**Annotation of the Fox Chase Cancer Center Digital Pathology Corpus**

*D.M. Hackel, M. Bagritsevich, I. Obeid and J. Picone*

Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA

{dmitry.hackel, maria.bagritsevich, iobeid, picone}@temple.edu

The Neural Engineering Data Consortium (NEDC) has previously developed two major open-source corpora to support machine learning research in digital pathology [1]: the Temple University Digital Pathology Corpus (TUDP) and the Fox Chase Cancer Center Digital Pathology Corpus (FCDP). TUDP contains over 100,000 high resolution images scanned from a diverse array of tissue types, including normal and pathological specimens, providing a broad spectrum of data for various research applications. FCDP, with 14,276 images, offers a focused collection that emphasizes a variety of oncological studies. Recently, we have enhanced our FCDP dataset by releasing detailed annotations for the breast tissue subset, comprising 1,463 images [1]. These annotations were carefully selected to include a range of histological features and were processed to ensure precise classification of cancerous, non-cancerous, and pre-cancerous structures. This abstract will present a comprehensive analysis of the pathological and biomedical aspects of the FCCC annotation corpus, including highlighting variations in cancer prevalence, annotation challenges, and how images are prepared for machine learning applications.

An overview of the FCDP Corpus is provided in [1]. The organization of the data, including a detailed description of the metadata available, is also provided in this publication. The data is organized into subdirectories by subject, specimen, and tissue site identifiers. Information about the type and location of tumors is provided in the form of ICCO codes. All files are anonymized to ensure compliance with HIPPA regulations and ensure patient anonymity.

The annotation team utilized Aperio’s open source ImageScope [2] software for annotation. Slides used histological stains like H&E and immunohistochemistry, distinguishing benign from cancerous cells based on color variation, and antigen presence. Annotators identified five to ten regions of interest in each slide based on the classification described in [1] and shown in Table 1. Annotators conduct weekly team reviews to maintain consistent standards and achieve consensus. Unclear cases escalated to senior members or medical mentors for resolution.

It is important to understand that the data was partially annotated. Annotation of every region of interest in each slide is prohibitively expensive, so a decision was made to annotate a representative number of regions from each slide that captured significant phenomena occurring in the slide. Obviously, there is a bit of vagueness about this strategy, but our annotation team has become quite experienced at determining behavior in each slide that captures the variety in each category. This is important for the application and development of machine learning technology.

**Table 1.** A unified classification scheme for FCDP and TUDP [1]

|  |  |  |
| --- | --- | --- |
| **Class** | **Label** | **Description / Features** |
| Low Grade (lg) | Normal (norm) | normal ducts and lobules |
| Background (bckg) | stroma, no ducts or lobules |
| Artifact (artf) | grease pen marks, stitches, foreign bodies, etc. |
| Intermediate Grade (ig) | Non-Neoplastic (nneo) | fibrosis, hyperplasia, intraductal papilloma, adenosis, ectasia, etc. |
| Inflammation (infl) | areas of inflammation |
| Suspected (susp) | regions that are at risk of developing into cancerous regions |
| High Grade (hg) | Ductal Carcinoma in Situ (dcis) | ductal carcinoma in situ, and lobular carcinoma in situ |
| Invasive Ductal Carcinoma (indc) | invasive ductal carcinoma, invasive lobular carcinoma, and invasive mammary carcinoma |
| unknown | Indistinguishable (null) | indistinguishable tissue, normally due to issues with the cut/stain |

Cancer that presents itself within breast tissue is most commonly a result of abnormal polarized luminal epithelial cells that line the lobules and lactiferous ducts within the breast [3]. A normal duct is comprised of a single layer of epithelial cells that line the inside of the duct. Abnormal cells present themselves as layers of abnormal epithelial cells that can partial or completely fill the ducts. When abnormal cells begin to spread beyond the ducts and invade the surrounding tissue, these cells are classified as invasive carcinoma. This is a more progressed disease that has the potential to travel through the lymph nodes, blood stream, or invade other tissue.



Figure 1. A typical example of the annotations provided for FCDP

The analysis of these abnormal structures occurs through histological examination, which looks for characteristics of abnormal cells: degrees of atypia, mitotic activity, and presence of invasive growth. Histology examination also helps grading the cancer, which can reflect its aggressiveness, and indicate the spread of the disease. Both directly impact the course of treatment for an individual with breast cancer.

Annotators were tasked with identifying these cancerous, non-cancerous, and pre-cancerous cells. Once identified, these structures are annotated and labeled. An annotators’ responsibility includes labeling structures on the slides encompassing a range of structures (Table 1). Annotators ensure regions are correctly labeled and that the margins are within reason. The margin defines the boundary between different pathological features. Proper margins do not include a separate tissue type that is not defined by the label or overlapping annotations. An example of an annotation is shown in Figure 1.

A comparison of annotations between FCDP and TUDP datasets reveals significant differences in cancer prevalence. FCDP, with its emphasis on oncological cases, contains a greater proportion of cancerous slides, including more instances of non-neoplastic, ductal carcinoma in situ, and invasive carcinoma compared to TUDP. This variance in the breast tissue subset benefits the training of the machine learning models to detect and classify with more accuracy, a greater variety of cancerous conditions. In contrast, the TUDP offers a broader spectrum of normal tissue types. The two datasets allow for more nuanced understanding of cancerous and non-cancerous tissue characteristics.

One issue that the team had difficulties with was determining the difference between atypical ductal hyperplasia, a non-neoplastic structure, and ductal carcinoma in situ. Annotators realized that many atypical ductal hyperplasia were originally labeled as ductal carcinoma. Although atypical hyperplasia can be indicative of carcinogenic structures in the specimen, these structures are not cancerous and should not have been labeled as such. Atypical ductal hyperplasia is defined by atypical growth of cells within the duct [3]. In some cases, atypical hyperplasia can fill the duct completely making it difficult to differentiate from a ductal carcinoma. Atypical hyperplasia is often a precursor of low-grade ductal carcinoma in situ adding a greater challenge in differentiating between how to label these cellular structures. During our review of previously labeled slides, the annotation team realized these errors in the labeling on some slides which has led us to review both databases and further define the nneo and dcis labels.

When differentiating these structures, the National Cancer Institute grades tumors based on cellular structure and differentiation [4]. For annotation purposes, the definition of non-neoplastic structures has expanded to include irregular cell composition where ducts exhibit patches of varying nuclei sizes, lesions differing in size and shape, less densely packed cells within ducts, and inconsistently defined lumens. Ductal carcinoma is labeled when there is uniformity in cellular composition, nuclei of similar size, consistent lesions in cribriform subtype, presence of comedonecrosis or necrosis in surrounding tissue, and distinct luminal features. These expanded definitions for the nneo and dcis labels attempts to reduce error in annotators’ work and improve the quality of data.

The integration of new annotated breast tissue data from FCDP and TUDP offers a unique opportunity to thoroughly assess the robustness and generalizability of machine learning algorithms in digital pathology. The diverse and comprehensive database from both corpora can train and validate algorithms on a wide array of different tissue structures, diseases stages and pathological features. This combined approach enables the development of more resilient and accurate models that can effectively handle the variability of images, annotation style, and disease presentation. The dual-dataset strategy enhances should also address potential biases or limitations in the models. Ultimately, this comprehensive evaluation fosters advancement of machine learning systems that are both precise and adaptable.

Now that we have completed annotation of FCDP, we are reviewing the annotations of the TUDP Breast Tissue subset for consistency and accuracy. Both resources are being released in Summer 2024. Those interested in learning more about these resources, including how to gain access, can find additional information at the following URL: *https://isip.piconepress.com/projects/nedc/*. At this URL, there are instructions for how to sign up for access to the corpus, instructions on how to download the data, and software that demonstrates how to build machine learning systems based on the data.

Acknowledgements

This material is based upon work supported in part by several organizations including: the Pennsylvania Breast Cancer Coalition Breast and Cervical Cancer Research Initiative, the National Science Foundation under grants under grants nos. CNS-1726188 and 1925494, the Temple University Office of the Vice President for Research, and the Temple University College of Engineering Summer Research Experience for Undergraduates program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of these organizations.

We also wish to acknowledge a large team of investigators and students who have contributed to the collection and development of this data over the past six years.

References

1. Shalamzari, S. S., Bagritsevich, M., Melles, Anne-Mai, Obeid, I., Picone, J., Connolly, D., Wu, C., Brown, B., James, J., Gong, Y., & Wu, H. (2023). Big Data Resources for Digital Pathology. Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium, 1–19. doi: *10.1109/SPMB59478.2023.10372721*.
2. L. Biosystems, “Aperio ImageScope - Pathology Slide Viewing Software,” Leica Biosystems. 2018. [Online]. url: *https://www.leicabiosystems.com/digital-pathology/manage/aperio-image scope/*.
3. “ICD-10-CM, Official Guidelines for Coding and Reporting”, Centers for Medicare & Medicaid Services (CMS), January 01, 2020, url: [*https://www.hhs.gov/guidance/sites/default/files/hhs-guidance-documents/ICD-10-CM\_Guidelines-FY2020\_final.pdf*](https://www.hhs.gov/guidance/sites/default/files/hhs-guidance-documents/ICD-10-CM_Guidelines-FY2020_final.pdf).
4. National Cancer Institute, “Tumor Grade.” Oct. 19, 2024. url: *https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-grade#how-tumor-grade-is-determined*.