration.

## References

[1] S. Chung, S. Suh, C. Park, K. Kang, J. Choo, and B. C. Kwon. Revacnn: Real-time visual analytics for convolutional neural network. In *ICML Workshop on Visualization for Deep Learning*, 2016.

[2] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[3] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[4] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, June 2009.

[5] A. W. Harley. An interactive node-link visualization of convolutional neural networks. In *International Symposium on Visual Computing*, pp. 867–877. Springer, 2015.

[6] M. Kahng, P. Andrews, A. Kalro, and D. H. Chau. Activis: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics*, 24(1), 2018.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

[8] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[10] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. Towards better analysis of deep convolutional neural networks. *IEEE transactions on visualization and computer graphics*, 23(1):91–100, 2017.

[11] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[12] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094, 2016.

[13] A. Nguyen, J. Yosinski, and J. Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. In *ICML Workshop on Visualization for Deep Learning*, 2016.

[14] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim. An approach to supporting incremental visual data classification. *IEEE transactions on visualization and computer graphics*, 21(1):4–17, 2015.

[15] J. Qiu, H. J. Yoon, P. A. Fearn, and G. D. Tourassi. Deep learning for automated extraction of primary sites from cancer pathology reports. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–1, 2017. doi: 10.1109/JBHI.2017.2700722

[16] P. E. Rauber, S. G. Fadel, A. X. Falcao, and A. C. Telea. Visualizing the hidden activity of artificial neural networks. *IEEE transactions on visualization and computer graphics*, 23(1):101–110, 2017.

[17] D. Ren, S. Amershi, B. Lee, J. Suh, and J. D. Williams. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics*, 23(1):61–70, 2017.

[18] D. Smilkov, S. Carter, D. Sculley, F. B. Viégas, and M. Wattenberg. Direct-manipulation visualization of deep networks. In *ICML Workshop on Visualization for Deep Learning*, 2016.

[19] H. Strobelt, S. Gehrmann, B. Huber, H. Pfister, and A. M. Rush. Visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1), 2018.

[20] F.-Y. Tzeng and K.-L. Ma. Opening the black box-data driven visualization of neural networks. In *Visualization, 2005. VIS 05. IEEE*, pp. 383–390. IEEE, 2005.

[21] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy. Hierarchical attention networks for document classification. In *HLT-NAACL*, pp. 1480–1489, 2016.

[22] L. Yeager, G. Heinrich, J. Mancewicz, and M. Houston. Effective visualizations for training and evaluating deep models. In *ICML Workshop on Visualization for Deep Learning*, 2016.

[23] H.-J. Yoon, A. Ramanathan, and G. Tourassi. Multi-task deep neural networks for automated extraction of primary site and laterality information from cancer pathology reports. In *INNS Conference on Big Data*, pp. 195–204. Springer, 2016.

[24] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. In *ICML Deep Learning Workshop*, 2015.

[25] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

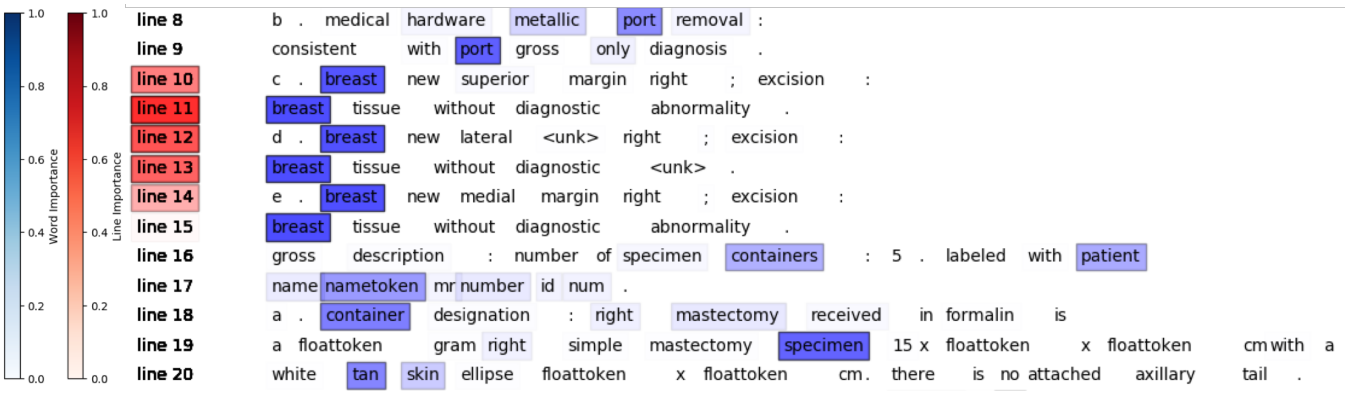


Figure 5: Detail View: Visualizing important words and sentences in a pathology report.

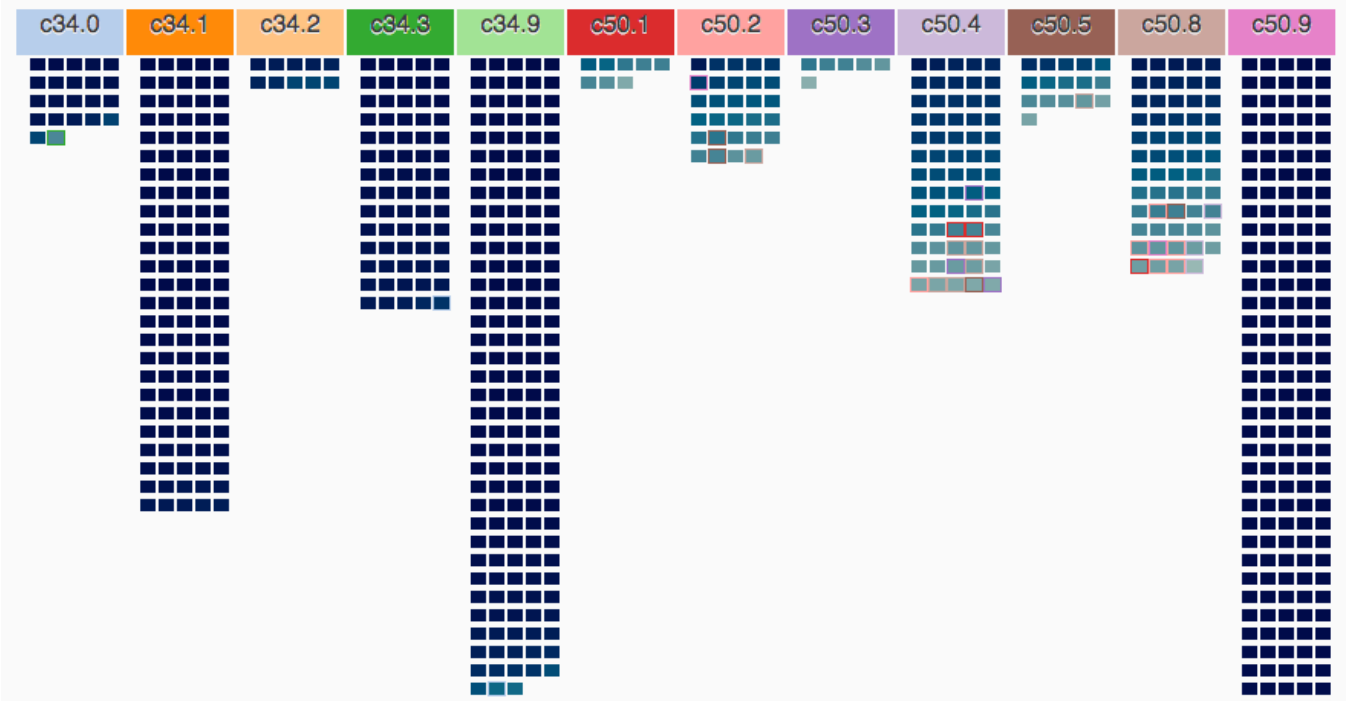


Figure 6: Grid-based classification view. This type can handle more number of samples.