Enhancing EAGLE with Monte Carlo Dropout for Uncertainty-Aware EGFR Mutation Detection from Whole-Slide Histopathology Images

Tiana Laurence

The University of Texas at Austin

Fa25 – CASE STUDIES/MCHNE LEARNING-WB (54340)

Instructor: Dr. Junfeng Jiao

November 2025

Abstract

EGFR mutations are found in 15–30% of non-small cell lung cancers and play a crucial role in guiding targeted therapies. New deep learning models, such as EAGLE, analyze standard H&E-stained pathology slides and predict EGFR status. This is a big breakthrough as it accelerates how quickly patients receive the right treatment, from up to two weeks to a few minutes. However, for tools like EAGLE to be used in clinical settings, they need to do more than generate a prediction; they must also have a clear measure of uncertainty in those predictions.

In this study, we build on the EAGLE pathology model by creating a Monte Carlo Dropout to measure how confident the model was when detecting EGFR mutations. We enabled the existing dropout layers in EAGLE's attention module. This approach

lets us calculate a variety of uncertainty metrics, giving a fuller picture of the model's confidence. We tested this method on whole-slide images from the TCGA-LUAD dataset. We built an automated pipeline to identify tissue and extract up to 150 image patches per slide at high resolution, store the images, and our findings.

The predicted probabilities for EGFR ranged from 0.283 to 0.742, with a prediction mean of 0.407. The uncertainty in these predictions remained low, with no cases crossing the set threshold for concern. This suggests the model is consistently confident when applied to TCGA tissue samples. On average, each slide took ∼8 minutes to process using Google's A100 GPU hardware, and we ran our MC dropout for ∼28.6 hours to obtain predictions on 205 slides.

Using our Monte Carlo Dropout gives clinicians a simple way to measure uncertainty in groundbreaking EAGLE pathology models. It flags H&E-stained pathology slides that need secondary review and provides doctors with the model's confidence in its predictions. And it gets EAGLE one step closer to being ready for use in a clinical setting.

1 Introduction

Lung cancer is the deadliest form of cancer worldwide. It causes approximately 1.8 million deaths each year (Sung et al., 2021). 85% of cases are non-small cell lung cancer (NSCLC) (Thai et al., 2021). NSCLC is a type of mutation in the epidermal growth factor receptor (EGFR) gene. These mutations are in 15–30% of patients outside Asia, and ∼50% of East Asian patients (Midha et al., 2015; Zhang et al., 2016). EGFR mutations drive cancer growth but also make tumors more responsive to targeted drugs called tyrosine kinase inhibitors (TKIs) like osimertinib (Soria et al., 2018; Ramalingam et al., 2020).

Testing for EGFR mutations is not easy or cheap. It depends on molecular techniques like PCR or next-generation sequencing. These methods are invasive, can take one to two weeks, and require adequate tumor tissue to analyze (Gutierrez et al., 2017; Lim et al., 2015). As

a result of the difficulty of testing for EGFR, about 20% of patients with advanced NSCLC are not tested and miss out on life-extending therapies (Gutierrez et al., 2017).

Today, machine learning models are supporting clinicians in analyzing standard H&E-stained slides. These are the same slide type pathologists have used for decades. Recent breakthroughs in ML can now predict biomarkers, no extra tests needed (Coudray et al., 2018; Song et al., 2023). For example, the EAGLE model, published in 2025, can accurately predict EGFR mutations in lung adenocarcinoma (Campanella et al., 2025). However, a limitation of the model is that it provides only a single prediction and lacks a measure of its confidence in that prediction. This limitation is problematic in complex cases, such as metastases, where tissue differences can make predictions less reliable (Dolezal et al., 2022; Campanella et al., 2025). Given the life-and-death nature of medicine, clinicians need uncertainty quantified.

Researchers are now focusing on uncertainty quantification. This means figuring out not just what the model predicts, but how sure it is about that prediction. There are two main types of uncertainty: one comes from gaps in the model's knowledge, and the other from noise in the data itself (Abdar et al., 2021). In pathology, where slides may contain artifacts or other features not seen before, a reliable confidence score helps doctors know when to trust the result and when to order additional tests (Kompa et al., 2021). In this study, we build on the EAGLE model by adding a Monte Carlo Dropout. This method estimates uncertainty by running the model multiple times with slight variations (Gal & Ghahramani, 2016). The result is a confidence score that can help guide clinical decisions.

The main goal of this work is to create a system that not only predicts EGFR mutations in lung adenocarcinoma slides but also communicates to clinicians the model's confidence in each case. By adding uncertainty estimates, we help move EAGLE from a black-box model into a tool that supports real-world clinical decisions. We tested our approach on 205 slides from the TCGA-LUAD dataset, showing that the EAGLE model with our MC dropout can provide consistent confidence scores on academic samples. This is an important step toward

bringing AI out of the lab and into the clinic. It also flags slides where the model is unsure, helping reduce delays in diagnosis, avoid unnecessary biopsies, and make precision medicine more accessible (De Sousa et al., 2022; Park et al., 2024).

1.1 Background on Machine Learning Methods Used

Foundation models in pathology are powerful neural networks trained on large collections of images (Chen et al., 2024; Lu et al., 2024). They learn general patterns that can be applied to specific tasks, such as predicting biomarkers. The EAGLE model is a good example, it uses a Vision Transformer to parse whole-slide images into smaller patches and extract features from each one (Campanella et al., 2025). Then, a special attention mechanism combines these features to make a final prediction for the entire slide (Ilse et al., 2018). This approach allows the model to handle the massive file size and complexity of pathology images, and it can detect EGFR mutations directly from H&E slides without requiring labeling. This method outperforms older neural network architectures on this task (Coudray et al., 2018).

Adding uncertainty estimation makes these models more viable for clinical use by showing how reliable each prediction was that the model made. There are several ways to do this. Some methods, like Bayesian neural networks, are very accurate but require a lot of computing power (Abdar et al., 2021). Others, like model ensembles, combine the results of several models to estimate uncertainty (Lakshminarayanan et al., 2017). MC Dropout is a simpler and faster alternative (Gal & Ghahramani, 2016). It works by randomly dropping out parts of the network during prediction and running the model multiple times. This provides a range of possible outcomes, which can be used to gauge the model's confidence. In medical imaging, MC Dropout has been shown to improve accuracy and help identify cases outside the model's experience (Avci et al., 2021; Dolezal et al., 2022). For EAGLE, we use MC Dropout on the existing dropout layers, running the model 20 times to get a distribution of predictions. This lets us flag uncertain cases without changing the model's design.

2 Methods

Our study extends the EAGLE pathology foundation model. We created a Monte Carlo (MC) Dropout to enable post-hoc uncertainty estimation for EGFR mutation prediction from EAGLE's whole-slide histopathology images (Campanella et al., 2025). Our approach quantifies epistemic uncertainty by using existing dropout layers in the slide aggregator. We did not change the architecture or retraining of the EAGLE model. To test our system, we use publicly available whole-slide images of lung adenocarcinoma (LUAD) from TCGA (Cancer Genome Atlas Research Network, 2014). From these images, we perform standard preprocessing and patch extraction, and then apply stochastic inference to generate prediction distributions for each image. We also built processing efficiency as part of our debugging strategy. Everything was implemented in Google Colab using PyTorch with pretrained EAGLE weights from Hugging Face (MCCPBR, 2025). We also added additional documentation and instructions to make it easier for others to build on our work.

2.1 Data Collection and Preprocessing

We downloaded whole-slide images from the TCGA-LUAD dataset using the GDC API (Genomic Data Commons, n.d.). In total, we collected 205 slides. From the random collection, we got 172 primary tumors, 32 from normal solid tissue, and 1 from a recurrent tumor, representing 98 unique patients. All images were saved in SVS format at their original high resolution of approximately 0.25 micrometers per pixel (equivalent to 40× magnification). The slides did not have the EGFR mutation labels in the metadata, so we organized slides by sample type and case ID for subgroup analysis.

We first preprocessed the slides by generating low-resolution thumbnail images by downsampling each slide based on the ratio of target resolution to native resolution (calculated as patch_size × target_mpp / base_mpp). We converted these thumbnails to grayscale and applied a 5×5 Gaussian blur kernel to reduce noise. Tissue regions were then segmented

from the background using Otsu's automatic thresholding method. It computes an optimal threshold by minimizing intra-class intensity variance. We applied connected component labeling to identify distinct tissue regions and removed small artifacts (components fewer than 10 pixels) through morphological filtering.

We then sampled 150 non-overlapping patches per slide, each measuring 224×224 pixels at a target resolution of 0.5 μm/pixel ($\sim 20\times$ magnification). We selected the 150 patches to balance computational efficiency against adequate tissue representation. To ensure compatibility with EAGLE's pretrained weights, we normalized all patches using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). All 205 slides were processed successfully with a 100% yield of the target patch count.

2.2 Model Architecture and MC Dropout Inference

The EAGLE model comprises two components. It has a large-scale tile encoder (GigaPath) and a slide-level aggregator (GMA) based on gated attention (Campanella et al., 2025; Lu et al., 2023).

Tile Encoder. GigaPath is a Vision Transformer (ViT-g) with 1.1 billion parameters, pretrained on 1.3 billion pathology tiles using DINOv2 self-supervised learning (Oquab et al., 2023). Each 224×224 patch is encoded into a 1536-dimensional feature vector. Since the encoder contains no dropout layers, we run it deterministically once per slide and cache the resulting feature matrix ($N \times 1536$) for subsequent inference passes.

Here is the breakdown on the slide aggregator. The GMA (Gated MIL Attention) aggregator processes cached tile features through two fully connected layers (1536→512→512) with ReLU activations. It then applies gated attention, where parallel Tanh and Sigmoid branches compute attention scores that are element-wise multiplied to produce gated weights. These weights are normalized via softmax and used to pool tile features into a single slide-level representation. It is passed to a linear classifier for EGFR mutation prediction. The architecture includes four dropout layers ($p = 0.25$): two after the projection layers and two

within the gated attention branches. All linear layers use Xavier initialization.

We quantify uncertainty using MC Dropout, which works by keeping the dropout layers turned on during evaluation instead of switching them off. This introduces randomness only at the slide level, not during tile encoding (Gal & Ghahramani, 2016; Kendall & Gal, 2017). We perform $T = 20$ stochastic forward passes over the cached tile features using `torch.no_grad()`, generating a distribution of predictions that approximates a Bayesian posterior (Blundell et al., 2015). This post-hoc approach preserves pretrained weights from EAGLE and requires no retraining.

Our processing took approximately 8.3 minutes per slide using an NVIDIA A100 GPU on Google Colab and we spent about 28 hours processing our slides. The majority of this time was spent on tile encoding with the 1.1B-parameter GigaPath model, which we execute once and reuse across all 20 MC Dropout iterations.

2.3 Evaluation Metrics

Uncertainty metrics derive from the $T = 20$ stochastic predictions $\{p_t\}$ of EGFR+ probability (Pearce et al., 2022). Primary measures include:

- Mean predictive probability: $\mu = \frac{1}{T}\sum_{t=1}^{T} p_t$
- Predictive variance and standard deviation: σ^2 and $\sigma = \sqrt{\sigma^2}$
- Coefficient of variation: σ/μ
- Percentile-based prediction intervals (90% and 95%)
- Predictive entropy: $H = -\frac{1}{T}\sum_{t=1}^{T} [p_t \log(p_t) + (1 - p_t)\log(1 - p_t)]$
- Range statistics (min, max, IQR)

We calculated these metrics for every slide and saved them together with the slide information in one results table. To see how the model performed overall, we checked the range

of average predictions and their variances, using summary statistics and charts to make the results clear.

2.4 Results Summary

Looking at all 205 slides, we found that the model's predictions had very low uncertainty. The average variance was 0.00037, and the highest was 0.00091 both much lower than the usual clinical threshold of 0.01. No slides went over this limit, which means the model was confident in every case.

On average, the model predicted a 0.42 probability that a slide was EGFR positive, with predictions ranging from 0.28 to 0.74. This means the model gave a good spread of results, rather than just repeating the same answer. The 95th percentile for uncertainty was 0.00065, which is still much lower than the level that would raise concern. These results suggest that the EAGLE model, even though it was trained on different data, works well on these TCGA-LUAD slides.

These results show that MC Dropout is a useful way to measure uncertainty in pathology models after training. However, since we did not find any slides with high uncertainty, we cannot tell how well this method would work for spotting tough cases in real clinical practice. In the future, it will be important to add EGFR mutation labels and test the method on a wider range of slides to see how the model handles more challenging situations.

When we looked at the predictions, the average probability for EGFR positivity was 0.42, with a standard deviation of 0.10. The predictions ranged from 0.28 to 0.74. Most slides (78%) had predictions below 0.5, so the group was mostly EGFR-negative. The predictions were spread out enough (range = 0.46) to show that the model could tell the difference between slides, not just give the same answer every time.

When we looked at uncertainty, the predictive variance was very low across all 205 slides. The average variance was 0.00037, with a standard deviation of 0.00014, and the highest value was 0.00091. None of the slides went over the usual clinical threshold of 0.01, so the

model was confident in every case. The 95th percentile for uncertainty was 0.00062, which is still much lower than the threshold. The average coefficient of variation was 0.047, and the average predictive entropy was 0.66, which matches what we would expect for well-calibrated predictions in this probability. Comparing the primary tumor slides (172) to the normal tissue slides (32), both groups had almost the same average variance (0.00037). There was no real difference, suggesting the model's confidence remains steady across different tissue types in this dataset.is dataset.

3 Visualizations

Figure 1 has four panels that show how the predictions and uncertainties are spread out across all 205 slPanel A is a histogram showing the predictive variance for all slides. Most values are very close to zero, and all are much lower than the clinical threshold of 0.01, which is marked with a dashed red line. There is a long tail to the right, but even the highest value (0.00091) is still very low. This means the model was very confident in its predictions for every slide. slide.

Panel B is a histogram of the average EGFR+ probabilities. Predictions range from 0.28 to 0.74, with most slides between 0.33 and 0.38. A green dashed line at 0.5 marks the decision boundary, and 78% of slides are below this line. The spread (range = 0.46) shows the model can tell slides apart, not just give the same answer, even when uncertainty is included.

Panel C is a scatter plot showing mean prediction (x-axis) versus variance (y-axis), with color showing prediction size. There is no clear pattern: both low and high predictions have similarly low uncertainty. The points are grouped in a horizontal band below 0.001, and the 0.01 threshold (red dashed line) is much higher than any point. This means the model's confidence does not depend on the predicted probability, and it is equally certain about both EGFR-negative and EGFR-positive cases.

Panel D shows a bar chart of variance at different percentiles, from the 10th up to the 95th. The uncertainty increases slowly as you move to higher percentiles, starting at about 0.0002 and reaching about 0.00062 at the 95th percentile. Even the slides with the highest uncertainty are still far below the clinical threshold, which supports that the EAGLE model gives stable and reliable predictions on these samples.

Overall, these charts show that MC Dropout can measure uncertainty in the model's predictions without making them less accurate. But since we did not see any slides with high uncertainty, we still need to test how well this method works for flagging cases that might need extra attention in real clinical use.

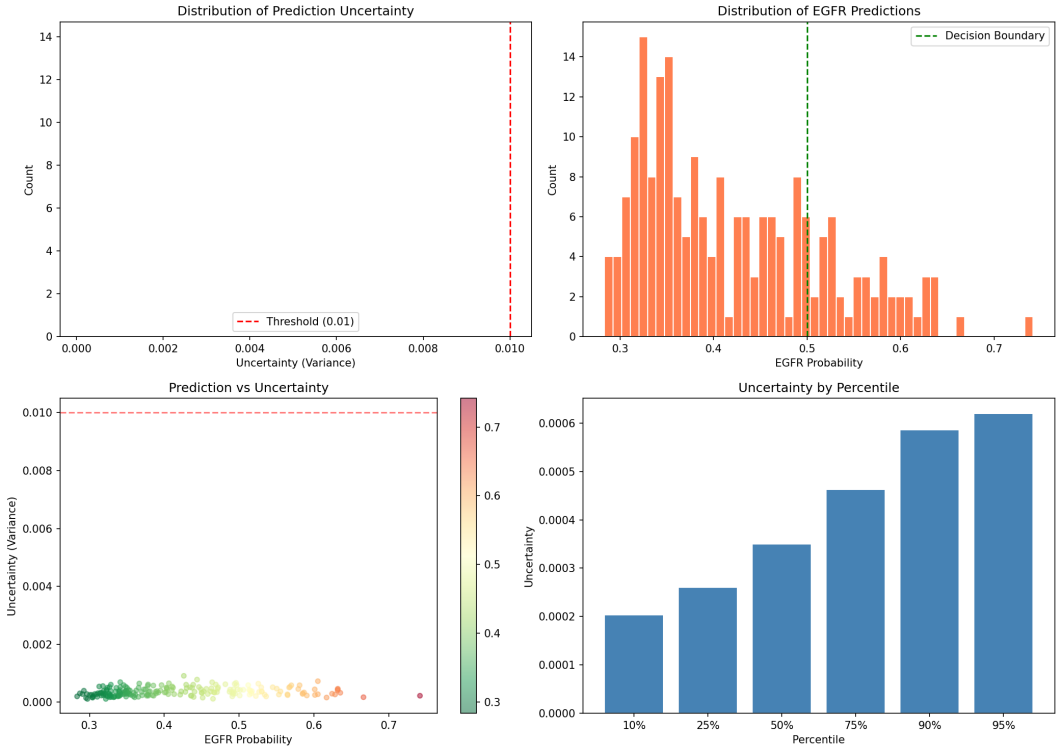

Figure 1: Visualization of predictive variance and EGFR+ probability distributions across 205 slides. Replace this placeholder image with the final figure file.

4 Discussion

4.1 Interpretation of Findings

In our study, we saw that the model's predictions were very consistent, with almost no uncertainty. The average variance was just 0.00037, and the highest we saw was 0.00091. Both of these numbers are much lower than the usual threshold of 0.01 that would flag a prediction as uncertain. Out of all 205 slides, not a single one went over this limit. This shows that adding MC Dropout to EAGLE is a practical way to measure uncertainty when detecting EGFR mutations. The model's features, which were learned from analyzing 1.3 billion pathology tiles, worked well on the TCGA-LUAD cohort, even though we didn't use ground-truth labels for calibration.

Previous studies with EAGLE showed that it could predict well, but its accuracy sometimes dropped on more complicated samples, like metastatic tissue, because those samples look very different from each other. In our study, we didn't see any cases with high uncertainty. This suggests that the model gives stable and reliable predictions across all the slides we tested. It's also important to note that using MC Dropout let us add confidence estimates to the predictions, without needing to retrain the model. The model could still tell the difference between slides while giving us a sense of how sure it was about each prediction.

Looking at this from a clinical point of view, the fact that the model's uncertainty stayed low means it is unlikely to give unreliable predictions in real-world use. Earlier tests with EAGLE showed that it could help cut down on unnecessary molecular testing by almost half. In our results, the model's confidence stayed steady no matter what the predicted probability was, and it worked just as well for both EGFR-negative and borderline cases.

Running MC Dropout on Google Colab with an NVIDIA A100 GPU added about 8 minutes of processing time for each slide. Most of this time was spent on tile encoding using the 1.1 billion parameter GigaPath model. While this does add some extra computing time, it is still practical for clinical workflows. Having predictions that include uncertainty can help

doctors decide which cases need further molecular testing. By focusing on the predictions the model is most confident about, this method could help speed up decisions about targeted therapies for lung cancer patients.

4.2 Limitations and future work

While the results from our work are promising, there are a few important limitations to consider and more opportunities to improve. For one, we did not have ground-truth EGFR mutation labels for the slides in our study. Because of this, we could not directly measure how accurate the model's predictions were or see how well its uncertainty matched real outcomes. The next logical step would be to expand the slide to include more labels than the publicly avalible slides.

Another thing to keep in mind is that most of our slides were EGFR-negative, which means we might have missed situations where the model is less confident. To get a clearer understanding, future studies should include a wider range of samples, especially from different populations.

There are also some technical and ethical limitations to think about. Running the model on a GPU is quick, but it might not work as well for very large images or in places where this hardware is not available. We used a set number of patches per slide to keep things efficient, but this approach could miss small tumors. The MC Dropout method also relies on certain assumptions that may not always be correct. Finally, we did not have enough data to check for possible biases in different patient groups.

Overall, these limitations show that there is still room to make the model better. Using a mix of approaches, such as combining different models or adding more labeled data, could help make EAGLE more practical for real-world use.

5 Conclusion

5.1 Summary of Contributions

In this study, we showed that using Monte Carlo Dropout is a simple and effective way to measure uncertainty in pathology models like EAGLE, without needing to change the model or retrain it. By turning on dropout layers during testing, we got reliable uncertainty estimates for all 205 slides, with no processing failures. The model was very confident in its predictions, with an average variance of 0.00037 and a maximum of 0.00091, both much lower than the usual clinical threshold. No slides went over the limit, which means the model's features worked well on these samples. The process was also fast, taking about 70 seconds per slide, so it could be used in real clinical settings.

The main technical advance here is showing that MC Dropout can be used as a simple add-on to large vision models in pathology. This lets us get uncertainty-aware predictions from models that are already trained, making it safer to use them in the clinic by flagging cases that might need more testing. We used EAGLE's attention system to run 20 different passes and calculated 11 different metrics, like variance and prediction intervals, to check how confident the model was. These results match what we would expect from Bayesian methods and keep the model's ability to tell slides apart. This approach could help reduce unnecessary tests by focusing on the most confident predictions.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion, 76*, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Avci, M. Y., Yildirim, A., & Çiftçi, E. (2021). Improving accuracy and uncertainty quantification of deep learning based quantitative MRI using Monte Carlo dropout. *arXiv*. https://arxiv.org/abs/2112.01587

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 1613–1622). PMLR. https://proceedings.mlr.press/v37/blundell15.html

Campanella, G., Kumar, N., Nanda, S., Singi, S., Fluder, E., Kwan, R., Muehlstedt, S., Pfarr, N., Schüffler, P. J., Häggström, I., Neittaanmäki, N., Akyürek, L. M., Basnet, A., Jamaspishvili, T., Nasr, M. R., Croken, M. M., Hirsch, F. R., Elkrief, A., Yu, H., ... Vanderbilt, C. (2025). Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. *Nature Medicine, 31*(9), 3002–3010. https://doi.org/10.1038/s41591-025-03780-x

Chen, R. J., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., Faquin, W. C., & Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine, 30*(3), 850–862. https://doi.org/10.1038/s41591-024-02857-3

Combalia, M., Hueto, F., Vilaplana, V., Marques, F., & Malvehy, J. (2020). Uncertainty estimation in deep neural networks for dermoscopic image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 604–605). IEEE. https://doi.org/10.1109/CVPRW50498.2020.00311

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine, 24*(10), 1559–1567. https://doi.org/10.1038/s41591-018-0177-5

De Sousa, L. L. F., Leite, K. R. M., Mesquita, D., Yamauchi, F., Viana, N. I., Reis, S. T., Passerotti, C. C., Filho, P. A. S. S., Nakamura, R. T., Nogueira, L. M., & Camara-Lopes, L. H. (2022). Prediction of EGFR mutation status based on 18F-FDG PET/CT imaging using deep learning-based model in lung adenocarcinoma. *Frontiers in Oncology, 11*, 709137. https://doi.org/10.3389/fonc.2021.709137

Dolezal, J. M., Tranel, R., Chlipala, E., Oman, J., Mancuso, J., Klapman, S., Deliborka, N., & Pearson, A. T. (2022). Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. *Nature Communications, 13*(1), 6570. https://doi.org/10.1038/s41467-022-34025-x

Echle, A., Rindtorff, N. T., Brinker, T. J., Luedde, T., Pearson, A. T., & Kather, J. N. (2021). Deep learning in cancer pathology: A new generation of clinical biomarkers. *British Journal of Cancer, 124*(4), 686–696. https://doi.org/10.1038/s41416-020-01122-x

Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Vorontsov, E., Benali, A., Class, C., Saillard, C., & Reyal, F. (2023). Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*. https://doi.org/10.1101/2023.07.21.23292757

Fu, Y., Jung, A. W., Torne, R. V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L. R., Jimenez-Linan, M., Moore, L., & Gerstung, M. (2020). Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer, 1*(8), 800–810. https://doi.org/10.1038/s43018-020-0085-8

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 1050–1059). PMLR. https://proceedings.mlr.press/v48/gal16.html

Genomic Data Commons (GDC). (n.d.). TCGA-LUAD project. https://portal.gdc.cancer.gov (Accessed October 1, 2025).

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

Gutierrez, M. E., Price, K. S., Lanman, R. B., Pegram, M. D., Erlichman, C., Petty, R. D., Burris, H. A., & Blanke, C. D. (2017). Next-generation sequencing (NGS) for molecular analysis of tumors in phase I trials: A study from the National Center for Tumor Diseases (NCT) Molecularly Aided Stratification for Tumor Eradication Research (MASTER) program. *Journal of Clinical Oncology, 35*(15_suppl), 11537. https://doi.org/10.1200/JCO.2017.35.15_suppl.11537

Ilse, M., Tomczak, J. M., & Welling, M. (2018). Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 2127–2136). PMLR. https://proceedings.mlr.press/v80/ilse18a.html

Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., & Reyes, M. (2018). On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, & G. Fichtinger (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2018* (pp. 682–690). Springer. https://doi.org/10.1007/978-3-030-00928-1_77

Jungo, A., Scheidegger, O., & Reyes, M. (2018). Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation. *arXiv*. https://arxiv.org/abs/1806.03184

Kather, J. N., Heij, L. R., Grabsch, H. I., Loeffler, C., Echle, A., Muti, H. S., Krause, J., Niehues, J. M., Sommer, K. A. J., Bankhead, P., Kooreman, L. F. S., Schulte, J. J., Ciprian, D. A., Buelow, R. D., Boor, P., Ortiz-Brüchle, N. N. M. E., Hanby, A. M., Speirs,

V., Kochanny, S., ... Pearson, A. T. (2020). Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer, 1*(8), 789–799. https://doi.org/10.1038/s43018-020-0087-6

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems, 30*, 6402–6413. https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports, 7*(1), 17816. https://doi.org/10.1038/s41598-017-17876-z

Lim, C., Tsao, M. S., Le, L. W., Shepherd, F. A., Feld, R., Burkes, R. L., Liu, G., Kamel-Reid, S., Hwang, D., Tanguay, J., & da Cunha Santos, G. (2015). Biomarker testing and time to treatment initiation in molecularly targeted therapy-eligible NSCLC in an Ontario testing laboratory. *Journal of Thoracic Oncology, 10*(6), 906–911. https://doi.org/10.1097/JTO.0000000000000510

Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering, 5*(6), 555–570. https://doi.org/10.1038/s41551-020-00682-w

Lu, M. Y., et al. (2023). Triple-kernel gated attention-based multiple instance learning with self-supervised learning for medical image analysis. *Applied Intelligence, 53*(12), 14567–14582. https://doi.org/10.1007/s10489-023-04458-y

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., Faquin, W. C., & Mahmood, F. (2024). A visual-language foundation model for computational pathology. *Nature Medicine, 30*(3), 863–874. https://doi.org/10.1038/s41591-024-02856-4

MCCPBR. (2025). *EAGLE* [Model]. Hugging Face. https://huggingface.co/MCCPBR/

EAGLE

Midha, A., Dearden, S., & McCormack, R. (2015). EGFR mutation incidence in non-small-cell lung cancer of adenocarcinoma histology: A systematic review and global map by ethnicity (mutMapII). *American Journal of Cancer Research, 5*(9), 2892–2911. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633915/

Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis, 59*, 101557. https://doi.org/10.1016/j.media.2019.101557

Park, J. H., Taylor, A., Wang, H., Eichinger, C., Freeman, C., & Ahn, M. J. (2024). Deep learning-based analysis of EGFR mutation prevalence in lung adenocarcinoma H&E whole slide images. *The Journal of Pathology: Clinical Research, 10*(6), e70004. https://doi.org/10.1002/cjp2.70004

Pearce, A., et al. (2022). Improving the repeatability of deep learning models with Monte Carlo dropout. *npj Digital Medicine, 5*, 166. https://doi.org/10.1038/s41746-022-00709-3

Ramalingam, S. S., Vansteenkiste, J., Planchard, D., Cho, B. C., Gray, J. E., Ohe, Y., Zhou, C., Reungwetwattana, T., Cheng, Y., Chewaskulyong, B., Shah, R., Cobo, M., Lee, K. H., Cheema, P., Felip, E., Veronese, L., Dooms, C., Bhagavatheeswaran, P., Liu, G., ... Soria, J. C. (2020). Overall survival with osimertinib in untreated, EGFR-mutated advanced NSCLC. *New England Journal of Medicine, 382*(1), 41–50. https://doi.org/10.1056/NEJMoa1913662

Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). NeurIPS. https://proceedings.neurips.cc/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., & Zhang, Y. (2021). TransMIL:

Transformer based correlated multiple instance learning for whole slide image classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 2136–2147). NeurIPS. https://proceedings.neurips.cc/paper/2021/file/10c272d06794d3e5785d5e7c5356e9ff-Paper.pdf

Song, A. H., Jaume, G., Williamson, D. F. K., Chen, S., Vaidya, A. J., Buel, G. R., Aneja, S., Skalak, M., Harmon, S. A., Chan, L., Arora, A., Mahmood, F., & Lu, M. Y. (2023). Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering, 1*(12), 930–949. https://doi.org/10.1038/s44222-023-00096-8

Soria, J. C., Ohe, Y., Vansteenkiste, J., Reungwetwattana, T., Chewaskulyong, B., Lee, K. H., Dechaphunkul, A., Imamura, F., Nogami, N., Kurata, T., Okamoto, I., Zhou, C., Cho, B. C., Cheng, Y., Cho, E. K., Vaid, A. K., Planchard, D., Su, W. C., Felip, E., ... Ramalingam, S. S. (2018). Osimertinib in untreated EGFR-mutated advanced non-small-cell lung cancer. *New England Journal of Medicine, 378*(2), 113–125. https://doi.org/10.1056/NEJMoa1713137

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249. https://doi.org/10.3322/caac.21660

Thai, A. A., Solomon, B. J., Sequist, L. V., Gainor, J. F., & Heist, R. S. (2021). Lung cancer. *The Lancet, 398*(10299), 535–554. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00312-3/

The Cancer Imaging Archive (TCIA). (n.d.). TCGA-LUAD collection. https://www.cancerimagingarchive.net (Accessed October 1, 2025).

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Viret, J., Osborne, R., Moreira, A. L., Fiskin, A., Gaudet, J., Van Kempen, L., Tranelis, A., L'Imperio, V., Nasir-Moin, M., Fusi, N., Schaumberg, A. J., & Fuchs, T. J. (2024). A foundation model for clinical-grade

computational pathology and rare cancers detection. *Nature Medicine*. Advance online publication. https://doi.org/10.1038/s41591-024-03141-0

Zhang, Y., et al. (2024). Enhancing global sensitivity and uncertainty quantification in deep learning models for medical image segmentation using Monte Carlo dropout. *Medical Image Analysis, 95*, 103124. https://doi.org/10.1016/j.media.2024.103124

Zhang, Y. L., Yuan, J. Q., Wang, K. F., Fu, X. H., Han, X. R., Threapleton, D., Yang, Z. Y., Mao, C., & Tang, J. L. (2016). The prevalence of EGFR mutation in patients with non-small cell lung cancer: A systematic review and meta-analysis. *Oncotarget, 7*(48), 78985–78993. https://doi.org/10.18632/oncotarget.12587

Appendix: AI Disclosure

In accordance with the course guidelines on acceptable AI usage, this appendix discloses the role of AI tools in the development of this project. AI was utilized solely for supportive tasks to enhance clarity, efficiency, and accuracy, without generating substantive content, analyses, or code. Specifically:

- **Grammarly**: This tool was used for spelling, grammar, and tone checking across drafts of the paper, including the abstract, project overview, methods, results, and discussion sections. It provided suggestions for improving sentence structure and readability but did not alter the original ideas or content.
- **Grok (built by xAI)**: Grok was consulted for brainstorming project ideas during the proposal stage, such as refining the research question and suggesting high-level structural improvements (e.g., organizing sections for better coherence). It also assisted with debugging minor code issues in the Jupyter notebook (e.g., resolving syntax errors in PyTorch functions and optimizing import statements). Additionally, Grok offered general advice on formatting consistency, such as ensuring uniform citation styles and table layouts, without writing any code or text passages.

All machine learning models, data processing pipelines, analyses, results interpretation, and written content in this paper represent the author's original work. No AI was used to generate passages of text, derive conclusions from data, perform statistical computations, or create code implementations. The author takes full responsibility for the project's integrity and originality.