Deep Learning-Based Model to Classify Histopathological Cancer Tissues from Whole-Slide Images

Aashik Sharif Basheer Ahamed
Washington State University
Pullman, USA
WSU ID: 011870531
Email: a.basheerahmed@wsu.edu

Sheheryar Ahmad Pirzada
Washington State University
Pullman, USA
WSU ID: 011869749
Email: sheheryar.pirzada@wsu.edu

***Abstract*—A deep learning-based method for classifying histopathological cancer tissues from whole slide images (WSI) into adenocarcinoma and adenoma categories is presented in this work. We investigate three different designs using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs): InceptionV3 with Max-Pooling, InceptionV3 with SimpleRNN, and InceptionV3 with LSTM. In order to improve generalizability, the models were trained on a sizable dataset of 96×96 pixel patches that were taken from WSIs using sophisticated data augmentation techniques. According to our research, the InceptionV3 with SimpleRNN architecture performs the best, showing promise for helping pathologists with clinical diagnosis. To enable the learned models to be deployed practically, an intuitive graphical user interface (GUI) program was created.**

***Index Terms*—Deep Learning, Histopathological Images, Whole-Slide Images, CNN, RNN, LSTM**

I. Introduction

In the healthcare industry, machine learning has become a potent instrument, especially for automating difficult processes like cancer diagnosis. Among the most common malignancies in the world, stomach and colon cancers mostly depend on pathologists manually reviewing histopathological slides, a laborious, error-prone, and expertise-dependent procedure. In order to increase the precision and effectiveness of diagnosis, this study uses deep learning to categorize Whole-Slide Images (WSIs) of malignant tissues into adenoma and adenocarcinoma.

Key machine learning problems are examined in this research, including the best neural network architecture for WSI classification, how data augmentation improves generalizability, and the trade-offs in performance between aggregation techniques like max-pooling and recurrent networks. Three InceptionV3-based designs that integrated Max-Pooling, SimpleRNN, and LSTM layers were used to investigate these topics.

Inspired by machine learning's potential to transform cancer diagnosis, this project offered a chance to apply cutting-edge methods to a significant real-world issue. Designing and assessing models that improve classification accuracy while tackling issues like high computing needs and picture data variability was the main objective.

The InceptionV3 with SimpleRNN performed the best out of all the models created, indicating its potential for real-world application. To demonstrate the wider applicability of the suggested method in clinical settings, a graphical user interface (GUI) program was also developed to help pathologists efficiently categorize photos.

II. Machine Learning Task

A. Input and Output

Classifying histopathological Whole-Slide Images (WSIs) into two groups—adenoma and adenocarcinoma—is the main machine learning job for this study. The task's input data is WSIs, which are usually saved in high-resolution TIFF format. With the fine features required for a precise diagnosis, each slide depicts a magnified picture of malignant tissue. To improve their quality, these photos are preprocessed, and to make computational analysis easier, they are separated into smaller pieces. These processed patches are fed into machine learning models, which then forecast the corresponding categories. This method produces a label that indicates whether a patch is associated with adenoma or adenocarcinoma, as well as a confidence score that indicates how certain the model is of its prediction. For example, a patch extracted from a colon tissue slide may be classified as adenocarcinoma with a confidence score of 0.87, signifying high confidence in the prediction.

The project's dataset is made up of histological Whole-Slide Images (WSIs) of colon tissues that have been classified as either adenocarcinoma or adenoma. These pictures came from publicly accessible repositories of preprocessed medical imaging data that are open source on Github and Kaggle, guaranteeing access to diverse and high-quality samples that are typical of actual clinical situations. To enable effective processing and model training, each WSI—typically a high-resolution image—was split into smaller patches, each measuring 96×96 pixels. The model may concentrate on specific tissue properties thanks to this patch-based method, which is essential for precise categorization. With more than 220,000

patches taken from 400 WSIs, the dataset offers a strong basis for deep learning model training and assessment. The dataset's size and diversity make it ideal for examining the effectiveness of various aggregation methods and neural network topologies.

B. Questions to be investigated

In order to properly tackle this challenge, this study explores a number of machine learning issues. Finding the best neural network architecture for this issue is one of the main issues, especially when it comes to examining the advantages of merging convolutional and recurrent neural networks, as discussed in the base study. Understanding how data augmentation methods affect the model's capacity to generalize across various histopathological samples is the subject of another query. The study also aims to assess the trade-offs between various aggregation strategies, including max-pooling and RNN-based approaches, in terms of performance indicators like accuracy and AUC-ROC. The function of hyperparameter optimization in improving model performance and guaranteeing robust learning is an additional query.

C. Challenges involved

The work poses a number of difficulties that need to be resolved in order to produce trustworthy outcomes. The computational complexity of processing big WSIs, which requires a substantial amount of GPU and memory resources, is one of the main obstacles. To make the data manageable for training, effective patch extraction and preprocessing techniques are required. Because histopathological images vary in staining, illumination, and tissue structure, it can be difficult to guarantee the model's robustness and generalizability. Careful model design and extensive data augmentation are needed to address this variability.

Finally, it's crucial to strike a balance between model complexity and training effectiveness, especially when dealing with time and hardware limitations. Another problem was developing a front-end application with appropriate documentation to enable users who are unfamiliar with computer science or medical technical fields to engage with the model. These difficulties highlight how carefully the machine learning models for this task must be designed, trained, and assessed.

III. Technical Approach

We created three deep learning models based on the InceptionV3 architecture to tackle the job of labeling histological pictures as either adenoma or adenocarcinoma. To examine their effects on classification accuracy and generalization performance, each model integrated several aggregation approaches. The algorithmic methodology, architectures, and dataset application to the models are described in this part.

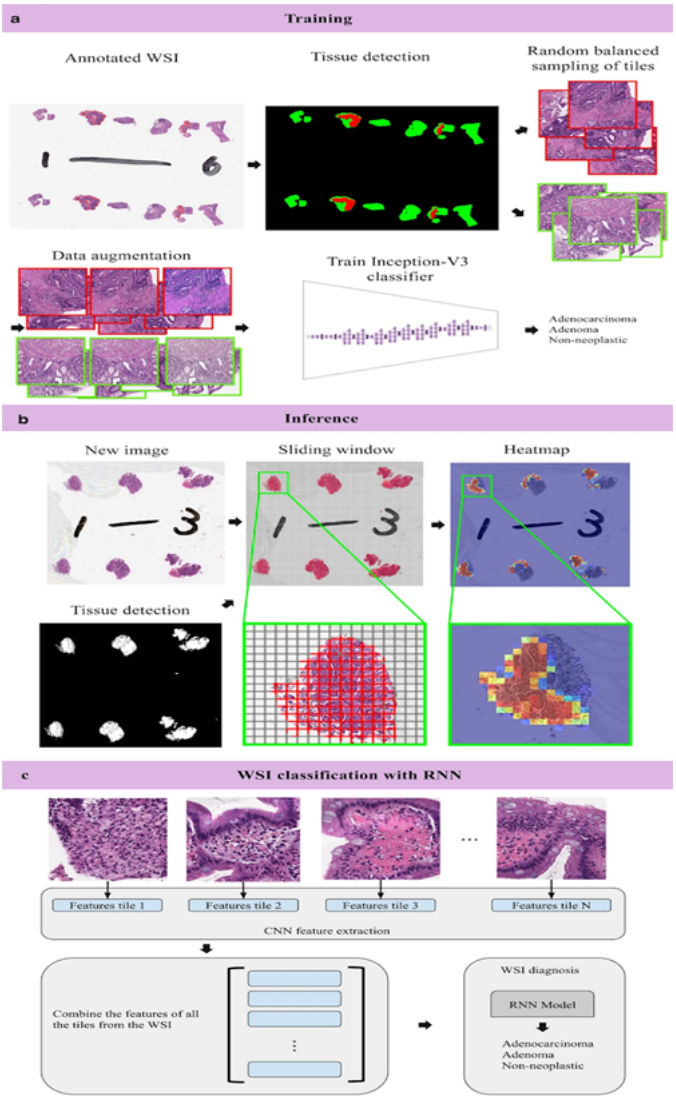


Fig. 1: An overview of the pipeline we attempted to implement from the base study (a) Augmentation and Inception v3 training are performed from WSI images of the training set. (a) A new Whole Slide Image (WSI)-based illustration of inference.(omitted since there were already preprocessed images) (c) RNN-based WSI.

A. Algorithmic Approach

To ensure that the input size is acceptable for neural networks, the approach starts by preprocessing the Whole-Slide Images (WSIs) to extract smaller patches measuring 96 × 96 pixels. To increase the models' capacity for generalization, these patches—which were classified as either adenoma or adenocarcinoma—were enhanced using methods like flipping, rotation, scaling, and contrast alterations.

The primary algorithmic steps for this task include:

1) Data Preprocessing: WSIs are divided into patches and augmented for diversity.
2) Model Development: Three neural network architectures were designed, incorporating different aggrega-

tion techniques—standard max-pooling, SimpleRNN, and LSTM.

3) Training and Validation: The models were trained using the Adam optimizer, with weighted cross-entropy as the loss function, to balance the class distributions. Hyperparameters, such as learning rate and batch size, were fine-tuned to optimize performance.
4) Evaluation: The models were evaluated using metrics such as accuracy, loss, and AUC-ROC to determine their effectiveness in classifying the patches.
5) User Interface: To make sure the model developed is easy to use for users without domain expertise in machine learning or Cancer Classification.

B. Model Architectures

To answer the ML issues and problems, three models were created, each based on the InceptionV3 backbone.

C. Pseudo-Code

The algorithmic process can be summarized in the following pseudo-code:

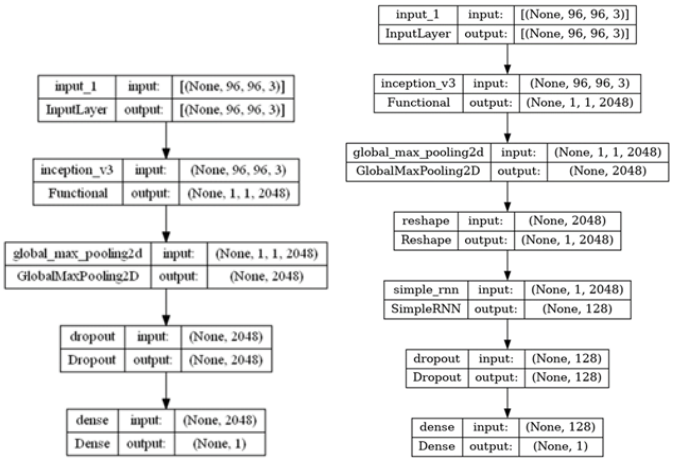
Algorithm 1 Deep Learning-Based Classification of WSIs

```
Load and preprocess Whole-Slide Images (WSIs).
Divide WSIs into 96 × 96 pixel patches.
Apply data augmentation techniques to the patches.
for each model (Max-Pooling, SimpleRNN, or LSTM) do
    Initialize the InceptionV3 backbone.
    Apply the respective aggregation method (Max-Pooling, SimpleRNN, or LSTM).
    Train the model using the Adam optimizer.
    Evaluate the model using metrics: accuracy, loss, and AUC-ROC.
end for
Compare the performance of all models.
Develop a GUI for image classification using trained models.
```

Model 1: InceptionV3 with Max-Pooling Aggregation The InceptionV3 architecture, a cutting-edge convolutional neural network intended for effective image classification, is used in this model. Max-pooling is used to aggregate the InceptionV3 layers' output, using the highest activation values from each patch to serve as the final classification determination. This method provides a baseline for comparison and is computationally efficient.

Model 2: InceptionV3 with SimpleRNN Aggregation This architecture builds on the first model by adding SimpleRNN layers following InceptionV3's convolutional blocks. The RNN captures spatial dependencies that max-pooling can miss by processing sequential patterns in the patch embeddings. To guarantee that the convolutional layers' 2D outputs are fed into the RNN correctly, a reshaping layer is included.

Model 3: InceptionV3 with LSTM Aggregation Long Short-Term Memory (LSTM) layers, which are more adept at resolving the vanishing gradient issue and capturing long-term dependencies, take the role of the SimpleRNN in the third model. In order to provide more detailed insights into the tissue structure, this model attempts to preserve important spatial patterns across patches.



(a) Model 1 architecture (b) Model 2 architecture

(c) Model 3 architecture

Fig. 2: Architectures of the three models: Model 1 (InceptionV3), Model 2 (InceptionV3 with Simple RNN), and Model 3 (InceptionV3 with LSTM).

D. Dataset Application to Models

To maximize memory use, the models were fed the preprocessed dataset of 96 × 96 pixel patches in batches. The patches were normalized and enhanced throughout training to guarantee consistency and improve the models' resilience. The optimization procedure was guided by the cross-entropy loss, which was calculated using the labels for each patch. Each patch's classification results were derived from the outputs of the max-pooling, SimpleRNN, or LSTM layers, and the overall WSI classification was deduced from these patch-level predictions.

E. Graphical User Interface (GUI)

Using Python's Flask framework, a graphical user interface (GUI) was created to make the model outputs accessible

and easy to use. Pathologists and other users can upload histopathological image patches for categorization using the GUI. After processing the uploaded image via the trained models and displaying the projected class (adenoma or adenocarcinoma) and confidence score, the interface offers a clear depiction of the image.

Additionally, the GUI has a model selection function that lets users pick between the LSTM, SimpleRNN, and Max-Pooling models. Real-time usage is ensured by the classification results being provided in a matter of seconds. The application is useful for clinical workflows because of its user-centric design, which guarantees that even those without technical expertise may use it efficiently.

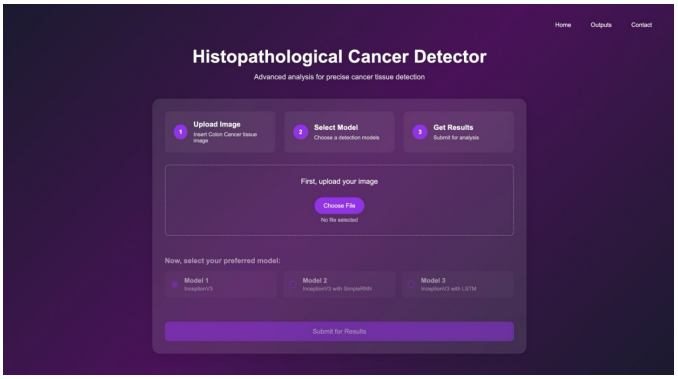


Fig. 3: The image shows the GUI application hosted in its home page and instructions to use them in the website.

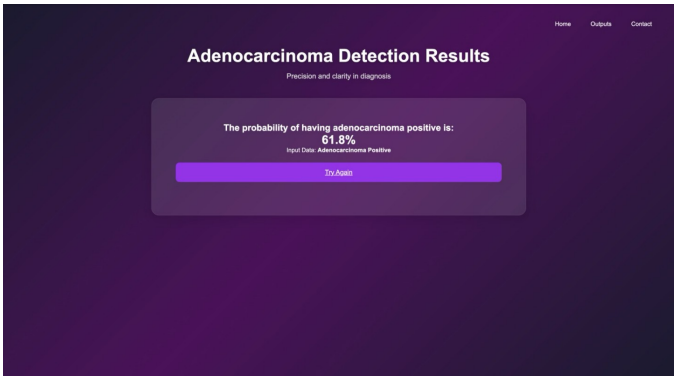


Fig. 4: In this snapshot the output of the classified image is shown as Adenocarcinoma and the probability of image being it shown..

F. Platforms and Packages Used

Google Colab and Kaggle were the main platforms used for the trials, utilizing their cloud-based GPUs to speed up training and lessen hardware limitations. Python-based frameworks and libraries were essential to the project's execution. The deep learning models were designed, trained, and evaluated using TensorFlow and Keras, while effective data management and preparation were made possible with NumPy and Pandas. Key findings, including confusion matrices, ROC curves, and training loss curves, were visualized using Matplotlib and Seaborn. Essential tools for calculating assessment measures, such as accuracy, precision, recall, F1-score, and AUC-ROC, were made available by Scikit-learn. Flask was used for the graphical user interface (GUI) development.

Depending on the architecture and dataset size, the models' entire training procedure took between three to five hours. Due to their higher computing demands, more complicated models—like those that integrate recurrent layers (SimpleRNN or LSTM)—needed somewhat longer periods. Each model's overall training duration varied, but it mostly fell within this range, emphasizing the trade-off between computational efficiency and architectural complexity. This, together with the plots' convergence, led us to use 16 epochs as the benchmark for comparing other models.

G. Addressing Challenges

By splitting the images into smaller patches and employing batch processing during training, the computational difficulties caused by the enormous WSIs were lessened. Numerous data augmentation strategies were used to address the diversity in the histopathology pictures. The models' robustness was increased by using sophisticated aggregation techniques like RNNs and LSTMs, which allowed them to better capture spatial patterns. Hyperparameter tweaking also made sure that the models struck a balance between efficiency and complexity.

Three alternative aggregation techniques are used to give a thorough grasp of how various topologies affect classification performance. This method guarantees a reliable and expandable solution for the classification task of histopathology images.

IV. Evaluation Methodology

A. Dataset and Challenges

Histopathological Whole-Slide Images (WSIs) of colon cancer tissues obtained from publically accessible medical imaging sources that have been preprocessed and posted on Github and Kaggle comprise the dataset utilized for this study. To aid with computational feasibility, these pictures, which are classified as either adenoma or adenocarcinoma, were separated into smaller patches of size 96×96 pixels. With more than 220,000 patches taken from 400 WSIs, the dataset offers a strong basis for testing and training.

This dataset presented a number of difficulties. Large, high-resolution pictures, known as WSIs, demand a lot of processing and analysis power. Furthermore, noise and possible bias were added by variations in tissue architecture, staining processes, and picture quality, which is why thorough preprocessing and data augmentation methods are essential. Another crucial factor was making sure the dataset reflected actual clinical situations, since this has a direct impact on how broadly the trained models can be applied.

B. Evaluation Metrics

To evaluate the models, the following metrics were employed:

- **Accuracy:** calculates the percentage of all patches that are successfully classified. This statistic is an excellent starting point for assessing model performance because it is straightforward and easy to understand.
- **Loss:** calculates the model's error during validation and training. Overfitting or underfitting during the learning process can be detected with the aid of loss monitoring.
- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** shows how well the model can differentiate between the two groups (adenocarcinoma and adenoma). Because it offers a thorough understanding of the trade-offs between sensitivity and specificity across many thresholds, it is especially significant in medical applications.
- **Confusion Matrix:** gives a thorough explanation of false positives, false negatives, true positives, and true negatives. When assessing the model's performance in unbalanced datasets and comprehending the many kinds of classification errors, this statistic is especially helpful. Additionally, it serves as the foundation for calculating F1-score, recall, and precision.

These measures were chosen because they are applicable to real-world situations. The loss metric helps to optimize the model during training, whereas accuracy offers a clear indicator of overall performance. In clinical settings, where false positives and false negatives can have serious repercussions, AUC-ROC is essential. The evaluation guarantees a thorough and impartial comprehension of model performance, both statistically and in real-world situations, by integrating various measures.

C. Real-World Justification

High accuracy reduces diagnostic errors in medical diagnostics by guaranteeing trustworthy predictions. AUC-ROC makes the models more applicable to real-world situations by taking into consideration different clinical circumstances where the costs of false positives and false negatives vary. Additionally, loss tracking during training guarantees the creation of well-generalizing models, lowering the possibility of subpar performance on unseen data in clinical settings.

V. Results and Discussion

A. Overview of Results

The project's outcomes show how well deep learning models discriminate between histological cancer tissues. InceptionV3 with Max-Pooling, InceptionV3 with SimpleRNN, and InceptionV3 with LSTM were the three designs that were assessed. The models' performance under various setups was evaluated using metrics such as accuracy, loss, and AUC-ROC. Key findings and patterns from the review process are shown in the subsections that follow.

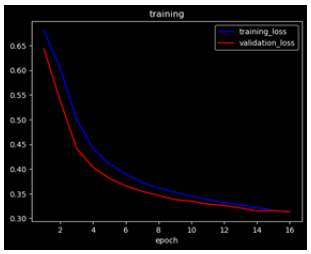
B. Model Performance

A summary of the findings for each of the three models can be seen in the table below, and more specific plots can be found in the figures.

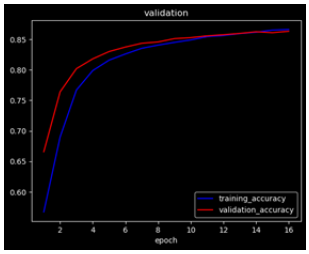
TABLE I: Performance Metrics of the Models

Metric	Training	Validation
Model 1 (InceptionV3)	Accuracy: 0.866 Loss: 0.313	Accuracy: 0.863 Loss: 0.313
Model 2 (InceptionV3 with Simple RNN)	Accuracy: 0.870 Loss: 0.308	Accuracy: 0.864 Loss: 0.310
Model 3 (InceptionV3 with LSTM)	Accuracy: 0.866 Loss: 0.315	Accuracy: 0.862 Loss: 0.319

C. Model 1 Figures

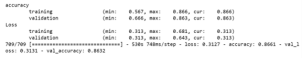


(a) Model 1 Loss Plot

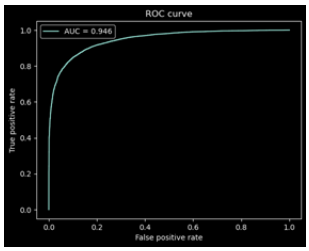


(b) Model 1 Accuracy Plot

Fig. 5: Model 1 Training Loss and Accuracy Plots.

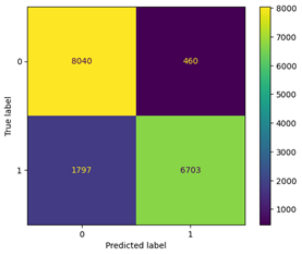


(c) Model 1 Training and Validation Statistics



(d) Model 1 ROC Curve

Fig. 5: Model 1 Training Statistics and ROC Curve.



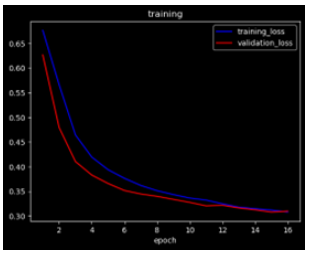
(e) Model 1 Confusion Matrix

	precision	recall	f1-score	support
0	0.82	0.95	0.88	8500
1	0.94	0.79	0.86	8500
accuracy			0.87	17000
macro avg	0.88	0.87	0.87	17000
weighted avg	0.88	0.87	0.87	17000

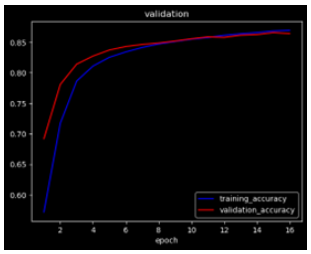
(f) Model 1 Precision, Recall, F1-Score and Support based on confusion matrix

Fig. 5: Model 1 Confusion Matrix and Precision-Recall Metrics.

D. Model 2 Figures

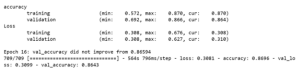


(a) Model 2 Loss Plot

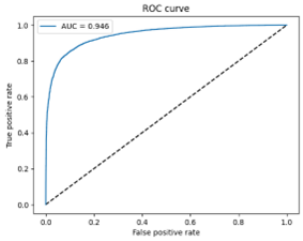


(b) Model 2 Accuracy Plot

Fig. 6: Model 2 Loss and Accuracy Plots.

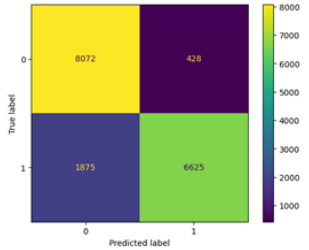


(c) Model 2 Training and Validation Statistics



(d) Model 2 ROC Curve

Fig. 6: Model 2 Training Statistics and ROC Curve.



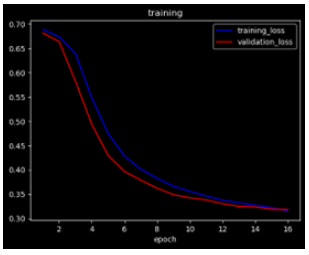
(e) Model 2 Confusion Matrix

	precision	recall	f1-score	support
0	0.81	0.95	0.88	8500
1	0.94	0.78	0.85	8500
accuracy			0.86	17000
macro avg	0.88	0.86	0.86	17000
weighted avg	0.88	0.86	0.86	17000

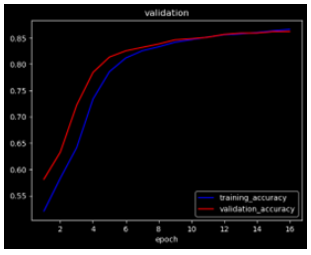
(f) Model 2 Precision, Recall, F1-Score, and Support

Fig. 6: Model 2 Confusion Matrix and Precision-Recall Metrics.

E. Model 3 Figures

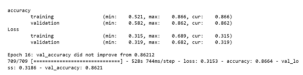


(a) Model 3 Loss Plot

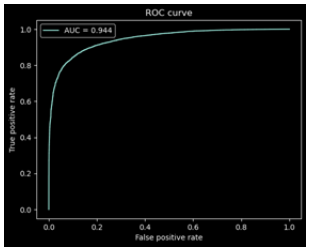


(b) Model 3 Accuracy Plot

Fig. 7: Model 3 Loss and Accuracy Plots.

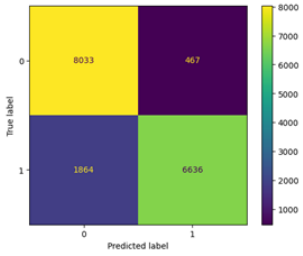


(c) Model 3 Training and Validation Statistics



(d) Model 3 ROC Curve

Fig. 7: Model 3 Training Statistics and ROC Curve.



(e) Model 3 Confusion Matrix

	precision	recall	f1-score	support
0	0.81	0.95	0.87	8500
1	0.93	0.78	0.85	8500
accuracy			0.86	17000
macro avg	0.87	0.86	0.86	17000
weighted avg	0.87	0.86	0.86	17000

(f) Model 3 Precision, Recall, F1-Score, and Support

Fig. 7: Model 3 Confusion Matrix and Precision-Recall Metrics.

F. What Worked and Why?

InceptionV3 with SimpleRNN outperformed the other two designs in terms of accuracy, validation loss, and AUC-ROC score. The capacity of SimpleRNN to recognize sequential patterns across image patches—which are essential for comprehending spatial dependencies in WSIs—is responsible for this better performance. The model's capacity to function successfully on unseen data was further improved by the use of data augmentation. The model's success was further aided by the effective preprocessing pipeline and hyperparameter tuning.

G. What Didn't Work and Why Not?

Despite having more sophisticated memory, the InceptionV3 with LSTM performed worse than the SimpleRNN architecture. This might be because of the dataset's comparatively short sequence length, which might not make the most of LSTM's capacity to identify long-term dependencies. The LSTM model's increased complexity also resulted in longer training times and a small chance of overfitting. Despite being computationally efficient, the Max-Pooling-based model was unable to capture the subtle spatial correlations seen in the data, which led to inferior accuracy and AUC-ROC scores when compared to the RNN-based models. Additionally, the accuracy and loss of this model are rather comparable to those of other models. The model using LSTM-simpleRNN is superior, in our opinion, based on the confusion matrix and reduced validation loss.

H. Discussion of Findings

These findings emphasize how crucial it is to choose the right aggregation methods for jobs involving the classification of medical images. Max-Pooling makes computations simpler, yet RNNs are able to recognize spatial patterns, which improves performance. The SimpleRNN model is a viable option for real-world use in clinical workflows because it achieves a compromise between efficacy and complexity. To further enhance performance, future research could investigate hybrid models that combine the advantages of both RNN and LSTM architectures.

VI. Lessons Learned

Insights into the use of deep learning methods for medical image classification, particularly for histopathological Whole-Slide Images (WSIs), were gained from this effort. By addressing a real-world issue, we were better able to comprehend the difficulties associated with handling sizable and intricate datasets, as well as the significance of choosing the right model architecture and efficiently augmenting data.

The importance of data preprocessing and augmentation in enhancing the generalizability of machine learning models was one of the main lessons learned. One important factor influencing model performance was the need for careful handling due to the variations in staining, illumination, and tissue architecture. Furthermore, testing several aggregation methods like Max-Pooling, SimpleRNN, and LSTM brought to light how crucial it is to strike a balance between computational complexity and the capacity to identify spatial links in the data.

In retrospect, there are a number of places where the project might have been enhanced by alternate choices. First off, the models might have been more robust if they had included data from more sources or a more varied dataset that included more forms of cancer. Second, it may have been possible to produce better-performing models more quickly by utilizing sophisticated hyperparameter optimization approaches like Bayesian optimization. Lastly, using explainability techniques like Grad-CAM would have improved the models' interpretability for clinical usage by offering insightful information about the models' decision-making process.

Ultimately, this study demonstrated the value of meticulous model performance evaluation, iterative experimentation, and practical considerations for creating machine learning solutions for healthcare applications.

VII. Conclusion

In order to categorize histological Whole-Slide Images (WSIs) into adenoma and adenocarcinoma groups, we investigated the use of deep learning algorithms in this study. We learned a lot about the efficacy of convolutional and recurrent neural networks for medical image classification by creating and assessing three model architectures: InceptionV3 with Max-Pooling, InceptionV3 with SimpleRNN, and InceptionV3 with LSTM. The InceptionV3 with SimpleRNN model, which balanced accuracy and computational economy, performed the best overall among the models.

In order to overcome the difficulties presented by big, high-dimensional datasets, the project underlined the need of preprocessing and data augmentation. The models were thoroughly evaluated using metrics including accuracy, AUC-ROC, and precision-recall, which guaranteed their robustness and dependability for practical clinical applications. Additionally, the creation of an intuitive graphical user interface (GUI) showed the possibility of using these models in real-world situations, allowing pathologists to diagnose patients more quickly and accurately.

Even if the outcomes were encouraging, there is still room for development, including the use of explainability techniques like Grad-CAM for improved interpretability, the incorporation of various datasets, and the investigation of hybrid architectures. These additions might strengthen machine learning's relevance and influence in the medical field.

This study concludes by demonstrating the revolutionary potential of deep learning for automating cancer diagnosis, providing a practical and scalable way to assist medical practitioners. The results highlight how crucial it is to carry out more research in this area in order to make breakthroughs that enhance patient outcomes and expedite diagnostic procedures.

VIII. Future Works

Although this effort has shown that deep learning may be used to categorize histopathology pictures, there are still a number of areas that need to be investigated further. Adding more datasets from various clinical contexts can improve the models' generalizability and resilience. In order to provide more thorough diagnostic insights, future research might also concentrate on integrating multimodal data, such as genetic and clinical information. Future work will focus on reducing the high number of false negatives while minimizing false positives to improve the reliability of the model in clinical applications.

Explainability techniques such as Grad-CAM can be used to depict the regions that contribute to the forecasts, improving the models' interpretability and building medical professionals' trust. Furthermore, the system's smooth integration into clinical workflows would be made possible by refining its real-time processing capabilities for Whole-Slide Images (WSIs) and implementing it as an on-premises or cloud-based application. With these improvements, the system should become more dependable, scalable, and useful in practical applications.

In order to capture long-term spatial dependencies and global context in histopathological pictures, future research should investigate hybrid architectures that integrate the advantages of both RNN and transformer-based models, such as Vision Transformers (ViT). Examining how self-supervised learning techniques could lessen reliance on labeled data might also be a fruitful avenue, especially in fields like medicine where annotating datasets requires a lot of resources. Furthermore, streamlining the workflow through the automation of the preprocessing pipeline—which includes patch extraction

and augmentation—would increase the solution's scalability and effectiveness for extensive clinical applications.

IX. Acknowledgements

Our profound appreciation goes out to Washington State University for their assistance and access to computational resources during this study. We would also like to thank Professor Janardhan Doppa of our course for his advice and helpful criticism during the project's development.

The authors and maintainers of the publicly accessible histopathology datasets included in this investigation are also acknowledged, as they played a crucial role in the development and assessment of our models. Their support of open data sharing has made medical image analysis research much easier. We also like to thank the forums, tutorials, and internet sites that offered insightful information about deep learning methods and application. In order to overcome technical obstacles and put cutting-edge solutions into practice, platforms like PyTorch documentation, TensorFlow, and Kaggle were essential.

Lastly, we would like to express our gratitude to our peers and colleagues for their helpful conversations and assistance, which helped us improve the concepts and techniques used in this project.

References

[1] O. Iizuka, F. Kanavati, K. Kato, et al., "Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours," *Scientific Reports*, vol. 10, no. 1504, 2020. Available: https://doi.org/10.1038/s41598-020-58467-9.

[2] Kaggle, "Histopathologic Cancer Detection Dataset," Available: https://www.kaggle.com/competitions/histopathologic-cancer-detection.

[3] Bas Veeling, "PatchCamelyon (PCam) Dataset," Available: https://github.com/basveeling/pcam.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778. Available: https://doi.org/10.1109/CVPR.2016.90.

[5] H. Sharma, N. Zerbe, I. Klempert, et al., "Deep Convolutional Neural Networks for Automatic Classification of Gastric Carcinoma Using Whole-Slide Images in Digital Histopathology," *Computerized Medical Imaging and Graphics*, vol. 61, pp. 2–13, 2017. Available: https://doi.org/10.1016/j.compmedimag.2017.06.001.

[6] B. Ehteshami Bejnordi, et al., "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer," *JAMA*, vol. 318, no. 22, pp. 2199-2210, 2017. Available: https://doi.org/10.1001/jama.2017.14585.

[7] R. R. Selvaraju, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618-626. Available: https://doi.org/10.1109/ICCV.2017.74.

[8] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. Available: https://doi.org/10.1016/j.patrec.2005.10.010.

[9] Y. Lecun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015. Available: https://doi.org/10.1038/nature14539.

[10] J. Deng, W. Dong, R. Socher, et al., "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255. Available: https://doi.org/10.1109/CVPR.2009.5206848.

[11] L. Taylor and G. Nitschke, "Improving Deep Learning Using Generic Data Augmentation," *Proceedings of the IEEE Symposium Series on Computational Intelligence*, 2018, pp. 154-160. Available: https://doi.org/10.1109/SSCI.2018.8628742.