www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 930-933

e-ISSN : 2583-1062
Impact Factor : 7.001

CANCER DETECTION IN HISTOPATHOLOGY IMAGES: INSIGHTS FROM EXPLORATORY DATA ANALYSIS

K. Anandhi[1], R. Sinduja[2]

[1]PG Student, RVS College Of Engineering And Technology, Kannampalayam, Sulur, India.

[2]Assistant Professor, RVS College Of Engineering And Technology, Kannampalayam, Sulur, India.

DOI: https://www.doi.org/10.58257/IJPREMS44307

ABSTRACT

Cancer is a leading cause of death worldwide, and the growing demand for early and accurate diagnosis has intensified research in automated histopathological image analysis. Histopathology, the microscopic examination of tissue samples, remains the gold standard in cancer detection. However, manual inspection is time-consuming, subjective, and prone to inter-observer variability. To address these challenges, this study focuses on exploratory data analysis (EDA) of the publicly available Histopathologic Cancer Detection dataset from Kaggle, which contains over 220,000 labeled tissue image patches. The EDA covers class distribution, pixel intensity histograms, box plots, and correlation analysis. The findings highlight balanced class representation, staining heterogeneity, and distinct intensity patterns that carry discriminative features. Furthermore, EDA reveals subtle inter-channel correlations and staining variability that are essential to understand tissue morphology. These insights not only characterize dataset quality but also provide a foundation for selecting effective preprocessing and augmentation strategies in downstream studies.

Keywords: Histopathology, Cancer Detection, Exploratory Data Analysis, Class Distribution, Correlation Analysis.

1. INTRODUCTION

Cancer remains a major global health challenge, accounting for millions of deaths each year. Early detection and accurate diagnosis play a critical role in improving treatment outcomes and survival rates. Histopathology, the microscopic examination of tissue biopsies, is regarded as the gold standard in cancer detection. However, manual inspection is time-consuming, prone to fatigue, and subject to inter-observer variability, leading to inconsistencies in diagnosis.

Before applying advanced deep learning techniques, it is essential to conduct Exploratory Data Analysis (EDA) to understand dataset characteristics. EDA provides valuable insights into class distribution, pixel intensity variation, staining differences, and overall dataset quality. These insights inform preprocessing strategies, augmentation choices, and model design. In this paper, we focus on performing a detailed EDA of the Histopathologic Cancer Detection dataset [1]. The findings from this analysis serve as a foundation for future work involving CNNs [4], Vision Transformers [3,5], and explainable AI techniques [2].

2. RELATED WORK

CNN-based Histopathology Analysis: Convolutional Neural Networks (CNNs) have been widely used in histopathology image analysis[4]. They excel at learning hierarchical spatial features and have achieved high accuracy in cancer detection tasks. The merit of CNNs lies in their ability to automatically extract discriminative features. However, they primarily capture local dependencies and may fail to capture global contextual relationships, limiting performance in complex tissue structures [7].

Transformer Models for Vision: Vision Transformers (ViTs) have recently emerged as powerful alternatives to CNNs in computer vision[3]. Their key strength is the ability to model global dependencies using self-attention mechanisms, which is beneficial for capturing long-range contextual features in histopathology images. The main drawback is their requirement for large-scale training data and computational resources, making them less effective in smaller medical datasets unless transfer learning is applied [6].

Hybrid CNN-Transformer Models: Recent research has explored hybrid architectures that combine CNNs and Transformers. CNNs handle low-level feature extraction effectively, while Transformers capture global dependencies. This combination offers superior performance, particularly in complex medical imaging tasks[5]. However, the added architectural complexity increases computational cost and training time, which may limit real-time clinical deployment.

Explainable AI (XAI): Explainability methods such as Grad-CAM provide heatmaps that highlight regions influencing model predictions [2]. The advantage of these techniques is that they improve transparency and trust in automated systems, which is critical for adoption in healthcare. Nevertheless, such methods may sometimes produce coarse or ambiguous heatmaps that do not perfectly align with clinical understanding [9].

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 930-933

e-ISSN : 2583-1062
Impact Factor : 7.001

3. DATASET DESCRIPTION

The study uses the publicly available Histopathologic Cancer Detection dataset from Kaggle [1]. It consists of over 220,000 labeled histopathology image patches, each of size 96x96 pixels, extracted from lymph node sections of breast cancer patients. The dataset is organized into two categories: Normal (benign) and OSCC (Oral Squamous Cell Carcinoma, malignant).

4. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is an essential step in understanding the structure, distribution, and characteristics of a dataset before building predictive models. In medical imaging, EDA provides valuable insights into class balance, staining variability, pixel intensity patterns, and tissue heterogeneity. Such observations guide preprocessing, augmentation, and model design strategies. The following subsections present the EDA carried out on the Histopathologic Cancer Detection dataset [1].

4.1 Class Distribution

Bar charts and pie charts are used to visualize the distribution of cancerous and non-cancerous patches.

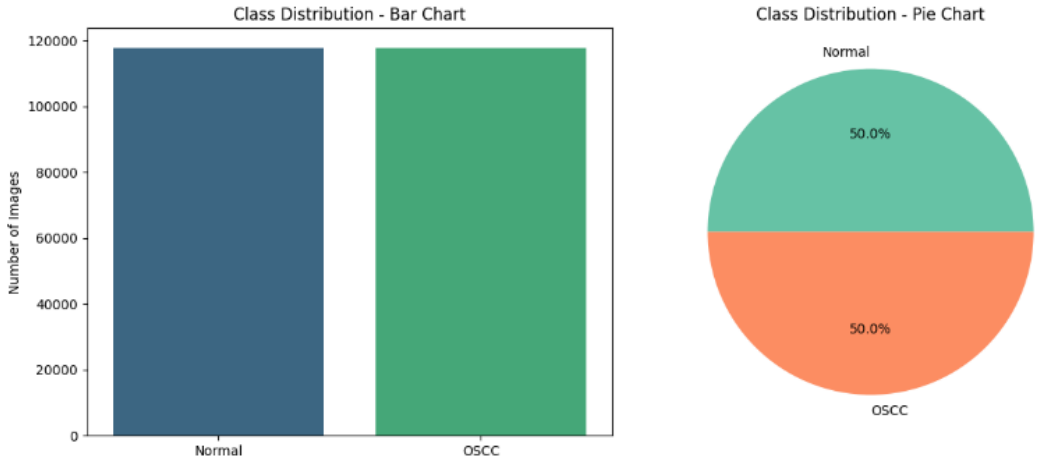

Figure 4.1: Class distribution of Normal vs. OSCC patches.

From the above figures, it is observed that the dataset is relatively balanced between cancerous and benign patches. This balance reduces the risk of bias during model training and ensures that classification models can generalize effectively to both classes.

4.2 Pixel Intensity Distribution – RGB Histograms

RGB histograms are plotted for both classes to analyze staining differences and pixel intensity ranges.

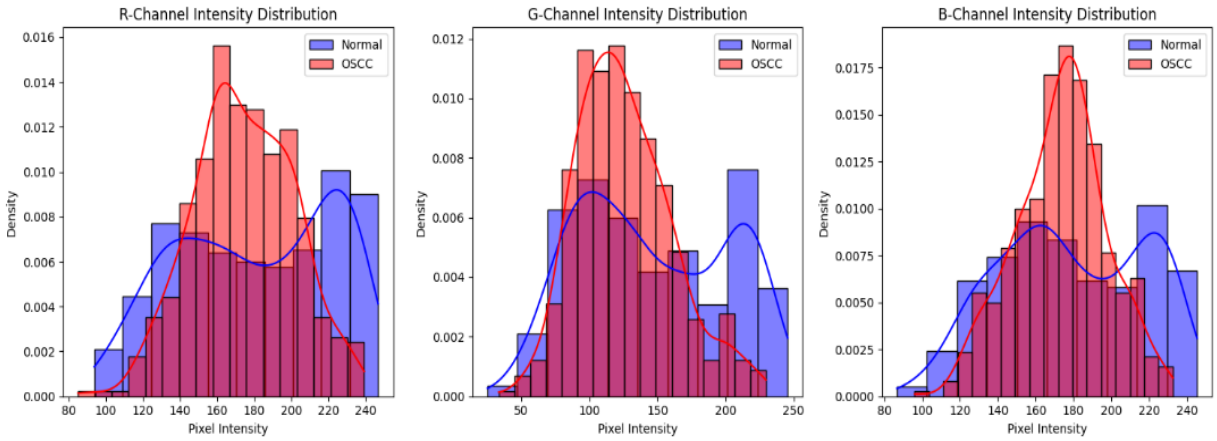

Figure 4.2: RGB intensity histograms for Normal vs. OSCC patches.

From the above figures, it is observed that cancerous patches exhibit higher variation in the red channel due to dense nuclear staining, whereas non-cancerous patches display more uniform intensity distributions. This suggests that nuclear staining patterns provide useful discriminative features for cancer detection.

4.3 Pixel Intensity Comparison – Box Plots

Box plots summarize the variability of pixel intensities for each RGB channel in both classes.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 930-933

e-ISSN : 2583-1062
Impact Factor : 7.001

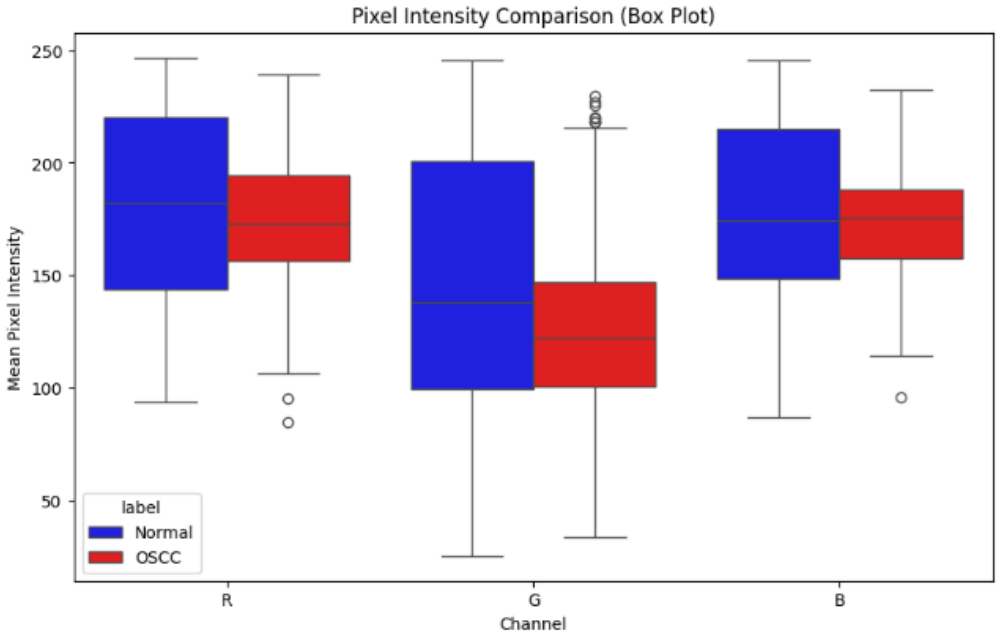

Figure 4.3: Box plots of RGB intensities in Normal vs. OSCC patches.

From the above figures, it is observed that cancerous patches show wider intensity variability, particularly in the red channel. This highlights the heterogeneous nuclear and cytoplasmic structures present in malignant tissue, compared to the more uniform patterns seen in benign samples.

4.4 Correlation Analysis

Correlation coefficients between the RGB channels are computed, and correlation heatmaps are plotted for cancerous and non-cancerous patches.

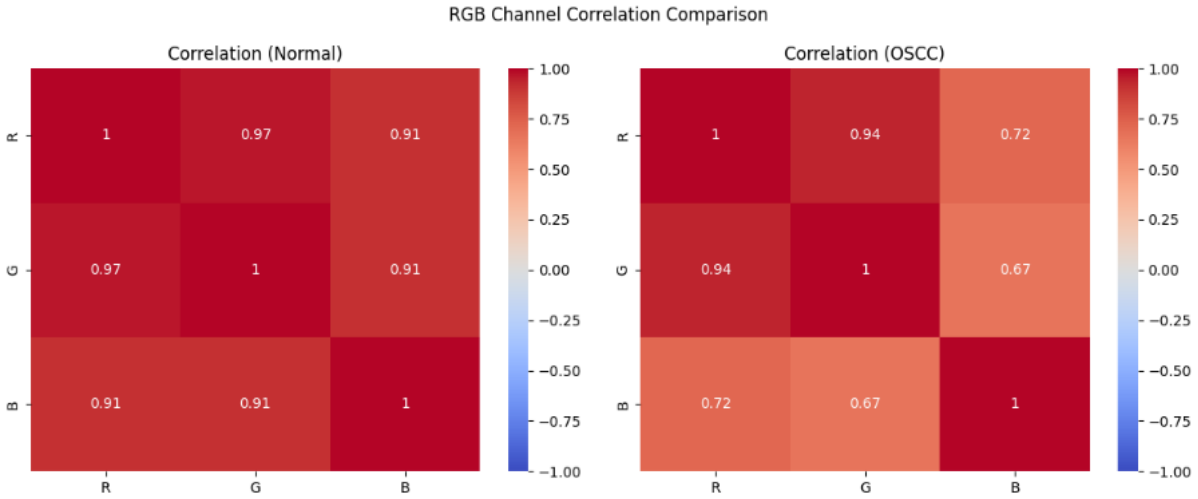

Figure 4.4: RGB channel correlations in Normal patches and correlations in OSCC patches.

From the above figures, it is observed that while RGB channels are strongly correlated across all patches, slight differences appear between cancerous and benign groups. This indicates that inter-channel relationships vary with tissue type, suggesting that color correlations may hold discriminative information for classification tasks.

5. FUTURE WORK

Future work will expand this exploratory data analysis to include brightness and contrast distributions, texture-based feature analysis, and dimensionality reduction techniques like PCA and t-SNE to visualize class separability. These extended analyses will provide a deeper understanding of tissue variability and structural differences.

Beyond EDA, research will focus on developing CNN baselines (e.g., ResNet50, EfficientNet), Vision Transformers with pretraining, and hybrid CNN–Transformer architectures (e.g., CTransPath) [6]. Self-supervised learning strategies such as Masked Autoencoders will be applied to improve feature learning [8]. Explainability methods like Grad-CAM, SHAP, and attention rollout will ensure model predictions align with histopathological understanding [9].

Additionally, domain adaptation will be explored to handle staining variability across institutions, moving toward clinically robust deployment.

The future work will also focus on developing and evaluating deep learning models such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). These models will be trained on the dataset using preprocessing and augmentation strategies guided by the current EDA findings. Explainable AI (XAI) methods such as Grad-CAM and attention map visualizations will then be integrated to provide interpretability, ensuring that automated predictions align with histopathological understanding and clinical trust.

6. CONCLUSION

This study presented an exploratory data analysis (EDA) of the Histopathologic Cancer Detection dataset. The analyses included class distribution, pixel intensity distribution through RGB histograms, pixel intensity comparison using box plots, and correlation analysis of RGB channels.

From these analyses, it is observed that the dataset is reasonably balanced, ensuring fair representation of both cancerous and benign patches. The RGB histograms revealed noticeable staining variations, particularly in the red channel, which is associated with nuclear features in cancerous tissue. Box plots further confirmed that malignant patches exhibit higher variability in pixel intensities, reflecting the heterogeneity of cancerous regions. Correlation analysis demonstrated subtle differences in inter-channel relationships between cancerous and non-cancerous patches, suggesting that color correlations carry useful discriminative information.

The findings from this EDA provide important guidance for preprocessing, normalization, and augmentation strategies, and at the same time create a strong basis for future exploration of CNN and Vision Transformer models integrated with explainable AI.

7. REFERENCES

[1] Kaggle. (2018). Histopathologic cancer detection dataset. Kaggle. https://www.kaggle.com/competitions/histopathologic-cancer-detection

[2] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618–626). IEEE.

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).

[4] Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 411–418). Springer.

[5] Chen, R. J., Lu, M. Y., Shaban, M., Chen, C., Yao, C., & Mahmood, F. (2022). Scaling vision transformers for histopathology with self-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12000–12010). IEEE.

[6] Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J. M., Ciompi, F., & van der Laak, J. (2022). Whole-slide transformers: Toward weakly supervised classification of gigapixel histopathology images. Medical Image Analysis, 81, 102530.

[7] Mahmood, F., Shaban, M., Zhang, Y., & Chen, R. J. (2023). Deep learning for cancer diagnosis and prognosis from histopathology: Advances and challenges. IEEE Transactions on Medical Imaging, 42(8), 2058–2074.

[8] Wang, X., Li, Y., Zhang, H., & Zhao, Y. (2023). Self-supervised vision transformers for histopathological image analysis. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 123–133). Springer.

[9] Li, H., Zhang, Z., Liu, Q., & Wang, Y. (2024). Explainable vision transformers for histopathology: Bridging accuracy and interpretability. Nature Machine Intelligence, 6(2), 145–156.