MultiVarNet - Predicting Tumour Mutational status at the Protein Level

Louis-Oscar Morel[1], Muhammad Muzammel[2], Nathan Vinçon[2], Valentin Derangère[3], Sylvain Ladoire[3], and Jens Rittscher[4]

[1] Medical Science Division, University of Oxford, Oxford, UK
[2] Ummon HealthTech, 21000 Dijon, France
[3] Centre Georges François Leclerc, 21000 Dijon, France
[4] Department of Engineering Science, University of Oxford, Oxford, UK
`louis-oscar.morel@linacre.ox.ac.uk`

Abstract. Deep learning research in medical image analysis demonstrated the capability of predicting molecular information, including tumour mutational status, from cell and tissue morphology extracted from standard histology images. While this capability holds the promise of revolutionising pathology, it is of critical importance to go beyond gene-level mutations and develop methodologies capable of predicting precise variant mutations. Only then will it be possible to support important clinical applications, including specific targeted therapies.

To address this need we developed MultiVarNet which allows us to decipher complex genomic patterns, facilitating precise predictions of hotspot alterations at the protein level. For the first time we demonstrate that we can achieve notable success in identifying over 20 mutation variants across major oncogenes. This study introduces a novel approach that underscores the importance of incorporating the underlying molecular biology of tumours to enhance algorithm accuracy, moving us towards more personalized and advanced targeted treatment options for patients.

Keywords: Digital Pathology · Biomedical Imaging · Image Based Phenotyping · Tumoral Mutation Variants · Genomic Subtypes.

1 Introduction

With an estimated 9.6 million deaths, or one in six deaths, in 2018 according to the WHO [5], cancer continues to be a significant global health challenge because of the heterogeneity and intricacy of the disease. There has been significant progress in precision medicine aimed to provide more efficient therapies by targeting specific molecular tumour profiles [12]. Molecular diagnostic tools, such as DNA sequencing, RNA quantification or methylation profiling, can reveal specific molecular characteristics of the tumours[16, 4]. These techniques are crucial as they condition the access to molecularly guided treatment options (MGTOs) which are specifically designed to target these precisely identified alterations [4, 8]. However, these tests require large tumour samples, have long waiting times, and are costly [6, 11]. For instance, Gondos and colleagues [6] found that almost a quarter of patients with newly diagnosed advanced non-small cell lung cancer (NSCLC) in their large study did not receive 'gold-standard' genomic testing for any of the four guideline-recommended therapeutic targets (ALK, BRAF, EGFR, and ROS1 alterations) before first-line treatment due to these limitations. Consequently, there is an increasing demand for alternative solutions to conventional molecular profiling methods, aiming to fulfill the pressing need for comprehensive testing of molecular alterations[11].

Meanwhile, a growing body of evidence supports the use of deep learning in analyzing hematoxylin and eosin (H&E) stained histopathology images to infer molecular information [9, 10, 14, 15], demonstrating state-of-the-art performance in predicting outcomes [9] and relevant biomarkers[10]. These methods have been applied to predict single somatic mutations, copy number variations, molecular subtypes, RNA expression, and prognosis [10, 15]. However, despite their success in identifying overall gene mutation statuses, the precise prediction of the protein consequences resulting from specific mutations within genes has yet to be explored. This gap is significant, as variations at the protein level often hold greater clinical relevance [7, 20]. These specific alterations, hereafter referred to as variants, dictate access to targeted therapies such as Sotorasib [7] and Adagrasib [22]. These drugs target specific protein variants like p.G12C in genes such as KRAS in NSCLC, highlighting the importance of analyzing mutations at a variant-specific level rather than solely at the gene level.

We address the need of predicting specific cancer-associated mutations and variants across 11 cancer types, leveraging deep learning analysis of H&E digitized slides. Our approach not only demonstrates the potential to accurately predict variant alterations, often surpassing gene-level mutation accuracy but also introduces a novel label-engineering paradigm to exploit unique morphological signatures of these variants. Indeed, current developments in deep learning methodologies primarily focus on enhancing architectures and training processes, often overlooking the significance of available biological information. Our MultiVarNet method illustrates that this information can be intelligently leveraged to enhance mutation prediction accuracy, offering a novel direction for advancement in the field of histopathology.

2 Novel architecture for predicting gene variants

In this study, we introduce MultiVarNet, a novel deep learning approach designed to predict genetic mutations, protein variants, and particularly, simplified gene mutations. MultiVarNet is grounded in the hypothesis that each variant exhibits a unique morphological signature, which can be leveraged to represent the mutation's morphological spectrum and improve the algorithm predictions. This method (illustrated in Fig. 1.C) targets simplified gene mutations, defined by the two most prevalent variants, utilizing a multi-modal strategy. Starting from the diagnostic slides contained in Aperio SVS files from 11 TCGA datasets, identified by the 'DX' label in their filenames, we employ an in-house trained

U-net for foreground extraction [18]. The slides are then segmented into 600x600 pixel patches at 5x resolution (Fig. 1.A). These image patches labeled with each variant and their corresponding simplified genes are then fed to three distinct ImageNet-based pretrained EfficientNetB7 neural networks [3, 21]. Following feature extraction via average pooling layers, the high-level variant features are directed into two separate multilayer perceptrons (MLPs), each consisting of a 128-dimension layer with dropout, culminating in unique linear and sigmoid outputs for each variant (V1 and V2). These variant features, along with high-level simplified gene features, are then concatenated and fed into another MLP, followed by linear and sigmoid layers for the classification of simplified gene mutations, where each of the three MLPs includes a 128-dimension layer with dropout and a ReLU activation layer.

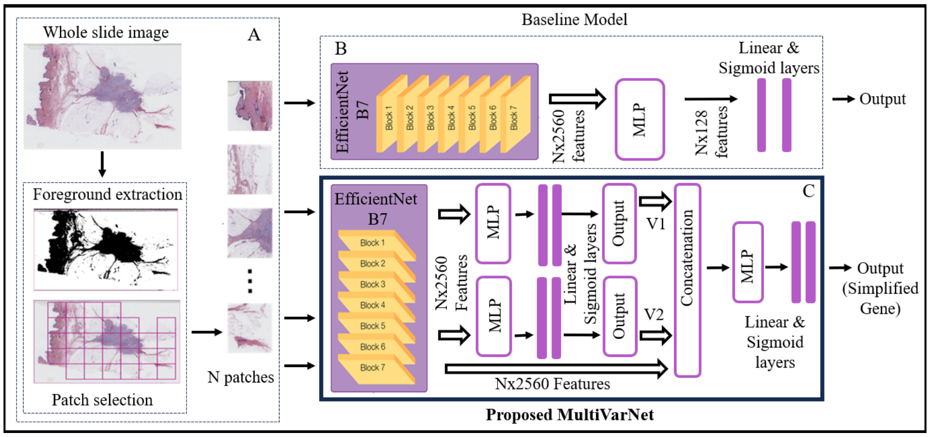

Fig. 1: Proposed methodologies to predict gene mutations, protein alterations and pseudo gene mutations.

3 Experimental Design

To validate MultiVarNet's effectiveness, we compared it against a baseline approach. The baseline method involves predicting gene mutations, single variant alterations, or simplified gene mutations using a conventional strategy. In this setting, the image patches obtained after the foreground extraction are classified using an EfficientNetB7 neural network pretrained on ImageNet [3, 21], proceeding through an average pooling layer to a MLP setup, identical in structure to those used in MultiVarNet, for final classification (as depicted in Fig. 1.B). Both approaches utilized MLPs trained for 5 epochs on the training set with Adam optimizer (learning rate: 1e-4, loss function: binary cross-entropy). Each assessment was conducted through a 3-fold cross-validation approach, where in slides originating from the same case were grouped together within the same fold for the same dataset. We assessed the discriminative power of the provided approaches by computing the mean AUC across 3 folds and assessed the statistical significance using Mann–Whitney U test [13]. Experiments were performed using NVIDIA RTX A4000 graphics card, with Python 3.8, TensorFlow 2.13.0, Keras 2.13.0, and CUDA 11.5 software stack.

4 Dataset and Data Processing

This study is based upon a retrospective analysis, employing de-identified scanned Whole Slide Images (WSIs) procured from The Cancer Genome Atlas (TCGA) project. TCGA dataset spans a diverse range of cancer types from multiple centers. Comprehensive information regarding the TCGA dataset and patient-related particulars can be found in [2, 17]. In order to assess the robustness of our methodology across a diverse spectrum of cancer sub-types, we curated a selection of 11 distinct TCGA datasets, as given in supplementary Table 1. Further, we utilized a public dataset, which will be essential for other teams to replicate our results.

To ensure adequate statistical power for quantifying significant effects, we restricted our analysis to protein variants occurring in at least 15 patients. Moreover, we did not take into account the fast frozen for slides selection. Consequently, this criterion yielded a final set of 20 gene mutations and 35 protein alteration/pathology pairs across all datasets, as detailed in Table 1 and supplementary Table 2.

We generated three types of labels to guide our analysis: genetic mutation, protein variant, and simplified genes. To begin, we extracted critical information from the MAF (Mutation Annotation Format) file using the MuTect2 algorithm [1]. This information includes 'IMPACT', indicating the level of pathogenicity (HIGH or MODERATE suggesting significant changes); 'tumor_Sample_Barcode', identifying the specific tumour samples; 'Hugo_Symbol', denoting the gene symbols; and 'HGVSp_Short', which specifies the somatic mutations at the protein level. We then labeled the presence of pathogenic mutations with '1' and the absence with '0'. Labels for protein variants were created based on gene symbols and specific mutation details, allowing us to track mutations down to their impact on protein structure. Gene mutations are flagged if any pathogenic variant is detected, encompassing any significant alteration within a gene.

The concept of simplified genes is developed to explore the impact of morphological signatures from gene variants. Specifically, our MultiVarNet method is tailored to incorporate just two variants per gene, combining them to enhance the predictability of the gene as a whole. These simplified genes are constructed from the two most prevalent variants of a gene, defining a simplified gene mutation as the presence of either variant. This approach helps us determine whether improvements in prediction accuracy come from the method itself or merely from reducing complexity by excluding less common variants. Further details on protein variant outcomes are discussed in Section 5.1. From 11 different pathologies, we selected 8 simplified genes for this analysis.

5 Results

5.1 Gene Mutations Status and Variants Alteration Prediction

To predict gene mutations and precise protein variants from WSIs, we used the established baseline deep learning setup. This model is key for predicting gene mutations and protein variants in various cancers by analyzing digital pathology slides on a large scale. We derived slide-level predictions by aggregating tile-level predictions, focusing on the 99th percentile of these values as an indicator of mutations [14]. We categorized clinically significant genes (KRAS, EGFR, IDH1, BRAF, NRAS, presented in Table 1) separately from others (PIK3CA, TP53, CDKN2A in Supplementary Table 2) due to spatial constraints. The model successfully predicted 15 out of 20 gene mutations and showed that 20 out of 35 specific variant mutations were identifiable, and sometimes with higher accuracy than their overall gene mutations. Among these mutations, the IDH1 gene mutation showed robust discrimination in GBM (AUC=81.17%, P=1.07E-12) and LGG (AUC=80.04% ,P=1.72E-34) (see Table 1). Similarly, for the NRAS gene mutation, the baseline model achieved an AUC of 56.72% (P=0.026) and an AUC of 77.58% (P=1.00E-08) for THCA for SKCM datasets. However, the model exhibited reduced discriminative capacity for TP53 in LUSC (AUC=57.25%, P=0.055), as detailed in supplementary Table 2. This variability underscores the intricate nature of morphological signatures associated with mutations, emphasizing the imperative for our ongoing research to refine and enhance predictive methodologies.

Some specific variant mutations were distinctly identifiable and achieved higher AUC scores compared to the overall gene mutations AUC scores. For instance, BRAF p.V600E (AUC=87.76%,P=6.54E-49) in THCA, as observed in Table 1, exhibited enhanced discernibility compared to the gene's overall mutation (AUC=82.81%, P=5.36E-37). However, not all variant mutations showed this pattern. Some, like p.Q61R for NRAS in THCA (AUC=70.88%, P=9.60E-05) or p.V600M for BRAF in SKCM (AUC= 54.48%, P=0.34), were less predictable than their overall gene mutation counterparts. Interestingly, the predictability of the same protein variant, such as PIK3CA p.E545K (refer to Supplementary Table 2), varied significantly across different types of cancer, suggesting distinct morphological signatures specific to each variant. This observation led us to evaluate whether we could take advantage of this fine-grained morphological signature to improve the predictability of overall gene mutations.

5.2 Simplified Gene Mutations Prediction

We designed the MultiVarNet proof-of-concept method to evaluate whether we could take advantage of these morphological variabilities to improve the predictability of overall gene mutations, and compared it to standard weakly supervised setups used in the literature, using two aggregation methods for slide-level

Table 1: Genes and protein variants prediction scores.

dataset	Gene	Variant	Mean AUC (%)	Mean P value
COAD	KRAS	All	66.31	4.26E-09
		p.G12V	65.29	0.003
		p.G12D	58.98	0.034
GBM	EGFR	All	65.00	4.02E-09
		p.A289V	67.33	0.0055
		p.G598V	65.08	0.0386
	IDH1	All	81.17	1.07E-12
		p.R132H	73.05	9.47E-06
LGG	IDH1	All	80.04	1.72E-34
		p.R132C	74.33	8.34E-06
		p.R132G	86.29	1.50E-06
LUAD	EGFR	All	62.823	0.0002
		p.L858R	75.67	1.57E-06
		p.E746_A750del	70.58	0.005
	KRAS	All	62.07	1.76E-05
		p.G12D	63.53	0.045
		p.G12C	55.12	0.2025
SKCM	BRAF	All	58.77	0.001
		p.V600E	64.64	3.67E-08
		p.V600M	54.48	0.3431
	NRAS	All	56.72	0.0258
		p.Q61L	58.62	0.2276
		p.Q61K	70.76	3.67E-06
		p.Q61R	63.46	0.0018
THCA	BRAF	All	82.81	5.36E-37
		p.V600E	87.76	6.54E-49
	NRAS	All	77.58	1.00E-08
		p.Q61R	70.88	9.60E-05

prediction accuracy: the 99th percentile and the mean of tile prediction values, as detailed in recent studies [14, 9] and considered as the best aggregators for molecular prediction.

As shown by our results displayed in Tables 2 and 3, MultiVarNet consistently outperformed the baseline models across various datasets and cancer types, revealing a significant improvement in predictive accuracy as measured by the Area Under the Curve (AUC). For example, in the bladder cancer (BLCA) dataset, MultiVarNet using 99th percentile aggregator achieved a higher mean AUC of 71.47% (P=2.39E-07) for the PIK3CA simplified gene mutation compared to the baseline's 69.14% (P=4.10E-06) as given in Table 2. Enhanced performance was observed across multiple datasets, including COAD, SKCM, and LUAD, underscoring the relevance of our approach in capturing the nuanced morphological features associated with these mutations. However, in the UCEC dataset for PIK3CA simplified gene mutation, the baseline model marginally outperformed MultiVarNet with the mean AUC of 65.12% (P=0.0031) compared to MultiVarNet's mean AUC of 64.75% (P=0.0039).

Table 2: Performance of baseline model and MultiVarNet for simplified gene mutations (using the 99th percentile of their tile prediction values).

Dataset	Gene Mutation	Number of cases	Proteins Pair	Baseline Model [14] mean AUC (%)/ Mean P value	MultiVarNet mean AUC (%)/ Mean P value
BLCA	PIK3CA	43/386	p.E545K, p.E542K	69.14/4.10E-06	**71.47/2.39E-07**
BRCA	PIK3CA	68/687	p.E545K, p.E542K	56.872/0.0587	**57.46/0.0403**
COAD	KRAS	81/451	p.G12D, p.G12V	66.17/3.99E-06	**67.21/9.10E-07**
GBM	EGFR	17/389	p.A289V, p.G598V	60.08/0.0355	**61.20/0.0195**
LGG	IDH1	25/491	p.R132C, p.R132G	84.10/2.47E-14	**84.48/1.28E-14**
LUAD	KRAS	72/478	p.G12C, p.G12D	57.67/0.0319	**60.30/0.004**
SKCM	NRAS	83/433	p.Q61K, p.Q61R	62.72/0.0001	**64.52/1.39E-05**
UCEC	PIK3CA	24/505	p.E542K, p.G118D	**65.11/0.0031**	64.76/0.0039

Table 3: Performance of baseline model and MultiVarNet for simplified gene mutations (using the mean of their tile prediction values).

Dataset	Gene Mutation	Number of cases	Proteins Pair	Baseline Model [9] mean AUC (%)/ Mean P value	MultiVarNet mean AUC (%)/ Mean P value
BLCA	PIK3CA	43/386	p.E545K, p.E542K	71.94/1.30E-07	**72.09/1.05E-07**
BRCA	PIK3CA	68/687	p.E545K, p.E542K	59.24/0.0110	**59.58/0.0084**
COAD	KRAS	81/451	p.G12D, p.G12V	65.75/7.05E-06	**66.21/3.75E-06**
GBM	EGFR	17/389	p.A289V, p.G598V	68.61/0.0001	**69.29/5.72E-05**
LGG	IDH1	25/491	p.R132C, p.R132G	85.62/1.68E-15	**86.59/2.84E-16**
LUAD	KRAS	72/478	p.G12C, p.G12D	**61.45/0.0013**	60.94/0.0022
SKCM	NRAS	83/433	p.Q61K, p.Q61R	62.99/0.0001	**63.15/8.34E-05**
UCEC	PIK3CA	24/505	p.E542K, p.G118D	69.96/9.41E-05	**70.48/6.17E-05**

Consistent with the 99th percentile aggregator, the MultiVarNet also outperformed the baseline model using the mean aggregation (Table 3) for simplified genes defined in various cancer subtypes. For instance, in BLCA (72.09%, P=1.05E-07) compared to 71.94%, P=1.30E-07), BRCA (59.58%, P=0.0084 vs 59.24%, P=0.0110), COAD (66.21%, P=3.75E-06 vs 65.75%, P=7.05E-06), and LGG (86.59%, P= 2.84E-16 vs 85.62%, P=1.68E-15), respectively. These findings underscore MultiVarNet's superior predictive accuracy for simplified gene mutations across diverse cancer types.

6 Conclusion and future work

The goal of precision oncology is to tailor therapeutic strategies to individual molecular tumour profiles. Identifying tumour gene mutations is essential for prescribing targeted therapies effectively [8]. Several deep learning methods [9, 10, 15] on WSIs have shown effectiveness in predicting gene mutations, offering a cost-effective alternative. We demonstrate that is is possible to predict specific protein variants from WSI with high accuracy, suggesting each variant may have a distinct morphological signature, promising refinement in precision oncology approaches.

Our experimental results show that MultiVarNet can improve existing methods in predicting gene mutations, highlighting the potential for enhancing deep learning methods with variant-specific morphological signatures. These results offer valuable insights for oncological research. For example, in breast cancer (BRCA), our algorithms reliably identify PIK3CA gene variants p.E545K and p.E542K, targeted by Alpelisib, hinting at pharmaceutical applications [19]. This specificity underscores the morphological consequences of mutations and suggests predictability differences among variants may reflect their biological impact, and could guide therapeutic strategies using morpho-molecular correlations. In future work we will validate these results in external datasets to overcome some of the known limitations of TCGA. In addition, we plan to explore a deeper biological integration, e.g. the prediction of signalling pathway disruptions, to provide a more holistic view of tumoural behavior and therapeutic opportunities.

In this study, we focus on the innovative application of our MultiVarNet proof-of-concept method, emphasizing its potential to strengthen the prediction of molecular alterations by harnessing distinct variant morphological signatures. This approach is grounded in the integration of detailed molecular biology information and label engineering, marking a departure from the traditional emphasis on deep learning architecture and training methodologies. Given the exploratory and pioneering nature of MultiVarNet, our comparison is limited to establish its foundational capabilities and demonstrate its unique contribution to enhancing gene-wide mutation predictability. This study spotlights the method's novelty and its potential to open new avenues for research, rather than positioning it as another incremental improvement in model architecture or training processes. The scope of comparative analysis is tailored to underscore the conceptual advancement MultiVarNet represents. We demonstrate that even with a simpler model, we can achieve performance improvements. Our results are significant, and future efforts will explore more sophisticated architectures to leverage these findings effectively. One factor contributing to the relatively modest improvement is that we benchmark our results against two state-of-the-art methods ([9, 14]) and focus exclusively on variants, which reduces the number of examples and complicates the demonstration of enhanced performance. Analyzing more complex architectures will require a systematic study of robustness on larger datasets. TCGA, a multicentric and public dataset, allows for benchmarking new methods, which is essential for other teams to replicate our results and build on our findings. However, external validation will further strengthen our demonstration.

With this first paper we establish the principal feasibility and highlight the need of this important biological context. In this paper we have demonstrated a very exciting possibility of developing a new class of biomarkers. This paper lays the groundwork for further research to enhance our understanding of fine-grained tissue morphology. While exploring various architectures is important, this paper sets a foundational baseline for a new set of challenges. The MultiVarNet method is designed to show that each variant mutation within a single gene has a unique signature that can be leveraged. Each MLP is tasked with predicting a specific variant, thereby directing them to identify distinct morphological signatures, rather than a composite of signatures. Finally, we do acknowledge that further validation will be necessary before this technology can be translated into the clinical setting.

Compliance with Ethical Standards Our study strictly adhered to the Declaration of Helsinki and International Ethical Guidelines. Ethics oversight for the TCGA study is outlined at the provided link https://www.cancer.gov/ccg/ research/genome-sequencing/tcga/history/ethics-policies. All participants provided informed consent. Methodologies and outcomes were reported following STARD guidelines for transparency.

Disclosure of Interests. Louis-Oscar Morel and Nathan Vinçon own shares in the Ummon HealthTech company. Other authors have no conflicts of interest to declare.

Acknowledgements This work was supported by Cancer Research UK.

References

1. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G.: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology **31**(3), 213–219 (2013)
2. Collisson, E., Campbell, J., Brooks, A., Berger, A., Lee, W., Chmielecki, J., Beer, D., Cope, L., Creighton, C., Danilova, L., et al.: Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. Nature **511**(7511), 543–550 (2014)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. Ding, J., Adiconis, X., Simmons, S.K., Kowalczyk, M.S., Hession, C.C., Marjanovic, N.D., Hughes, T.K., Wadsworth, M.H., Burks, T., Nguyen, L.T., et al.: Systematic comparison of single-cell and single-nucleus rna-sequencing methods. Nature biotechnology **38**(6), 737–746 (2020)
5. GLOBOCAN, W.H.O.: International agency for research on cancer. (2018), https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf. [Accessed: 10/02/2023]

6. Gondos, A., Paz-Ares, L.G., Saldana, D., Thomas, M., Mascaux, C., Bubendorf, L., Barlesi, F.: Genomic testing among patients (pts) with newly diagnosed advanced non-small cell lung cancer (ansclc) in the united states: A contemporary clinical practice patterns study. Journal of Clinical Oncology **35**(15) (2020)
7. Hong, D.S., Fakih, M.G., Strickler, J.H., Desai, J., Durm, G.A., Shapiro, G.I., Falchook, G.S., Price, T.J., Sacher, A., Denlinger, C.S., et al.: Krasg12c inhibition with sotorasib in advanced solid tumors. New England Journal of Medicine **383**(13), 1207–1217 (2020)
8. Horak, P., Heining, C., Kreutzfeldt, S., Hutter, B., Mock, A., Hüllein, J., Fröhlich, M., Uhrig, S., Jahn, A., Rump, A., et al.: Comprehensive genomic and transcriptomic analysis for guiding therapeutic decisions in patients with rare cancers. Cancer discovery **11**(11), 2780–2795 (2021)
9. Laleh, N.G., Muti, H.S., Loeffler, C.M.L., Echle, A., Saldanha, O.L., Mahmood, F., Lu, M.Y., Trautwein, C., Langer, R., Dislich, B., et al.: Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Medical image analysis **79**, 102474 (2022)
10. Lee, S.H., Jang, H.J.: Deep learning-based prediction of molecular cancer biomarkers from tissue slides: A new tool for precision oncology. Clinical and Molecular Hepatology **28**(4), 754 (2022)
11. Mateo, J., Steuten, L., Aftimos, P., André, F., Davies, M., Garralda, E., Geissler, J., Husereau, D., Martinez-Lopez, I., Normanno, N., et al.: Delivering precision oncology to patients with cancer. Nature Medicine **28**(4), 658–665 (2022)
12. McCann, K.E., Hurvitz, S.A., McAndrew, N.: Advances in targeted therapies for triple-negative breast cancer. Drugs **79**, 1217–1230 (2019)
13. McKnight, P.E., Najab, J.: Mann-whitney u test. The Corsini encyclopedia of psychology pp. 1–1 (2010)
14. Morel, L.O., Derangère, V., Arnould, L., Ladoire, S., Vinçon, N.: Preliminary evaluation of deep learning for first-line diagnostic prediction of tumor mutational status. Scientific Reports **13**(1), 6927 (2023)
15. Murchan, P., Ó'Brien, C., O'Connell, S., McNevin, C.S., Baird, A.M., Sheils, O., Ó Broin, P., Finn, S.P.: Deep learning of histopathological features for the prediction of tumour molecular genetics. Diagnostics **11**(8), 1406 (2021)
16. Nichols, R.V., O'Connell, B.L., Mulqueen, R.M., Thomas, J., Woodfin, A.R., Acharya, S., Mandel, G., Pokholok, D., Steemers, F.J., Adey, A.C.: High-throughput robust single-cell dna methylation profiling with scimetv2. Nature communications **13**(1), 7627 (2022)
17. Raju, B..W.H..H.M.S.C.L...P.P.J.K., data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, G., for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, I., et al.: Comprehensive molecular portraits of human breast tumours. Nature **490**(7418), 61–70 (2012)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
19. Rugo, H.S., Lerebours, F., Ciruelos, E., Drullinsky, P., Ruiz-Borrego, M., Neven, P., Park, Y.H., Prat, A., Bachelot, T., Juric, D., et al.: Alpelisib plus fulvestrant in pik3ca-mutated, hormone receptor-positive advanced breast cancer after a cdk4/6 inhibitor (bylieve): one cohort of a phase 2, multicentre, open-label, non-comparative study. The Lancet Oncology **22**(4), 489–498 (2021)

20. Skoulidis, F., Li, B.T., Dy, G.K., Price, T.J., Falchook, G.S., Wolf, J., Italiano, A., Schuler, M., Borghaei, H., Barlesi, F., et al.: Sotorasib for lung cancers with kras p. g12c mutation. New England Journal of Medicine **384**(25), 2371–2381 (2021)
21. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
22. Yaeger, R., Weiss, J., Pelster, M.S., Spira, A.I., Barve, M., Ou, S.H.I., Leal, T.A., Bekaii-Saab, T.S., Paweletz, C.P., Heavey, G.A., et al.: Adagrasib with or without cetuximab in colorectal cancer with mutated kras g12c. New England Journal of Medicine **388**(1), 44–54 (2023)