High-Resolution Interpretable Classification of Artifacts versus Real Variants in Whole Genome Sequencing Data from Archived Tissue

Dylan Domenico [*1,2] **Georgios Asimomitis** [*1,3] **Gunes Gundem** [1,2] **Emily Stockfisch** [2] **Cassidy Cobbs** [4] **Neeman Mohibullah** [4] **Leonidas G Alexopoulos** [3] **Neerav Shukla** [2] **Andrew Kung** [2] **Elli Papaemmanuil** [1,2]

Abstract

Whole genome sequencing (WGS) has demonstrated substantial benefits in computational genomics and oncology, with costs rapidly decreasing. A major limitation in WGS is the reliance on fresh frozen (FF) tissue specimens which pose challenges for long-term storage and transportation, contrary to formalin-fixed paraffin-embedded (FFPE) preserved samples which have a long-standing history of use in the clinical domain. However, ex vivo processes involved with FFPE preparation and extraction lead samples to acquire non-biological, artifact mutations presenting challenges in WGS analyses. Here we present an interpretable deep learning-based workflow designed to classify mutations from FFPE samples as either artifacts or real variants on the basis of pileup image visualizations. The model boasts an AUC of 0.9 and yields explainable outcomes that lead to the identification of high-resolution input traits representative of key concepts in variant assessment.

1. Introduction

FFPE preparations represent a widely utilized method for preserving biospecimens. Institutions have accrued hundreds of millions of FFPE samples housed in bio-archives across the globe, with the number increasing daily (Sah et al., 2013). While this approach offers benefits in terms of storage and transportation compared to resource-intensive, costly FF alternatives, it also presents the challenge of compromised quality. Deamination and fragmentation resulting from FFPE may give rise to misleading sequence alterations or "artifact mutations," leading to convoluted results in molecular assays such as WGS (Mathieson & Thomas, 2020). Numerous studies have attempted to optimize fixation and extraction processes prior to sequencing to address this issue, albeit with varying levels of success (Do & Dobrovic, 2015; Robbe et al., 2018). From a computational perspective, a limited number of tools have been developed to tackle this concern. Guo et al. (2022) identify a set of FFPE Single Nucleotide Variant (SNV) signatures by utilizing signature extraction techniques without classifying artifacts at the mutation level per se, while Dodani et al. (2022) provide a logistic regression model coupled with feature importance analysis to distinguish real variants from artifacts. However, these approaches do not model the spatial relationships of the reads across loci nor do they yield explainable classification outcomes at the mutation level or even the sample level.

To address these issues, this study, to the best of our knowledge, is the first to present a novel deep learning-based strategy for classifying FFPE single SNVs as artifacts or real variants, from WGS data, while offering per-variant interpretability. To do this: 1) we convert a set of high confidence variant calls to pileup images by utilizing Google's DeepVariant workflow (Poplin et al., 2018); 2) we train a convolutional neural network (CNN) with the ResNet-50 architecture (He et al., 2015) to distinguish artifact variants from genuine mutations; 3) we apply Guided Grad-CAM and provide high resolution interpretable visual maps for each variant (Selvaraju et al., 2019). This allows us to offer insights into the model's behavior as well as localize traits associated with the decision-making processes employed by clinicians and researchers during manual variant review.

2. Data & Methods

Our study cohort consists of spatially adjacent and matched FF and FFPE tumor resections acquired for ten pediatric patients with varying disease types from Memorial Sloan Kettering Cancer Center (MSKCC) (Shukla et al., 2022).

[*]Equal contribution [1]Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA [2]Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, NY, USA [3]Biomedical Systems Laboratory, Department of Mechanical Engineering, National Technical University of Athens, Athens, Greece [4]Integrated Genomics Operation Core, Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. Correspondence to: Dylan Domenico <domenicd@mskcc.org>.

The 2023 ICML Workshop on Computational Biology. Honolulu, Hawai'i, USA, 2023.

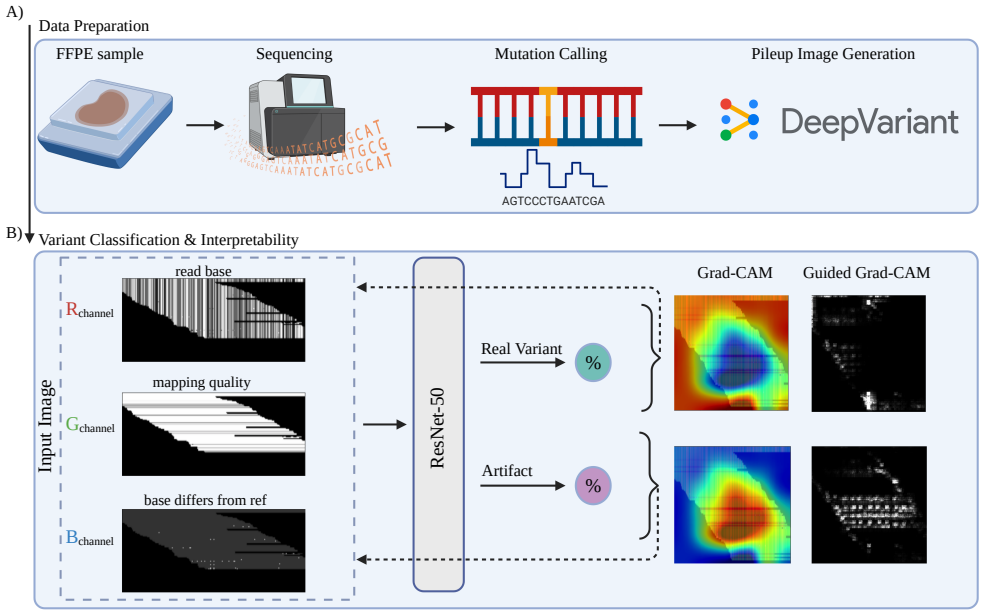

Figure 1. **Specimen Sequencing to Interpretable Classification Framework.** Workflow describing sample processing to model output and interpretation results. Divided into two subsections: A) Data preparation consisting of sample sequencing, variant identification, and image generation; B) Image classification and human-understandable Grad-CAM and Guided Grad-CAM overlays. RGB channels are shown in grayscale and represent read base with different pixel intensities denoting each base pair, mapping quality with intensity corresponding to higher values, and bases that differ from the reference where intensity denotes a base that mismatches. The variant itself is centered on the x-axis. Grad-CAM results originate from the feature maps of the last convolutional layer of the ResNet-50 architecture.

To analyze this data we developed a multi-step framework that takes raw sequencing data as input and produces variant classification paired with high-resolution interpretability maps (Figure 1).

2.1. Data Preparation

Firstly, FF/FFPE tumor samples along with a matched normal from peripheral blood were concurrently processed and sequenced for each patient with an intended coverage of 80x for the tumor samples and 30x for the normals. Utilizing the Isabl platform (Medina-Martínez et al., 2020), sequencing data underwent alignment to reference genome build GRCh37, quality control, and somatic SNV calling using the consensus approach described in the online methods section of Shukla et al. (2022) (Figure 1A).

In the context of labeling the output of the somatic variant calling, independent call sets from the matched FF/FFPE specimens were examined, and the mutations common to both specimens were identified as definitive, real variants. Conversely, unique mutation calls originating solely from FFPE samples were categorized as potential artifact variants, a method employed by previous studies mentioned to assess sensitivity and precision of detection from FFPE (Robbe et al., 2018; Dodani et al., 2022). Our final dataset consists of 45,815 real variants and 218,431 artifacts.

Next, the binary alignment map (BAM) files for the FFPE tumors in conjunction with the previously described call sets in variant call format (VCF) are used as input for Google's DeepVariant `make_examples` module to generate a set of 6-channel pileup images per variant (Figure 1A). These channels represent specific characteristics at each locus including: 1. read base; 2. base quality; 3. mapping quality; 4. strand; 5. read supports variant; and 6. bases differing from the reference. To enable the visualization of each variant in a more human-readable format as well as to enhance the understanding of subsequent model interpretability outcomes, we map the read base, mapping quality, and bases diverging from the reference channels (of each variant) to each of the RGB channels of an image respectively (Figure 1B). We note that the selection of read base, mapping quality, and bases diverging from the reference as the pileup channels that formed the RGB image for each variant, was based on evaluating the performance of the model (described in the next paragraph) across all three-way combinations of pileup channels (see Appendix A, Table 2) as well as on the biological significance of the channels. The same model and interpretability strategy can be used for any triplet of pileup channels.

2.2. Variant Classification and Interpretability

Representing each variant as an RGB image allows us to employ the widely-applied ResNet-50 Convolutional Neural Network architecture (He et al., 2015) as the core of our

classification model that receives as input one such RGB image per variant and produces two probability estimates, one for each of the two candidate classes, namely real variant and artifact (Figure 1B). These probabilities are the result of a softmax operation performed after the last linear layer of ResNet-50 and denote how likely it is for a variant to be either a real one or an artifact. To localize the image regions that are most significant for the model's classification decision as well as to highlight the fine-grained pixel importances on the input image, we apply Grad-CAM and Guided Grad-CAM on the trained model, respectively (Selvaraju et al., 2019) for each class per variant.

3. Results

To form an input set for the task of classifying real variants from artifacts, we randomly sourced variants from all available samples. To prevent class imbalance, we obtained up to 2,500 real variants and 2,500 artifacts from each sample. This approach was grounded in the observation that samples harbored a median of 2,442 real variants and 9,040 artifacts. We split the total 45,034 variant calls into training, validation, and test sets by randomly assigning 70%, 15%, and 15% respectively from each class. The ResNet-50 model was trained and optimized based on the training and validation sets and tested for its performance on the holdout test data. Particularly, it was trained for 15 epochs in batches of 32 images using the Adam optimizer (Kingma & Ba, 2017) with a learning rate of 0.001 and a weight decay of 0.001. The class with the highest output probability was regarded as the output class. The evaluation of the model performance on the test set showed that the model has a remarkable capacity (accuracy 90%, precision 89%, recall 90%, AUC 90%) to separate real variants from artifacts (Table 1).

Grad-CAM and Guided Grad-CAM outputs were overlaid onto each respective image for comprehensive interpretation. To understand the model's decision-making process we examined the results for the variants from the test set. For instance, Figure 2A shows a characteristic example of an artifact. This can be justified by the presence of low mapping quality regions, visually demonstrated by the two prominent horizontal sequences in rows 90 and 120, as well as the lack of supporting alternate alleles as indicated by the very few and scarce white pixels in the column 110 of the 'base differs from ref' channel. The model has the capacity to correctly classify this case with confidence (assignment probability to the artifact class is 99.9%) by identifying both traits mentioned above. Particularly, the low mapping quality stands out in the Guided Grad-CAM of the artifact class as displayed by the horizontal sequences in rows 90 and 120. The few good quality reads of the image that are representative of alternate allele support are highlighted in the Guided Grad-CAM of the real variant class in column 110 between rows 40-50. On the contrary, the reads supporting alternate alleles located in the low mapping quality sequences, are disregarded in the Guided Grad-CAM of the real variant class and are identified as important in the Guided Grad-CAM of the artifact class.

Figure 2B displays a real variant noted by supporting alternate alleles of high mapping quality in column 110 between rows 100-190. These characteristics are captured as important for the correct assignment of the case as a real variant. This is shown through the stretch of white pixels in column 110 in the Guided Grad-CAM of the real variant class. However, as displayed in the original RGB image, this case also includes several truncated sequences demonstrated by reads with premature stops (i.e. rows 100-110) that would typically be indicative of an artifact. The model identifies this as support of a potential artifact as presented by the increase in pixel intensity for these regions in the Guided Grad-CAM of the artifact class. This reduces the confidence assigned to the real class (57.4%), but not at a level that would lead to misclassification.

Table 1. Confusion matrix of model performance on the test set.

		Output class	
		Artifact	Real variant
True class	Artifact	3100	332
	Real variant	373	2952

4. Discussion

In this study, we develop an interpretable classification framework that differentiates artifacts from real variants in FFPE samples from WGS data. The initial part of the framework includes an automated pipeline to align and curate the raw sequencing data, identify SNVs, convert them to pileup images and subsequently to RGB ones. The classification backbone of the framework utilizes the ResNet-50 architecture and deploys Grad-CAM and Guided Grad-CAM in the context of explaining the model behavior at the pixel level.

The model shows notable performance, misclassifying only 10% of the test set. As expected, the model output probabilities of these misclassified cases for the incorrect class (e.g. predicted artifact class for real variant cases and vice versa) are skewed, in their majority, towards the 50% assignment threshold, indicating lower confidence of classification (Figure 3A, Appendix B). However, for artifacts predicted to be real variants only, the output probability distribution contains an additional peak in the lower range percentage values (high confidence of case being a real variant). Examination of these respective artifacts suggests that these could be true variants owing to intratumor heterogeneity (ITH) and therefore that their misclassification is not a product of poor model performance, but of a caveat in the labeling

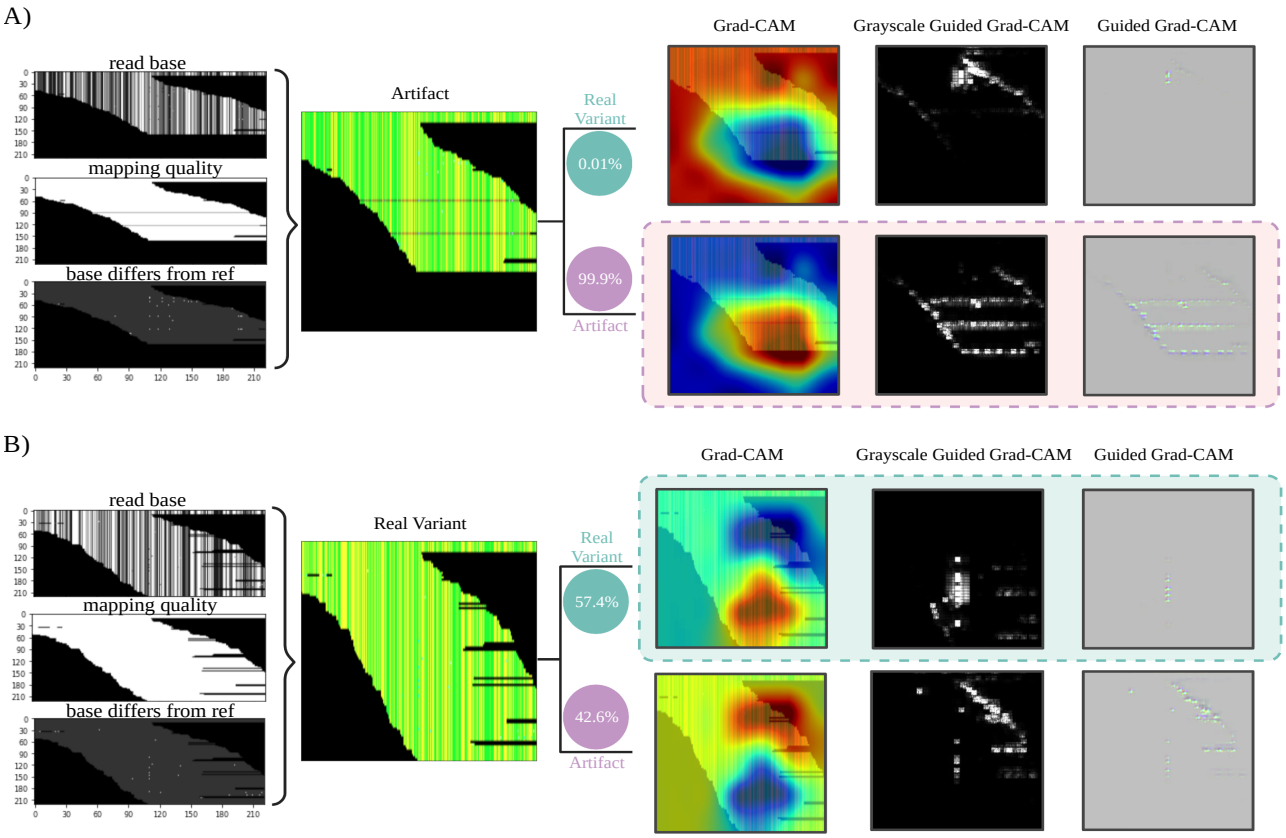

Figure 2. **Interpretable Examples. (A)** Example 3-channel RGB image of an artifact properly classified as such by model. Bubbles denote class percentages based on model output. Righthand plots show Grad-CAM overlay (class probability increasing from blue to red), Grayscale Guided Grad-CAM, and Guided Grad-CAM results for each output class. **(B)** Same as **(A)** but for a real variant properly classified as such by the model.

process (Figure 3B, Appendix B).

Analyzing the model behavior and the interpretability of its outcomes, we note that Grad-CAM results alone provide a coarse yet comprehensible explanation, highlighting the areas of the pileup images that held particular significance for the model's prediction probabilities for each class. Moreover, the results of Guided Grad-CAM allow for a higher resolution, class-discriminative visualization that helps understand which areas and structures within the pileup images are used during classification. For a majority of artifacts, this involves areas inclusive of truncated and/or low-quality mapped reads (Figure 2A). On the other hand, the model gains confidence for the prediction of the real variant class by relying on regions of high mapping quality with mutant allele support (Figure 2B). These characteristics of the input are also typically invoked during manual variant review, showing that the model closely mirrors the key concepts decisive for the experts' annotation. In particular, the different pixel-level importances highlight reads typically of interest to experts, outlining the advantage of accounting for the heterogeneity across reads over the lower-resolution feature averages per variant employed in past approaches (Figure 4, Appendix C).

Through Guided Grad-CAM, it also becomes apparent that the model specifically emphasizes both the pixels corresponding to the exact reads of highest importance (e.g. specific reads characterized by low mapping quality in Figure 2A presented as prominent horizontal lines in the visual maps) as well as the subsequent sequences that make up the context surrounding the variant (e.g. in Figure 2B highlighted by high intensity pixels neighboring the central column in the real variant class). The latter property springs from the representation of the variants as RGB images and the application of convolution operations that enables the classification process to account for the spatial relationship between aligned reads and their overall qualities.

Concluding, we outline that this classification framework is extendable to any type of labeled WGS variants, while its inherent aspect of interpretability empowers its broad utilization by researchers and clinicians. Incorporation of images that enable rapid visualization and recognition of artifactual regions as compared to high quality variants can provide an additional information layer to support clinical reporting workflows of molecular data.

References

Do, H. and Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: Causes and strategies for minimization. *Clinical Chemistry*, 61(1):64–71, 2015. doi: 10.1373/clinchem.2014.223040. URL https://doi.org/10.1373/clinchem.2014.223040.

Dodani, D. D., Nguyen, M. H., Morin, R. D., Marra, M. A., and Corbett, R. D. Combinatorial and machine learning approaches for improved somatic variant calling from formalin-fixed paraffin-embedded genome sequence data. *Frontiers in Genetics*, 13:834764, 2022. doi: 10.3389/fgene.2022.834764. URL https://doi.org/10.3389/fgene.2022.834764.

Guo, Q., Lakatos, E., Bakir, I. A., et al. The mutational signatures of formalin fixation on the human genome. *Nature Communications*, 13:4487, 2022. doi: 10.1038/s41467-022-32041-5. URL https://doi.org/10.1038/s41467-022-32041-5.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Mathieson, W. and Thomas, G. Why formalin-fixed, paraffin-embedded biospecimens must be used in genomic medicine: An evidence-based review and conclusion. *Journal of Histochemistry and Cytochemistry*, 68(8):543–552, 2020. doi: 10.1369/0022155420945050. PMID: 32697619; PMCID: PMC7400666.

Medina-Martínez, J. S., Arango-Ossa, J. E., Levine, M. F., et al. Isabl platform, a digital biobank for processing multimodal patient data. *BMC Bioinformatics*, 21:549, 2020. doi: 10.1186/s12859-020-03879-7. URL https://doi.org/10.1186/s12859-020-03879-7.

Poplin, R., Chang, P.-C., Alexander, D., et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36:983–987, 2018. doi: 10.1038/nbt.4235. URL https://doi.org/10.1038/nbt.4235.

Robbe, P., Popitsch, N., Knight, S. J. L., et al. Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 genomes project. *Genetics in Medicine*, 20:1196–1205, 2018. doi: 10.1038/gim.2017.241. URL https://doi.org/10.1038/gim.2017.241.

Sah, S., Chen, L., Houghton, J., Kemppainen, J., Marko, A. C., Zeigler, R., and Latham, G. J. Functional DNA quantification guides accurate next-generation sequencing mutation detection in formalin-fixed, paraffin-embedded tumor biopsies. *Genome Medicine*, 5(8):77, 2013. doi: 10.1186/gm481. URL https://doi.org/10.1186/gm481.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019. doi: 10.1007/s11263-019-01228-7. URL https://doi.org/10.1007%2Fs11263-019-01228-7.

Shukla, N., Levine, M. F., Gundem, G., et al. Feasibility of whole genome and transcriptome profiling in pediatric and young adult cancers. *Nature Communications*, 13:2485, 2022. doi: 10.1038/s41467-022-30233-7. URL https://doi.org/10.1038/s41467-022-30233-7.

A. Channel Benchmarking

Three out of six of the available channels from DeepVariant's pileup image structure were used in order to adhere to an RGB image format and maintain human readability. The selection of these channels was primarily driven by the results of benchmarking all possible combinations from the six available channels (reference table). The channels were labeled as follows: 0 - read base; 1 - base quality; 2 - mapping quality; 3 - strand; 4 - read supports variant; and 5 - base differs from reference. All possible channel combinations were evaluated using the same training and validation split percentages previously described (70% and 15% of mutations respectively) and trained for 15 epochs. We compared the validation accuracy of these combinations to determine the most performative arrangements.

The results demonstrate that the channelizations with validation accuracy greater than 90% were 025, 234, and 345. We opted for the combination denoted by 025, which integrates read base, mapping quality, and bases that differ from the reference as we believe that the actual biological sequence (which is explicitly mapped by the pixel values for channel 0 - read base) is crucial for a comprehensive assessment at each locus. We acknowledge that a trade-off was made for channel 3 (strand) which also depicts potentially important information for artifact detection. This was not the case for channel 4 (read supports variant) as this information is captured through the combinatorial aspects of channels 0 and 5.

Table 2. Benchmarking of all combinations of pileup channels as input to the classification framework.

Benchmarking	
Channel Combination	Validation Accuracy %
012	82.5
013	77.3
014	77.6
015	89.4
023	82.4
024	89.2
025	**90.3**
034	82.6
035	89.1
045	89.7
123	82.6
124	89.7
125	89.8
134	85.5
135	87.3
145	86.2
234	**90.4**
235	89.5
245	90
345	**90.1**

B. Supplementary Figures

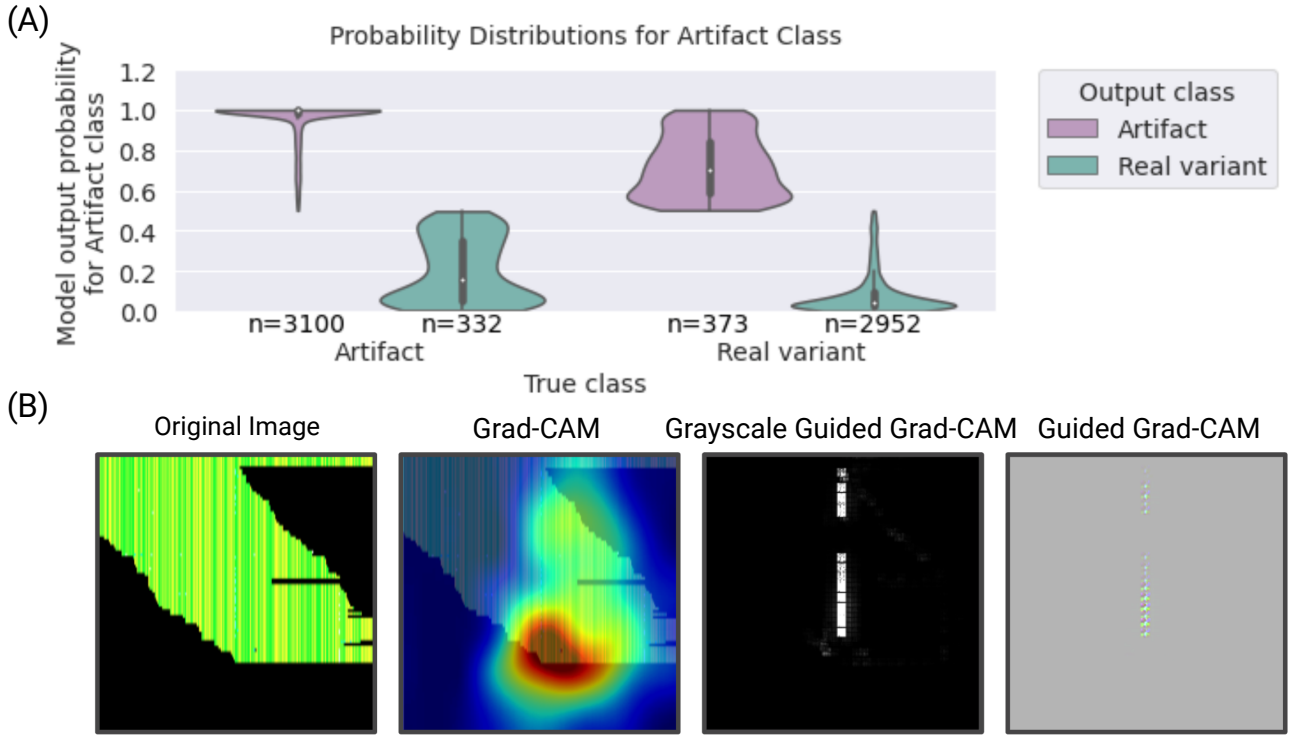

Figure 3. **Probability Distributions. (A)** Violinplot showing the distribution of probabilities attributing to the artifact class for all variants from the test set. **(B)** Example of a mutation labeled as an artifact but classified as a real variant with a very low artifact class probability of 0.4% that is likely due to ITH. Panel shows original RGB image, Grad-CAM and Guided Grad-CAM results for the real variant class.

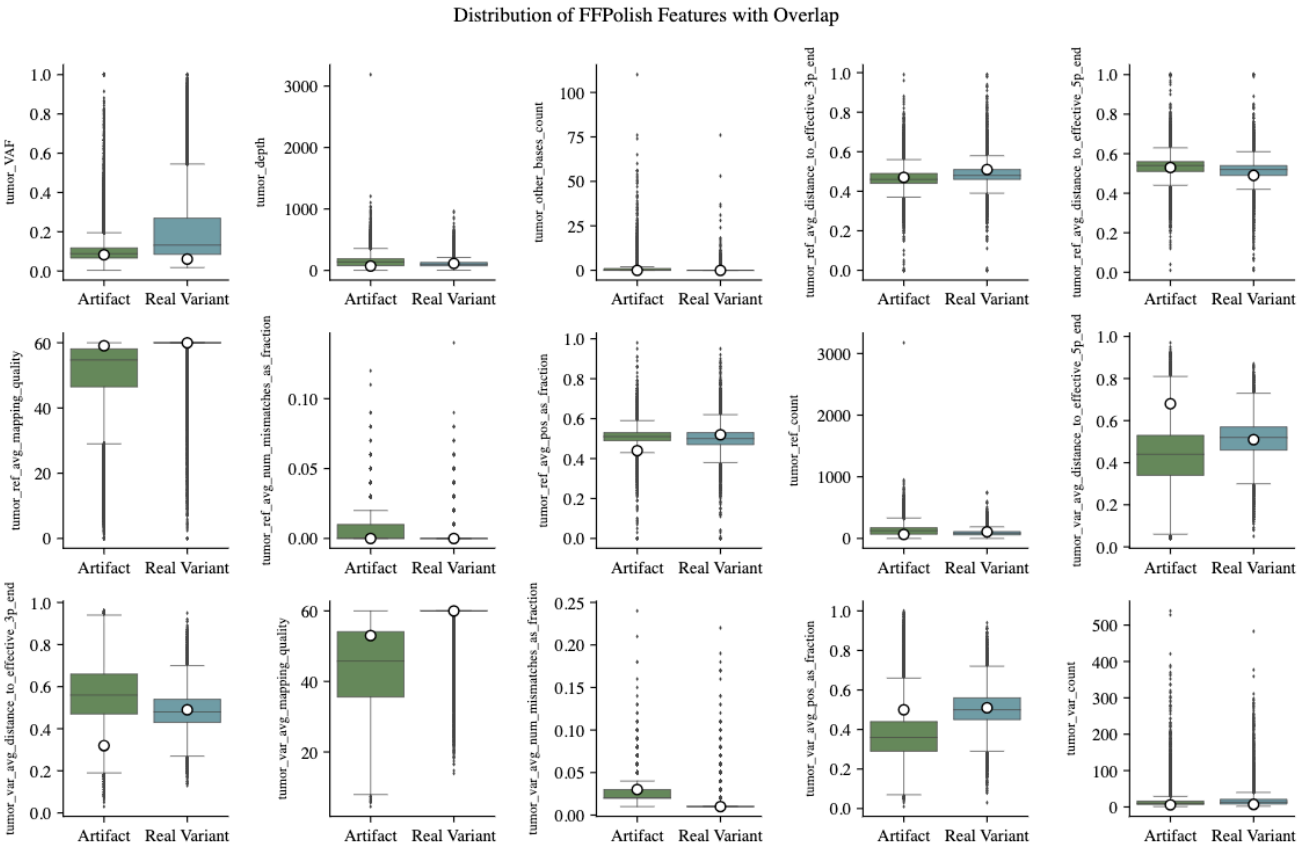

Figure 4. **Feature Comparisons.** Distributions in all available mutations for 15 features from FFPolish (Dodani et al., 2022) logistic regression model. The features displayed are the ones that can be observed visually in the pileup images used in our implementation. Mutations are split into their associated true class. The feature values for the artifact and real variant highlighted in Figure 2 are represented by the single white points for each class. FFPolish classifies both mutations as artifacts.

C. Use Case Comparison with Variant Metric Based Approach

The presented framework resembles the manual variant review process by taking read-specific characteristics into account instead of lower-resolution averages across all reads per variant. The real variant presented in Figure 2A showcases the usefulness of considering for read heterogeneity. In particular, FFPolish (Dodani et al., 2022), a logistic regression based model that uses averages across read level statistics as features, classifies this case as an artifact in contrast to our framework that correctly recognizes this case as a real variant. Specifically, almost all of the FFPolish feature values of this mutation displayed in Figure 4, are close to the median values of the real variants. However the average tumor variant allele frequency (VAF) is low, particularly close to the median level of VAF for artifacts. This might mislead to the classification of this real variant as an artifact. Our framework, by processing read-specific traits, identifies the few variant alleles of this case, highlights them as important in the context of interpretability and correctly classifies this mutation as a real variant.

D. Code Availability

Source code of the model and interpretability approach will become available in the Papaemmanuil lab github upon publication.