Exploring Genetic-histologic Relationships in Breast Cancer

Ruchi Chauhan[1,2]*, PK Vinod*[1]*, CV Jawahar*[2]

[1]Center for Computational Natural Sciences & Bioinformatics (CCNSB)
[2]Center for Visual Information Technology (CVIT)
International Institute of Information Technology, Hyderabad, India

ABSTRACT

The advent of digital pathology presents opportunities for computer vision for fast, accurate, and objective solutions for histopathological images and aid in knowledge discovery. This work uses deep learning to predict genomic biomarkers - TP53 mutation, PIK3CA mutation, ER status, PR status, HER2 status, and intrinsic subtypes, from breast cancer histopathology images. Furthermore, we attempt to understand the underlying morphology as to how these genomic biomarkers manifest in images. Since gene sequencing is expensive, not always available, or even feasible, predicting these biomarkers from images would help in diagnosis, prognosis, and effective treatment planning. We outperform the existing works with a minimum improvement of 0.02 and a maximum of 0.13 AUROC scores across all tasks. We also gain insights that can serve as hypotheses for further experimentations, including the presence of lymphocytes and karyorrhexis. Moreover, our fully automated workflow can be extended to other tasks across other cancer subtypes.

***Index Terms*—** Genomic Biomarkers, Cancer, Imaging Genomics, Mutation Prediction, Histopathology images

1. INTRODUCTION

Histopathological evaluation involving microscopic examination of Hematoxylin & Eosin (H&E) stained specimen on the glass slide is considered a gold standard for cancer diagnosis. However, the manual assessment may be subjected to error, human bias, inter-intra pathologist variability, and low throughput. There have been remarkable advances through deep learning in cancer detection, mitosis detection [1], cancer metastasis detection [2], etc. These works focus on developing automated, fast, and accurate solutions for analysis routinely performed by pathologists.

There have been significant developments in individualized diagnosis, prognosis, and treatment planning based on genomic biomarkers. However, despite the plummeting cost of genome sequencing, it can still be inaccessible, time-consuming, expensive, or infeasible due to the tissue being insufficient for excision. The association of histopathology findings from Whole Slide Images (WSI) and genomic alterations remains mostly unknown in different cancers. It is based on the hypothesis that image features encode the tumor's underlying genotype. It is a challenging problem since genetic changes can manifest as subtle patterns in the images that are undetectable in an unaided approach to histopathology. Deep learning has shown promise in this aspect [3, 4].

We endeavor to improve classification and get insights into morphological features associated with genomic biomarkers. We seek to explore the histological influence of mutation on the nuclei shape and size at the cellular level vs. its impact on the tumor microenvironment involving spatial aspects. Moreover, we examine any visual differences in terms of staining that could potentially be used by pathologists.

This work focuses on breast cancer, the most common cancer in women worldwide. Breast cancer is characterized by molecular features, therapeutic responses, disease progression, and preferential organ sites of metastases [5]. Cancer can be caused by unrepaired alterations known as a mutation in the DNA sequences that encode for genes. These mutations occur due to DNA replication errors or environmental factors. TP53 is a tumor suppressor gene that regulates uncontrolled growth and cell division. It is mutated across cancer types and is an independent risk factor for determining survival. Mutations in PIK3CA oncogene leads to increased signaling for cell proliferation, which may result in a tumor. Further, the presence of Estrogen Receptor (ER) and Progesterone Receptor (PR) is examined in the cancer cells. These hormone biomarkers [6] are predictive of the efficacy of hormone therapy- the treatment strategy of modifying the tumor's hormonal milieu. HER2 proto-oncogene encodes for a growth-promoting protein in the breast cells whose overexpression may lead to a tumor. Such HER2 positive cases respond to therapies targetting HER2 protein.

There has been a rising interest in classifying histopathology images with biomarkers in lung cancer, bladder cancer, prostate cancer, and breast cancer [7, 8, 9, 10, 11]. While these works establish that phenotype is predictive of the genotypic features using deep learning, they do not explore the features that were or could be used for classification.

2. METHODS

This work classifies the WSIs according to six biomarkers: mutations in TP53, PIK3CA, ER status, PR status, HER2 status, and intrinsic subtypes - Basal vs. non-Basal. The discriminative patches from these classifications were further analyzed to understand their distinguishing features in terms of intensity, morphology, spatial arrangement, and cell types. Random forest classification, nuclei annotations, and statistical analysis on gene expressions were used for this purpose. Fig 1 shows the workflow in detail.

Data Preparation: The data used in this work is taken from TP53 [12], which has a repository of 1054 anonymized WSI of breast invasive carcinoma patients with their genomic, pathologic, and de-identified clinical information. Images with highlighters or other artifacts were excluded, resulting in 708 cancer and 100 normal slides. For patients with multiple slides, the last biopsy slide was used. The standard practice of patch extraction resulted in an average of 3,000 patches per WSI of size 512×512 pixels at $20\times$ resolution. Color normalization using [13] accounted for the variations due to staining reagents and scanners by different manufacturers, protocols for slide preparation, etc.

Classification: Our work used InceptionNet-v3 pretrained on ImageNet, which by design, processes the images using multiple

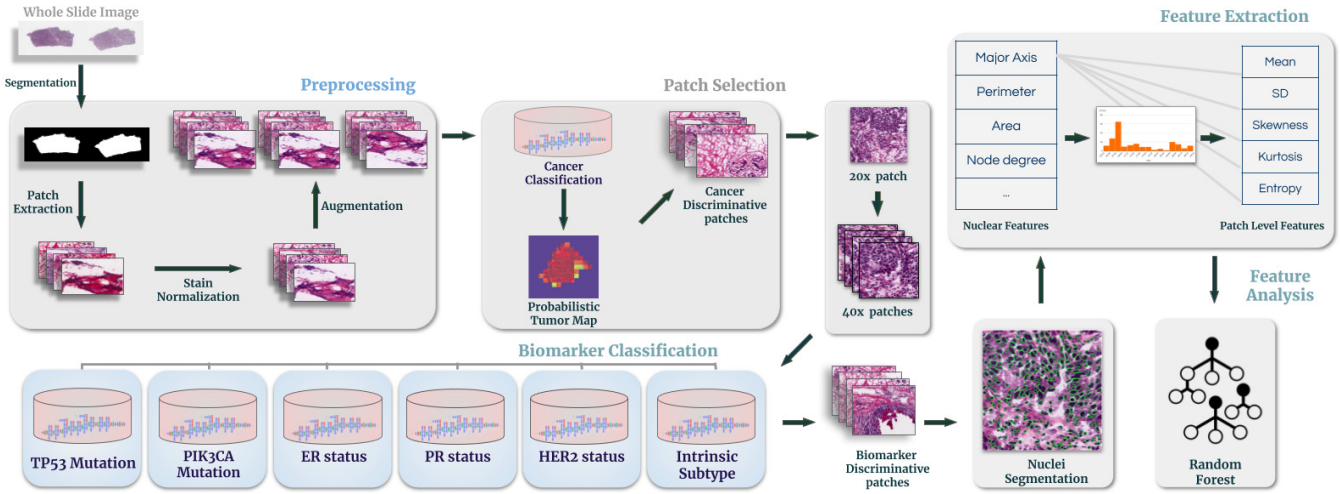

Fig. 1: **Workflow:** The whole slide image is segmented to identify the tissue regions and divided into patches. The patches undergo colour normalization and augmentation to be used for cancer detection. The discriminative patches are converted to a higher resolution and used for biomarker prediction. The biomarker discriminative patches are analyzed by extracting features at the nuclei level and aggregated at the patch level. The features are tested using statistical analysis and random forest classifier.

kernel filter sizes. Such architecture helps process pathology images from multiple fields of view, capturing the cellular along with glandular structures, befitting for our tasks. To handle the class imbalance (18-48% in our dataset), we used random undersampling (with replacement) followed by random augmentation. Once a patch was sampled, a random augmentation was chosen with a random value and then inputted into the model. This approach introduced ample variation in the training dataset without multiplying its size and provided the desired robustness at the expense of convergence time. Colour normalization with stain augmentation is proven effective by a systematic analysis [14]. We used light HSV transformation, HED transformation, additive Gaussian noise, flip, and rotations. Test-time augmentation (without randomness) and mixed-precision training were used. [1]

It must be noted that the labels for the data are available at the patient level, i.e., for slides, and may not be true for the individual patches. The tumor within a slide may be localized, and not all patches from a slide may be cancerous containing the genomic information. Hence there is a mislabelling in the training set that can deteriorate the performance. To overcome this, we identify diagnostically salient regions on the WSI. To keep our workflow fully automated, we obtain machine-generated annotations employing deep learning instead of pathologists' annotations. Moreover, manual annotation on a slide is not viable for mutations as the pathologists do not look for mutations in the histopathological images. To that end, cancer detection - a binary classification of cancer versus normal patches was used to find the *discriminative patches*. This task used all the available normal (non-cancer) slides and an equal number of Cancer slides. The remaining cancer slides were kept as an external test set, later used for biomarker prediction. The model gets uncorrupted labels from the normal slides and can understand a non-cancer patch. This seems to be sufficient even though the problem mentioned above persists for the cancer slides, i.e., not all cancer slide patches are cancerous. The cancer detection is done at 20× resolution, and the discriminative patches are translated to 40× resolution, the highest magnification, to capture finer level details. This translation is done by zooming using Lanczos interpolation, and cropping giving four 40× patches for every 20× patch. These discriminative patches are used for the main tasks of genomic biomarker prediction. Roughly 150,000 patches were obtained at 40× resolution.

Discriminative patches are the patches that were correctly classified by the model with high confidence. The softmax probabilities of the predictions from the classification model were used to obtain the confidence scores. Despite achieving better performance, modern neural networks tend to be overconfident. This can be attributed to the increase in width and depth compared to the older neural networks like LeNet, and methods like batch normalization and weight decay. Thus, the probabilities of prediction are not representative of the true correctness likelihood and calls for a calibrated confidence score. To this end, we use a straightforward calibration technique: Temperature Scaling [15]. The confidence score is calculated as

$$\hat{q} = \sigma_{SM}(z/T) \tag{1}$$

where σ_{SM} is softmax operation, z refers to the logits, and T is the temperature computed by logistic regression. These confidence scores were thresholded at 0.9, above which the patch is deemed 'discriminative'. Filtering discriminative patches from the cancer detection model reduced the dataset to roughly 45%. These patches are termed as Cancer Discriminative (CD) patches and used for biomarker classification tasks. The discriminative patches from each biomarker classification model are called Biomarker Discriminative (BD) patches.

Feature Analysis: Nuclei segmentation [16] followed by ellipse fitting on each nucleus was used to examine the discriminative patches. For each of the nuclei, the following features were computed: *Morphological features:* minor axis, major axis, ratio of major axis to minor axis, area, perimeter, circularity, eccentricity, and solidity. *Intensity Features:* mean pixel value for red, green, and blue channels. Note that the patches are colour normalized. *Spatial Features:* For each patch, a Delaunay triangulation graph connecting the centroids of the nuclei as nodes were constructed. From this graph, the following were calculated for each nucleus: minimum distance from a neighbor, the maximum distance from a neighbor, the average of distances with all neighbors, and the number of neighbors (node degree). All these features calculated for nuclei were aggregated at the patch level. The distribution statistics: mean, standard

[1] All implementation details & data summary can be found at: https://github.com/theRuchiChauhan. Codes will be made available upon publication.

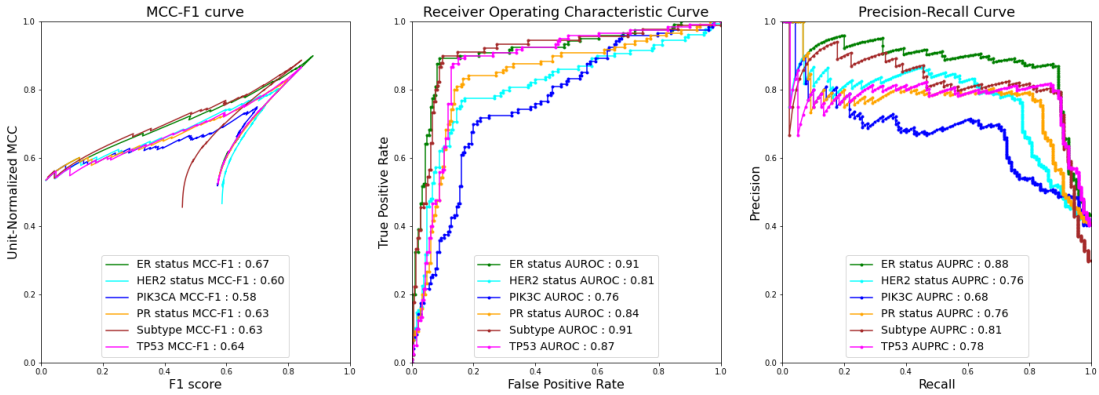

Fig. 2: Classification Results of Genomic Biomarkers using InceptionNet. [L-R] MCC-F1 curve, Receiver Operating Characteristic Curve, Precision-Recall Curve. Overall, ER performed the best, while PIK3CA performed the worst

deviation, entropy, skewness, and kurtosis were taken for each patch using a ten-binned histogram. The histogram was $L1$ normalized to mitigate the effects of an unequal number of nuclei across patches. Thus, we obtained a total of 75 features.

To analyze the features characteristic of a biomarker class, we classify the patches using Random Forest. To further explore the relative importance of features, an ablation study was performed using combinations of features. The classification is done on biomarker discriminative patches and also on the cancer discriminative patches for each task. The complete pipeline is shown in Fig. 1. To further explore the tumor microenvironment at the cellular level, nuclei annotation was done using the tool provided by [17], which employs mask-RCNN for nuclei segmentation.

3. RESULTS AND DISCUSSION

The model achieved slide level 0.99 AUROC for Cancer vs. Normal classification, comparable to that reported in the literature on the TCGA dataset [18]. Our models outperform the results reported elsewhere in the literature for biomarker prediction Table 1. The slide level results are calculated by aggregation using the average of probabilities. We observed an expected improvement in the biomarker classification by training only on our cancerous discriminative patches. The baseline approach used all the patches from the WSIS. AUROC has been tabulated for benchmarking. Additional metrics - AUPRC, MCC-F1 are shown in Fig. 2. The MCC-F1 curve reported is more appropriate for unbalanced datasets than other metrics as it provides a complete summary of the confusion matrix [19]. Qualitative results are shown in Fig. 3

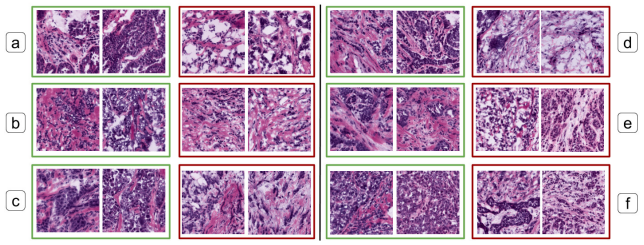

Fig. 3: Top Biomarker Discriminative patches. a: ER status, b: PR status, c: HER2 status, d: TP53, e: PIK3CA, f: Intrinsic Subtype. Green Box: positive/mutated/Basal subtype, Red Box: negative/not-mutated/non-basal subtypes

Table 2 presents the results of the biomarker classification on biomarker and cancer discriminative patches using random forest.

Table 1: Classification Results of Genomic Biomarkers using InceptionNet: † : [4], ‡ : [10], § : [11]

AUROC	Level	TP53	PIK3CA	ER	PR	HER2	Subtype
Ours	patch	0.829	0.721	0.866	0.820	0.798	0.877
	slide	**0.875**	**0.765**	**0.910**	**0.839**	**0.811**	**0.909**
Baseline	patch	0.677	0.565	0.665	0.614	0.666	0.703
	slide	0.643	0.541	0.632	0.578	0.622	0.685
Best	slide	0.75†	0.63†	0.89‡	0.81‡	0.79‡	0.826§

The same set of features did not work as well on all cancer discriminative patches compared to biomarker discriminative patches. This substantiates our claim that the generated patches are indeed discriminative of the genomic information. It might be possible that not all cancer regions contain the manifestation of biomarkers.

We observe that for TP53, the intensity features alone outperform morphological and spatial features combined. Moreover, intensity and spatial features together perform comparably to all features taken together. This contrasts with other biomarkers, where the spatial features do not seem to be contributing much. Morphology features worked reasonably well for almost all the tasks except TP53.

Interestingly, we observed a large number of lymphocyte nuclei in TP53 mutated discriminative patches. Lymphocytes are the white blood cells whose presence indicates an immunological response. Further investigation using gene expression and immune scores obtained from ESTIMATE [20] suggest the correlation between TP53 mutation and increased immunologic activities. The immune scores signify the infiltration of immune cells in tumor tissues. A P-value of 1.76e-05 was obtained from the Mann Whitney U test between the immune scores of TP53 mutated and nonmutated gene expression samples (Fig. 5). These observations go along with the demonstrated involvement of TP53 in crucial aspects of tumor immunology and the homeostatic regulation of the immune responses [21]. Moreover, we found cells undergoing karyorrhexis, nuclei fragmentation during a stage of cell death, in PIK3CA mutated discriminative patches, as shown in Fig. 4.

4. CONCLUSION

This work presented the classification of breast cancer genomic biomarkers from histopathological images. Conventionally, such classification is performed using gene expression data. Automated pipelines such as ours aim to augment pathological workflow while the pathologists may handle higher-level decisions. Despite the remarkable performance of deep learning solutions in computer-aided

Table 2: Results from random forest classifier on Biomarker Discriminative (BD) & Cancer Discriminative (CD) Patches. Best performance in bold and next best underlined.

AUPRC	TP53		PIK3CA		ER		PR		HER2		Subtype	
	BD	CD	BD	CD	BD	CD	BD	CD	BD	CD	BD	CD
All Features (n=75)	**0.934**	**0.657**	0.882	**0.618**	**0.837**	**0.655**	0.862	0.640	0.927	0.584	**0.828**	**0.696**
Intensity Features (n=15)	0.877	0.611	0.821	0.585	0.733	0.613	0.729	0.587	0.865	0.569	0.763	0.661
Spatial Features (n=20)	0.838	0.592	0.741	0.571	0.674	0.590	0.681	0.577	0.799	0.541	0.690	0.629
Morphology Features (n=40)	0.799	0.581	0.828	0.605	0.74	0.599	0.832	0.612	0.909	0.558	0.742	0.635
Morphology + Spatial Features (n=60)	0.866	0.623	0.851	0.603	0.768	0.622	0.845	0.613	0.904	0.561	0.765	0.662
Spatial + Intensity Features (n=35)	0.933	0.654	0.840	0.601	0.788	0.640	0.773	0.601	0.891	0.555	0.797	0.676
Morphology + Intensity Features (n=55)	0.922	0.624	**0.883**	0.617	0.820	0.648	**0.864**	**0.642**	**0.929**	**0.588**	0.825	0.683

diagnosis, there is still legitimate skepticism for widespread clinical adoption. Hence, there is a need for understanding the classification done by a deep neural network into human interpretable features to help reduce the opacity of the black box models and generate knowledge. It is worth noting that the characteristic features are from the network's perspective and the biological significance derived herein warrants further validation.

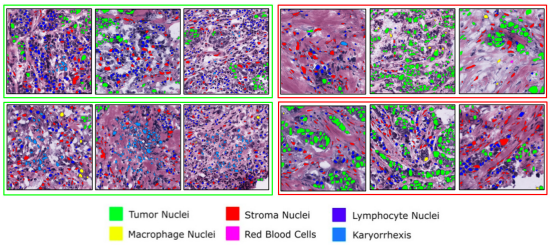

Fig. 4: Prevalence of lymphocytic nuclei in TP53 mutated [Top-left, green] vs TP53 non-mutated [Top-right, red] patches. Presence of Karyorrhexis in PIK3CA mutated [Bottom-left, green] vs PIK3CA non-mutated [Bottom-right, red] patches. Tumor nuclei can be observed in both TP53 & PIK3CA.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study used data from TCGA following the data access policies. Ethical approval was not required.

6. ACKNOWLEDGMENTS

No funding was received for this work, and the authors have no relevant financial or non-financial interests to disclose.

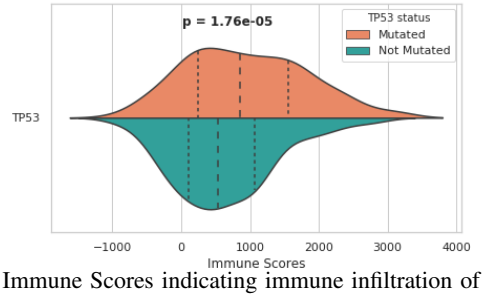

Fig. 5: Immune Scores indicating immune infiltration of the TP53 mutated gene expression samples are significantly higher than that of non-mutated genes. P-value from Mann Whitney U test.

7. REFERENCES

[1] Saha et al., "Efficient deep learning model for mitosis detection using breast histopathology images," *Computerized Med. Imag. and Graph.*, vol. 64, pp. 29–40, 2018.

[2] Pham et al., "Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach," *Amer. J. pathology*, vol. 189, no. 12, pp. 2428–2439, 2019.

[3] Couture et al., "Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype," *NPJ breast cancer*, vol. 4, no. 1, pp. 1–8, 2018.

[4] Kather et al., "Pan-cancer image-based detection of clinically actionable genetic alterations," *Nature Cancer*, pp. 1–11, 2020.

[5] et al. Reis-Filho, "Gene expression profiling in breast cancer: classification, prognostication, and prediction," *The Lancet*, vol. 378, no. 9805, pp. 1812–1823, 2011.

[6] Dunnwald et al., "Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients," *Breast cancer res.*, vol. 9, no. 1, pp. R6, 2007.

[7] Schaumberg et al., "H&e-stained whole slide image deep learning predicts spop mutation state in prostate cancer," *BioRxiv*, p. 064279, 2018.

[8] Coudray et al., "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature med.*, vol. 24, no. 10, pp. 1559, 2018.

[9] Xu et al., "Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients," *bioRxiv*, p. 554527, 2019.

[10] Rawat et al., "Deep learned tissue "fingerprints" classify breast cancers by er/pr/her2 status from h&e images," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

[11] Jaber et al., "A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival," *Breast Cancer Research*, vol. 22, no. 1, pp. 12, 2020.

[12] Grossman et al., "Toward a shared vision for cancer genomic data," *New England J. of Med.*, vol. 375, no. 12, pp. 1109–1112, 2016.

[13] Vahadane et al., "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Trans Med. Imag.*, vol. 35, no. 8, pp. 1962–1971, 2016.

[14] Tellez et al., "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Med. image anal.*, vol. 58, pp. 101544, 2019.

[15] Guo et al., "On calibration of modern neural networks," *ICML 2017*, 2017.

[16] Phoulady et al., "Nucleus segmentation in histology images with hierarchical multilevel thresholding," in *Med. Imag. 2016: Digit. Pathology*. Int. Soc. for Opt. and Photonics, 2016, vol. 9791, p. 979111.

[17] Wang et al., "Computational staining of pathology images to study the tumor microenvironment in lung cancer," *Cancer Research*, vol. 80, no. 10, pp. 2056–2066, 2020.

[18] Noorbakhsh et. al, "Pan-cancer classifications of tumor histological images using deep learning," *bioRxiv*, p. 715656, 2019.

[19] Cao et al., "The mcc-f1 curve: a performance evaluation technique for binary classification," *arXiv preprint arXiv:2006.11278*, 2020.

[20] Wong et al., "Characterization of cytokinome landscape for clinical responses in human cancers," *Oncoimmunology*, vol. 5, no. 11, pp. e1214789, 2016.

[21] Muñoz-Fontela et al., "Emerging roles of p53 and other tumour-suppressor genes in immune regulation," *Nature Reviews Immunology*, vol. 16, no. 12, pp. 741–750, 2016.