A Comparison of Machine Learning Methods in the Analysis of Lymphocyte Patterns in Cancer Research

Preethi Krishnamoorthy

A comparison of machine learning methods in the analysis of lymphocyte patterns in cancer research

Abstract

Cancer is a widespread disease that impacts many people. Tumor infiltrating lymphocytes (TILs) are important biomarkers for cancer that can be used for diagnosis and treatment. Current methods of cancer diagnosis, including manual evaluations of tumors by pathologists, are very time consuming. To make this process more efficient and effective, deep learning methods are being used in cancer diagnosis. In this research, we compare the effectiveness of machine learning techniques in classifying lymphocyte probability maps into immune subtypes. First, the dice coefficient was implemented to compare two sets of probability maps generated from two different networks. Then, several machine learning methods were developed—a fully connected network, a CNN, modified pretrained models, Random Forest, SVM, and K means clustering. These methods classified lymphocyte probability maps and cluster indices. Dice values between the sets of probability maps were high, showing good agreement between them. The machine learning methods implemented in this study had an accuracy ranging from 50-60%. Random Forest classification had the highest accuracy, with and accuracy of 63.5%. This research shows a strong proof of concept for future studies, and with a greater amount of data, this research can be continued and higher accuracies can be obtained to bring machine learning models as a tool in mainstream cancer diagnostics and analysis.

1 Introduction

1.1 Rationale

Cancer has a major impact on many people's lives all around the world. In the United States alone, approximately 38.4% of men and women will be diagnosed with cancer at some point during their lifetimes [1]. Many cases of cancer are either misdiagnosed or diagnosed very late, so it is important that we develop a better understanding of cancer to be able to diagnose patients with more accuracy and develop better treatments for patients [2].

In order to fight cancer, we first need to develop a better understanding of cancer; one way of doing this is by using new sources of information on cancer and the tumor to expand our knowledge of cancer. The FDA recently approved the use of whole slide images (WSI) for primary diagnostic use, which is a new and important source of data that can be used in order to further our understanding of cancer. WSI contain a lot of data that can be extracted and used to diagnose, evaluate, and treat cancer. Tumor infiltrating lymphocyte (TIL) information is found in WSI of tumors, and can be used to further analyze tumors [3].

TILs play an important role in the study, diagnosis, and treatment of cancer. TILs are a type of immune cell that has moved from the blood into a tumor to try and attack the cancer. Lymphocyte patterns and distribution in tissue specimens and their relation to tumor regions is very important, as studies have shown that tumor infiltrating lymphocyte patterns can be used as biomarkers to predict disease outcome and response to treatment [4]. Utilizing this information obtained from whole slide images can expand our knowledge of cancer and provide a better way of evaluating tumors in order to diagnose and treat patients in better ways.

One powerful tool that has begun to be used is artificial intelligence and deep learning and machine learning methods in this field. AI can be used in conjunction with both numerical and image TIL data obtained from whole slide images to analyze and evaluate tumors for better cancer diagnosis and treatment.

1.2 Background

Cancer plays a major role in our lives, and in order to fight cancer, we first need to develop a better understanding of cancer. One way of doing so is through the classification of patients based on certain characteristics of the tumor. Immune subtype classification is a method recently developed that sorts cancer patients based on immune expression signatures which are correlated with other important features of tumors. These criteria the patients are classified on are very important to further understand a patient's individual case and develop treatments for patients. There are six different immune subtypes—C1, C2, C3, C4, C5, and C6. Immune subtype C1 is wound healing, C2 is IFN-g dominant, C3 is inflammatory, C4 is lymphocyte depleted, C5 is immunologically quiet, and C6 is TGF-b dominant [5].

Immune subtypes are an important way of classifying patients, and it is important we come up with faster, more efficient ways to categorize patients into these subtypes, as the methods used in the paper (ex. exome sequencing, DNA methylation) were time consuming and expensive [5]. To do so, we can use two tools—tumor infiltrating lymphocyte (TIL) information and machine learning models.

Tumor infiltrating lymphocytes (TILs) are immune cells that have moved from the blood to a tumor to try to attack the cancer. The role of TILs in cancer diagnosis and treatment has become more important, as these lymphocytes can provide important information about the cancer's prognosis and can help predict patient responses to treatments [6]. Increased levels of TILs in women receiving neoadjuvant chemotherapy were associated with improved prognosis in HER2-positive and triple-negative breast cancers but poorer outcome in luminal HER2-negative breast cancer [4]. Additionally, TILs can be used as biomarkers to reflect immune response to tumors [7]. Being able to utilize information about lymphocyte distribution when evaluating images of tumors is important, as it can provide valuable information about a patient and how patients will react to certain treatments.

Contemporary digital microscopes can capture high resolution images of tissue specimens, called whole slide images. Current methods of evaluating and analyzing WSI include manual evaluations of tumors by pathologists in order to diagnose cancer and determine treatments. These reports include information about the tumor's appearance and tumor type and grade [8]. However, these methods are inadequate because this method is time consuming, labor intensive, and imprecise, and unable to evaluate complex TIL distribution patterns. The use of computational tools to augment human evaluation would allow us to unlock the full potential of information in TIL distributions contained in WSI to help patients.

Recently, Saltz et al utilized whole slide images of tumors from a range of cancer types to generate TIL maps using a deep learning model. These maps were reviewed and refined by a group of pathologists. Then, five different summary statistics were generated from the probability maps—ball hall, c index, Banfeld Raftery, determinant ratio, and leukocyte fraction. Each index is calculated in a different way from the probability maps and represents different information about TIL distribution and other things in the WSI [3].

In recent years, machine learning and deep learning have emerged as powerful tools for data analysis. Machine learning is a subfield of AI which uses algorithms to learn from data. Deep learning is a subfield of machine learning that uses deeper and more complex artificial neural networks. Deep learning-based analysis methods have demonstrated impressive results in image classification and segmentation compared to other machine learning approaches. Several deep learning models (such as ResNet, VGG, and Inception) have been developed by the imaging and computer vision communities [3, 9, 10]. Previously existing networks, such as Inception_V3, have been adapted to analyze images of breast cancer images [11]. Deep learning is rapidly becoming a core image analysis tool in biomedical imaging research, and in conjunction with TIL data from whole slide images, could serve as a powerful tool for cancer analysis, diagnosis, and developing treatments [12].

1.3 Objective

The objective of this research is to develop various deep learning and machine learning models to classify patients into immune subtypes C1 and C2. To do this, TIL probability maps and summary statistics from tumors of various cancer types were used with deep learning and machine learning methods.

2 Methodology

2.1 Data

The probability maps and summary statistics were previously generated by Saltz et al. and Abousamra et al. [3,13] from de-identified publicly available images of cancer tissue from The Cancer Genome Atlas by using three different neural networks—a customized convolutional neural network [3] and Inception V4 and VGG16 [13]. Figure 1 is an image of a probability map.

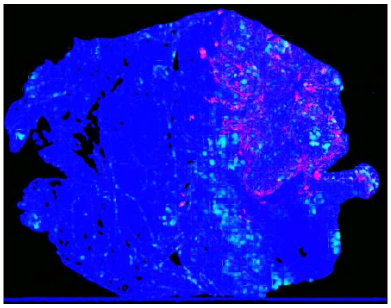

Figure 1: TIL probability map

2.2 Basic workflow

2.2.1 Prior Work

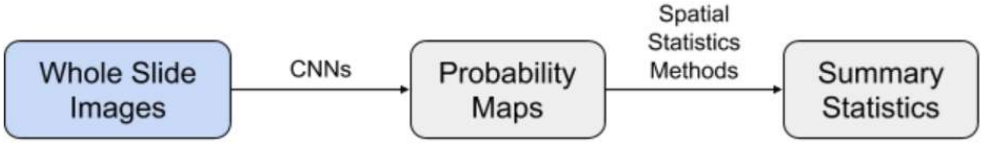

Figure 2: Prior Workflow

Figure 2 is a diagram of the work done by Saltz et al. and Abousamra et al. [3,13]. Whole slide tissue images of cancer were first used in three different CNNs—Inception, VGG, and Cell Reports—to generate probability maps. Summary statistics were then generated from these probability maps.

2.2.2 Current work

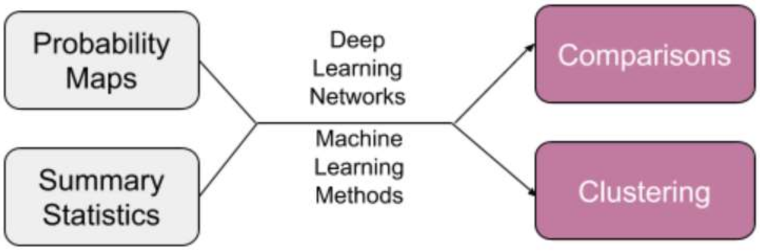

Figure 3: Current Workflow

Figure 3 is a diagram of the work done that is presented in this paper. The probability maps and summary statistics generated in prior work by Saltz et al. and Abousamra et al. [3,13] were then used for clustering using deep learning and machine learning methods. The probability maps were also used to compare the networks used to generate them (Inception, VGG).

2.3 Algorithms for classification

2.3.1 Fully connected network

First, a fully connected network with three dense layers was made using Keras. This network classifies images into two categories—immune subtypes C1 and C2. TIL Probability maps were used to train and test the network. 80% of the images were used to train the network, and 20% of the images were used to test the network.

2.3.2 CNN

Next, a convolutional neural network using Keras was made. This network has a Conv2D layer, a maxpooling layer, three Conv2D layers, one flatten layer, and two dense layers. 80% of the images were used for training, and 20% were used for testing. To test this network, color and greyscale images with a resolution of 224x224 were used. This network separated data into two immune subtypes: C1 and C2.

2.3.3 Pretrained Networks

Pretrained networks were modified and implemented using Keras. The networks used were VGG16, VGG19, ResNet50, Inception V3, MobileNet and DenseNet121. A final layer was added to these networks so that the network could classify the image into one of two immune subtypes. All networks took color images that had a size of 224x224, and 80% of the images were used for training, and 20% were used for testing.

2.3.4 Random Forest

Random forest is a classification method that combines many decision trees and finds the classification that occurs most often in the decision trees, and uses it as the final classification [14]. Figure 4 is a diagram of a generic random forest classifier.

Random Forest was implemented with scikit-learn used to classify patients into immune subtypes. Five different indices were used in random forest classification: ball hall, c index, Banfeld Raftery, determinant ratio, and leukocyte fraction. 80% of the data was used for training, and 20% was used for testing. Each set of data was sorted into one of two immune subtypes.

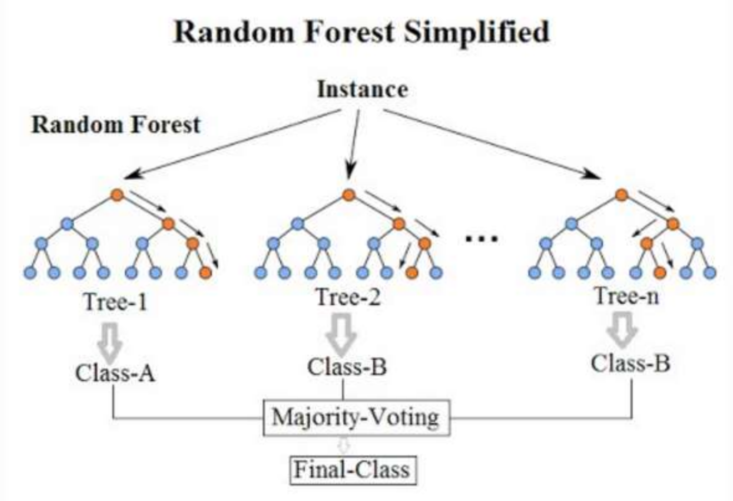

Figure 4: Diagram of generic random forest classifier [14]

2.3.5 SVM

SVM is a classifier that outputs a hyperplane that classifies new examples [15]. Figure 5 is a sample SVM classifier diagram with a hyperplane separating the two groups.

SVM classification was implemented with scikit-learn used to classify patients into immune subtypes. Five different indices were used in SVM classification: ball hall, c index, Banfeld Raftery, determinant ratio, and leukocyte fraction. 80% of the data was used for training, and 20% was used for testing. Each set of data was sorted into one of two immune subtypes.

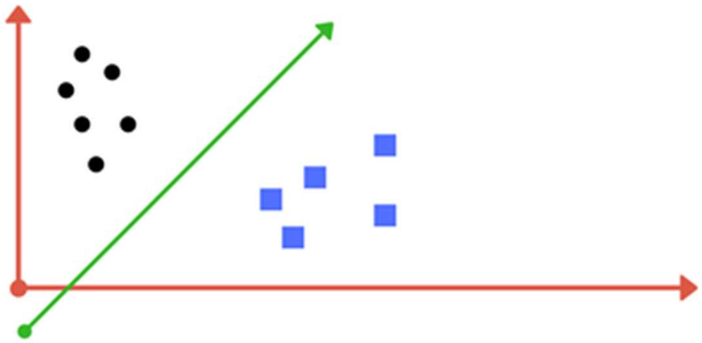

Figure 5: Sample SVM classifier [15]

2.3.6 K means clustering

K means clustering was implemented using the four indices previously stated. Each data point was put into one of five clusters. These clusters were then compared with the immune subtype of each data point (C1 or C2), and a confusion matrix was created. The rand index was then calculated to determine the accuracy of this model. Figure 6 is a sample confusion matrix.

		Actual Clusters	
		1	2
K means clustering: generated clusters	1	a	c
	2	d	b

Figure 6: Sample Confusion Matrix

The values a and b are the number of elements that have the same k-means clustering generated clusters and actual clusters. The values c and d are the number of elements that have different k-means clustering generated clusters and actual clusters.

The Rand index is a measure of accuracy of the clustering algorithm. The equation for calculating the Rand index from this confusion matrix is

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

2.4 Dice Coefficient

The Dice coefficient computes similarity between two sets. In our case, given a whole slide tissue image, one of the sets was the set of image patches classified as TIL positive by one deep learning network and the other set was the set of patches classified as TIL positive by the other deep learning network. Dice scores were computed for images generated by the Inception, VGG and cell reports networks. The dice coefficient is calculated using the equation below:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

Here, X and Y are two sets being compared.

3 Results

3.1 Experimental setup

All programs were written in Jupyter. Multiple notebooks were used in this project, with each notebook corresponding to a certain classification method or task. The programs were run on a virtual machine with multiple CPUs.

3.2 Deep Learning models and accuracy

Network	Resolution	Color	Test Accuracy
Fully connected	256 x 256	Color	0.532
Fully Connected	256 x 256	Grayscale	0.502
CNN	256 x 256	Color	0.515
CNN	256 x 256	Grayscale	0.515
Densenet refined	224 x 224	color	0.549
Inceptionv3 refined	224 x 224	color	0.505
Mobilenet refined	224 x 224	color	0.539
resnet50 refined	224 x 224	color	0.539
vgg16 refined	224 x 224	Color	0.549
vgg19 refined	224 x 224	color	0.549

Table 1: Accuracy of deep learning models

The table above shows the test accuracy of different networks with different inputs. The resolution of the images used were 256x256 in the fully connected network and CNN, and the resolution of the images used for the refined models was 224x224. Both color and greyscale images were tested for the fully connected network and CNN, and only color images were tested for the refined models.

3.3 Machine learning models used and accuracy

Model	Indices Used	Test Accuracy
SVM	Ball Hall, C index, Banfeld Raftery, Det Ratio, Leukocyte Fraction	0.567
Random Forest	Ball Hall, C index, Banfeld Raftery, Det Ratio, Leukocyte Fraction	0.635

Table 2: Accuracy of Machine Learning Models

The table above shows the test accuracy of two different machine learning models—Random Forest and SVM. Ball Hall, C index, Banfeld Raftery, Det Ratio, and leukocyte fraction were the indices used for classification.

3.4 Clustering results

		Immune Subtype	
		C1	C2
K-means clustering generated clusters	1	6	16
	2	102	79

Table 3: Confusion Matrix of K-means clustering results

The confusion matrix above shows the number of elements in each cluster that correspond to a certain immune subtype. The columns are immune subtypes and rows are clusters created by the algorithm.

Rand index: 0.576

3.5 Comparison of machine learning and deep learning models' accuracy

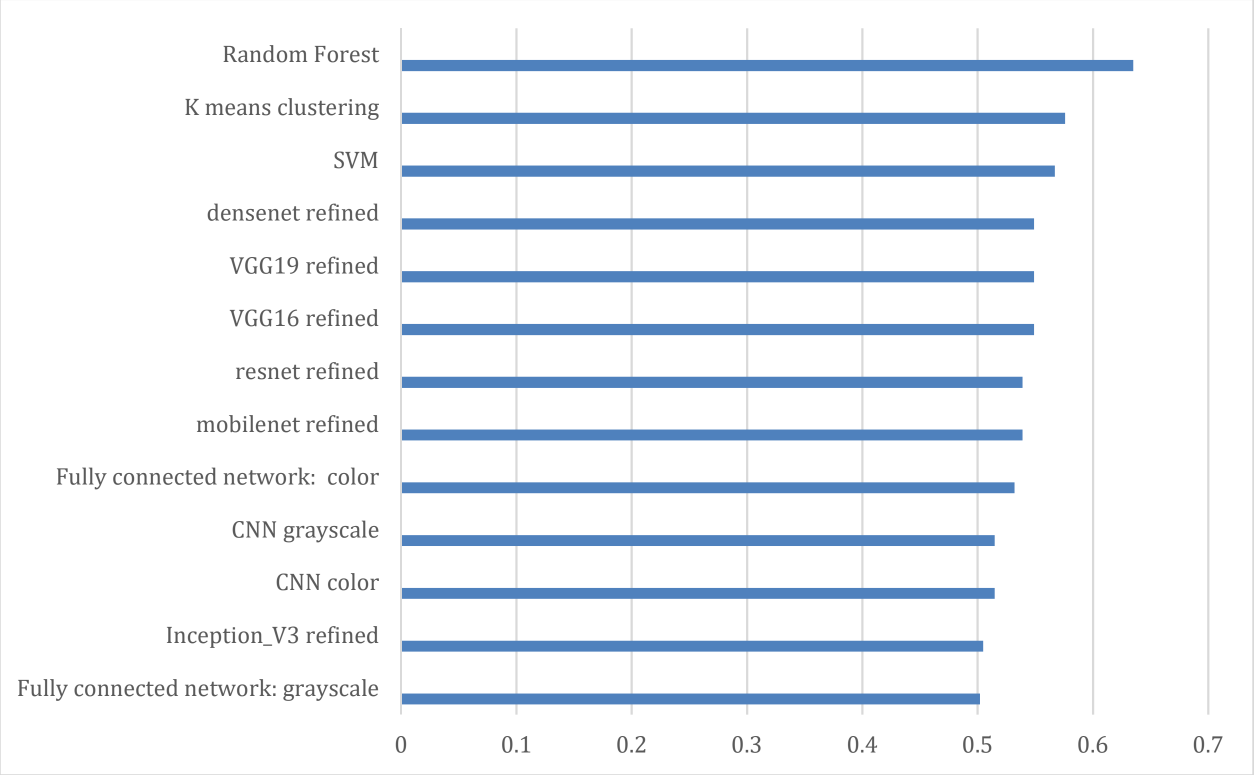

Figure 7: Comparison of classification accuracies of machine learning methods

The graph above shows the accuracies of the different models in decreasing order. Random Forest is the most accurate model.

3.6 Dice coefficient results

Probability maps generated by Inception and VGG were compared using the dice coefficient. A dice coefficient value of one indicates that the images are the same, and a value of zero indicates that the images are completely different.

Mean dice coefficient value: 0.7377412020052028
Median dice coefficient value: 0.7736537223771722

4 Discussion and Limitations

4.1 Discussion

There were three different classification methods used—deep learning, machine learning, and clustering. These methods classified probability maps and different indices into different immune subtypes. All methods had roughly the same accuracy (50%-60% for methods classifying between two groups, and 20% for methods classifying between 5 groups).

The dice coefficient was used to compare the similarity of two different sets of data (probability maps generated by Inception, and probability maps generated by VGG). The dice coefficient showed that both sets of data were similar.

4.1.1 Deep Learning models

All deep learning networks classified probability maps into two immune subtypes: C1 and C2. The accuracies ranged from 50% to 55%, with VGG having the highest accuracy.

The fully connected network had similar accuracies regardless of the color of the input, and the CNN had the same accuracy for both color and grayscale input, showing how the architecture fo the network has a greater impact on the accuracy of the network.

Pretrained models were them modified and trained with probability maps from the Inception network with a resolution of 224x224 in color. Because these pretrained models were already trained with other data, we hypothesized that the test accuracy would be increased. However, this was not the case as all the refined networks has roughly the same accuracy as the CNN and fully connected network.

Overall, the test accuracies were around 50%-55% when classifying images into one of two categories. However, pretrained models performed better than the fully connected network and CNN. This could mean that the pretrained weights and biases led to a more accurate classification.

4.1.2 Machine Learning Models

SVM and Random Forest both classified images into one of two immune subtypes. They used four different indices (Ball Hall, C index, Banfeld Raftery, Det Ratio, Leukocyte fraction) to classify each point into an immune subtype. Random forest had the highest accuracy, with an accuracy of 64%. This accuracy is very high for the amount of data available for this research, and is a strong proof of concept for future research.

4.1.3 Clustering

The clustering algorithm has a high accuracy as well, with an accuracy of 58%. For the amount of data available, this accuracy is a good start.

4.1.4 Dice Coefficient

When dice values were calculated for sets of probability maps generated by Inception and VGG, the mean dice value was 0.74 and the median dice value was 0.77. The dice values were close to 1, showing good agreement between probability maps generated by both networks. This is also a method of validating the data used—because the probability maps are very similar, there is a greater probability of them being accurate.

4.2 Limitations

One main limitation encountered in this research was the small sample size. We only had about 2500 images to use for both training and testing. For training and testing deep learning and machine learning algorithms, a larger sample size is needed to improve accuracy.

Additionally, there was no data for immune subtype C5 and there was limited data for immune subtypes C3, C4, and C6, so the networks trained cannot recognize other immune subtypes. However, with more data, the networks can be trained with all immune subtypes.

5 Conclusion and Future Studies

5.1 Conclusion

In this research, the effectiveness of various machine learning techniques in classifying TIL probability maps and cluster indices were tested. A fully connected network and CNN were created, and pretrained deep learning models were modified. These methods were trained and tested multiple times with different formats of data (different image resolution, color, and augmentation), and test accuracies were compared. Additionally, machine learning methods (Random Forest, SVM, K means clustering) were tested using cluster indices and compared. Random forest had the highest accuracy, with an accuracy of 64%. With the data available, this is a strong proof of concept and this research can be repeated with more data to get higher accuracies.

5.2 Future Studies

There are many possible future steps to take with this research. One thing we can try next is getting more data (probability maps and cluster indices), and train and test the same models used using a larger dataset. The increase in data will help these methods improve their accuracy. Also, different methods of data augmentation can be applied in order to increase the amount of data, possibly improving the network's accuracy. Additionally, the same methods can be used to classify data into five immune subtypes rather than two. Originally, we classified the data into two subtypes because there was not enough data for immune subtypes C3, C4, and C6. By getting more data for these subtypes, we can classify patients into one of six, rather than two, subtypes.

We can also try to improve the performance of the networks by using different types of input data. By coming summary statistics with probability maps, networks will have more data per patient, which could result in better classification. Additionally, different resolutions and colors can be tested for all the networks used to see if higher resolution or color corresponds to better test accuracies. Additionally, we can use different indices besides the ones used in this research to see if that will improve accuracy.

Another possible area of study is to modify the networks themselves, and to modify the layers. The modification of layers, or adding and removing layers, can improve the efficiency and accuracy of the networks.

These networks can also be used for other classifications besides immune subtypes. It is possible that these deep learning and machine learning methods will be able to classify probability maps and cluster indices into different types of groups better.

References

1. Cancer Statistics. (2018, April 27). Retrieved from NIH: National Cancer Institute website: https://www.cancer.gov/about-cancer/understanding/statistics
2. Siegel, R. L., Miller, K. D. and Jemal, A. (2019), Cancer statistics, 2019. CA A Cancer J Clin, 69: 7-34. doi:10.3322/caac.21551
3. Saltz, J., Gupta, R., Hou, L., Kurc, T., et al. 2018. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell reports, 23(1), pp.181-193
4. Dieci, M. V., Radosevic-Robin, N., Fineberg, S., Eynden, G. V. D., Ternes, N., Penault-Llorca, F., . . . Salgado, R. (2018). Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ: A report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. *Seminars in Cancer Biology*, *52*. https://doi.org/10.1016/j.semcancer.2017.10.003
5. Thorsson, V., Gibbs, D., & Brown, S. (2018). The Immune Landscape of Cancer. Immunity.
6. Hendry, S., Salgado, R., & Gevaert, T. (2017). Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group. *Advances In Anatomic Pathology*. https://doi.org/10.1097/PAP.0000000000000162
7. Badalamenti, G. (2019). Role of tumor-infiltrating lymphocytes in patients with solid tumors: Can a drop dig a stone? *Cellular Immunology*. https://doi.org/10.1016/j.cellimm.2018.01.013

8. Pathology Reports. (n.d.). Retrieved from NIH National Cancer Institute website: https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/pathology-reports-fact-sheet

9. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence 2017 Feb 12.

10. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In Proceedings of International Conference on Learning Representations, 2015.

11. Deep Learning Based Analysis of Histopathological Images of Breast Cancer. (2019). *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2019.00080

12. Bi, W. L., Hosny, A. , Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A. , Allison, T. , Arnaout, O. , Abbosh, C. , Dunn, I. F., Mak, R. H., Tamimi, R. M., Tempany, C. M., Swanton, C. , Hoffmann, U. , Schwartz, L. H., Gillies, R. J., Huang, R. Y. and Aerts, H. J. (2019), Artificial intelligence in cancer imaging: Clinical challenges and applications. CA A Cancer J Clin, 69: 127-157. doi:10.3322/caac.21552

13. Abousamra S, Hou L, Gupta R, Chen C, Samaras D, Kurc T, Batiste R, Zhao T, Kenneth S, Saltz J. Learning from Thresholds: Fully Automated Classification of Tumor Infiltrating Lymphocytes for Multiple Cancer Types. arXiv preprint arXiv:1907.03960. 2019 Jul 9

14. Koehrsen, W. (2017). Random Forest Simple Explanation. Retrieved from Medium website: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

15. Patel, S. (2017). Chapter 2 : SVM (Support Vector Machine) — Theory. Retrieved from Medium website: https://medium.com/machine-learning-101/

chapter-2-svm-support-vector-machine-theory-f0812effc72