Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.47750/jptcp.2023.30.09.010

Cancer Prognosis & Stratification with Sentimental Analysis using Deep Learning and Machine Learning Techniques

R.Yamini[1]*, Shiven Sharma[2], Ayush Sachdeva[3]

[1]Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur-603203, Tamil Nadu, India

[2,3]Final year, B.Tech., CSE, SRM Institute of Science and Technology, Kattankulathur-603203, Tamil Nadu, India

***Corresponding author:** R.Yamini, Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur-603203, Tamil Nadu, India, Email: yaminir@srmist.edu.in

Submitted: 21 February 2023; Accepted: 19 March 2023; Published: 10 April 2023

ABSTRACT

For therapy and monitoring, it is crucial to provide prognostic information at the time of cancer diagnosis. Even while factors including cancer staging, histopathological evaluation, genetic characteristics, and clinical variables might offer helpful prognostic clues, risk stratification still has to be improved. All these data generate defined patterns and those patterns can be examined with the help of Machine Learning and Deep Learning. The most promising algorithm for this use case is artificial neural networks. Decision trees might be used to the best extent as they provide an adequate balance of speed and accuracy. An ideal approach would be through the effective combination of ANN and Random Forests. Ensembling models would also be able to boost the performance of the system. The metrics and scores for the project must be in-scope of the development and at the same time extendable.

Keywords: *Machine learning, deep learning, multiple cancer prediction, data augmentation, analysis, data visualization, decision tree, random forest, artificial neural networks, supervised machine learning, ensemble models*

INTRODUCTION

Prognostic information plays a crucial role in oncology clinical management decisions, including treatment and monitoring. The widely used "TNM" cancer staging method by the American Joint Committee on Cancer

(AJCC) assesses primary tumor size/extent (T), lymph node metastasis (N), and the absence of distant metastases (M), and is commonly utilized for this purpose. The TNM staging approach has been investigated and is effective, but there is still potential for improvement when it comes to various circumstances and data, such as clinical variables [1,2], genetic data [3,4], histomorphologic data, etc. The development of forecasting techniques that take into account features such as tumor grade is underway [5]. This is because tumor histopathology can use computer-assisted image analysis to discover intricate and potentially novel tumor traits that are important for patient survival, thus enhancing patient outcome prediction. Deep learning has recently been demonstrated to be extremely accurate at object recognition [6] and disease diagnosis from medical images [7,8].

Deep learning models have been shown in earlier studies to perform equally well as human specialists in pathology on diagnostic tasks like tumor detection and histological grading [8–10]. Deep learning-based approaches offer a significant advantage over earlier methods that rely on manually constructed features, such as core size/shape, as they do not require any prior assumptions or a large number of well-known characteristics. These approaches have the ability to automatically learn predictive features without relying on a predefined set of variables. However, a limitation of deep learning is that it often requires large annotated datasets to effectively train the models. In the context of histopathology, these models are commonly trained on millions of small image fields extracted from digitized whole-slide images (WSI) of pathology slides, which are annotated with specific features of interest to pathologists.

These photos have thorough handwritten annotations that identify particular features of interest to pathologists. The use of expert input has two significant disadvantages. First off, these annotations are time-consuming for specialists, taking hundreds to thousands of hours to complete each relevant predictive assignment. This constrains our capacity to quickly expand to new applications, such as various malignancies and histological features. Examples of these annotations include identifying the locations of metastatic tumors and designating the correct tumor grade for each region (such as a gland) within the sample [8–10]. The generated image patches for each category of interest can then be made using these annotated regions. Second, annotation directly enforces the correlation between newly acquired morphological characteristics and recognized annotated patterns. This can be particularly challenging when prognostic labels are currently unknown or there is a desire to learn new prognostic features. It focuses on direct learning of morphological features relevant to survival without relying on expert commentary on subjects or areas of interest. Such an approach gives the machine learning model one "global" tag per slide or case. For example B. Sample mutational status or patient clinical result, for instance. Due to the scale of these pictures (about 100,000 by 100,000 pixels at maximum resolution) and the fact that survival-related morphological features can theoretically appear anywhere in WSI, it is difficult to predict clinical outcomes using WSI. It's because we believe it to be achievable. The fabric displayed is a very challenging challenge. A challenging 'weakly supervised' learning problem is created by the large amount of visual data, morphological variation, and unidentified patterns of discrimination.

The Cancer Genome Atlas (TCGA), the biggest publicly accessible database to our knowledge of digitized WSIs combined with clinical and genetic information, has been the source of data for a number of earlier studies employing machine learning and WSIs to solve the survival prediction challenge [11–17]. These earlier studies focused on learning known histologic features [17], employed feature-engineering techniques [13,16], made use of annotated regions of interest [12,18,19], used feature-engineering approaches, and/or created models that directly predict survival for a specific cancer type. By creating an end-to-end deep learning system (DLS) to predict patient survival directly across a range of cancer types and training it on whole-slide histopathology pictures without the use of expert annotations or known elements of interest, we expand on and extend earlier work. We evaluate a convolutional neural network that is directly optimized to extract prognostic characteristics from raw image data, an image subsampling technique, and multiple loss functions to solve the issue of right-censored patient outcomes.
For 10 cancer types from the TCGA, we assessed our DLS's capacity to enhance risk stratification in comparison to the baseline data of TNM stage, age, and sex. For several cancer types, we saw enhanced risk stratification based on model predictions, but it was difficult to assess effect sizes due to the TCGA's sparse case and clinical event data (350–1000 cases and 60–300 occurrences per cancer type). The findings presented here demonstrate the viability of creating low-supervision deep learning models to predict patient prognosis from whole-slide images across a variety of cancer types, but more study is required to fully understand and support the potential of these deep learning applications.

J Popul Ther Clin Pharmacol Vol 30(9):e80–e86; 10 April 2023.

Proposed system architecture

We have used artificial neural networks and random forest algorithms to a great extent along with XGBoost algorithm.

Artificial Neural Networks (ANN)

At least three interconnected layers make up an artificial neural network. Neurons in the input layer make up the first layer. These neurons communicate with lower layers, which then communicate with the terminal output layer with the final output data.

All internal layers are concealed and created by entities that convert the information that is passed from layer to layer in an adaptive manner. The ability of each layer to serve as both an input layer and an output layer enables the ANN to comprehend increasingly intricate objects. These deeper levels are referred to as neural layers collectively.

By allocating weights to the acquired data in accordance with the underlying architecture of the ANN, neural layer units attempt to learn about the data. These rules enable the unit to generate a transformed outcome and deliver it as the output for the following layer.

Backpropagation, a technique used by another learning rule set, enables the ANN to modify its output results in order to correct for faults. Every time an output during the supervised training phase is labelled as erroneous, backpropagation sends data backward. According to the amount of mistake each weight accounts for, it is updated.

To account for the discrepancy between the expected and actual results, we use this inaccuracy to modify the unit connection weights of the ANN. ANNs "learn" how to reduce the likelihood of mistakes and undesired outcomes over time.

Random Forest Algorithm

Random Forest is a widely used machine learning algorithm that falls under the category of supervised learning. It can be applied to various machine learning problems, including classification and regression tasks. The key concept behind Random Forest is ensemble learning, which involves combining multiple classifiers to tackle challenging problems and improve model performance.

Random Forest works by constructing a collection of decision trees, each trained on a different subset of the input dataset. These decision trees are then used to make predictions, and their results are averaged to obtain the final predicted outcome. This ensemble approach helps to improve the accuracy and robustness of the model by reducing the risk of overfitting and increasing the generalization capability.

In classification tasks, Random Forest can predict the class labels of input samples based on the majority vote of the decision trees, while in regression tasks, it can estimate the numerical values based on the average prediction of the decision trees. Random Forest is known for its ability to handle high-dimensional data, noisy datasets, and complex interactions among variables, making it a popular choice in many machine learning applications.

XGBoost

XGBoost, short for Extreme Gradient Boosting, is a distributed gradient boosting library that has been specifically optimized for fast and scalable machine learning model training. It is an ensemble learning technique that combines the predictions of multiple weak models to create a stronger prediction. XGBoost has gained widespread popularity due to its ability to handle large datasets and deliver state-of-the-art performance in various machine learning tasks such as classification and regression.

One of the key features of XGBoost is its effective handling of missing values, which allows it to handle real-world data with missing values without requiring extensive pre-processing. This makes it a practical choice for dealing with real-world datasets that often contain missing information.

Another notable characteristic of XGBoost is its integrated parallel processing capability, which enables it to train models on large datasets quickly. This makes it well-suited for big data scenarios where computational efficiency is crucial.

J Popul Ther Clin Pharmacol Vol 30(9):e80–e86; 10 April 2023.

Overall, XGBoost has become a popular choice among machine learning practitioners and researchers due to its ability to handle large datasets, handle missing values effectively, and deliver high performance in various machine learning tasks, making it a powerful tool in the field of machine learning.

Decision Tree Algorithm

A decision tree is a supervised learning method commonly used for solving both classification and regression problems. It is a tree-structured classifier, where internal nodes represent features of the dataset, branches represent the decision-making process, and each leaf node represents the classification or regression result.

There are two types of nodes in a decision tree: decision nodes and leaf nodes. Decision nodes are used to make decisions and have multiple branches, while leaf nodes represent the final results of the decisions and do not have any further branches.

The features of the given dataset are used to perform tests or make decisions at the decision nodes, leading to different branches and ultimately reaching the leaf nodes that provide the classification or regression results. The decision tree can be visualized as a graphical representation of all possible paths to a decision or solution based on predefined conditions.

The name "decision tree" comes from its tree-like structure, starting from the root node and branching out to subsequent nodes, resembling the branches of a tree. Decision trees are widely used in machine learning and data mining due to their interpretability, ease of use, and ability to handle both categorical and numerical data.

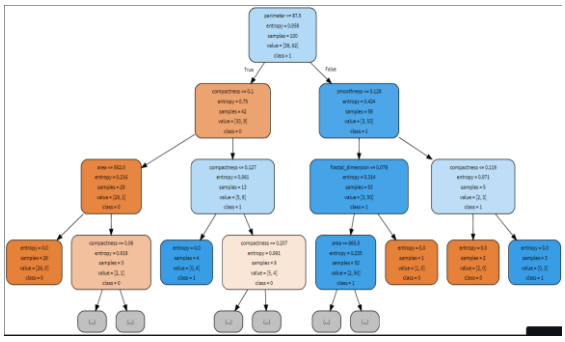

FIG 1: Decision Tree for Cancer Prediction

METHODOLOGY

Hematoxylin and eosin (H&E) stained specimens with digitized full-slide photos can be seen on the Genomic Data Commons Data Portal (http://gdc.cancer.gov) thanks to TCGA [20]. Images from frozen specimens as well as formalin-fixed paraffin-embedded (FFPE) diagnostic slide images were included. Early research and variations in the proportion of FFPE pictures available for various cancer types (i.e., TCGA studies) led to the adoption of cryo-WSI and FFPE as subject-level training and prediction methods for each patient. There were 1–10 slides in each case (median: 2). Clinical information, including an approximation of disease-specific survival, was sourced from Genomic Data Commons and the TCGA Pan-Cancer Clinical Data Resource [21]. We chose the 10 studies with the highest number of patients and survival events from among the TCGA trials with available cancer-stage data. Only serous cystadenocarcinoma (OV) of the ovary was staged clinically; pathological staging data were not available but were included anyway due to a large number of occurrences noted. Cutaneous melanoma (SKCM) was disregarded because it was not restricted to initial tumors that had not been treated [14, 22]. Thyroid cancer (THCA) was ruled out because it manifested in only 14 out of 479 instances. Only instances lacking disease-specific survival were omitted from model creation (training and adjustment), whereas cases with missing illness stage, age, sex, or disease-specific survival data were excluded from evaluation. Comparison of the values observed and expected cancer prognosis.

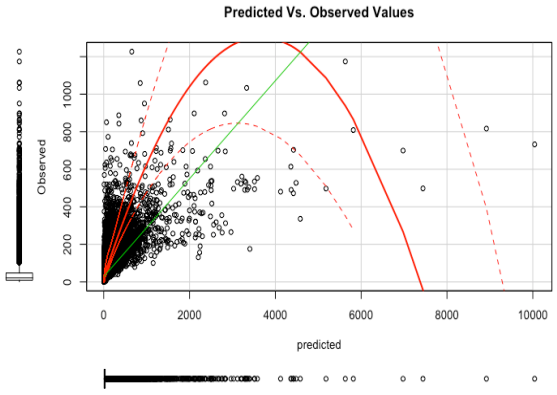

FIG 2: Predicted vs Observed values Cancer Prediction

J Popul Ther Clin Pharmacol Vol 30(9):e80–e86; 10 April 2023.

Multiple convolutional neural networks (CNN) modules with shared weights and an average pooling layer that combines the visual information calculated by these modules make up the core components of our deep learning system, or DLS. (Fig. 1). Our CNN has layers of depth wise separable convolutional layers, which was similar to the CNN design used by MobileNet [23]. By using a random grid search, the number of slices and slice size were matched in each research (see S2 table and S1 method). We selected this architecture family because it has fewer parameters than other cutting-edge CNN architectures, accelerates training, and lowers the possibility of overfitting. Each CNN module used a different set of randomly chosen image spots from the slide as input, so that if numerous locations were sampled, at least one was likely to influence the outcome. In particular, the likelihood of n patches not sampling an informative patch drops exponentially with n if the frequency of informative patches on a slide is p.

Even for modest levels of n, (1-p)n decreases to zero. In light of the weak marker nature of survival prediction in huge images and the unknown locations of information regions within an image or image set, our approach overcomes this issue. Of course, this method can be expanded to include numerous slides for each situation. To further guarantee that informative patches were chosen across training iterations, n patches were randomly selected for each training iteration.

To train the DLS, we experimented with three distinct loss functions. The truncated cross-entropy described below, which is used for training the final model, has been found through preliminary trials (tested with melody splitting) to produce the best results.

On Cox's partial probability, the first-loss function under test was built [24]. In addition to being used to train neural networks, it may also be modified to fit the Cox proportional hazards model:

$$\max \prod_{i:O_i=1} \frac{e^{f(X_i)}}{\sum_{j:T_j \geq T_i} e^{f(X_j)}}$$

For the i-th example, Ti represents the time of the incident or the time of the most recent follow-up examination, Oi the index variable for whether the event was observed, Xi the whole set of slide images, and f(Xi) the DLS risk score. To address bound event times in our approach, we adopted Breslow's [25] estimate. The loss of specific examples is, in theory, a function of every case in the training data. Instead of analyzing the whole training set, we evaluated small batches (n 128) of samples in order to approximation the loss at each optimization step. The exponential lower bound of the coincidence index served as the second loss function [26]. A typical performance indicator for survival models is the concordance index, which measures how likely it is for a randomly chosen pair of subjects to be appropriately ranked by the model with regard to event time. Raykar et al. [27] provided the following differential lower bound that can be used for model optimization despite the coincidence index itself not being differentiable.

$$E := \{(i,j) | O_i = 1 \text{ and } T_j > T_i\}$$

$$\max \sum_{(i,j) \in E} 1 - e^{f(X_i) - f(X_j)}$$

If Tj > Ti and E is the collection of pairs of occurrences (i,j) where the i-th event is observed. We approximate this lower bound on the coincidence index at each phase of the optimization process, analogous to Cox partial probabilities. An illustration of using a short batch (n 128) as opposed to the whole training data set. In order to train survival prediction models using right-censored data, the final loss function, censored cross-entropy, is an extension of the regular cross-entropy loss used for classification models. By discretizing time into intervals and training a model to predict discrete time intervals in which events occur rather than continuous event times or risk scores Did, survival prediction can be modelled as a classification problem as opposed to a regression or ranking problem.

The standard cross-entropy for the observed event example was determined. However, it is unclear when the events in the censored example take place. In order to maximize the log-likelihood that the event occurred during or after the censoring interval, we employ the knowledge that the event did not occur prior to the censoring

time. This is how the whole loss function may be expressed:

$$max \sum_{i} \left[O_i * \log(f(X_i)[Y_i]) + (1 - O_i) * \log\left(\sum_{y>Z_i} f(X_i)[y] \right) \right]$$

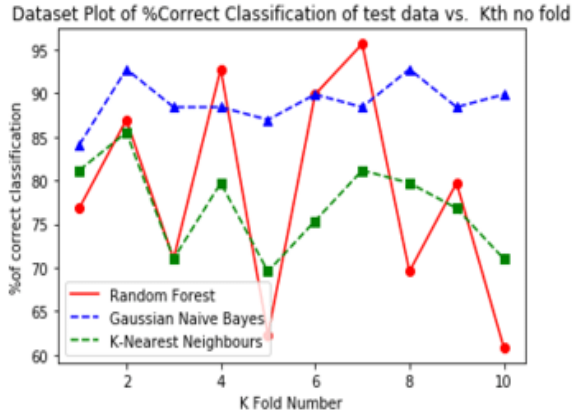

FIG 3: Algorithmic performance for Cancer Prediction

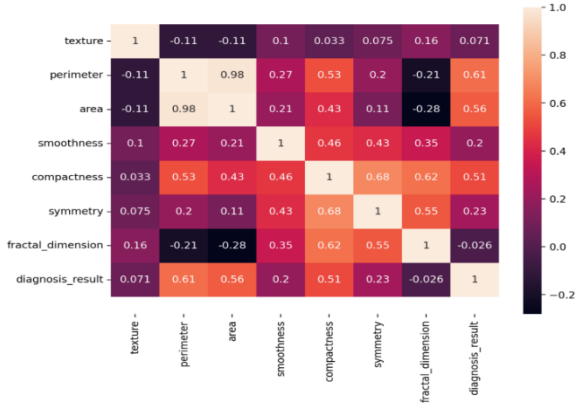

FIG 4: Correlation Heat map for Cancer Prediction

RESULTS AND DISCUSSION

The output of a deep learning-based risk score can be used as a continuous feature in survival analysis. In this study, cases were divided into risk quartiles using the risk ratings obtained from the deep learning system (DLS) to identify low and high risk groups. To ensure that the distribution of cancer types within each risk group was similar, binning was done separately for each type of cancer.

A log rank test was performed to compare the Kaplan-Meier (KM) survival curves for the high and low risk groups, and a statistically significant result of $p<0.001$ was obtained. This indicates that there is a significant difference in survival between the high and low risk groups based on the DLS risk ratings.

Furthermore, the researchers investigated whether the DLS could further stratify patients' risk within each stage, considering the established prognostic significance of stage in cancer. The resulting KM curves showed that, except for stage I and stage IV cancers, the DLS was able to further sub-stratify patients into low and high risk groups for stage II ($p<0.05$) and stage III cancers ($p<0.001$). This suggests that the DLS may provide additional prognostic information within these specific cancer stages, which could potentially aid in clinical management decisions for patients with stage II and stage III cancers

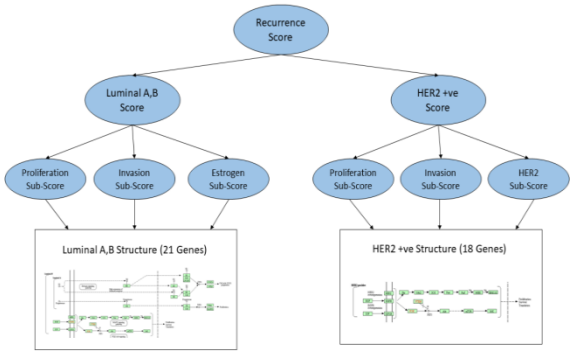

FIG 5: Parametric tree representation for Cancer Prediction features

LIMITATIONS

Our study has some significant limits, despite the fact that we have demonstrated promising outcomes for challenging deep learning circumstances. The test data set for each cancer type includes about 250 cases and about 100 disease-specific data sets, despite using information for 10 cancer kinds from the largest publicly accessible data set (TCGA). Wide confidence intervals were produced as a result of the inclusion of survival events, highlighting the significance of statistical conclusions (and of disclosing confidence intervals for model performance, if reported in this field). As a result, our work serves as a proof-of-concept investigation to improve techniques and gain a deeper comprehension of the viability of direct prediction by lax surveillance of clinical

J Popul Ther Clin Pharmacol Vol 30(9):e80–e86; 10 April 2023.

outcomes. Despite the fact that the model learned prognostic signals, these findings still need to be developed further on larger datasets in order to further enhance predictions, more precisely estimate effect sizes, and show therapeutic utility. and it is necessary to verify. Second, only TCGA datasets are used in our methodology and analysis. The tumor purity in each image is higher in TCGA cases and there are often fewer photos [14]. In real-world clinical settings, where tumor purity may be more varied, sectioning techniques may differ, and typically numerous slides are available in each instance, the random "patch sampling" approach outlined here may therefore prove effective. I'm not yet certain if it is. These findings do not account for a potential confounding effect of patient treatment variations, despite the fact that baseline characteristics were employed and all patients in these studies were treatment-naive at the time of tissue sample. Despite potential treatment differences, risk stratification reveals expected patterns.

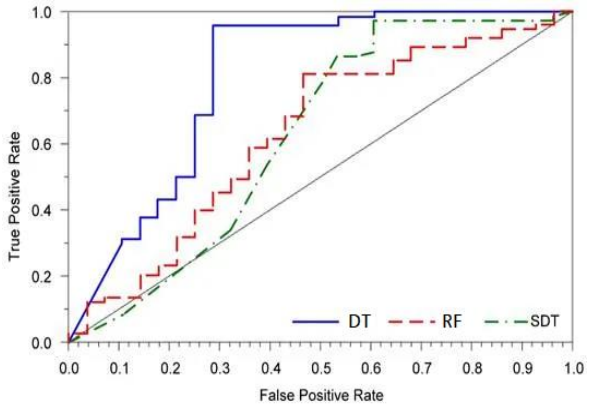

FIG 6: True Positive Rate v/s False Positive Rate

Finally, this report lacks specific molecular data from his TCGA and might need larger datasets and cancer-type-specific molecular analyses.

REFERENCES

1. Weiser MR, Gönen M, Chou JF, Kattan MW, Schrag D. Predicting survival after curative colectomy for cancer: individualizing colon cancer staging. J Clin Oncol. 2011;29: 4796–4802. pmid:22084366
2. Cooperberg MR, Pasta DJ, Elkin EP, Litwin MS, Latini DM, Du CHANE J, et al. The University of California, San Francisco Cancer of the Prostate Risk Assessment score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy. Journal of Urology. 2005. pp. 1938–1942. pmid:15879786
3. Sparano JA, Gray RJ, Ravdin PM, Makower DF, Pritchard KI, Albain KS, et al. Clinical and Genomic Risk to Guide the Use of Adjuvant Therapy for Breast Cancer. N Engl J Med. 2019. pmid:31157962
4. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. N Engl J Med. 2018;379: 111–121. pmid:29860917
5. Rakha EA, El-Sayed ME, Lee AHS, Elston CW, Grainge MJ, Hodi Z, et al. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. J Clin Oncol. 2008;26: 3153–3158. pmid:18490649
6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521: 436–444. pmid:26017442
7. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316: 2402–2410. pmid:27898976
8. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017;318: 2199–2210. pmid:29234806
9. Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, et al. Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection. Arch Pathol Lab Med. 2018. pmid:30295070
10. Nagpal K, Foote D, Liu Y, Chen P-HC, Wulczyn E, Tan F, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. npj Digital Medicine. 2019;2: 48. pmid:31304394
11. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. Cell Rep. 2018;23: 181–193.e7.

J Popul Ther Clin Pharmacol Vol 30(9):e80–e86; 10 April 2023.