



Xavier Fresquet, Jürgen Hesser (eds.)

1st. Joint DFH/UFA workshop on AI in Medicine: Optimised Trials with Machine Learning

Sep. 11 - 13, 2024

Abstract

Sorbonne University and Heidelberg University join forces to organise the first workshop on AI in medicine under the 4EU+ European University Alliance.

The aim of the three-day workshop, held from September 11th to 13th 2024, is to bring together researchers in the field of machine learning for an information exchange in the field of Artificial Intelligence in Medicine. The topic of the first workshop is "Optimised Trials with Machine Learning". This involves the use of machine learning methods in randomised clinical studies in order to obtain better statistical information about interventions or diagnostic procedures. Recent progress in this field is presented by internationally recognized researchers from Sorbonne and Heidelberg but also from other institutions such as the 4EU+ Alliance member universities.

The audience comprise a diverse group of PhD and Master students from the fields of medicine, biology, and computer science.

Acknowledgements

The editors would like to thank the Deutsch-Französische Hochschule / Université franco-allemande (DFH/UFA) to fund this workshop, and all the staff from both universities who contributed to the success of the event with their commitment.

Cite as

Xavier Fresquet, Jürgen Hesser (2024): 1st. Joint DFH/UFA workshop on AI in Medicine: Optimised Trials with Machine Learning. Heidelberg University Library.
<https://doi.org/10.11588/heidok.00035481>

Published by Heidelberg University Library, 2024

The electronic version of this work is permanently available on:

<https://archiv.ub.uni-heidelberg.de/volltextserver/>

doi: <https://doi.org/10.11588/heidok.00035481>

Text © 2024, Xavier Fresquet, Jürgen Hesser

Contents

1	Christoph Blattgerste, Jürgen Hesser, Cleo-Aron Weis (Heidelberg): Extending the margin of pathological tissue using a computer vision pipeline	4
2	Marcus Buchwald, Pascal Memmesheimer, Arash Dooghaie Moghadam, Ines Tuschner, Laura Alejandra Santamaria Suarez, Jimmy Daza, Timo Itzel, Christoph Antoni, Catharina Gerhards, Micheal Neumaier, Christoph Brochhausen, Peter R. Galle, Matthias Ebert, Arndt Weinmann, Jürgen Hesser, Vincent Heuveline, Andreas Teufel (Heidelberg): Discovering predictive biomarkers for liver staging using machine learning methods	9
3	Anna Kabo (Prague): Ethical issues in the use of artificial intelligence in medicine	15
4	Elaheh Mosaieby, Michael Michal (Prague): Epigenetic Signatures in Sarcoma: Unveiling Subtype-Specific Methylation Patterns	17
5	Sara Monji Azad, Marvin Kinz, Jyot Makadiya, David Mäannle, Claudia Scherl, Jürgen Hesser (Heidelberg): An Assessment of Non-Rigid Point Cloud Registration Methods on Two Novel Soft Tissue Deformation Datasets	22
6	Tobias Meißner, Werner Nahm, Jürgen Hesser (Heidelberg): Accuracy and Precision of 3D Reconstruction Methods for Localizing Point-Like Gamma Sources with a Coded Aperture Camera	26
7	Bacem K. Othman, Ilmo Leivo, Alena Skálová (Prague): Artificial Computational Pathology May Refine and Redefine Salivary Gland Tumor Entities: A Molecular Bioinformatic Perspective	31
8	Pooja Mali Rai (Milan): WIRELESS INTRAORAL ELECTRONICS INTEGRATED WITH AI FOR DISEASE DIAGNOSIS	34
9	Andrei Sirazitdinov, Jonas Tesarz, Jürgen Hesser (Heidelberg): Treatment Group Assignment using Data from PerPain RCT	39
10	Ishita Singhal, Bocca Giorgio, Giuseppe Maurizio Facchi, Jacopo Burger (Milan): Revolutionizing Dental Care: Harnessing Artificial Intelligence and Deep Learning for Early Caries Detection and Pathology Analysis	43
11	Abhinay Krishna Vellala, Moritz Schnitzer, Alexander Hertel, Florian Haag, Freba Grawe, Frank Gerrit Zöllner, Stefan Schönberg, Matthias Frölich (Heidelberg): Automated Radiology Report Generation from Radiologist Impressions using Fine-Tuned LLAMA-3 Model: A Novel Approach	46
12	Marcin Wierzbński, Damian Wójtowicz (Warsaw): Enhancing Tumor Site Prediction through Integrated Somatic Mutation Sequencing and Genomic Positional Encoding	49

Extending the margin of pathological tissue using a computer vision pipeline

Christoph Blattgerste*, Jürgen Hesser^{††}, Cleo-Aron Weis*[‡]

[‡]*Interdisciplinary Center for Scientific Computing, Heidelberg University, Heidelberg, Germany*

[†]*Mannheim Institute for Intelligent Systems in Medicine (MIISM),*

Medical faculty Mannheim, Heidelberg University, Data Analysis and Modeling in Medicine, Mannheim, Germany

**Institute of Pathology, Section Computational Pathology, Heidelberg University Hospital, Heidelberg, Germany*

Email: christoph.blattgerste@uni-heidelberg.de

Abstract—Surgical pathology examines diseases at a microscopic level by analyzing thin slices of extracted tissue. However, this process fails to consider the macroscopic context, such as the 3D structure of tumors, which could be essential for accurate diagnosis and treatment.

This article presents a method to extend the visible tissue area around whole slide images without additional context, aiming to improve the accuracy of determining resection margins. A deep learning approach involving four generative adversarial networks accurately predicts and extends adjacent tissue based on texture and shape. The final output, after an upscaling step, provides enhanced tissue continuity as shown in figure 1 and gives pathologists a more comprehensive view of tissue adjacency.

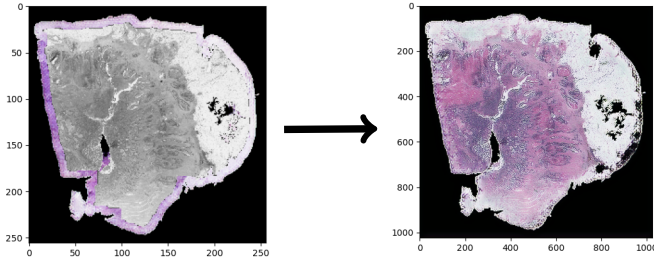


Figure 1. Low-resolution colorful example tissue extension for a whole slide image depicted in gray on the left. The 4x scaled up and adjusted high-resolution extension of the WSI on the right allows for a more detailed analysis of the tissue adjacency and might facilitate tissue reconstruction.

I. INTRODUCTION

Histology, the study of tissue at a microscopic level, is crucial for diagnosing diseases like cancer and determining treatment strategies. However, this process is time-consuming and relies heavily on the expertise of pathologists. Recent advances in high-resolution Whole Slide Image (WSI) scanners [1] have enabled computer vision algorithms for possible assistance in tasks such as classification and segmentation of histopathological images.

These tools might for example help pathologists identify regions of interest or quantify disease extent providing a more comprehensive view of tissue structure and cell distribution [2], [3].

Despite these advancements of computer vision in pathology, the application of Generative Adversarial Networks (GANs) for extending WSI boundaries remains underexplored, a gap this work seeks to address. Additional information about the adjacency of tissue in extended WSIs can be used to study larger tissue segments or even reassembling 3D volumes out of single WSI.

II. STATE OF THE ART

GANs have demonstrated significant capabilities in generating high-quality synthetic images and enhancing existing ones [4]. These networks have been used for tasks such as harmonization of manual slice preparation [5], correcting missing information [6], and domain adaptation [7], [8]. Khan et al. [9] even generated synthetic stained images.

While histopathologists focus on enhancing single WSIs, image extension is a common task in other fields, e.g. photography. While classical methods like pixel diffusion [10] and patch-based approaches [11] have been used in other fields, they struggle with large-scale image extension. Advanced machine learning architectures, including UNet [12], vision transformers [13], and GANs [14], [15], offer more robust solutions. In medical imaging, GANs are also used to enhance WSI resolution from low-resolution data [16]. As image extension has shown limitations in other disciplines especially for high resolution images with a high percentage of inpainted pixels, it is combined with image enhancement used frequently in pathology.

III. METHODS

The proposed tissue extension pipeline for pathological WSIs involves two primary components: WSI extraction (Figure 3) and fragment extension including low-resolution image extension and upscaling by a factor of 4 as illustrated in Figure 2.

WSIs consist of high-resolution tiles arranged in a pyramidal structure for efficient storage, depicted in Figure 2. For our method, WSIs are first converted into a single two-dimensional image. This requires extracting the WSIs as a colorized pixel matrix and then squaring the tissue fragments by extending the background, ensuring compatibility with the GAN input.

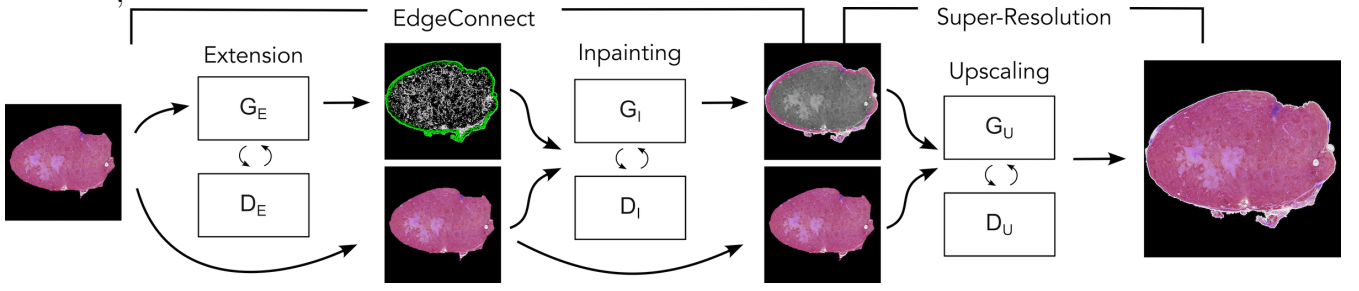


Figure 2. The proposed tissue extension pipeline for pathological WSIs consists of three subsequent GANs for image extension (G_E/D_E), -inpainting (G_I/D_I) and upscaling (G_U/D_U) combining two architectures [15], [17].

A. WSI extraction

WSIs are extracted from the digital pathology archive at the University Hospital Heidelberg. Due to the size of WSIs larger than $10k \times 10k$ pixels, preprocessing is required to generate two-dimensional images with masked backgrounds. We use the *OpenSlides* library [18] for reading and converting WSIs.

To focus on tissue, the background is suppressed using a Canny edge detection approach [19], adapted from the *Histolab* library [20], with segmentation performed in LAB color space [21].

While deep learning approaches are able to enhance tissue detection [22], [23], they do not significantly improve the result for our extension task, but decrease the pipeline performance.

After masking, the tissue is cropped and downsampled to a maximum edge length of 1024 pixels. For GAN training, images are extended to required squares by adding black background pixels.

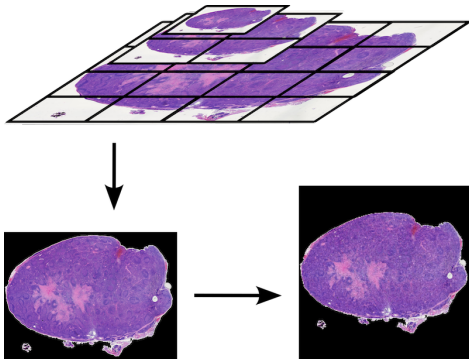


Figure 3. Whole slide image (top) is exported as high resolution pixelated thumbnail with masked tissue (left). The tissue thumbnail has to be cropped and downsampled for subsequent processing (right).

B. Fragment extension

After extracting the tissue fragments from the WSIs, the fragments are extended with a border of fixed diameter following the Edge Connect architecture [15] as shown in figure 2. In contrast to various available image extension tools or explored image completion removing artifacts, also

the shape itself has to be detected and enlarged. Therefore, the extension pipeline consists of a chain of several generative adversarial networks (GANs) which are trained on a supervised dataset of tissue fragments with eroded borders. The first GAN in the first part of the network with generator G_E and discriminator D_E continues the edges registered by a Canny edge detector. Its result is then colorized by the second GAN named G_I and D_I in the middle of figure 2. The second part of the chain consists of G_U and D_U as super-resolution GAN which increases the resolution of the image by a factor of 4 for in-depth analysis.

For the image extension relying on G_E/D_E as well as G_I/D_I , the Edge-Connect GAN framework [15], trained on general purpose image datasets *Places2* [24], *Paris StreetView* [25] and *CelebA* [26], is fine-tuned for pathological images. The training procedure used for the publication is closely adopted for fine-tuning the published model. With detailed description in the original paper [15], several losses are combined to account for the different characteristics of the image extension task. The adversarial loss ensures that the edge image is indistinguishable from the ground truth on pixel level while the feature-matching loss ensures that the activation maps within feature space are similar. For image inpainting, the perceptual loss is used diverging from the feature-matching loss by using a pretrained VGG network [27] instead of the discriminator itself for a sophisticated feature extraction. An additional style loss introduced in the same paper also takes into account the covariance of the activation maps inside the network. To adopt the existing knowledge of the model, an initial learning rate of 10^{-5} is used together with an increased discriminator to generator learning rate of 0.2. The first networks GANs are trained on images of a size of 256×256 pixels in order to achieve a feasible prediction accuracy. However, all initial images share a resolution of 1024×1024 pixels as training labels of the proposed second part of the network increasing the image resolution by a factor of 4.

The algorithm and training routine including its parameters is based on a super-resolution GAN [17] visualized as G_U and D_U in figure 2. The authors make use of

a two-step approach firstly training a ResNet on image prediction which is then fine-tuned as generator G_U in a GAN architecture to achieve a photo-realistic image. Due to the unique characteristics of pathological WSIs, which differ from general photography datasets used in the original paper, the model was retrained from scratch. Here, the used loss functions are accumulated according to Ledig et al. [17]: The overall loss is defined as weighted sum of the adversarial loss quantifying the discriminator D_U ability to distinguish between real and fake images and the so called content loss which is identical to the perceptual loss described above. For supervised training, the used dataset consists of image pairs of to 256×256 pixels downsampled high resolution images as well as predictions from the first network with predicted boundary and their corresponding 1024×1024 pixels originals. Hence, disturbing artifacts from the prediction process are removed beside better image quality because of the supervised training on original WSIs.

IV. RESULTS

Extraction, extension and upscaling are trained separately at first before combining the pipeline. Results from preceding steps are included into subsequent datasets for training adjustments and optimizing the interference.

A. Tissue masking

The preprocessing step as described in Section III-A did not involve machine learning but required significant tuning, particularly in tissue masking. A dynamic thresholding approach was necessary to handle the variability in tissue types and the effects of slide preparation and digitization. Initially, a fixed threshold led to incorrect detection of lighter fatty tissues. Switching to dynamic Canny thresholding [19] in RGB space improved edge detection, but converting the image to LAB color space and applying the same thresholding method further enhanced performance, particularly in distinguishing tissue from the background. Because of the combination of different characteristics, the results were compared manually instead of a quantitative metric. This refined process allowed us to create a large dataset suitable for GAN fine-tuning.

B. Fine-tune fragment extension

Using the masked WSI thumbnails, the fragment extension pipeline was trained as outlined in Section III-B. The task required the neural network to recognize geometric structures using G_E/D_E and textural content with G_I/D_I [15, Figure 3] to extend irregularly shaped tissue in the foreground. We fine-tuned the open-source *Edge-Connect* architecture on models pretrained on datasets like *Places2* [24], *Paris Street View* [25] and *CelebA* [26]. The evaluation metrics including *PSNR*, *SSIM*, and *FID* [28] are summarized in Table I and quantify pixel-level reconstruction as well as semantics.

While the original *Edge-Connect* models performed well on general-purpose images, their performance on

histopathological slices was initially lower, with *PSNR* dropping to 14.33 and *SSIM* to 0.794. Freshly training the GANs on a WSI dataset significantly improves the quality of the generated images as determined in the top row of table I.

Despite the limited real-world dataset size of less than 5500 WSI thumbnails, the fine-tuned models include the general knowledge of the pretrained network gathered over many iterations on a large general purpose dataset. The model pretrained on the *CelebA* dataset outperformed others, likely due to the similarity in the shapes and structures of facial images to tissue fragments, resulting in better PSNR and FID scores.

Table I
FINE-TUNING OF JOINTLY TRAINED G_E/D_E & G_I/D_I GANs BASED ON DIFFERENT DATASETS. THE HIGHER *PSNR* AND *SSIM*, THE BETTER THE PREDICTION. THE *FID* METRIC IS REVERSED WITH BEST RESULT PRINTED IN BOLD SHOWING THE RELEVANCE OF DEDICATED TRAINING DATA.

Pretrained dataset	PSNR	SSIM	FID
No pretraining	21.49	0.887	51.88
CelebA	21.53	0.881	35.44
Places2	20.13	0.871	42.87
PSV	19.69	0.867	43.13

C. Training of WSI upscaling

The final step of the proposed method is the upscaling of the extended tissue fragments based on the SRGAN architecture [17]. In order to tune the training output according to section III-B, various dataset compositions are compared in table II. The original model weights trained on the general purpose *ImageNet* database [29] are available for prediction and provide an optimization reference. The initial training using WSI_{orig} refers to the direct upscaling of original low- and high-resolution WSI images showing better results. The advanced method includes the extended tissue fragments WSI_{ext} resulting from G_E & G_I and trains the image correction as well. Thus, the model should not only be able to upscale pathological tissue but also enhance the predicted WSI contour.

Quantitative comparison shows that training the model with a specialised WSI dataset results in a higher PSNR and SSIM compared to the general purpose *ImageNet* dataset. However, a slight correction in the predicted images from the image extension GAN cannot be achieved by the upscaling model.

Table II
WSI UPSCALING RESULTS USING DIFFERENT TRAINING DATASETS. SPECIALIZED PATHOLOGICAL TRAINING DATA DOES SIGNIFICANTLY IMPROVE NEITHER THE UPSCALING QUALITY NOR THE EXTENSION ADJUSTMENT.

Dataset	PSNR	SSIM
ImageNet	18.13	0.721
WSI_{orig}	18.99	0.751
$WSI_{orig} + WSI_{ext}$	17.73	0.648
ImageNet + WSI_{orig}	19.12	0.750

V. DISCUSSION

In this project, we developed a pipeline of multiple GANs capable of extending WSIs by predicting adjacent tissue at high resolution, effectively increasing the field of view. Each component was optimized and trained on pathological tissue slices, incorporating anatomical and cellular knowledge. The extended images provide additional insights into the slide and surrounding tissue, such as continuous lesions.

Future improvements could involve integrating a more advanced upscaling network that combines information from the original WSI and the extended prediction, akin to multi-frame super-resolution techniques [30].

This pipeline has potential value for surgeons in verifying safety margins and creating surgical maps. Additionally, the extended WSIs can be recombined with neighboring tissue fragments, using texture and color information for more accurate macroscopic analysis.

VI. ACKNOWLEDGMENT

This publication was supported through state funds approved by the State Parliament of Baden-Württemberg for the Innovation Campus Health + Life Science alliance Heidelberg Mannheim. Furthermore, the authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

REFERENCES

- [1] M. G. Hanna, O. Ardon, V. E. Reuter, S. J. Sirintrapun, C. Englund, D. S. Klimstra, and M. R. Hameed, "Integrating digital pathology into clinical practice," *Modern Pathology*, vol. 35, pp. 152–164, Feb. 2022.
- [2] P. Pati, G. Jaume, A. Foncubierta, F. Feroce, A. M. Anniciello, G. Scognamiglio, N. Brancati, M. Fiche, E. Dubruc, D. Riccio, M. Di Bonito, G. De Pietro, G. Botti, J.-P. Thiran, M. Frucci, O. Goksel, and M. Gabrani, "Hierarchical Graph Representations in Digital Pathology," Mar. 2021. arXiv:2102.11057 [cs].
- [3] R. Deng, C. Cui, Q. Liu, T. Yao, L. W. Remedios, S. Bao, B. A. Landman, L. E. Wheless, L. A. Coburn, K. T. Wilson, Y. Wang, S. Zhao, A. B. Fogo, H. Yang, Y. Tang, and Y. Huo, "Segment Anything Model (SAM) for Digital Pathology: Assess Zero-shot Segmentation on Whole Slide Imaging," Apr. 2023. arXiv:2304.04155 [cs, eess].
- [4] S. A. Alajaji, Z. H. Khoury, M. Elgharib, M. Saeed, A. R. H. Ahmed, M. B. Khan, T. Tavares, M. Jessri, A. C. Puche, H. Hoorfar, I. Stojanov, J. J. Sciubba, and A. S. Sultan, "Generative Adversarial Networks in Digital Histopathology: Current Applications, Limitations, Ethical Considerations, and Future Directions," *Modern Pathology*, vol. 37, p. 100369, Jan. 2024.
- [5] M. Runz, D. Rusche, S. Schmidt, M. R. Weihrauch, J. Hesser, and C.-A. Weis, "Normalization of HE-stained histological images using cycle consistent generative adversarial networks," *Diagnostic Pathology*, vol. 16, p. 71, Dec. 2021.
- [6] S. Ali, N. K. Alham, C. Verrill, and J. Rittscher, "Ink Removal from Histopathology Whole Slide Images by Combining Classification, Detection and Image Generation Models," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 928–932, Apr. 2019. ISSN: 1945-8452.
- [7] T. Wollmann, C. S. Eijkman, and K. Rohr, "Adversarial domain adaptation to improve automatic breast cancer grading in lymph nodes," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 582–585, Apr. 2018. ISSN: 1945-8452.
- [8] C.-A. Weis, J. N. Bindzus, J. Voigt, M. Runz, S. Hertjens, M. M. Gaida, Z. V. Popovic, and S. Porubsky, "Assessment of glomerular morphological patterns by deep learning algorithms," *Journal of Nephrology*, vol. 35, pp. 417–427, Mar. 2022.
- [9] U. Khan, S. Koivukoski, M. Valkonen, L. Latonen, and P. Ruusuvaari, "The effect of neural network architecture on virtual H&E staining: Systematic assessment of histological feasibility," *Patterns*, vol. 4, p. 100725, May 2023.
- [10] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Transactions on Image Processing*, vol. 10, pp. 1200–1211, Aug. 2001. Conference Name: IEEE Transactions on Image Processing.
- [11] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: a randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, pp. 24:1–24:11, July 2009.
- [12] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image Inpainting for Irregular Holes Using Partial Convolutions," Dec. 2018. arXiv:1804.07723 [cs].
- [13] Y. Deng, S. Hui, S. Zhou, D. Meng, and J. Wang, "T-former: An Efficient Transformer for Image Inpainting," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6559–6568, Oct. 2022. arXiv:2305.07239 [cs].
- [14] P. Teterwak, A. Sarna, D. Krishnan, A. Maschinot, D. Belanger, C. Liu, and W. T. Freeman, "Boundless: Generative Adversarial Networks for Image Extension," Aug. 2019. arXiv:1908.07007 [cs].
- [15] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning," Jan. 2019. arXiv:1901.00212 [cs].
- [16] U. Upadhyay and S. P. Awate, "A Mixed-Supervision Multilevel GAN Framework for Image Quality Enhancement," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, eds.), (Cham), pp. 556–564, Springer International Publishing, 2019.
- [17] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," May 2017. arXiv:1609.04802 [cs, stat].
- [18] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "OpenSlide: A vendor-neutral software foundation for digital pathology," *Journal of Pathology Informatics*, vol. 4, p. 27, Jan. 2013.
- [19] J. Canny, "A Computational Approach To Edge Detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-8, pp. 679–698, Dec. 1986.
- [20] R. Colling, H. Pitman, K. Oien, N. Rajpoot, P. Macklin, CM-Path AI in Histopathology Working Group, D. Snead, T. Sackville, and C. Verrill, "Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice," *The Journal of Pathology*, vol. 249, pp. 143–150, Oct. 2019.
- [21] 14:00-17:00, "ISO/CIE 11664-4:2019."
- [22] R. Rong, H. Sheng, K. W. Jin, F. Wu, D. Luo, Z. Wen, C. Tang, D. M. Yang, L. Jia, M. Amgad, L. A. D. Cooper, Y. Xie, X. Zhan, S. Wang, and G. Xiao, "A Deep Learning Approach for Histology-Based Nucleus Segmentation and Tumor Microenvironment Characterization," *Modern Pathology*, vol. 36, p. 100196, Aug. 2023.
- [23] B. Banner, A. Mahbod, I. Ellinger, R. Ecker, and G. Dorffner, "Improving Mask R-CNN for Nuclei Instance Segmentation in Hematoxylin & Eosin-Stained Histological Images," in *Proceedings of the MICCAI Workshop on Computational Pathology*, pp. 20–35, PMLR, Sept. 2021. ISSN: 2640-3498.
- [24] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, "Places: An Image Database for Deep Scene Understanding," Oct. 2016. arXiv:1610.02055 [cs].
- [25] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, "What makes paris look like paris?," *ACM Transactions on Graphics (SIGGRAPH)*, vol. 31, no. 4, pp. 101:1–101:9, 2012.
- [26] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *2015 IEEE International Conference*

on *Computer Vision (ICCV)*, pp. 3730–3738, Dec. 2015. ISSN: 2380-7504.

- [27] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” Mar. 2016. arXiv:1603.08155 [cs].
- [28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.
- [29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” Jan. 2015. arXiv:1409.0575 [cs].
- [30] M. M. Afrasiabi, R. Hosseini, and A. Abbasfar, “Combination of Single and Multi-frame Image Super-resolution: An Analytical Perspective,” Mar. 2023. arXiv:2303.03212 [cs, eess].

Discovering predictive biomarkers for liver staging using machine learning methods

Marcus Buchwald^{1,2,3}, Pascal Memmesheimer^{1,2,3}, Arash Dooghaie Moghadam⁴, Ines Tuschner⁴, Laura Alejandra Santamaria Suarez⁴, Jimmy Daza⁴, Timo Itzel⁴, Christoph Antoni⁵, Catherina Gerhards⁶, Micheal Neumaier⁶, Christoph Brochhausen⁷, Peter R. Galle⁸, Matthias Ebert^{5,9}, Arndt Weinmann⁸, Jürgen Hesser¹, Vincent Heuveline^{2,3}, Andreas Teufel^{4,9}

Abstract—BACKGROUND: The progression of liver fibrosis to cirrhosis is a common pathway in chronic liver diseases. The development of liver cirrhosis, in particular, can lead to significant functional impairments and complications. Therefore, early detection of fibrotic changes is essential. However, existing serological markers remain inadequate, prompting the analysis of explainable machine learning-based models compared against a standard cohort of liver biopsies.

MATERIALS & METHODS: Between 06.23.1999 and 03.04.2022, 655 patients underwent liver biopsies at Mannheim University Hospital. The corresponding clinical and laboratory parameters were retrospectively extracted from the respective hospital information systems. For the analysis of machine learning methods used to classify liver stages as a binary task as well as a three-stage classification, a variety of classical machine learning models and state-of-the-art deep neural networks were implemented, hypertrained, and tested on the collected data as well as an independent evaluation dataset.

RESULTS: The performance of the selected machine learning models for predicting moderate liver fibrosis, severe fibrosis, and cirrhosis using blood markers was robust, with accuracies reaching up to 88.46%, 92.31% and 82.69% respectively. The tree-based ML models, namely LightGBM, XGBoost, and Random Forest performed the best considering all classification tasks with an accuracy range of 82.69% to 92.31% for binary classification, and an accuracy of 76.95% for three-stage classification, significantly outperforming the traditional FIB-4 score. SHAP analysis of the best ensemble models identified platelets, MCV, and INR as the most influential biomarkers. Models using only these three

parameters achieved comparable performance to those using the full set of biomarkers for the *Cirrhosis* classification (ACC up to 86.52%). These findings suggest that simplified diagnostic tools using established ML algorithms focusing on these key biomarkers could be highly effective for liver fibrosis prediction.

CONCLUSION: Machine learning models can significantly improve liver staging compared to serum-based tests alone. Platelets, MCV, and INR are significantly more important than previously thought. Our code is made publicly available at: https://github.com/mbuchwa/NAFLD_Prediction/

Index Terms—Liver Staging, Medical Machine Learning, Explainable AI

I. INTRODUCTION

HEPATOLOGICAL care faces challenges due to the rising prevalence of liver diseases, particularly fatty liver disease, affecting 20-30% of the global population [1]. Fatty liver disease is a leading chronic condition and a risk factor for liver cirrhosis [2], liver failure [3], liver cancer [4], and cardiovascular diseases [5]. The limited number of specialized hepatologists and increasing patient numbers may make current detection, diagnosis, and treatment approaches unsustainable, particularly in rural areas [6]. Accurate liver fibrosis assessment is essential for prognosis, treatment decisions, and monitoring disease progression [7]. Although liver biopsy is the gold standard, it is invasive and resource-intensive [8]. Non-invasive serological markers are gaining interest due to their simplicity, safety, cost-effectiveness, and ability to track fibrosis progression over time [9]. These markers include ALT, AST, the AST to Platelet Ratio Index (APRI), and Fibrosis-4 (FIB-4) score, alongside newer markers like cytokines and extracellular matrix proteins [9]. However, non-invasive tests have limitations, such as moderate accuracy and variability based on patient demographics [9]. Improved models for predicting liver fibrosis are necessary. Machine learning (ML) offers promising potential, but challenges exist, including the need for extensive patient data and well-defined control groups [10]. To address this, a biopsy-controlled dataset was developed to evaluate the potential of ML in liver staging, aiming to enhance disease prognosis and management strategies.

¹ Data Analysis and Modeling in Medicine, Mannheim Institute for Intelligent Systems in Medicine (MIISM), Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

² Engineering Mathematics and Computing Lab, Interdisciplinary Center for Scientific Computing, Heidelberg University, Heidelberg, Germany

³ Data Mining and Uncertainty Quantification Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

⁴ Division of Hepatology, Division of Clinical Bioinformatics, Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

⁵ Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, University, Mannheim, Germany

⁶ Institute for Clinical Chemistry, Medical Faculty Mannheim of the University of Heidelberg, Mannheim, Germany

⁷ Department Pathology, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

⁸ Department of Internal Medicine I, University Medical Center, Mainz University, Mainz, Germany

⁹ Clinical Cooperation Unit Healthy Metabolism, Center for Preventive Medicine, and Digital Health (CPD), Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

The authors gratefully acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG. This work is funded by the German federal project LeMeDaRT (01ZZ2105A).

II. MATERIALS & METHODS

A. Data

A retrospective study was conducted on 655 patients who underwent laparoscopies between 06.23.1999 and 03.04.2022 at Mannheim University Hospital (*UMM* data). Clinical data were extracted from the hospital information systems of Mannheim University Hospital. All analyses were conducted following approval from the Ethics Committee II of Heidelberg University. The complete list of biomarkers included in the study is given in appendix table VIII.

Furthermore, an independent dataset from the university medicine of the Johannes Gutenberg-University in Mainz has been collected with the same patient information and diagnostic data. This *MAINZ* dataset encompasses 284 patients with a definite liver staging between *F0* and *F4*.

Multiple imputation by chained equations (MICE) [11] was utilized to impute any missing values in the *UMM* data set using Bayes Ridge Regression as an estimator. With this method, $m=10$ differently imputed data sets were generated.

B. Patient Characteristics

After preprocessing according to the section Data, a total of 497 patients from the *UMM* data were included in the machine learning training, validation and testing. The statistical attributes of the scalar patient features are given in table VIII. Regarding fibrosis severity, a significant portion of patients exhibited liver cirrhosis or higher fibrosis grades in histological evaluation, while no changes were observed in a minority. Cirrhosis, indicated by *F4*, and moderate fibrosis (*F2/3/4*) were found in 35.73% and 46.63% of patients using serological staging markers, respectively. No cirrhosis (*F0/1/2/3*) and no moderate fibrosis (*F0/1*) were serologically detected in 64.27% and 54.37% of patients respectively.

Regarding the independent *MAINZ* dataset for evaluation, moderate fibrosis and cirrhosis were found in 8.80% and 55.28% of patients respectively.

C. Models

For predicting the liver stage, multiple different machine learning methods were implemented, hyper-tuned, and evaluated. The considered models in this work are shown in Table I. The mentioned models are chosen to ensure a sufficient range of classical machine learning approaches, gradient boosting tree-based techniques as well as current state-of-the-art deep learning methods for tabular data analysis.

	Machine Learning Models
1	Support Vector Machines (SVM) [12]
2	Random Forest Models (RF) [13]
3	Extreme Gradient Boosting (XGBoost) [14]
4	Light Gradient Boosting Machine (Light GBM) [15]
5	Multi-Layer Perceptrons (MLP) [16]
6	Variational Inference Bayesian Neural Network (VI-BNN) [17]
7	Tabular Transformer Model (TabTransformer) [18]

TABLE I: Summary of applied Machine Learning Models

III. EXPERIMENTS

In this experimental setup, we focus on the binary classification task of four separate cases (see Table II):

	Tasks
1	Moderate Fibrosis based on <i>F0/1</i> vs. <i>F2/3/4</i> (FIB-4 1.45)
4	Severe Fibrosis based on <i>F0/1/2</i> vs. <i>F3/4</i> (FIB-4 cut-off at 2.67)
2	Cirrhosis based on <i>F0/1/2/3</i> vs. <i>F4</i> (FIB-4 cut-off 3.25)
3	Three-Stage (<i>F0/1</i> vs. <i>F2/3</i> vs. <i>F4</i> - FIB-4 at 1.45 & 3.25)

TABLE II: Summary of Classification Tasks

All models described above are trained and analyzed on each of the $m = 10$ different data sets using 5-fold cross-validation. A random split of approximately 70% / 18% / 12% for training, validation, and testing data was done on each of the generated datasets and each model was trained using the early stopping technique on the validation data to circumvent overfitting. Furthermore, using a random search algorithm, each training implemented hyperparameter tuning to find optimal parameter settings (e.g., learning rate, number of layers, dropout rate, number of neurons per layer). The resulting m models were each independently tested on the $m = 1$ imputed test datasets. Hence, for the test dataset, m separate model predictions are given. Soft majority voting was implemented to merge all models to achieve a final ensemble prediction. The results in Tables III to VI are given by calculating the mean and standard deviation of the ensemble prediction and the FIB-4-based prediction on the single test dataset with a 5-fold stratified split from which the larger split (80% of the test dataset) is used for the metric calculation in each fold.

IV. RESULTS

A. Classification Performance

Tables III to VI detail the classification performance metrics. In binary moderate fibrosis classification (*F0-1* vs. *F2-4*), *LightGBM* leads with $88.46 \pm 2.82\%$ accuracy, followed by *Random Forest* at $86.53 \pm 3.62\%$ and *XGBoost* at $84.61 \pm 1.94\%$. For severe fibrosis (*F0/1/2* vs. *F3/4*) with a FIB-4 cut-off of 2.67, *Random Forest* excelled at $92.31 \pm 1.80\%$, and *XGBoost* as well as *LightGBM* followed at $88.47 \pm 2.80\%$ and $86.55 \pm 2.39\%$ respectively. For cirrhosis classification (*F0-3* vs. *F4*), *Random Forest* had an accuracy of $82.69 \pm 1.75\%$, with *Light GBM* and *Tab Transformer* at $82.68 \pm 2.83\%$ and $80.78 \pm 1.39\%$, respectively. In three-stage classification, *Light GBM* and *Random Forest* were best with $76.95 \pm 3.46\%$ and $76.95 \pm 4.33\%$ accuracy respectively, with *XGBoost* following up at $75.03 \pm 3.39\%$. These results highlight the effectiveness of *LightGBM*, *Random Forest*, and *XGBoost* in predicting liver fibrosis stages using blood-based markers, while modern deep learning methods, such as *Tab Transformer*, prove to be partially competitive. However, it should be noted that the best-performing models in all four tasks belong to the class of tree-based models.

B. Feature Importance

The SHAP (SHapley Additive exPlanations) analysis evaluated the influence of parameters on predicting the classification tasks with the trained models. Figures 3 - 6 show the SHAP value plots for the top performing model (*LightGBM*, *XGBoost*, and *Random Forest*) at every task. Thrombocyte count proves to be the most influential parameter. All models identified Platelets/Thrombocytes, Mean Corpuscular Volume (MCV), International Normalized Ratio (INR), and Quick as key predictors, unlike the FIB-4 equation. The only exception is given by the ASAT Marker which has a slight margin (+0.08) compared to MCV and Quick (+0.07) for the *LightGBM* model in the moderate fibrosis classification. The SVM ensemble, though similar to FIB-4, emphasized ASAT and ALAT more, while disregarding MCV, Quick and INR which the top-performing models focused on. This highlights a shift toward these key markers for liver fibrosis prediction.

C. Reduced data models

To confirm the importance of the top-ranked biomarkers, models were trained using only Thrombocytes/Platelets, MCV, INR. Quick was disregarded, as the INR represents a standardized Quick value. These reduced-biomarker models performed comparably to the full-biomarker models, ranging only slightly below their counterpart models which were trained on all markers. Notably, two of these reduced-biomarker models even outperformed the models that utilized all available biomarkers, as highlighted in table VII. *LightGBM* demonstrated even higher accuracy for cirrhosis classification, with $86.52 \pm 2.57\%$. *XGBoost* also showed better accuracy with $80.77 \pm 1.52\%$.

D. Comparison to FIB-4 & Evaluation on MAINZ data

The best-performing models in table III - VI outperform the FIB-4 score in predicting all four classification tasks for the *UMM* test dataset. FIB-4 reached an accuracy of 73.1 ± 3.32 , 76.9 ± 2.66 , 73.05 ± 2.09 , and 57.7 ± 1.65 in moderate fibrosis, cirrhosis, severe fibrosis, and three-stage classification respectively, trailing behind between 9.64% and 15.41% in accuracy for the binary classification tasks. For the three-stage classification, FIB-4 performs 17.33% worse in accuracy and 9.72% worse in F1 score.

Table VII shows the performance metrics for the reduced-marker models being trained on only Platelets, MCV, and INR regarding the *Cirrhosis* task. The *Cirrhosis* task was chosen for comparison as the *MAINZ* dataset is characterized by a high class imbalance for the fibrosis tasks and shows the best class balance for the binary classification of cirrhosis (see figure 2). The results indicate that although the difference between the FIB-4 and the selected-marker models is significant for the *UMM* test dataset, the results are considerably less diverging for the *MAINZ* data.

This can be explained due to the inherent shift in the patient selection criteria, as the *MAINZ* data was solely utilized for inference. Figure 1 depicts the different clustering when visualized using a two-dimensional principal component analysis.

When selecting the closest patients from the *MAINZ* dataset to the entire *UMM* test dataset, using a 1-nearest-neighbor approach including all biomarkers, the accuracies of the ML models surpass the FIB-4 score. However, regarding the F1-score, the models outperform the FIB-4 score only within the error of the standard deviation between 1-2%.

V. DISCUSSION & CONCLUSION

Using a biopsy-based dataset, it was proven that machine-learning approaches outperform the standard FIB-4 model by a significant margin. Tree-based models (*Random Forest*, *XGBoost*, *LightGBM*) generally outperformed deep learning methods on tabular datasets due to their ability to handle diverse, uninformative features effectively [19]. Moreover, the best-performing models identified the most predictive biomarkers for classifying the liver stage: Thrombocytes, MCV, and INR. These biomarkers ranked the highest across all models and tasks, suggesting an alternative approach for clinical diagnosis. Interestingly, the SVM model, with an inferior accuracy similar to the FIB-4 score, replicated the FIB-4 model for moderate fibrosis detection by analogously emphasizing ASAT and ALAT, disregarding the more crucial MCV, and INR.

However, it is important to note that when applying the reduced models that were trained on one dataset to an independent cohort, the performance might change as indicated by the results. This behavior likely occurs due to the shift in patient characteristics and due to diverging patient selection criteria, allowing for confounders to impede the model performance. Hence, the authors strongly emphasize the need to fine-tune externally trained models in the medical setting for liver staging before applying them in the internal clinical setting.

In summary, modern machine learning models enhance liver staging beyond traditional serum tests, with platelets, MCV, and INR proving significant importance in inferring the liver stage.

APPENDIX

TABLE III: Performance metrics of the ensemble models for the *Moderate Fibrosis* task on the *UMM* Test Dataset.

Model	ACC (%) $\uparrow$	F1 (%) $\uparrow$
SVM	73.09 ± 3.08	73.08 ± 3.09
Random Forest	86.53 ± 3.62	86.4 ± 3.65
XGBoost	84.61 ± 1.94	84.57 ± 1.94
Light GBM	88.46 ± 2.82	88.38 ± 2.82
MLP	76.92 ± 3.26	76.86 ± 3.24
Tab Transformer	82.68 ± 3.88	82.61 ± 3.87
VI BNN	72.13 ± 4.94	71.9 ± 4.88
FIB 4	73.1 ± 3.32	72.71 ± 3.48

REFERENCES

- [1] S. E. Baumeister, H. Völzke, P. Marschall, U. John, C.-O. Schmidt, S. Flessa, and D. Alte, "Impact of fatty liver disease on health care utilization and costs in a general population: a 5-year observation," *Gastroenterology*, vol. 134, no. 1, pp. 85–94, Jan. 2008.

TABLE IV: Performance metrics of the ensemble models for the *Severe Fibrosis* task *UMM* Test Dataset.

Model	ACC (%) $\uparrow$	F1 (%) $\uparrow$
SVM	76.91 $\pm$ 2.06	75.08 $\pm$ 2.47
Random Forest	92.31 $\pm$ 1.80	91.68 $\pm$ 1.99
XGBoost	88.47 $\pm$ 2.80	87.53 $\pm$ 3.12
Light GBM	86.55 $\pm$ 2.39	84.91 $\pm$ 2.83
MLP	78.85 $\pm$ 2.78	77.83 $\pm$ 2.99
Tab Transformer	74.99 $\pm$ 3.30	73.23 $\pm$ 3.77
VI BNN	76.45 $\pm$ 2.78	74.14 $\pm$ 3.42
FIB 4	76.9 $\pm$ 2.66	76.31 $\pm$ 2.45

TABLE V: Performance metrics of the ensemble models for the *Cirrhosis* task *UMM* Test Dataset.

Model	ACC (%) $\uparrow$	F1 (%) $\uparrow$
SVM	76.95 $\pm$ 2.22	70.36 $\pm$ 3.53
Random Forest	82.69 $\pm$ 1.75	79.19 $\pm$ 2.45
XGBoost	78.84 $\pm$ 1.85	76.26 $\pm$ 2.01
Light GBM	82.68 $\pm$ 2.83	79.2 $\pm$ 3.73
MLP	78.86 $\pm$ 2.21	74.5 $\pm$ 3.39
Tab Transformer	80.78 $\pm$ 1.39	76.34 $\pm$ 2.58
VI BNN	75.45 $\pm$ 3.08	71.5 $\pm$ 4.16
FIB 4	73.05 $\pm$ 2.09	70.16 $\pm$ 2.51

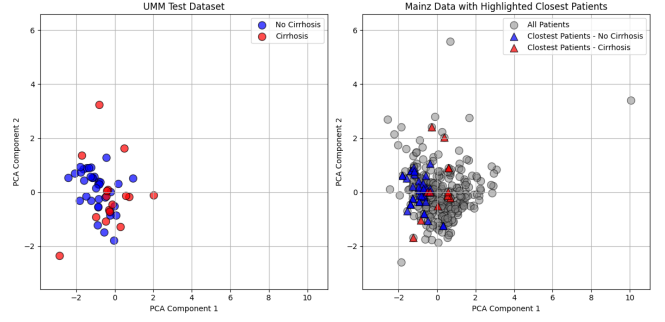
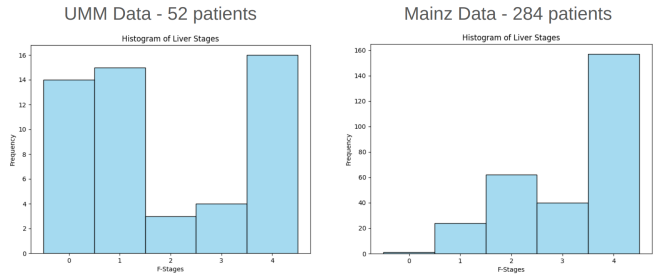
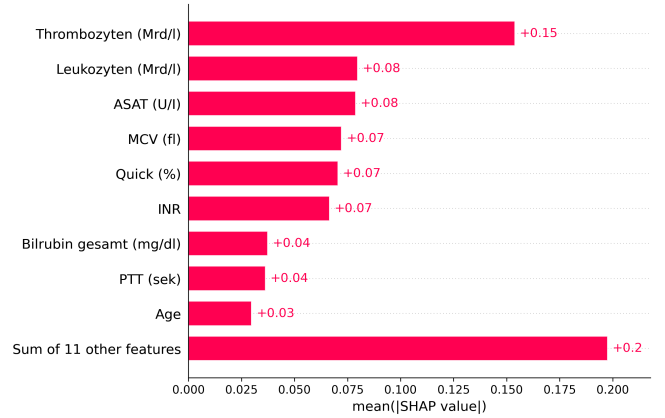
TABLE VI: Performance metrics of the ensemble models for the *Three Stage* task *UMM* Test Dataset.

Model	ACC (%) $\uparrow$	F1 (%) $\uparrow$
SVM	69.25 $\pm$ 3.43	55.05 $\pm$ 4.94
Random Forest	76.95 $\pm$ 4.33	65.22 $\pm$ 5.3
XGBoost	75.03 $\pm$ 3.39	62.2 $\pm$ 4.38
Light GBM	76.95 $\pm$ 3.46	66.22 $\pm$ 5.4
MLP	69.25 $\pm$ 5.09	61.55 $\pm$ 5.26
Tab Transformer	71.16 $\pm$ 4.28	60.03 $\pm$ 5.33
VI BNN	68.76 $\pm$ 1.32	54.34 $\pm$ 4.82
FIB 4	57.7 $\pm$ 1.65	52.58 $\pm$ 1.38

TABLE VII: Selected-Features-Model Test Performance for the textitCirrhosis Task: **F1** and **ACC** scores of SHAP top-3 trained best ensemble models *LightGBM*, *XGBoost*, and *Random Forest*, along with FIB-4. All models were only trained on Platelets, MCV & INR. Green-highlighted rows indicated higher performance metrics compared to the models trained with all biomarkers.

Model	Cirrhosis ACC % $\uparrow$		
	UMM	MAINZ	Selected MAINZ
Random Forest	80.75 $\pm$ 2.71	76.76 $\pm$ 1.59	88.48 $\pm$ 1.68
XGBoost	80.77 $\pm$ 1.52	80.28 $\pm$ 0.88	88.48 $\pm$ 1.68
LightGBM	86.52 $\pm$ 2.57	81.69 $\pm$ 1.32	90.39 $\pm$ 2.09
FIB-4	73.05 $\pm$ 2.09	79.99 $\pm$ 0.69	82.71 $\pm$ 2.7

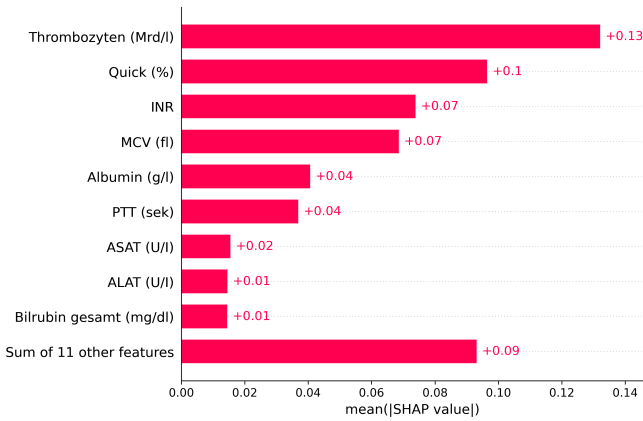
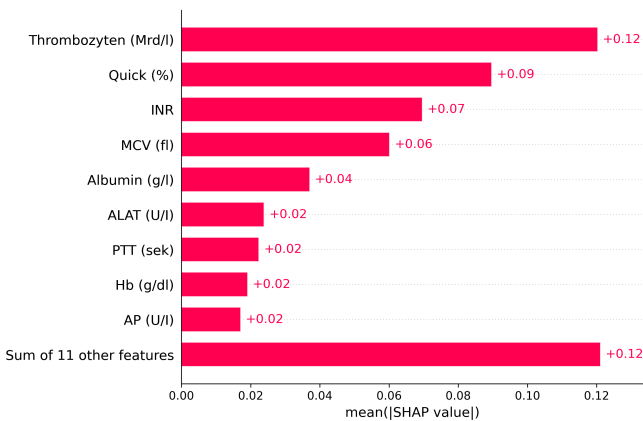
Model	Cirrhosis F1 % $\uparrow$		
	UMM	MAINZ	Selected MAINZ
Random Forest	78.72 $\pm$ 3.04	76.37 $\pm$ 1.57	82.72 $\pm$ 2.42
XGBoost	79.23 $\pm$ 1.69	78.99 $\pm$ 1.08	83.77 $\pm$ 2.29
LightGBM	83.11 $\pm$ 3.68	81.53 $\pm$ 1.32	83.78 $\pm$ 3.91
FIB-4	70.16 $\pm$ 2.51	79.99 $\pm$ 0.69	82.39 $\pm$ 2.73

Fig. 1: Clustering of the patients regarding the *Cirrhosis* task in both datasets. The closest patients between both datasets were estimated based on the 1-NN approach and mapped to the two-dimensional PCA for visualization purposes.Fig. 2: Histogram of the diagnosed liver stages between the test dataset from the *UMM* and *MAINZ*.Fig. 3: Summary of the feature importance of the best-performing ensemble model (*Light GBM*) on the *Moderate Fibrosis* task on the *UMM* data.

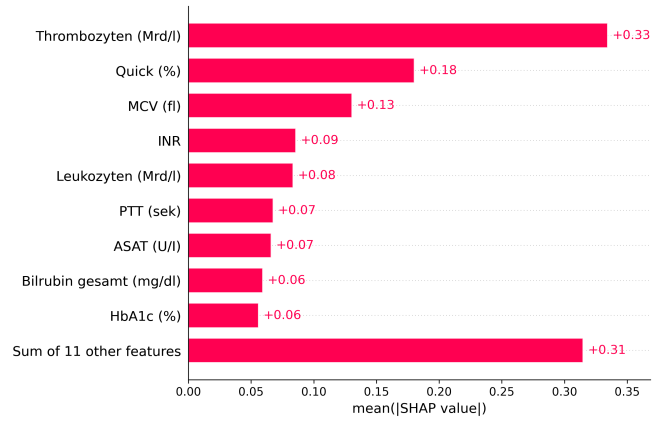
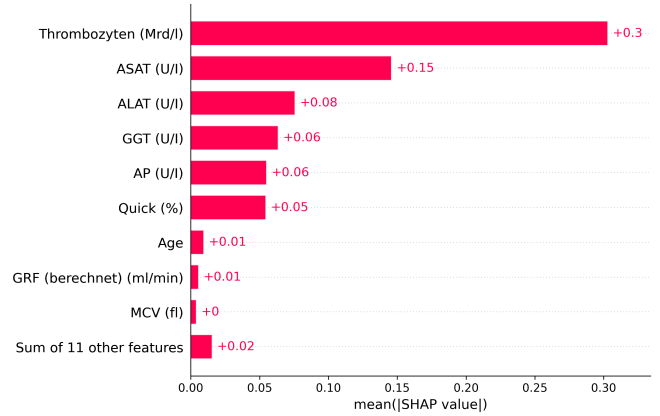
- [2] Y.-H. Huang, C. Chan, H.-W. Lee, C. Huang, Y.-J. Chen, P.-C. Liu, S.-N. Lu, W.-L. Chuang, J.-F. Huang, M.-L. Yu, J. Koshiol, and M.-H. Lee, "Influence of nonalcoholic fatty liver disease with increased liver enzyme levels on the risk of cirrhosis and hepatocellular carcinoma," *Clin. Gastroenterol. Hepatol.*, vol. 21, no. 4, pp. 960–969.e1, Apr. 2023.
- [3] R. T. Stravitz and W. M. Lee, "Acute liver failure," *Lancet*, vol. 394, no. 10201, pp. 869–881, Sep. 2019.
- [4] V. Katzke, T. Johnson, D. Sookthai, A. Hüsing, T. Kühn, and R. Kaaks, "Circulating liver enzymes and risks of chronic diseases and mortality in the prospective EPIC-Heidelberg case-cohort study," *BMJ Open*, vol. 10, no. 3, p. e033532, Mar. 2020.
- [5] J. Rahmani, A. Miri, I. Namjoo, N. Zamaninour, M. B. Maljaei, K. Zhou, R. Cerneviciute, S. M. Mousavi, H. K. Varkaneh, A. Salehisahlabadi, and Y. Zhang, "Elevated liver enzymes and cardiovascular mortality: a

TABLE VIII: Comparison of *UMM* and *MAINZ* Data Sets

Characteristics	UMM Data	MAINZ Data	P-Value
Platelets (Mrd/l)	216.8±115.7	169.3±91.2	0
MCV (fl)	89.0±7.5	94.1±9.2	0
Quick (%)	87.6±21.1	76.4±24.8	0
INR	1.1±0.2	1.3±0.6	0
Leukocytes (Mrd/l)	7.0±3.2	7.3±3.8	0.369
ASAT (U/l)	112.1±207.1	132.0±336.2	0.368
PTT (sek)	26.7±7.4	33.6±7.4	0
ALAT (U/l)	133.9±224.1	117.8±328.3	0.467
Age	63.3±15.8	54.6±12.3	0
IgG (g/l)	14.0±6.2	16.4±8.0	0.001
Albumin (g/l)	35.2±7.1	32.8±8.6	0
HbA1c (%)	5.6±1.4	6.0±3.4	0.153
Bilirubin (mg/dl)	2.6±6.7	4.1±7.3	0.006
AP (U/l)	171.2±192.3	150.8±117.0	0.074
Urea	31.4±21.2	18.5±16.4	0
Hb (g/dl)	13.6±4.4	12.8±2.5	0.002
Potassium	4.0±0.4	4.2±0.8	0
GGT (U/l)	236.3±303.8	222.3±332.0	0.562
Creatinine (mg/dl)	1.0±0.8	1.1±0.8	0.035
GFR (ml/min)	85.7±25.0	75.1±31.2	0

Fig. 4: Summary of the feature importance of the best-performing ensemble model (*Random Forest*) on the *Severe Fibrosis* task on the *UMM* data.Fig. 5: Summary of the feature importance of the best-performing ensemble model (*Random Forest*) on the *Cirrhosis* task on the *UMM* data.

systematic review and dose-response meta-analysis of more than one million participants,” *Eur. J. Gastroenterol. Hepatol.*, vol. 31, no. 5, pp. 555–562, May 2019.

Fig. 6: Summary of the feature importance of the best-performing ensemble model (*XGB*) on the *Three Stage* task on the *UMM* data.Fig. 7: Summary of the feature importance of the *SVM* ensemble model on the *Fibrosis* task on the *UMM* data.

- [6] A. Teufel, A. Geier, C. Sarrazin, J. M. Schattenberg, A. Kautz, R. Dörner, J. Kramer, K. Jerysiak, T. Baars, B. Hönscheid, D. Müller-Wieland, S. Rossol, C. Trautwein, F. Tacke, and A. Canbay, “Intersektorales management von patienten mit unklarer leberwerterhöhung und nichtalkoholischer fettlebererkrankung (NAFLD),” *Z. Gastroenterol.*, vol. 61, no. 8, pp. 1028–1036, Aug. 2023.
- [7] P. Angulo, D. E. Kleiner, S. Dam-Larsen, L. A. Adams, E. S. Björnsson, P. Charatcharoenwitthaya, P. R. Mills, J. C. Keach, H. D. Lafferty, A. Stahler, S. Haflidadottir, and F. Bendtsen, “Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease,” *Gastroenterology*, vol. 149, no. 2, pp. 389–97.e10, Aug. 2015.
- [8] K. S. Nallagangula, S. K. Nagaraj, L. Venkataswamy, and M. Chandrappa, “Liver fibrosis: a compilation on the biomarkers status and their significance during disease progression,” *Future Sci. OA*, vol. 4, no. 1, p. FSO250, Jan. 2018.
- [9] J. C.-T. Lai, L. Y. Liang, and G. L.-H. Wong, “Noninvasive tests for liver fibrosis in 2024: are there different scales for different diseases?” *Gastroenterol. Rep. (Oxf.)*, vol. 12, p. goae024, Apr. 2024.
- [10] J. Z. Ting Sim, Q. W. Fong, W. Huang, and C. H. Tan, “Machine learning in medicine: what clinicians should know,” *Singapore Med. J.*, vol. 64, no. 2, pp. 91–97, Feb. 2023.
- [11] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in r,” *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.
- [12] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [13] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

- [14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, Aug. 2016. [Online]. Available: <http://dx.doi.org/10.1145/2939672.2939785>
- [15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, pp. 3146–3154, 2017.
- [16] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.
- [17] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.
- [18] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "Tabtransformer: Tabular data modeling using contextual embeddings," 2020.
- [19] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," 2021. [Online]. Available: <https://arxiv.org/abs/2106.03253>

Ethical issues in the use of artificial intelligence in medicine

Anna Kabo
Department of Medical Ethics and Humanities
The Second Faculty of Medicine of Charles University
Prague, Czech Republic
kaboanna6@gmail.com

Abstract— Artificial intelligence already occupies an important place in many areas of human life, but not all medical professionals accept its use. What is the reason for this concern? This work is describing the ethical problems associated with the use of artificial intelligence in medicine. Since the concept of ‘ethics’ is very broad, it is worth paying attention to the ethical concepts associated with AI, adopted by law. The European Commission and other associations are trying to create a list of restrictions on the use of AI, but the problem is that each sector uses AI differently, so the restrictions required are different. Healthcare is a field where all changes are very slow, since any mistake will affect the lives and health of a huge number of people. The risks in this case are not justified. Are we ready at this point to entrust our health to AI? Ethical questions arise not only for health workers, students and people in related fields, but also for patients.

Keywords— Artificial Intelligence, ethics, healthcare, health workers

I. INTRODUCTION

Artificial intelligence in medicine is currently used primarily as an auxiliary tool. The main application is the analysis of large amounts of data. A human can also analyze images, symptoms or the dynamics of the disease, but not as quickly as a machine can. However, there is a risk that AI may lack the necessary information about rare cases. Additionally, a moral and ethical question arises: can we completely replace human specialists in working with patients?

II. MATERIALS AND METHODS

Within the process of the preparation of this work, the norms and guidelines of the European Commission were taken as a basis; a cross-analysis of previously published laboratory research results in the field was carried out. This thesis formulates extended reflections and conclusions on the topic of Practical use of AI in medicine stated by Michael R. King in the «The Future of AI in Medicine: A Perspective from a Chatbot» article.

III. ANALYSIS OF ISSUES

Empirical research indicates that artificial intelligence can be successfully integrated into the healthcare system. AI is advanced enough to analyze images and make a diagnosis (*Lancet Oncol*, 2023). However, there is a need of huge amount of diagnoses to expand the databases and train artificial intelligence. This poses a problem when artificial intelligence replaces human in some profession, which creates a situation where doctors' knowledge is used to deprive them of their jobs. This is evident in the case of radiology. Recent researches indicate that radiology will soon be completely devoid of human medical workers.

Ethical principles for working with artificial intelligence include:

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability

(European commission, 2019)

The situation can be seen as a violation of the first principle. Human specialists aren't fully autonomous if their work depends on AI-powered data analysis.

It is also impossible to prevent harm completely when working with AI in medicine. Due to the variety of approaches and opinions about diseases and diagnosis, the same situation can be considered both a cure and a harm. Examples of such medical procedures are euthanasia, abortion, chemotherapy, and more. Where exactly should the line be drawn between harm and benefit, and how can we explain this feature to artificial intelligence?

Additionally, the fundamental principle of working with AI is fairness. It is not possible in the field of medicine. Various methods and schools exist for constructing diagnoses. Alternative medicine is also employed by specialists in with classical approaches. Consequently, there is no single international standard that would be applicable to every patient. The most prominent example is the use of traditional Chinese medicine. Knowledge of medicinal herbs is not a compulsory program in most medical schools, but it is effectively used by many specialists. Bai-Yun Zeng and Kaicun Zhao write about effective methods of helping patients with neurological diseases in their book. That does not mean that one of the approaches is wrong. Medicine is just a field with many variations of methods, where doctors always make decisions based on many factors, such as the patient's financial situation, location, the patient's psychological state, and more.

DISCUSSION

At the moment, it is impossible to completely ignore artificial intelligence and not use this tool in medicine. However, it is worth paying attention to how deeply it will be introduced into the work of healthcare. Is it ethically correct to replace workers with AI? How trustworthy AI is?

REFERENCES

- [1] King, M. R. (2022). The Future of AI in Medicine: A Perspective from a Chatbot. *Annals of Biomedical Engineering*, 51(2), 291–295. <https://doi.org/10.1007/s10439-022-03121-w>
- [2] Lång, K., Josefsson, V., Larsson, A., Larsson, S., Högberg, C., Sartor, H., Hofvind, S., Andersson, I., & Rosso, A. (2023). Artificial intelligence-supported screen reading versus standard double reading

in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, 24(8), 936–944. [https://doi.org/10.1016/s1470-2045\(23\)00298-x](https://doi.org/10.1016/s1470-2045(23)00298-x)

- [3] Neurobiology of Chinese Herb Medicine. (2017b). Academic Press.
- [4] Smuha, N. A. (2019). The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence. *Computer Law Review*

International, 20(4), 97–106. <https://doi.org/10.9785/cri-2019-200402Y>. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” *IEEE Transl. J. Magn. Japan*, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

- [5] M. Young, *The Technical Writer’s Handbook*. Mill Valley, CA: University Science, 1989.

Epigenetic Signatures in Sarcoma: Unveiling Subtype-Specific Methylation Patterns

Elaheh Mosaieby, Department of Pathology, Charles University, Faculty of Medicine in Plzen, Czech Republic

Supervisor: Assoc.prof. Michael Michal, MD, PhD. Department of Pathology, Charles University, Faculty of Medicine in Plzen, Czech Republic

Abstract

Sarcomas are a diverse group of mesenchymal tumors, with a wide range of genetic and epigenetic changes, making precise classification and therapeutic targeting difficult. This study aimed to clarify the epigenetic landscape of specific sarcoma subtypes employing Methylation EPIC arrays.

We conducted methylation profiling using the Illumina EPIC microarray on a previously unexplored group of sarcoma subtypes, including malignant peripheral nerve sheath tumor (MPNST) and undifferentiated pleomorphic sarcoma (UPS) of 18 samples, creating a combined dataset of 107 samples representing six distinct sarcoma subtypes. We observed unique methylation patterns that defined sarcoma subtypes using t-SNE visualization, revealing underlying epigenetic differences. The integration of clinical and genomic data aided in the discovery of methylation signatures linked with clinical outcomes. Notably, our findings identified new epigenetic markers, top biomarkers affected, with potential diagnostic and prognostic value in sarcoma therapy.

Finally, our study demonstrates the value of Methylation EPIC arrays in understanding the epigenetic landscape of sarcoma subsets. The observed methylation signatures provide important insights into sarcoma biology and may aid in the development of personalised therapy methods to enhance patient outcomes.

Introduction and State of the Art

Sarcomas are a rare and heterogeneous group of malignancies arising from a variety of mesenchymal tissues, including bone, muscle, fat, and connective tissues. They encompass over 100 different histological subtypes with varying clinical behaviors and responses to therapy. (Grünewald 2020) Due to their rare occurrence and heterogeneity, sarcomas pose challenges in accurate classification, diagnosis, and treatment. Traditional diagnostic methods rely heavily on histopathological examination, which often fails to account for the complexity of these tumors. (Crombé 2022)

Epigenetic modifications and DNA methylation are important targets for cancer research and provide biomarkers for early detection, diagnosis, and prognosis in cancer patients. Methylation patterns can affect gene expression and are often abnormal in cancer, contributing to tumor growth and progression. Unlike genetic mutations, epigenetic changes are reversible, which makes them an attractive target for therapeutic intervention. (Pan 2018)

Although genetic alterations in sarcomas have been extensively studied, the epigenetic landscape has not been thoroughly explored. Recent advances in microarray technology such as the Illumina methylation EPIC array have enabled detailed profiles of DNA methylation on a genome-wide scale. This study seeks to fill the gap by exploring the methylation patterns specific to different sarcoma subtypes and their potential clinical implications.

Materials and Methods

We analyzed 125 sarcoma tissue samples, 107 of which were randomly selected from the Cancer Genome Atlas (TCGA) Sarcoma project came from six major sarcoma subtypes: dedifferentiated liposarcoma (DDLPS), leiomyosarcoma from gynecologic and soft tissue sites (LMS), undifferentiated pleomorphic sarcoma (UPS), myxofibrosarcoma (MFS), malignant peripheral nerve sheath tumor (MPNST), and synovial sarcoma (SS), also 18 from our institutional biobank. At this stage we only have UPS and MPNST samples, as other samples couldn't pass the criteria for being included in the analysis.

For our archived samples, DNA was extracted from formalin-fixed paraffin-embedded. Methylation profiling was performed according to the Illumina's protocol. (Illumina n.d.) Enzymatic fragmentation, hybridization, bisulfite conversion, and whole-genome amplification were among the preliminary processing methods.

Methylation data from both the 450K and 850K arrays (Version 1 and Version 2) were combined using R's different packages, like minfi which includes normalization processes to reduce batch effects and technical differences between the two devices. Quality control was performed to eliminate any low-quality or outlier samples. We extracted beta values from our 850k IDAT files and aligned them with TCGA's 450k data using a manifest file mapping common probes between both platforms. Filtered 3 datasets(450k,850KV1, and 850KV2) in order to retain only shared CpGs. After ensuring consistency in probe names, we merged the beta values from both platforms into a combined dataset. Duplicate transcripts were removed, and t-SNE analysis was performed to generate integrated methylation profiles, to compare the two data sets

To visualize the high-dimensional methylation data, we utilized t-distributed Stochastic Neighbor Embedding (t-SNE), which reduced it to two dimensions and revealed clustering patterns among sarcoma subtypes. The t-SNE analysis was carried out with the Rtsne software. The t-SNE plot showed how samples from different sarcoma subtypes cluster based on their methylation profiles.

We employed the MethSurv tool (Modhukur et al., 2018) to perform multivariable survival analyses on CpG methylation data. MethSurv provides a platform to assess the prognostic significance of individual CpG sites across multiple cancers.

Results

The t-SNE analysis, revealed clusters for our 6 subtypes of TCGA sarcomas, indicating subtype-specific methylation patterns. As it is clear samples from the same subtype cluster close together, which would indicate unique methylation patterns to each subtype. However, it is notable that some of these subtypes showed significant overlap; this could suggest similar epigenetic profiles or potential challenges in distinguishing these subtypes based on methylation alone.

As shown in Figure 1, the DDLPS samples in our dataset exhibit a robust, distinct methylation pattern that is specific to this subtype, as indicated by their general clustering together. The cluster's homogeneity indicates a possibly consistent epigenetic signature that may be useful for focused treatment approaches. While not as firmly as DDLPS, LMS samples tend to form a distinct cluster. Although LMS samples may be distinguished from other sarcoma subtypes by their generally consistent methylation profile, there may still be some overlapping methylation features between them and neighboring groups. MFS samples exhibit a loose grouping despite being slightly scattered. This might point to differences in the methylation profiles of the subtype, pointing to

subgroups or a transitory methylation pattern that might coincide with traits of closely related sarcoma types. Overlapping areas in synovial sarcomas, especially spindle cell and biphasic forms, may indicate that these subtypes share common methylation patterns. Because of the complexity and variability within synovial sarcomas, the clustering is not very distinct. This overlap could indicate shared pathways or diagnostic indicators, which would be clinically meaningful. MPNST and UPS show moderate clustering with some spread, suggesting that while each group samples share some genomic similarities, there is also variability, potentially reflecting different subtypes or stages within the UPS classification.

Our archived samples, 850K, shown by triangles on the plot are interleaved throughout existing clusters, indicating that these samples have similar methylation patterns to their TCGA categories. Notably, triangles 5, 6, and 7, which are classed as MPNST, indicating broader similarity in methylation patterns across other sarcoma types as well. Other triangular samples, representing UPS, are scattered around the plot, interacting with other sarcoma clusters. This pattern illustrates potential methylation overlap with other forms of sarcomas. The most noticeable point is that the number of samples as tests and controls are limited and it is not possible to see a significant distinction between different subtypes. We need to extend the number of samples for the future study.

Among the top biomarkers identified in TCGA-Sarcoma, several CpG sites demonstrated a strong association with survival. For instance, cg17658634 (NRN1) exhibited a hazard ratio of 3.854 (CI: 2-7.426, $p = 5.5e-05$), indicating a significant risk factor. Similarly, cg14128735 (ALX4) and cg13218710 (SHROOM1) were associated with increased mortality risk, with HRs of 3.716 (CI: 1.931-7.15, $p = 8.5e-05$) and 3.62 (CI: 1.93-6.787, $p = 6.1e-05$), respectively.

Survival analysis further confirmed the prognostic significance of these biomarkers. The high hazard ratios associated with CpG sites like cg17658634 suggest that hypermethylation at these loci is linked to poorer survival outcomes. Conversely, several CpG sites, such as cg26068283 (GRM5), displayed low hazard ratios (HR = 0.318, CI: 0.21-0.483, $p = 7.4e-08$), suggesting a protective effect against mortality.

Discussion

In conclusion, this study uses public database and archived samples methylation profiling to gain insight into the epigenetic landscape of different sarcoma subtypes. We were able to use state-of-the-art techniques such as the Illumina EPIC to identify unique methylation signals that characterize specific sarcoma subtypes. These subtype-specific methylation patterns identify promising biomarkers for diagnostic, prognostic, and therapeutic targeting in addition to providing important insights into the biology underlying these malignancies.

Our findings suggest that some sarcoma subtypes such as dedifferentiated liposarcoma (DDLPS) and leiomyosarcoma (LMS) exhibit complex and distinct methylation patterns, which may facilitate accurate classification and potentially inform personalized treatment strategies. However, the observed overlap of methylation profiles in some sarcoma subtypes, particularly synovial sarcomas and undifferentiated pleomorphic sarcomas (UPS), highlights the complexity of these tumors and the need for further studies to enhance subtype differentiation. Despite the valuable insights gained from this study, it is important to acknowledge the limitations, particularly the small sample size. Future studies with larger cohorts are necessary to validate these findings and further explore the clinical utility of the identified methylation markers. Expanding the sample size

and integrating additional molecular data may enhance our ability to distinguish between sarcoma subtypes and improve patient stratification for personalized therapy.

This study emphasizes how methylation profiling might help us learn more about the biology of sarcomas and promote the creation of diagnostic and treatment strategies that are guided by epigenetic signature. Identifying new epigenetic signals advances our ability to develop personalized treatments for sarcoma patients and provides hope for better prognoses in this difficult and varied class of cancers.

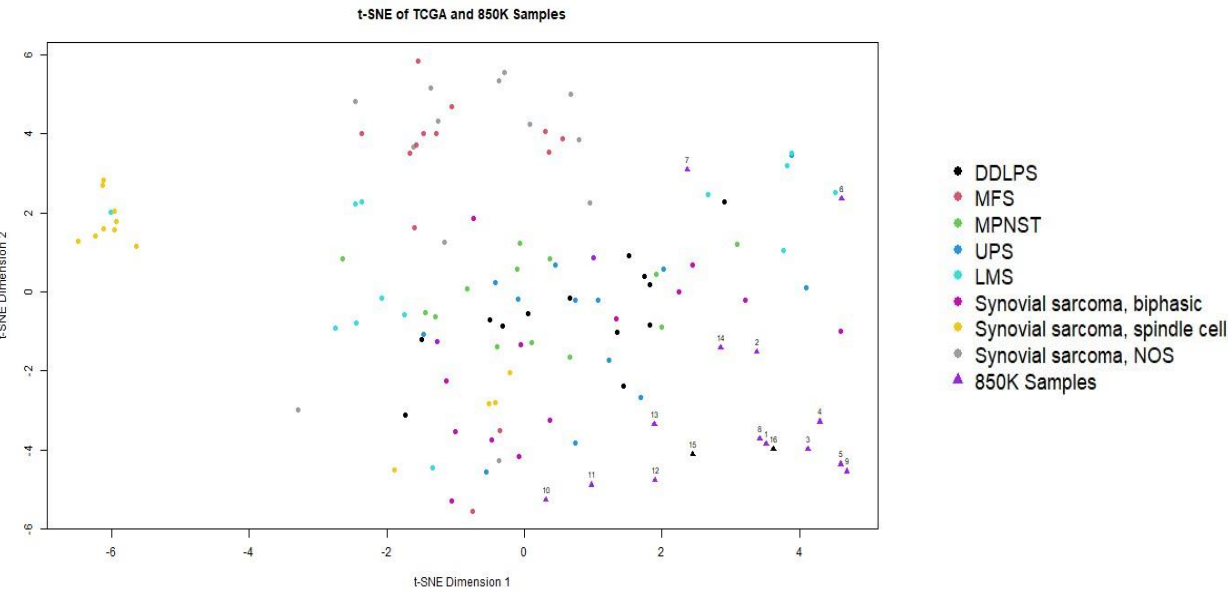


Figure 1: t-SNE Visualization of TCGA and 850K Sarcoma Samples

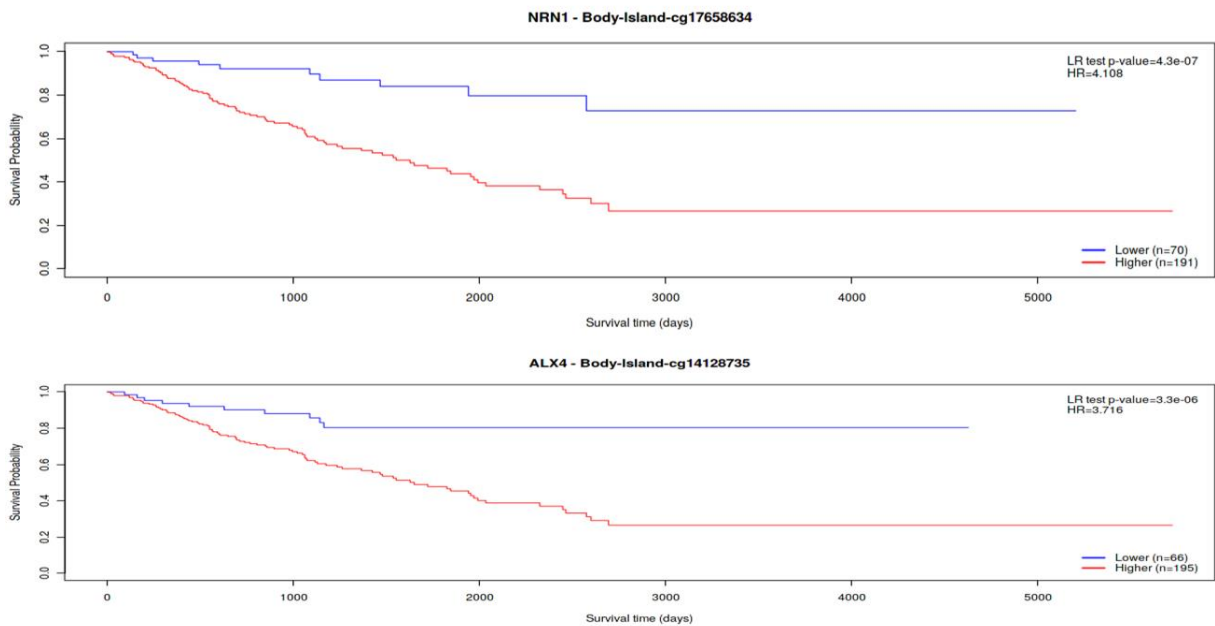


Figure 2: Kaplan-Meier Survival Analysis of ALX4 (cg14128735) and NRN1 (cg17658634) Methylation Status

References

- Crombé, A., Roulleau-Dugage, M., & Italiano, A. 2022. "The diagnosis, classification, and treatment of sarcoma in this era of artificial intelligence and immunotherapy." *Cancer Communications* 42: 1288 - 1313. doi:<https://doi.org/10.1002/cac2.12373>.
- Grünewald, T., Alonso, M., Avnet, S., Banito, A., Burdach, S., Cidre-Aranaz, F., Pompo, G., Distel, M., Dorado-García, H., García-Castro, J., González-González, L., Grigoriadis, A., Kasan, M., Koelsche, C., Krumbholz, M., Lecanda, F., Lemma, S., Longo, D. 2020. "Sarcoma treatment in the era of molecular medicine." *EMBO Molecular Medicine* 12. doi:<https://doi.org/10.15252/emmm.201911131>.
- Illumina. n.d. "Infinium MethylationEPIC v2.0 BeadChip." Accessed August 20, 2024. <https://emea.illumina.com/products/by-type/microarray-kits/infinium-methylation-epic.html#tabs-990a8871b5-item-3e9fb612fc-documentation>.
- Modhukur, Vahid, Toomas Iljasenko, Tambet Metsalu, Kaupo Lokk, Margus Heinmaa, Jaak Vilo, and Sulev Kõks. 2018. "MethSurv: A Web Tool to Perform Multivariable Survival Analysis Using DNA Methylation Data." *Epigenomics* 10, No 3 277-288.
- Pan, Y., Liu, G., Zhou, F., Su, B., & Li, Y. 2018. "DNA methylation profiles in cancer diagnosis and therapeutics." *Clinical and Experimental Medicine* 18: 1-14. doi:<https://doi.org/10.1007/s10238-017-0467-0>.

An Assessment of Non-Rigid Point Cloud Registration Methods on Two Novel Soft Tissue Deformation Datasets

1st Sara Monji Azad

*Mannheim Institute for Intelligent
Systems in Medicine (MIISM)
Medical Faculty Mannheim,
Heidelberg University
Mannheim, Germany*

sara.monji@medma.uni-heidelberg.de

2nd Marvin Kinz

*Department of Radiation Oncology
Brigham and Women's Hospital,
Dana-Farber Cancer Institute,
Harvard Medical School
Boston, MA, USA*

mkinz@bwh.harvard.edu

3rd Jyot Makadiya

*Mannheim Institute for Intelligent
Systems in Medicine (MIISM)
Medical Faculty Mannheim,
Heidelberg University
Mannheim, Germany*

jyot.makadiya@stud.uni-heidelberg.de

4th David Männle

*Department of Otorhinolaryngology,
Head and Neck Surgery,
Medical Faculty Mannheim,
Heidelberg University
Mannheim, Germany*

david.maennle@medma.uni-heidelberg.de

5th Claudia Scherl

*Department of Otorhinolaryngology,
Head and Neck Surgery,
Medical Faculty Mannheim,
Heidelberg University
Mannheim, Germany*

Claudia.Scherl@umm.de

6th Jürgen Hesser

*Mannheim Institute for Intelligent
Systems in Medicine (MIISM)
Medical Faculty Mannheim,
Heidelberg University
Mannheim, Germany*

juergen.hesser@medma.uni-heidelberg.de

Abstract—This paper presents a comprehensive evaluation of three non-rigid point cloud registration methods—Deep-Geo-Reg, GP-Aligner, and our proposed Robust-DefReg—across two newly introduced datasets, SynBench (synthetic) and DeformedTissue (real-world). These datasets were specifically designed to simulate soft tissue deformation challenges during surgery, with the aim of providing a realistic and robust benchmark for non-rigid point cloud registration methods. Our evaluation covers varying levels of deformation and noise, highlighting the strengths and weaknesses of each method. Robust-DefReg consistently outperforms the other methods, maintaining the lowest mean distances across both synthetic and real datasets, demonstrating superior accuracy and robustness to noise. The study highlights the usefulness of the proposed datasets for testing non-rigid point cloud registration methods and provides insights into how well the evaluated methods perform.

Index Terms—Soft Tissue Deformation, Non-Rigid Registration, Point Cloud Registration, Deformed Objects Datasets

I. INTRODUCTION AND RELATED WORK

A point cloud is a collection of data points within a coordinate space, where each point has unique coordinates that determine its position. These points usually represent the external surface of a 3D object or environment. Point cloud registration involves calculating a spatial transformation to align a source point cloud with a target point cloud, ensuring that corresponding points in both clouds are as closely matched as possible [1].

Point cloud registration can involve two types of transformations: rigid and non-rigid. A rigid transformation maintains the object's shape and size, involving only rotation and translation.

On the other hand, a non-rigid transformation allows for deformation of the object, where points within the cloud can move relative to each other. This type of transformation can involve stretching, bending, twisting, or scaling, making it suitable for aligning point clouds that represent objects or surfaces that may undergo shape changes, such as biological tissues or flexible materials. Non-rigid transformations are more complex to model and compute, as they require estimating the spatial variations in the object's structure while maintaining the overall correspondence between the point clouds. Given two sets of points $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^3 \mid i = 1, \dots, N\}$ and $\mathbf{Y} = \{\mathbf{y}_j \in \mathbb{R}^3 \mid j = 1, \dots, M\}$, the goal of 3D point cloud non-rigid registration is to find a non-rigid transformation $\mathcal{T}$ that maps the source point cloud $\mathbf{X}$ onto the target point cloud $\mathbf{Y}$. This transformation can be represented as $\mathcal{T}(\mathbf{x}_i) = \mathbf{x}'_i$, where $\mathbf{x}'_i$ are the transformed points in $\mathbb{R}^3$. The objective is to minimize a cost function $E(\mathcal{T})$, typically defined as:

$$E(\mathcal{T}) = \sum_{i=1}^N \|\mathbf{x}'_i - \mathbf{y}_{j(i)}\|^2 + \lambda \text{Reg}(\mathcal{T}),$$

where $\|\cdot\|$ denotes the Euclidean distance, $\mathbf{y}_{j(i)}$ is the corresponding point in $\mathbf{Y}$ for $\mathbf{x}'_i$, λ is a regularization parameter, and $\text{Reg}(\mathcal{T})$ is a regularization term that ensures smoothness and prevents excessive deformation of the point cloud.

This paper evaluates three non-rigid point cloud registration methods: Deep-Geo-Reg [4], GP-Aligner [5], and our proposed method, Robust-DefReg [6], using two newly introduced datasets, SynBench [3] and DeformedTissue [2].

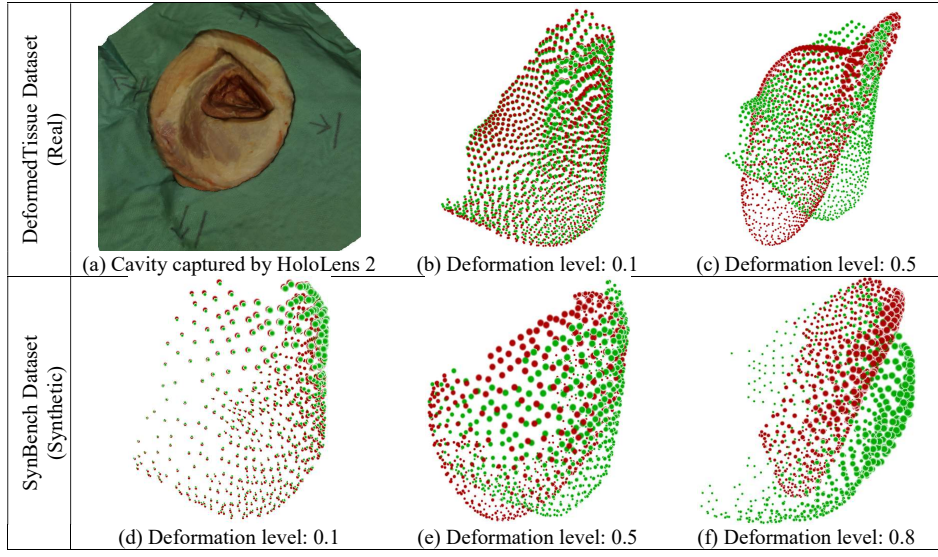


Fig. 1. Visualization of the DeformedTissue (real) [2] and SynBench (synthetic) [3] datasets used in the study. (a) shows a cavity captured by HoloLens 2 from the DeformedTissue dataset. (b) and (c) depict different deformation levels (0.1 and 0.5) within the DeformedTissue dataset. (d), (e), and (f) illustrate various deformation levels (0.1, 0.5, and 0.8) within the SynBench dataset. In the point cloud visualizations, the source point clouds are shown in green, and the target point clouds are shown in red, representing different degrees of deformation.

SynBench is a synthetic dataset created with our SimTools [7], designed to simulate tissue deformation during surgery. DeformedTissue is a real dataset derived from experiments on pig heads to mimic tissue deformation [2]. The main goal of this study is to show the practical usefulness of our datasets by testing these baseline non-rigid point cloud registration methods and comparing their performance on both synthetic and real datasets related to soft tissue deformation.

Several learning architectures are commonly utilized in learning-based registration methods, including CNNs, RNNs, GNNs, and MLPs. Numerous survey articles have explored these architectures from various perspectives. For example, [8] reviews pairwise and groupwise registration methods, while [9] and [10] focus on learning-based methods. Non-learning approaches are discussed in [11], and [12] examines both learning and non-learning methods for non-rigid point clouds, addressing same-source and cross-source domains. Other reviews cover 3D multi-modal registration [13], learning-based registration with diverse inputs [14], mobile robotics [15], and laser scanner point clouds [16], [17]. Given these existing studies, the article [1] focuses on learning approaches and non-rigid transformations, offering a comprehensive overview that includes existing datasets, evaluation metrics, and challenges.

In this paper, we discuss three baseline methods and the experimental datasets in Section II. The experimental results and evaluation performance are presented in Section III. A discussion on the potential and limitations of the investigated methods is provided in Section IV.

II. MATERIALS AND METHODS

As previously mentioned, we have developed two new soft tissue deformation datasets, which are described in the

materials section. Following that, three baseline methods will be discussed.

A. Materials

Various datasets have been developed to assess point cloud registration methods, but a comprehensive benchmark that covers all relevant challenges is still lacking. To address this, the authors introduced SynBench as a synthetic dataset and DeformedTissue as a real dataset. Some samples of SynBench and DeformedTissue datasets are shown in Figure 1.

1) *SynBench*: In a previous study, the authors presented SimTool [7], a toolbox for simulating soft body deformation and generating deformable point clouds. We used SimTool to create SynBench, a benchmark designed for evaluating non-rigid point cloud registration methods. SynBench, available at <https://doi.org/10.11588/data/R9IKCF>, is based on 30 primitive objects from SimTool outputs, featuring various challenges like outlier ratios, noise levels, and incompleteness, across small to large deformation levels. The dataset includes five main sets: "Data," "Deformation Level," "Incompleteness," "Noise," and "Outlier," containing thousands of object samples, each with corresponding source and target point clouds.

2) *DeformedTissue*: Tissue deformation is a critical issue in soft-tissue surgery, particularly during tumor resection, as it causes landmark displacement, complicating tissue orientation. The authors conducted an experimental study on 45 pig head cadavers to simulate tissue deformation, approved by the Mannheim Veterinary Office (DE 08 222 1019 21) [2] [18]. We used 3D cameras and head-mounted displays to capture tissue shapes before and after controlled deformation induced by heating. The data was processed using software

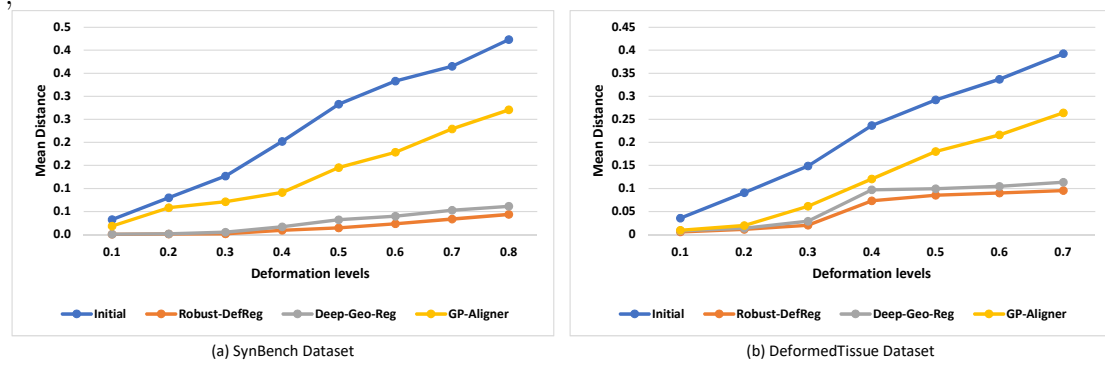


Fig. 2. Mean distances of three non-rigid point cloud registration methods (Robust-DefReg [2], Deep-Geo-Reg [4], and GP-Aligner [5]) compared to the unregistered state across different deformation levels on SynBench (synthetic) and DeformedTissue (real) datasets. Robust-DefReg consistently outperforms other methods, maintaining the lowest distances.

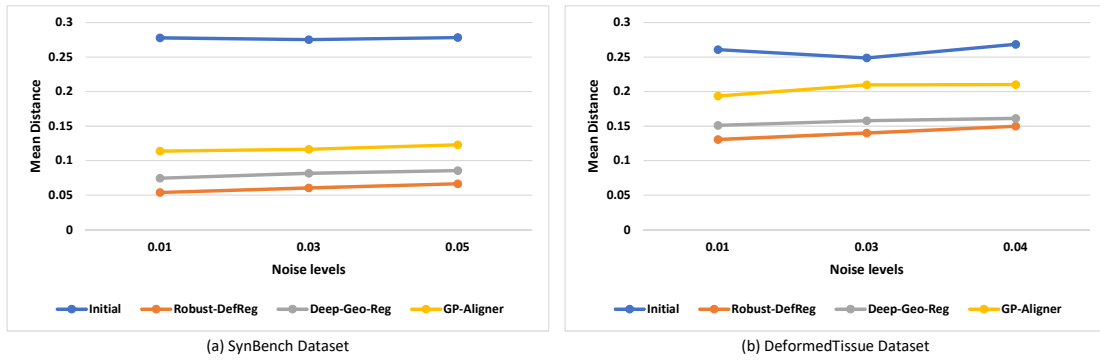


Fig. 3. Performance comparison of three non-rigid point cloud registration methods (Robust-DefReg [2], Deep-Geo-Reg [4], and GP-Aligner [5]) on SynBench (synthetic) and DeformedTissue (real) datasets under varying noise levels. Robust-DefReg consistently achieves the lowest mean distances, demonstrating strong resilience to noise across both datasets.

like Meshroom, MeshLab, and Blender to create and evaluate 2½D meshes. The dataset is available upon request to readers.

B. Methods

The first studied method, a point cloud registration approach for deformed lungs proposed by Hansen et al. [1], begins by generating feature vectors for each point in the source cloud using the Geometric Feature Net (GFN). Next, the method finds the closest points (k-nearest neighbors) in the target cloud for each point in the source cloud. It then calculates possible shifts (displacements) for these points and estimates how much each shift will cost, based on how similar the points' features are. A k-NN graph is created in the source cloud to understand the spatial relationships between points. To reduce errors from noisy data, a process called Loopy Belief Propagation (LBP) is used to refine these costs by repeatedly updating the estimates of which points match. Finally, a Softmax function is used to standardize the costs, and these adjusted costs are combined with the calculated shifts to generate the final transformation that aligns the source and target point clouds.

The second method introduces an unsupervised approach called GP-Aligner [5] for groupwise non-rigid point set registration, focusing on aligning multiple highly deformed 3D shapes. Unlike traditional methods, GP-Aligner leverages deep

neural networks for more effective and efficient alignment without requiring extensive training data. Instead of using explicit feature encoding networks, GP-Aligner employs a model-free learnable latent descriptor to capture shape correlations within a group. This is achieved by initializing a group latent descriptor (GLD) from a Gaussian distribution, which is then combined with the coordinates of each point in the group. A neural network-based decoder predicts flow fields that align the shapes within the group. During optimization, both the GLDs and the decoder's parameters are updated to minimize an unsupervised alignment loss, aligning the point sets toward a mean shape without designating a specific target shape.

The third studied method is Robust-DefReg [6], a robust non-rigid point cloud registration approach that leverages Graph Convolutional Networks (GCNNs). It employs a coarse-to-fine registration strategy within an end-to-end pipeline, combining the advantages of both coarse and fine methods. Initially, Robust-DefReg learns global features to identify correspondences between source and target point clouds for initial alignment, followed by fine registration. The method is designed to achieve high accuracy and robustness across various challenges, addressing gaps often left by existing approaches. Despite the increased complexity of its feature

descriptor network, Robust-DefReg maintains computational efficiency through the integration of T-Net, GCNNs, and Loopy Belief Propagation (LBP).

III. RESULTS

As shown in Figure 2, the performance of three non-rigid point cloud registration methods—Robust-DefReg [6], Deep-Geo-Reg [4], and GP-Aligner [5]—is compared to the initial unregistered state across various deformation levels on two datasets: SynBench (synthetic) and DeformedTissue (real). Robust-DefReg consistently achieves the lowest mean distances, maintaining values close to 0.1 in the SynBench dataset and between 0.05 and 0.1 in the DeformedTissue dataset, indicating superior performance. Deep-Geo-Reg also performs well, with mean distances slightly higher than Robust-DefReg but still significantly lower than the initial state, particularly staying around 0.1 to 0.15. In contrast, GP-Aligner shows a more noticeable increase in mean distance, especially in the SynBench dataset, where it climbs from about 0.1 to 0.35, making it less effective as deformation increases. Overall, Robust-DefReg emerges as the most robust method, effectively handling both synthetic and real-world deformations across all levels. As mentioned, noise is one of the most important challenges in non-rigid point cloud registration. In this regard, the three mentioned methods are also studied on the two datasets with varying levels of noise added to the point clouds. In the SynBench dataset, Robust-DefReg performs the best, maintaining the lowest mean distance consistently around 0.05, indicating strong resilience to noise. Deep-Geo-Reg also performs well, with mean distances slightly above Robust-DefReg, hovering around 0.1. GP-Aligner, however, shows higher mean distances, around 0.15 to 0.2, indicating that it is less effective at handling noise in the synthetic dataset. In the DeformedTissue dataset, the initial mean distance stays relatively stable, close to 0.3 across all noise levels. Robust-DefReg again achieves the lowest mean distances, around 0.05 to 0.1, demonstrating its robustness to noise even in real-world data. Deep-Geo-Reg performs similarly to its performance on the synthetic dataset, with mean distances around 0.1 to 0.15. GP-Aligner shows higher mean distances, similar to the SynBench results, ranging from 0.15 to 0.2.

IV. DISCUSSION

The results of this study clearly show that Robust-DefReg is the most effective method among the ones studied for non-rigid point cloud registration, particularly in scenarios involving high levels of deformation and noise. Its ability to maintain low mean distances across both synthetic and real datasets demonstrates its robustness and generalizability. Deep-Geo-Reg also performs well, though it is slightly less resilient to noise and deformation compared to Robust-DefReg. GP-Aligner, while still reducing mean distances compared to the unregistered state, shows greater sensitivity to noise and deformation, particularly in the synthetic dataset. The results suggest that Robust-DefReg's coarse-to-fine registration strategy and its use of Graph Convolutional Networks

contribute significantly to its superior performance. Overall, this study highlights the importance of using diverse tissue deformation datasets, such as SynBench and DeformedTissue, for evaluating non-rigid registration methods, and suggests that Robust-DefReg is well-suited for practical applications in complex, deformable environments like soft tissue surgeries.

REFERENCES

- [1] S. Monji-Azad, J. Hesser, and N. Löw, "A review of non-rigid transformations and learning-based 3d point cloud registration methods," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 58–72, 2023.
- [2] S. Monji-Azad, D. Männle, J. Hesser, J. Pohlmann, N. Rotter, A. Afholter, C. A. Weis, S. Ludwig, and C. Scherl, "Point cloud registration for measuring shape dependence of soft tissue deformation by digital twins in head and neck surgery," *Biomedicine hub*, vol. 9, no. 1, pp. 9–15, 2024.
- [3] S. Monji-Azad, M. Kinz, C. Scherl, D. Männle, J. Hesser, and N. Löw, "Synbench: A synthetic benchmark for non-rigid 3d point cloud registration," *arXiv preprint arXiv:2409.14474*, 2024.
- [4] L. Hansen and M. P. Heinrich, "Deep learning based geometric registration for medical images: How accurate can we get without visual features?" in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 18–30.
- [5] L. Wang, N. Zhou, H. Huang, J. Wang, X. Li, and Y. Fang, "Gp-aligner: Unsupervised groupwise nonrigid point set registration based on optimizable group latent descriptor," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.
- [6] S. Monji-Azad, M. Kinz, and J. Hesser, "Robust-defreg: A robust deformable point cloud registration method based on graph convolutional neural networks," *arXiv preprint arXiv:2306.04701*, 2023.
- [7] S. Monji-Azad, M. Kinz, J. Hesser, and N. Löw, "Simtool: A toolset for soft body simulation using flex and unreal engine," *Software Impacts*, vol. 17, p. 100521, 2023.
- [8] H. Zhu, B. Guo, K. Zou, Y. Li, K.-V. Yuen, L. Mihaylova, and H. Leung, "A review of point set registration: From pairwise registration to groupwise registration," *Sensors (Basel, Switzerland)*, vol. 19, no. 5, 2019.
- [9] Z. Zhang, Y. Dai, and J. Sun, "Deep learning based point cloud registration: an overview," *Virtual Reality & Intelligent Hardware*, vol. 2, no. 3, pp. 222–246, 2020.
- [10] W. Tang, D. Zou, and P. Li, "Learning-based point cloud registration: A short review and evaluation," in *2021 2nd International Conference on Artificial Intelligence in Electronics Engineering*. New York, NY, USA: ACM, 2021.
- [11] B. Maiseli, Y. Gu, and H. Gao, "Recent developments and trends in point set registration methods," *Journal of Visual Communication and Image Representation*, vol. 46, pp. 95–106, 2017.
- [12] X. Huang, G. Mei, J. Zhang, and R. Abbas, "A comprehensive survey on point cloud registration," *arXiv preprint arXiv:2103.02690*, 2021.
- [13] E. Saiti and T. Theoharis, "An application independent review of multimodal 3d registration methods," *Computers & Graphics*, vol. 91, pp. 153–178, 2020.
- [14] V. Villena-Martinez, S. Oprea, M. Saval-Calvo, J. Azorin-Lopez, A. Fuster-Guillo, and R. B. Fisher, "When deep learning meets data alignment: A review on deep registration networks (drns)," *Applied Sciences*, vol. 10, no. 21, p. 7524, 2020.
- [15] F. Pomerleau, F. Colas, R. Siegwart *et al.*, "A review of point cloud registration algorithms for mobile robotics," *Foundations and Trends® in Robotics*, vol. 4, no. 1, pp. 1–104, 2015.
- [16] Z. Dong, F. Liang, B. Yang, Y. Xu, Y. Zhang, J. Li, Y. Wang, W. Dai, H. Fan, J. Hyypä *et al.*, "Registration of large-scale terrestrial laser scanner point clouds: A review and benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 163, pp. 327–342, 2020.
- [17] L. Cheng, S. Chen, X. Liu, H. Xu, Y. Wu, M. Li, and Y. Chen, "Registration of laser scanning point clouds: A review," *Sensors*, vol. 18, no. 5, p. 1641, 2018.
- [18] D. Männle, J. Pohlmann, S. Monji-Azad, J. Hesser, N. Rotter, A. Afholter, A. Lammert, B. Kramer, S. Ludwig, L. Huber *et al.*, "Artificial intelligence directed development of a digital twin to measure soft tissue shift during head and neck surgery," *Plos one*, vol. 18, no. 8, p. e0287081, 2023.

Accuracy and Precision of 3D Reconstruction Methods for Localizing Point-Like Gamma Sources with a Coded Aperture Camera

1st Tobias Meißner

Mannheim Institute for Intelligent Systems in Medicine
Heidelberg University
Mannheim, Germany
ORCID: [0000-0002-0435-5103](https://orcid.org/0000-0002-0435-5103)

2nd Werner Nahm

Institute of Biomedical Engineering
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
ORCID: [0000-0001-8095-5090](https://orcid.org/0000-0001-8095-5090)

3rd Jürgen Hesser

Mannheim Institute for Intelligent Systems in Medicine
Interdisciplinary Center for Scientific Computing (IWR)
Central Institute for Computer Engineering (ZITI)
CZS Heidelberg Center for Model-Based AI
Heidelberg University
Mannheim, Germany
ORCID: [0000-0002-4001-1164](https://orcid.org/0000-0002-4001-1164)

Abstract—Gamma cameras with coded aperture collimators are under investigation for guiding surgeries like sentinel lymph node biopsy. Coded aperture collimators promise a better trade-off between sensitivity and spatial resolution, but require image reconstruction. Additionally, they offer the possibility of 3D-localization, where two 3D reconstruction methods have been used: MURA Decoding and a convolutional Maximum Likelihood Expectation Maximization (3D-MLEM) algorithm. The goal of this paper is to assess the accuracy of the 3D-MLEM algorithm, the more precise 3D reconstruction method, in localizing point-like gamma sources. With the Monte Carlo (MC) simulation TOPAS MC, a dataset of 17 images was generated where the field of view (FOV) of an experimental gamma camera was systematically sampled with a point-like ²⁴¹Am source. The previously enhanced 3D-MLEM algorithm and the developed Iterative Source Localization method (ISL) is used to estimate the source position by alternating between a lateral and an axial search. Additionally, we assessed the axial resolution as the method's precision by measuring the Full Width at Half Maximum (FWHM) of the Contrast-to-Noise Ratio (CNR) profile. The average localization error is 0.63mm and 0.77mm for 3D-MLEM and MURA Decoding. The FWHM increases for 3D-MLEM from 2.76mm to 12.14mm for a 20mm and 100mm source-to-mask distance. In conclusion, it is demonstrated that that despite its superior precision, the 3D-MLEM algorithm only provides marginally smaller localization errors than MURA Decoding. The selection of reconstruction planes remains an important issue for further research.

Index Terms—coded aperture imaging, gamma imaging, image reconstruction, nuclear medicine, 3d-localization.

I. INTRODUCTION

With 2.3 million new cases worldwide in 2020, female breast cancer is now the most commonly diagnosed cancer [1]. The biopsy of pathological examination of Sentinel

Lymph Nodes (SLNs), the first lymph nodes to receive lymph-borne metastatic cells, are crucial for staging breast cancer due to their role in predicting recurrence and survival [2], [3]. SLNs are dissected by injecting a radiotracer that accumulates in the SLNs emitting gamma photons, which are detected during Sentinel Lymph Node Biopsy (SLNB). Recently, Intraoperative Gamma Cameras (IGCs) are under investigation to improve SLNB by providing visual guidance during surgery, offering more comprehensive data than counting gamma probes [4]–[6]. Because of the high-energy photons involved, refractive lenses cannot be used to produce an image of the scene. Instead, parallel-hole or pinhole collimators are utilized to capture the required spatial information, but suffer from a sub-optimal trade-off between photon sensitivity (acquisition time) and spatial resolution (high-resolution images). [7]. Coded aperture collimators have been proposed since they offer a better compromise between those key imaging characteristics. [8], [9]. However, Coded Aperture Imaging (CAI) requires image reconstruction which adds a layer of complexity to the imaging procedure [5], with multiple reconstruction methods proposed [10]. Another advantage of CAI, is that the source-to-mask distance can be estimated from a single acquisition [11]. Fig. 1 illustrates, how the lateral position of a point source is encoded by the shift of the mask's shadow, while the source-to-mask distance is related to the size of the shadow. Thus, in theory a 3D-localization of point-like sources is feasible. As Fig. 2 illustrates, on one hand, the accuracy of a reconstruction method can be described as how close the obtained source position is to its true position. On the other hand, the precision describes the spread of multiple measurements and is, thus, usually identified as the resolution. In the case of 3D-localization especially the axial resolution is of interest, because it is multiple times worse than the lateral resolution [12].

In a previous work, the axial resolution of two 3D reconstruction methods were compared and a major difference

The authors gratefully acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG. This paper was partially funded by the Zentrales Innovationsprogramm Mittelstand (ZIM) under grant KK5044701BS0.

between MURA Decoding and convolutional 3D Maximum Likelihood Expectation Maximization (3D-MLEM) algorithm discovered [12]. On average 3D-MLEM yields a 0.3 times lower, and thus, better axial resolution than MURA Decoding. In a consequent study, it was shown that the monomodal Contrast-to-Noise Ratio (CNR) profile, extracted from the 3D reconstruction, can be used to estimate the 3D position of point-like sources in the near-field. On a simulated dataset a localization accuracy of below 1 mm and on experimental data below 3 mm was obtained. However, the localization accuracy achievable with the 3D-MLEM algorithm remained an open question. Thus, the goal of this study is to address the following research question: How accurately can the 3D-MLEM algorithm, the more precise 3D reconstruction method, localize point-like gamma sources? In this paper, we show that the 3D-MLEM algorithm is capable of localizing point-like gamma sources with approximately the same accuracy as MURA Decoding.

II. MATERIAL & METHODS

First, the Monte Carlo Simulation is described that was used to generate a dataset of images based on our experimental IGC. Then, the 3D-MLEM algorithm is briefly presented, before the localization method is outlined. For the remainder of this paper, the simulated detector image is denoted as $p(x, y)$, its 3D reconstruction as $\hat{f}_z(x, y)$, and the estimated source position as $[\hat{x}, \hat{y}, \hat{z}]$.

A. Monte Carlo Simulation

To analyze the accuracy and precision, a dataset consisting of 17 images was simulated with the Monte Carlo (MC) simulation framework TOPAS MC [13]. More details about the camera setup can be found in [14]. In summary, the modeled IGC was composed of a rank 31 Not-Two-Holes-Touching (NTHT) Modified Uniform Redundant Array (MURA) mask having 0.08 mm diameter holes in a 0.11 mm thick tungsten sheet and a semiconductor-based detector unit. The basic MURA pattern was duplicated to form a 2×2 arrangement leading to a total mask size of $D_m \times D_m = 9.92 \times 9.92 \text{ mm}^2$. The detector of our experimental IGC is a Timepix3[®] chip with a pixelation of 256×256 , bump-bonded to a 0.5 mm thick silicon crystal and a sensitive area of $D_d \times D_d = 14.08 \times 14.08 \text{ mm}^2$. However, modeling an active gamma

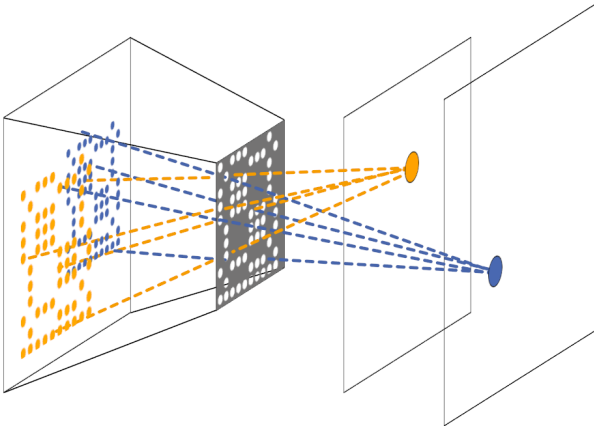


Fig. 1: The basic principle of 3D Coded Aperture Imaging (CAI): the lateral position of a point source is encoded by the shift of the mask's shadow, while the source-to-mask distance determines the size of the shadow.

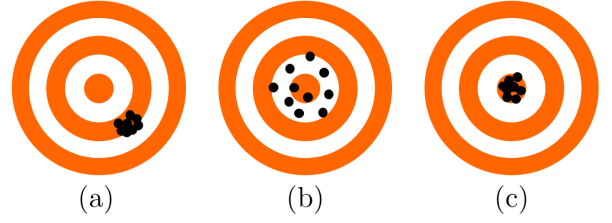


Fig. 2: The difference between accuracy and precision explained on a dartboard: (a) has a high precision but low accuracy, (b) has a high accuracy but a low precision and (c) has a high accuracy and precision.

detector in TOPAS MC is a field of ongoing research [15], it was simplified as a so-called *phase space*. The phase space is a virtual plane as large as the detector area, where every photon crossing this plane is stored along with its position, direction and photon energy. The detector-to-mask distance was set to $b = 20 \text{ mm}$. The source was modeled as a sphere of ^{241}Am , because it is the smallest source available to use for real-world experiments. The volumetric source had a diameter of 1 mm, and the emitted energy was set to 59.5 keV. The simulation was run until a total of 10^9 photons have been emitted. Finally, the phase space files were converted into images of 256×256 pixels where the intensity represents the photon count per pixel.

All in all, 17 detector images were acquired, where the field of view (FOV) was sampled at four different source-to-mask distances: 20 mm, 50 mm, 75 mm, and 100 mm. While the x-position was kept constant at 0 mm, we moved the source along the y-axis in 2 mm increments from the center outward to 8 mm or until the end of the FOV was reached. Additionally, an image the source position $[0, 14, 100] \text{ mm}$ was captured. Table I shows all 17 positions.

B. 3D Reconstruction

The axial resolution of two reconstruction methods for 3D imaging have been analyzed in a previous work [12]: sequential MURA Decoding and an enhanced 3D-MLEM algorithm. While the 3D reconstruction from MURA Decoding consist of a sequence of planar reconstructions at multiple in-focus planes, the 3D-MLEM algorithm provides a 3D reconstruction where all slices in axial direction are reconstructed simultaneously. For a detailed description, the reader is referred to the publication mentioned above. Since this study focuses on the localization accuracy of the 3D-MLEM algorithm its equation is presented. The 3D-MLEM algorithm is an iterative reconstruction method developed by Mu et al. [16].

The process involves a forward and backward projection of the current estimate, where instead of using a transfer matrix, the projections are performed via convolution with the point-spread function (PSF) $h_z[x, y]$. To improve readability, the lateral coordinates $[x, y]$ are omitted, and the axial coordinate z is represented as a subscript. The 3D reconstruction $\hat{f}_z^{(k+1)}$ at iteration k can then be defined as follows:

$$\hat{f}_z^{(k+1)} = \frac{\hat{f}_z^{(k)}}{s_z} \cdot \left[\frac{p - \sum_{\forall z' \neq z} \mathbf{F}(\hat{f}_{z'}^{(k)}, h_{z'})}{\mathbf{F}(\hat{f}_z^{(k)}, h_z)} \times h_z \right] \quad (1)$$

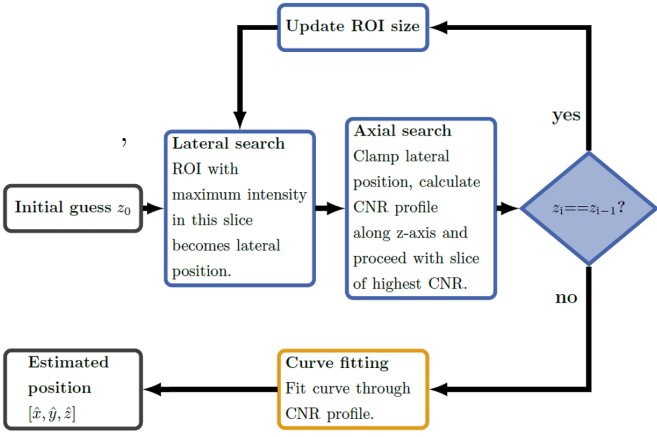


Fig. 3: A flowchart representing the ISL algorithm with its alternating lateral and axial search and the final curve fitting to estimate the 3D position of the source. Fig. from [14].

with the forward projection operator $\mathbf{F}$

$$\mathbf{F}(\hat{f}_z^{(k)}, h_z) = (1 - T) (\hat{f}_z^{(k)} * h_z) + T \sum_{x,y} \hat{f}_z^{(k)}, \quad (2)$$

where “ $\times$ ” denotes the 2D correlation operator, “ $*$ ” the 2D convolution operator, and T the transmission probability of the mask and the photon energy involved. The sensitivity term s_z is defined as

$$s_z[x, y] = h_z[x, y] \times \mathbf{1}. \quad (3)$$

No preprocessing of the detector images took place prior to reconstruction. From a visual inspection we found 40 iterations to be a good compromise between noise amplification and reconstruction quality. Each detector images was reconstructed with a tight z -range around the actual source position in 0.5 mm steps resulting in image stacks with between 50 and 101 images. The range for MURA Decoding was larger and the stacks contain between 60 to 240 images. Both algorithms were implemented in Python (3.8) using the NumPy library (1.24), where all convolutions are performed as multiplications in the frequency domain via 2D Fourier transformation. These, and other reconstruction methods (planar and 3D) are publicly available on Github [17].

C. Iterative Source Localization Method

The Iterative Source Localization Method (ISL) was developed in a previous work to obtain the source position from a 3D reconstruction [14]. It is not based on the reconstruction intensity itself, but on the Contrast-to-Noise Ratio (CNR), a common metric in nuclear imaging [18].

The ISL method iteratively alternates between a lateral and an axial search, followed by fitting a curve through the obtained CNR profile. In the last step, an Exponentially Modified Gaussian (EMG) curve is fitted through the final CNR profile, because it takes into account the unsymmetrical behavior in z -direction discovered in [14]. A flowchart of the ISL is presented in Fig. 3. The algorithm requires the 3D reconstruction, the FOV for each slice, an initial guess denoted as z_0 for the source-to-mask distance, and the source size as input. The output of the ISL is the estimated source center $[\hat{x}, \hat{y}, \hat{z}]$ in millimeters. A thorough description can

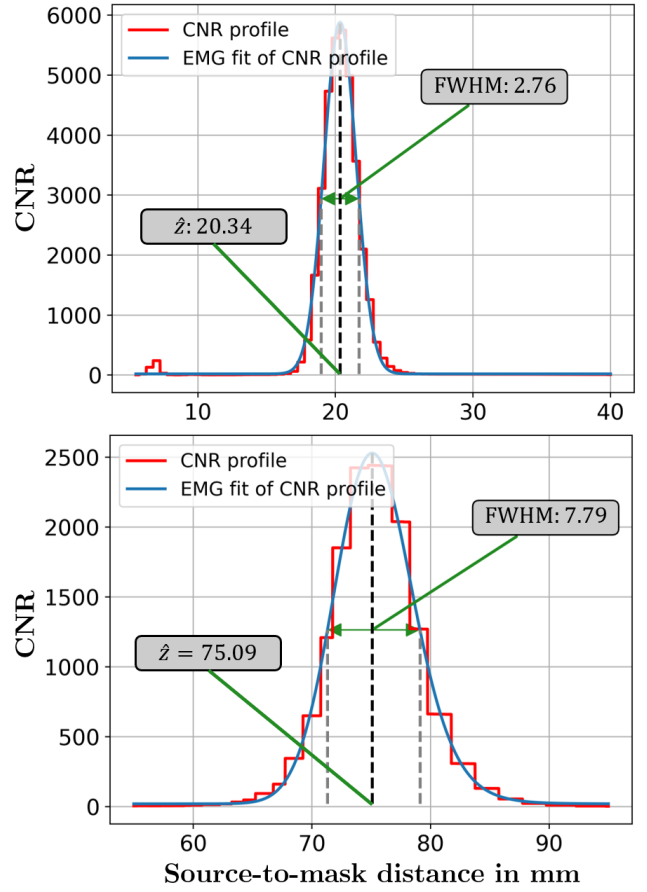


Fig. 4: Two exemplary CNR profiles and the EMG fit obtained by the ISL method from 3D-MLEM. Top: source at $[0, 0, 20]$ mm. Bottom: source at $[0, 0, 75]$ mm.

be found in the publication mentioned above. We used the correct z -coordinate as the initial guess, because both axial as well as lateral ghost sources impeded the localization.

D. Axial Resolution

The axial resolution indicates the precision with which a point source can be localized along the z -axis, i.e. in depth direction. It is typically represented by the Full Width at Half Maximum (FWHM) of the discrete reconstruction intensity across the source. In this work, we measure the axial resolution of the CNR profile of the obtained 3D reconstruction for reasons that were previously explained [12].

Instead of using an equation to directly convert the fitted EMG curve to the FWHM [19], we employed a numerical method to determine the FWHM. First, the fitted curve was finely sampled with a $\Delta z = 0.01$ mm. Then, the two points where the curve intersects half the maximum CNR were identified by finding the z -values that is minimal regarding the squared distance to the horizontal line. The difference between these two values give the FWHM, which corresponds to the axial resolution. A graph of the original CNR profile, its fitted EMG curve and the obtained axial resolution is depicted in Fig. 4. The axial resolution was calculated for all 17 3D reconstructions and both reconstruction methods.

III. RESULTS

The full localization results of the 3D-MLEM algorithm together with the results obtained from MURA Decoding are presented in Tab. I. Additionally, the axial res-

TABLE I: This table shows the Ground Truth (GT) of the 17 source positions and the localization results with the ISL-EMG method based on 3D reconstructions from both the 3D-MLEM algorithm and MURA Decoding. In addition to the estimations, the 3D localization error (L_2), the coefficient of determination (R^2) and the Full Width at Half Maximum (FWHM) of the EMG fit as axial resolution are presented.

Ground Truth			3D-MLEM							MURA Decoding							
x	y	z	$\hat{x}$	$\hat{y}$	$\hat{z}$	L_2	R^2	FWHM	$\hat{x}$	$\hat{y}$	$\hat{z}$	L_2	R^2	FWHM			
0	0	20	0.03	-0.03	20.34	0.34	1.00	2.76	-0.02	0.14	20.36	0.39	0.98	8.28			
0	0	50	0.07	-0.07	48.60	1.40	0.97	5.84	-0.03	0.30	50.28	0.41	0.94	27.89			
0	2	50	0.07	1.84	49.41	0.61	0.99	5.19	-0.10	2.23	49.36	0.69	0.92	25.66			
0	4	50	0.07	3.86	49.06	0.95	0.99	5.05	-0.03	4.19	49.80	0.28	0.99	22.92			
0	6	50	0.07	5.94	49.35	0.66	0.99	5.58	0.03	6.21	49.81	0.28	0.98	23.77			
0	8	50	0.07	7.78	49.13	0.90	0.99	5.16	-0.03	8.20	50.02	0.20	0.99	24.80			
0	0	75	0.10	-0.10	75.09	0.17	0.99	7.79	0.04	0.40	74.65	0.54	0.93	40.13			
0	2	75	0.10	1.93	74.94	0.14	0.99	7.66	-0.14	2.27	75.55	0.63	0.97	42.14			
0	4	75	0.10	3.83	75.01	0.20	0.99	8.52	-0.14	4.30	74.31	0.77	0.98	38.02			
0	6	75	0.10	5.78	74.61	0.46	0.99	7.41	-0.05	6.24	74.33	0.71	0.98	37.85			
0	8	75	0.10	7.81	74.82	0.28	0.99	8.19	-0.05	8.28	75.18	0.34	0.99	38.37			
0	0	100	0.14	-0.14	100.58	0.61	0.98	12.11	-0.17	0.29	100.07	0.35	0.94	48.03			
0	2	100	0.14	1.79	101.15	1.18	0.98	12.60	-0.17	2.26	102.05	2.07	0.96	48.59			
0	4	100	0.14	3.73	100.40	0.50	0.98	11.81	-0.17	4.35	102.15	2.19	0.98	50.04			
0	6	100	0.14	5.94	101.05	1.06	0.98	12.66	-0.06	6.32	100.48	0.58	0.99	51.88			
0	8	100	0.14	7.87	100.28	0.34	0.98	12.90	-0.17	8.29	100.60	0.69	0.98	50.51			
0	14	100	0.14	13.94	100.93	0.94	0.98	12.14	-0.06	14.21	101.91	1.92	0.98	51.67			
Mean						0.63	0.99	8.43							0.77	0.97	37.09
STD						0.38	0.01	3.31							0.64	0.02	12.85

olution in FWHM, and the coefficient of determination (R^2) are reported. The average localization error (denoted as L_2), as the mean of the Euclidean norm between Ground Truth (GT) position and estimated position, of 3D-MLEM is (0.63 ± 0.38) mm and of MURA Decoding (0.77 ± 0.64) mm. On one hand, the FWHM grows from 2.76 mm to 12.14 mm for the central source at 20 mm and at 100 mm for 3D-MLEM. On the other hand, for 3D reconstructions from MURA Decoding, the FWHM is generally larger and increases from 8.28 mm to 51.67 mm for the same range.

IV. DISCUSSION

The localization accuracy of the 3D-MLEM algorithm is comparable to that of MURA decoding in terms of the mean error and standard deviation. However, the 3D-MLEM method demonstrates superior axial resolution, with an average FWHM that is 0.23 ± 0.04 smaller than that of MURA Decoding. This factor is even smaller than that observed in the previous study mentioned above, where a factor of 0.3 ± 0.1 was determined [12]. The better axial resolution, however, comes at the costs of a higher computational cost. Depending on the number of iterations and number of slices along the z -axis, the 3D-MLEM took around 4000 times longer (not shown here) to generate the 3D reconstructions [12]. It is conceivable that the difference to the previous study comes from the higher amount of photons that contributed to the simulated detector images. Based on a rough estimation, the detector images from our MC simulation contain on average 4.8 times more photons than the images captured with our experimental IGC that were previously used [14]. This implies a dependency between the number of captured photons and the axial resolution. All in all, this means that the 3D-MLEM algorithm is not substantially better in 3D-localization despite its superior axial resolution. Because MURA Decoding is so fast, it is affordable the finely sample in z -direction to accurately locate a source. However, the 3D-MLEM algorithm and its high computational cost, might benefit from a more carefully

chosen set of in-focus planes. The high precision might impede the choice of optimal planes, because it requires to reconstruct at planes that are already comparably close to the true source position. Therefore, the optimal number of planes and the distances for 3D-localization remain to be investigated for both reconstruction methods, but particularly for the more precise 3D-MLEM algorithm.

An important issue with CAI, that has already been encountered, are ghost sources [20], [21]. Because the basic MURA pattern is arranged to form a 2×2 pattern, a source that is just outside the left-hand side of the FOV, will appear as source on the right-hand side [22]. For the 3D-MLEM algorithm this manifests in not just one visible source, but up to 7 additional ghost sources in a grid-like pattern. This is the reason why the GT z -coordinate was used as initial guess for the ISL method. Further investigations regarding ghost sources might make the 3D-MLEM algorithm more beneficial. A further limitation of the study presented here is its exclusive reliance on a MC simulation. Imperfect detectors with cracks in the substrate and defective pixels are known to impair reconstruction quality [19] and, thus, both the axial resolution and localization accuracy. Nevertheless, the methods employed in this study can provide guidance and serve as a benchmark for further experimental studies.

V. CONCLUSION & OUTLOOK

In conclusion, this study reveals that despite its substantially better axial resolution, the 3D-MLEM algorithm can only provide a marginally more accurate localization than MURA Decoding. Furthermore, we confirmed the better axial resolution of 3D-MLEM. Even though, there might be a more efficient way to use the computationally expensive but precise algorithm to locate a source by carefully choosing the number and distances of the slices to reconstruction the detector images.

The results of the presented study reveal more avenues of future research. First, there is a need to eliminate ghost sources. This question is linked to the 2×2 arrangement of the mask pattern and the question rises on how a large FOV can be maintained while avoiding ghost sources. Second,

SLNB and other radio-guided surgeries require the imaging of not just point-like sources but also extended sources. Other experiments have indicated a deterioration of the localization accuracy with extended sources [23]. With a solution to both issues, intraoperative gamma imaging might be able exploit the advantages of CAI and improve the outcome of radio-guided surgeries.

REFERENCES

- [1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021. DOI: [10.3322/caac.21660](https://doi.org/10.3322/caac.21660).
- [2] J. M. Chang, J. W. T. Leung, L. Moy, S. M. Ha, and W. K. Moon, "Axillary Nodal Evaluation in Breast Cancer: State of the Art," *Radiology*, vol. 295, pp. 500–515, jun 2020. DOI: [10.1148/radiol.2020192534](https://doi.org/10.1148/radiol.2020192534).
- [3] M. Georgiou, G. Loudos, E. Fysikopoulos, E. Lamprou, K. Mikropoulos, A. Shegani, and P. Georgoulas, "λ-Eye: A high-sensitivity γ imaging probe for axillary sentinel lymph node mapping," *Nuclear Medicine Communications*, vol. 37, no. 10, pp. 1001–1009, 2016. DOI: [10.1097/MNM.0000000000000546](https://doi.org/10.1097/MNM.0000000000000546).
- [4] S. Heller and P. Zanzonico, "Nuclear Probes and Intraoperative Gamma Cameras," *Seminars in Nuclear Medicine*, vol. 41, pp. 166–181, may 2011. DOI: [10.1053/j.semnuclmed.2010.12.004](https://doi.org/10.1053/j.semnuclmed.2010.12.004).
- [5] A. L. Farnworth and S. L. Bugby, "Intraoperative Gamma Cameras: A Review of Development in the Last Decade and Future Outlook," *Journal of Imaging*, vol. 9, p. 102, may 2023. DOI: [10.3390/jimaging9050102](https://doi.org/10.3390/jimaging9050102).
- [6] A. Gonzalez-Montoro, C. D. Vera-Donoso, G. Konstantinou, P. Sopena, M. Martinez, J. B. Ortiz, M. Carles, J. M. Benlloch, and A. J. Gonzalez, "Nuclear-medicine probes: Where we are and where we are going," *Medical Physics*, vol. 49, pp. 4372–4390, jul 2022. DOI: [10.1002/mp.15690](https://doi.org/10.1002/mp.15690).
- [7] T. E. Peterson and L. R. Furenlid, "SPECT detectors: The Anger Camera and beyond," *Physics in Medicine and Biology*, vol. 56, no. 17, 2011. DOI: [10.1088/0031-9155/56/17/R01](https://doi.org/10.1088/0031-9155/56/17/R01).
- [8] R. H. Dicke, "Scatter-hole cameras for x-rays and gamma rays," *The astrophysical journal*, vol. 153, p. L101, 1968.
- [9] J. G. Ables, "Fourier Transform Photography: A New Method for X-Ray Astronomy," *Publications of the Astronomical Society of Australia*, vol. 1, pp. 172–173, dec 1968. DOI: [10.1017/s1323358000011292](https://doi.org/10.1017/s1323358000011292).
- [10] T. Meißner, V. Rozhkov, J. Hesser, W. Nahm, and N. Loew, "Quantitative comparison of planar coded aperture imaging reconstruction methods," *Journal of Instrumentation*, vol. 18, p. P01006, jan 2023. DOI: [10.1088/1748-0221/18/01/P01006](https://doi.org/10.1088/1748-0221/18/01/P01006).
- [11] P. Russo, F. Di Lillo, V. Corvino, P. M. Frallicciardi, A. Sarno, and G. Mettivier, "CdTe compact gamma camera for coded aperture imaging in radioguided surgery," *Physica Medica*, vol. 69, no. June 2019, pp. 223–232, 2020. DOI: [10.1016/j.ejmp.2019.12.024](https://doi.org/10.1016/j.ejmp.2019.12.024).
- [12] T. Meißner, L. A. Cerbone, P. Russo, W. Nahm, and J. Hesser, "Assessment of the axial resolution of a compact gamma camera with coded aperture collimator," *EJNMMI Physics*, vol. 11, p. 30, mar 2024. DOI: [10.1186/s40658-024-00631-5](https://doi.org/10.1186/s40658-024-00631-5).
- [13] J. Perl, J. Shin, J. Schumann, B. Faddegon, and H. Paganetti, "TOPAS: An innovative proton Monte Carlo platform for research and clinical applications," *Medical Physics*, vol. 39, no. 11, pp. 6818–6837, 2012. DOI: [10.1118/1.4758060](https://doi.org/10.1118/1.4758060).
- [14] T. Meißner, L. A. Cerbone, P. Russo, W. Nahm, and J. Hesser, "3D-localization of single point-like gamma sources with a coded aperture camera," *Physics in Medicine & Biology*, vol. 69, p. 165004, aug 2024. DOI: [10.1088/1361-6560/ad6370](https://doi.org/10.1088/1361-6560/ad6370).
- [15] R. Nabha, O. Van Hoey, C. Granja, A. Parisi, M. De Saint-Hubert, L. Struelens, C. Oancea, E. Sterpin, V. Zach, J. Stursa, A. Rucinski, J. Gajewski, P. Stasica, and F. Vanhavere, "A novel method to assess the incident angle and the LET of protons using a compact single-layer Timepix detector," *Radiation Physics and Chemistry*, vol. 199, p. 110349, oct 2022. DOI: [10.1016/j.radphyschem.2022.110349](https://doi.org/10.1016/j.radphyschem.2022.110349).
- [16] Z. Mu and Yi-Hwa Liu, "Aperture collimation correction and maximum-likelihood image reconstruction for near-field coded aperture imaging of single photon emission computerized tomography," *IEEE Transactions on Medical Imaging*, vol. 25, pp. 701–711, jun 2006. DOI: [10.1109/TMI.2006.873298](https://doi.org/10.1109/TMI.2006.873298).
- [17] T. Meißner, J. Hesser, and W. Nahm, "Reconstruction methods for coded aperture imaging," *GitHub repository*, 2024. <https://github.com/tomeiss/cai-reconstruction>.
- [18] T. Buzug, "Algebraic and Statistical Reconstruction Methods," in *Computed Tomography*, pp. 201–240, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. DOI: [10.1007/978-3-540-39408-2_6](https://doi.org/10.1007/978-3-540-39408-2_6).
- [19] V. Rozhkov, G. Chelkov, I. Hernández, O. Ivanov, D. Kozhevnikov, A. Leyva, A. Perera, D. Rastorguev, P. Smolyanskiy, L. Torres, and A. Zhemchugov, "Visualization of radiotracers for SPECT imaging using a Timepix detector with a coded aperture," *Journal of Instrumentation*, vol. 15, no. 06, pp. P06028–P06028, 2020. DOI: [10.1088/1748-0221/15/06/p06028](https://doi.org/10.1088/1748-0221/15/06/p06028).
- [20] R. Willingale, M. Sims, and M. Turner, "Advanced deconvolution techniques for coded aperture imaging," *Nuclear Instruments and Methods in Physics Research*, vol. 221, pp. 60–66, mar 1984. DOI: [10.1016/0167-5087\(84\)90180-7](https://doi.org/10.1016/0167-5087(84)90180-7).
- [21] J. Braga, "Coded Aperture Imaging in High-energy Astrophysics," *Publications of the Astronomical Society of the Pacific*, vol. 132, no. 1007, p. 12001, 2020. DOI: [10.1088/1538-3873/ab450a](https://doi.org/10.1088/1538-3873/ab450a).
- [22] Q. Liu, Y. Cheng, X. Tuo, Y. Mu, Y. Xiao, Y. Xiong, and T. Zhu, "Neural network method for localization of radioactive sources within a partially coded field-of-view in coded-aperture imaging," *Applied Radiation and Isotopes*, vol. 170, no. November 2020, p. 109637, 2021. DOI: [10.1016/j.apradiso.2021.109637](https://doi.org/10.1016/j.apradiso.2021.109637).
- [23] I. Kaissas, C. Papadimitropoulos, C. Potiriadis, K. Karafasoulis, D. Loukas, and C. P. Lambropoulos, "Imaging of spatially extended hot spots with coded apertures for intra-operative nuclear medicine applications," *Journal of Instrumentation*, vol. 12, no. 1, pp. 1–9, 2017. DOI: [10.1088/1748-0221/12/01/C01059](https://doi.org/10.1088/1748-0221/12/01/C01059).

Artificial Computational Pathology May Refine and Redefine Salivary Gland Tumor Entities: A Molecular Bioinformatic Perspective

Bacem K. OTHMAN

Department of Pathology Faculty of
Medicine in Pilsen, Charles University,
Czech Republic
bacemkhalele@gmail.com

Ilmo Leivo

Institute of Biomedicine, Pathology
Turku University, Turku, Finland
ilmo.leivo@utu.fi

Alena Skálová

Department of Pathology, Faculty of
Medicine in Pilsen, Charles University,
Czech Republic; Bioptic Laboratory,
Ltd, Pilsen, Czech Republic
skalova@fnplzen.cz

Abstract—This study proposes an artificial neural network-based model for clustering salivary gland carcinomas using molecular profiling. Genetic alterations from next-generation sequencing studies, gene databases, and a collection of 212 molecularly labeled cases were analyzed using Orange's data mining toolkit. The computational model, configured with 64 neurons in one hidden layer and Sigmoid/Softmax activation functions, was validated through 10-fold cross-validation. The results suggested cribriform adenocarcinoma and polymorphous adenocarcinoma are separate entities due to molecular heterogeneity. Intercalated intraductal carcinoma showed distinct molecular alterations from intraductal carcinoma, while genetic mutations occurring secondary to the canonical fusions may indicate high-grade transformation at the molecular level. This perspective has prognostic and therapeutic potentials and may help refine the molecular pathogenetic understanding.

Keywords— salivary gland carcinomas, sequencing, computational

I. INTRODUCTION

Artificial neural network (ANN) algorithms have been applauded in bioinformatics for use in classifying heterogeneous spectra of skin cancers, chronic renal diseases, IgG4-diseases and lymphoepithelial lesions[1], [2]. When integrated with a support vector machine network, conventional neural networks can be used to generate computational models with high accuracy of classification[1]. Herein we propose an ANN-based model for the classification of salivary gland carcinomas (SGCs).

II. METHOD

A. Identifying oncogenes

Genetic alterations occurring in salivary gland carcinomas were retrieved from three sources: (a) information from next-generation sequencing studies undertaken to gain molecular information for classification of SGCs, (b) gene information on SGCs retrieved from *PubMed*, *OMIM* and *Genecards* databases and (c) gene data from the SGC registry of the senior author (AS, $n=212$). All SGCs were annotated at a subordinate level, following the taxonomy of the 2024 WHO Classification of Head and Neck Tumors.

B. Maintaining Specifications

Criteria for eligibility included cases of primary SGCs with NGS data and an immunohistochemical workup.

Sinonasal salivary-type neoplasms were excluded. For example, cases of intraductal carcinoma (superordinate) were treated as four entries according to histological subtype (oncocytic, apocrine, intercalated duct-like, and mixed). IHC marker information was embedded into the annotation sheet.

C. Computational Clustering

Data mining toolkit (Orange 3.37.0, Ljubljana, Slovenia[3]) was employed for the analysis of genetic data, focusing on ANN modeling via Keras, cross-validation for hyperparameter optimization, and K-Means clustering visualization. The process was initiated with data preparation from Gene Expression Omnibus, formatting samples as rows and gene expression as columns, including class labels. The File widget imported the dataset, followed by the Select Columns widget for feature and target class selection. Pre-processing utilized the Impute widget for missing values and the Normalize widget for standardization of gene expression level. Feature selection employed the Rank widget with Information Gain and ANOVA methods, retaining top-ranked genes via the Select Attributes widget. The Neural Network (Keras) widget configured the ANN, experimenting with architectures (64 neurons in one hidden layer), activation functions [Rectified Linear Unit for hidden: $f(x)=\max(0,x)$]. Sigmoid/Softmax for output was calculated, as follows:

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \quad (1)$$

where x_i is the input to neuron i , and the denominator is the sum of exponentials of all the inputs in the layer.

Learning rates (initially 0.001), and optimizers were also used. The Test & Score widget performed 10-fold cross-validation, evaluating accuracy, AUC, precision, recall, and F1 score. The Confusion Matrix, ROC Analysis, and Lift Curve widgets provided visual performance insights. K-Means clustering was applied to selected features, and the Scatter Plot widget to visualize clusters using principal components and the Silhouette Plot widget to assess cluster quality.

III. RESULTS

The weighted number of links and total interactions score values were exported and processed. Furthermore, the ANN model, trained on the selected gene expression data, demonstrated high accuracy in distinguishing between different salivary gland carcinoma entities. However, cases with secondary gene mutations in addition to canonical

fusions showed interesting classification patterns. In the scatter plot visualization, these cases often clustered near the boundary between high-grade and low-grade tumor entities.

The ROC analysis for these cases showed a distinct curve pattern, with a notably higher Area Under the Curve (AUC) compared to cases without secondary mutations. This suggests that the model was able to detect subtle molecular differences associated with secondary mutations. The confusion matrix further supported this observation, showing a higher rate of misclassification of these cases into high-grade categories. The distance on the visualization charts suggested the following:

- Cribriform adenocarcinoma and polymorphous adenocarcinoma of salivary gland are too heterogenous to be clustered under the same entity. The demonstration of *ARID1A-PRKD1*, *KTN1-PRKD1*, *DDX3X-PRKD1*, *STRN3-PRKD1*, *YLPM1-PRKD1* fusions in cribriform adenocarcinoma suggests a different molecular mechanism from polymorphous adenocarcinoma's singular *PRKD1/2/3* mutation
- Intercalated intraductal carcinoma differs from the other variants remarkably. This is supported by the presence of unique molecular alterations such as *PIK3CA* mutations, *NCOA4-RET* fusions in

intraductal carcinoma, and *TRIM27-RET* fusions in oncocyctic intraductal carcinoma, distinguishing it from other salivary gland carcinomas.

- Squamoglandular differentiation in acinic cell carcinoma (AciCC) and adenoid cystic carcinomas (AdCC) warrants special attention and subtyping. AciCC shows a variety of molecular alterations including *NR4A2* and *NR4A3* rearrangements, *SCPP-NR4A3*, *HTN3-MSANTD3*, *SYN2-PPARG*, and *PON3-LCN1* fusions, indicating molecular heterogeneity within this entity. AdCC also demonstrates heterogeneity with *MYBL1-NFIB*, *MYB-NFIB*, and *NOTCH* mutations.
- High grade Myoepithelial carcinoma show specific *LIFR-PLAG1* or *CTNNB1-PLAG1* fusions.
- Gene mutations arising secondary to canonical fusions may represent high-grade transformation at the molecular level. Clustering of secondary mutations (*HRAS*, *PIK3CA*, *TP53*, *KMT2A*) alongside primary canonical fusions may reveal patterns associated with high-grade transformation. *RUNX1* mutation is observed as a secondary mutation in metatypical AdCC. (**Figure 1**).

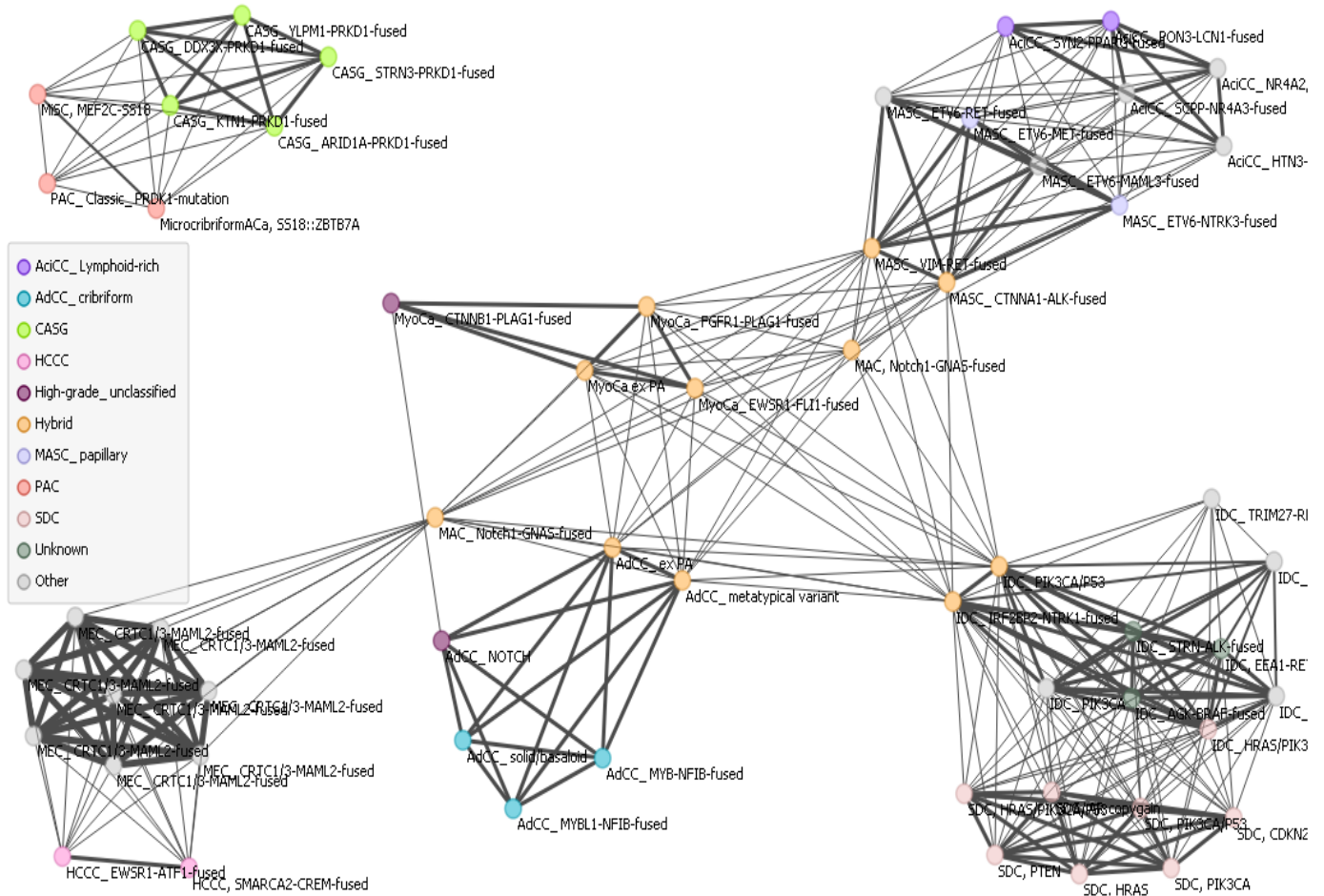


Fig. 1. Network analysis showing clustering of SGCs based on immuno-morpho-molecular profiling

IV. DISCUSSION

SGCs tend to show several canonical fusions although their genomic profiling is dynamically changing[4]. This study employed a computational approach to cluster salivary gland carcinomas using molecular profiling retrieved from medical literature with next-generation sequencing data, gene databases, and molecularly labeled 212 cases. The ANN-based clustering method differed from traditional approaches in using 10-fold cross-validation, evaluating metrics such as accuracy, AUC, precision, recall, and F1 score that can compare molecular heterogeneity mathematically. The study revealed that gene mutations arising secondary to canonical fusions may explain high-grade transformation at the molecular level. Unlike the progression of tumors in other topographies (e.g. transformation of low-grade gliomas to higher-grade tumors being associated with specific mutation patterns where the Cancer Genome Atlas has identified both genetic and epigenetic alterations that contribute to tumor development and progression[5]), SGCs did not show a particular epigenetic pattern.

Although *HRAS*, *PIK3CA*, *TP53*, *KMT2A* mutations are non-specific diagnostically[6], [7], they are still actionable genes therapeutically[8]. Future research should focus on identifying actionable fusion genes for molecular subtypes of SGCs, potentially following the successful model of personalized medicine seen in lung cancer treatment.

The study primarily focuses on canonical molecular fusions, as these are diagnostic criteria for classifying salivary carcinomas within AS' registry, using Archer's sequencing panels and, to a lesser extent, Illumina's TruSight Oncology 500. This approach may introduce bias against new molecular clusters of unknown clinical significance that could be revealed using more comprehensive next-generation sequencing panels. Including more detailed clinical information on survival outcomes would strengthen the clinicopathological understanding of SGCs.

To conclude, studies featuring next-generation sequencing data on SGCs present a promising source for data mining that can enhance the validity of research findings when compared to standard real-time registries. While IT-based pathological knowledge complements the histopathological WHO

classification, it does not supplant traditional tumor classifications; instead, it opens up new avenues for cancer research with valuable molecular pathogenetic insights.

ACKNOWLEDGMENT

B. K. O. thanks Linos Nchena, PhD (Faculty of Applied Informatics, Tomas Bata University in Zlín, Czechia) for the tips on improving the model.

FUNDING

This study was supported by study grant SVV 260652 from the Ministry of Education, Czech Republic, 4EU+, the European Union—Next Generation EU and the Finnish Cancer Society, Finland..

REFERENCES

- [1] U. A. Lyakhova and P. A. Lyakhov, "Systematic review of approaches to detection and classification of skin cancer using artificial intelligence: Development and prospects," *Comput. Biol. Med.*, vol. 178, p. 108742, 2024, doi: 10.1016/j.compbimed.2024.108742.
- [2] J. Carreras and R. Hamoudi, "Artificial Neural Network Analysis of Gene Expression Data Predicted Non-Hodgkin Lymphoma Subtypes with High Accuracy," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 3, 2021, doi: 10.3390/make3030036.
- [3] J. Demšar *et al.*, "Orange: Data mining toolbox in python," *J. Mach. Learn. Res.*, vol. 14, 2013.
- [4] A. Skálová *et al.*, "Molecular pathology in diagnosis and prognostication of head and neck tumors," *Virchows Arch.*, vol. 484, no. 2, pp. 215–231, 2024, doi: 10.1007/s00428-023-03731-2.
- [5] M. M. Ershadi, Z. R. Rise, and S. T. A. Niaki, "A hierarchical machine learning model based on Glioblastoma patients' clinical, biomedical, and image data to analyze their treatment plans," *Comput. Biol. Med.*, vol. 150, 2022, doi: 10.1016/j.compbimed.2022.106159.
- [6] Y. S. Chang *et al.*, "Combined mutational analysis of RAS, BRAF, *PIK3CA*, and TP53 genes in Taiwanese patients with oral squamous cell carcinoma," *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.*, vol. 118, no. 1, 2014, doi: 10.1016/j.oooo.2014.03.016.
- [7] D. J. Zaccarini *et al.*, "Comprehensive Molecular Profiling of Oncocytic Salivary Gland Malignancies," *Appl. Immunohistochem. Mol. Morphol.*, vol. 30, no. 9, 2022, doi: 10.1097/PAI.0000000000001048.
- [8] D. Helbig *et al.*, "Oncogene and therapeutic target analyses in Atypical fibroxanthomas and pleomorphic dermal sarcomas," *Oncotarget*, vol. 7, no. 16, 2016, doi: 10.18632/oncotarget.7845.

WIRELESS INTRAORAL ELECTRONICS INTEGRATED WITH AI FOR DISEASE DIAGNOSIS

Pooja Mali Rai

dept. Department of Biomedical, Surgical and Dental Sciences,

University of Milan

Milan, Italy

poojarai93@gmail.com

Abstract—

Artificial intelligence (AI) and wearable sensors are two essential fields to realize the aim of tailoring the precision medicine treatment for individual patients. As high-quality healthcare and well-being services is significantly increasing. Integration of these two fields has attracted extensive academic and technological attention which enables better acquisition of patient data and improved design of powerful high-dimensional data processing of wearable biosensing networks for monitoring the wearers' health, fitness, and their surroundings. With the development of sensor and operating system hardware technology, the functions of wearable devices have been gradually enriched with more diversified forms and more accurate physiological indicator. The present study reviews the recent developments in AI-assisted wearable biosensing devices in disease diagnostics and fatigue monitoring demonstrating the trend towards personalised medicine with highly efficient, cost-effective, and accurate point-of-care diagnosis. Currently, as the Internet of Things (IoT), big data and big health move from concept to implementation, AI-biosensors with appropriate technical characteristics are facing new opportunities.

Keywords— *Wearable biosensors Artificial intelligence personalized health monitoring, machine learning (ML)*

Introduction

With the growing demand for real-time and continuous health monitoring and disease diagnostics in the healthcare sector, there has been a significant shift towards digital health technology. The rising prevalence of chronic diseases, such as diabetes, heart disease, and cancer, coupled with an aging population, has increased the need for remote and continuous health monitoring. [1–3]. Wearable sensors become the key players. Non/minimally invasive analytical devices have been manufactured for providing point-of-care systems for the continuous and regular monitoring of individuals' physiological as well as biological markers. [4] Clinically important biomarkers can be detected by conventional point-of-care devices, wearable sensors that can utilize easy-to-sample bio-fluids such as saliva, tears, and sweat. [5]

However increasing prevalence of chronic diseases such as cancer, diabetes, heart disease and coupled with an aging population has increased the need for remote and real time health monitoring. This has led to the emergence of artificial intelligence (AI)-based wearable sensors that can collect, analyze, and transmit real-time health data to healthcare providers so that they can make efficient decisions based on patient data. [6-7]. For chronic illness patients wearable devices can lower the number of hospitalizations and help in daily monitoring of effect of diet and lifestyle modifications. Development of AI in the field of health monitoring sensors, provides algorithms and powerful tools for healthcare data processing and analysis which can provide solutions for limitations faced by health care management systems. [8]

Using AI algorithms, the data collected by health care management systems can be processed and analysed. Meanwhile, machine learning algorithms enable mine potential health problems from big data, providing more accurate diagnosis and treatment plans. This can achieve intelligent health monitoring [9]. The wearable devices can be modified to provide personalized healthcare. The large amount of data obtained through these devices can be used to train machine learning algorithms, which can be helpful for decision-making through artificial intelligence.

The concept of a wearable device used in 1960s provided the foundation for today's next-generation wearable sensors and it's not new. Integrating Internet of Things (IoT) technology, AI technology and health care management systems, a close loop with real-time monitoring, data collection, online analysis. It applications include, disease diagnosis and treatment recommendations, monitoring disease as a self-healing route and other precautionary routes. IoT is the sub-category of a computer having sensors, microcontrollers, and transceivers that are interconnected to provide information to aid the user with detailed information.

The goal of this review is to revolutionize health care and to provide insight for researchers to design and develop novel integrated wearable sensors. Hence, aim of the current review is to highlight the state-of-the-art integrated wearable sensors, methods of manufacturing, describing their important components, and to discuss the challenges being faced while designing and scaling up these sensors.

Types of biosensors

On the basis of Design and Utility		On the basis of Design and Utility	Wearables(existing and advanced forms)	
<i>Wrist mounted</i>	<i>Head and face mounted</i>			
Wrist band	Eyeglasses	Self Healing Flexible Wearable Sensors	Smart Garments	Sensors patches
Wrist watches	Cavities	Biocompatible Wearable Sensors	Smart Watches	E-tattoo
	Helmets	Biodegradable Flexible Sensors	Foot and Hand worn	E-skin
				Implant

Components of AI integrated wearable sensors

Knowledge on different components of wearable sensors can help in understanding the functionality of wearable sensors. The significant parts of wearable sensors are sensors for detection of signals, for signal improvement there is a conditioning circuits and there are analog to digital converters to convert analogue signals to digital format, microprocessors or microcontrollers, which process the data, memory, which stores the data, wireless communication modules, which transmit the data, power sources, which supply energy, enclosures, which protect the sensors, and user interfaces, which allow for interaction. [10-11] Different components that make up the integrated wearable sensors.

While designing the sensing platform interactions between a target and a bio-recognition element are the guiding principle. The targets present in the bio-fluid are transported to the sensing surface, where the interactions with the bio-recognition element cause a change in the electrical signal, thereby enabling their detection. The electrochemistry-based transducer is the most reliable and flexible option available for the design of a wearable sensor because of its sensitivity, ease of fabrication, no moving parts, mechanical robustness, and resistance to other conditions [12]. This transducer is the most widely used and trusted option for a biosensor. The electrical changes can be amperometric (change in current), voltammetric (change in potential), or impedimetric (change in impedance or resistance). For a typical biosensing surface, the surface of the electrode is modified with a bio-recognition element, which can be an antibody, aptamer, enzyme, peptide, or molecularly imprinted polymers (MIPs).[13] Using bio-recognition elements is challenging since these need to adapt to the extreme conditions that are generally faced by wearable sensors.

For the fabrication of a sensor, a substrate is first modified with one of the above-mentioned bio-recognition elements. Since most of the wearable sensors developed are electrochemical transduction-based, the substrate is an electrode. [14-16] The successful modification of the substrate is a critical part of biosensor design. The electrodes chosen should allow efficient and easy modification with the bio-recognition element, should not get fouled, be bio-compatible,

and provide excellent electron mobility for the detection of various biomarkers of interest. Various electrical components are required to make a wearable sensor work. The foremost is the battery to power up the sensing surface, integrated circuits, and read-out devices. For a typical sweat-based sensor, the induction of sweat through iontophoresis also requires power. The other important component is the integrated circuits (ICs), which manage the power delivery, potentiostat, data acquisition, and transmission.[17]

AI-assisted wearable biosensors

In the development of wearable biosensing devices, the functionalities of the devices have expanded to obtain physiological information from users, communicate wirelessly, process and store data, and provide an interactive user interface. AI-assisted wearable biosensors are usually composed of sensing modules, wireless communication components, processors, memory devices, displays, wireless charging, energy harvesting components, and power supplies [18]. Biosensors act as data collection units, which collect biochemical or biophysical data from body fluids and convert them into signals that can be recognised by data collecting and processing devices. Wearable biosensors can directly collect biofluids on the body surface to detect health-related biomarker levels [19].

The essential components of a typical biosensor include a bioreceptor (such as antibody, nucleic acid or glucose oxidase) and a transducer that converts physiological information into optical, electrochemical or mechanical signal. Based on different biofluids, biosensors could be integrated with various wearable platforms such as wristbands, contact lenses, and electronic skin [20]. Wireless communication devices then transfer the data collected by biosensors to personal smart readout devices processing terminals [21]. Current communication technologies potentially applied in wearable biosensors are Bluetooth, NFC, and the 5G cellular network. The raw sensing data is eventually processed and stored in local devices or cloud servers, where machine learning (ML) algorithms can be applied to assist in diagnosis [22]. The integration of the abovementioned components improves the accessibility of AI-assisted wearable biosensors providing a compelling alternative to invasive blood-based diagnosis.

AI for biosensing

The critical point of using AI-assisted wearable biosensors is to help users understand and interpret the data collected, which involves analysing data points and drawing conclusions correctly. Especially when continuous and multiplexed sensing becomes a common trend for wearable biosensing technologies, the dimension of data is higher and becomes more complex to extract useful information. Whereas ML algorithms can process a large amount of high-dimensional streaming biosensing data effectively, implying their great potential to be employed to translate raw sensing data into user-friendly information. In addition, ML algorithms can process biosensing data with noise or low resolution.[23]

Biosensing data pre-processing

Depending on the types of sensors, the raw sensing data are present in different forms, such as digital data sets from electrochemical biosensors and image data sets from optical

sensors. An ML model can be designed based on data sets to process the data effectively. In addition, the raw sensing data often requires different pre-processing methods to enhance the performance of ML models and ensure timely warning provided after the detection of abnormalities. The accuracy and computational time of AI models can heavily depend on data pre-processing. The accuracy of the early detection of skin cancer based on a camera was often affected by image enhancement, image restoration, and hair removal to improve the quality of images and remove the noise [24]. The pre-processed data was fed into a CNN for further processing, significantly enhancing sensitivities to metabolites or toxins. Moreover, data pre-processing has also been applied to wearable devices for posture-reminding system [25]. The system used Kalman filtering to denoise the data collected by sensors and improve the accuracy of posture classification to 99.58%. Additionally, a wearable wireless integrated interface was designed and classify gestures via surface electromyogram signals [26]. The signals were pre-processed using a pseudo-wavelet pre-processor to remove the noise caused by motion during the usage. The denoising data represents higher signal quality and therefore, less processing time will be needed.

Synthetic biosensing data

Synthetic data can mimic real-world properties and relationships by producing artificial data from scratch or using advanced data manipulation techniques to generate diverse training samples. Although the reliability of applying synthetic data in the medical field still needs further testing, applying synthetic data can deal with the missing pieces of information during data collection, supplement the data for subpopulations and rare diseases and improve the transparency and reproducibility of the model and minimise biases of using real patient data [27]. The Cycle Generative Adversarial Networks (CycleGAN) model was implemented into computerised tomography (CT) image segmentation task. The model significantly improved the segmentation performance in non-contrast CT test images. The application of the CycleGAN model in CT image segmentation is believed to reduce the manual segmentation effort and cost. The major challenge of using synthetic data in medical applications is keeping the synthetic data's representativeness and linking it to clinical/scientific evidence [28].

ML algorithms for biosensing data analysis

ML algorithms can learn from the existing biosensing data to analyse users' health status. ML can effectively process high dimensional biosensing data generated by wearable multiplexed sensors and find hidden patterns and relations between the data points. The powerful pattern identification capability can help users interpret and understand the biosensing data collected. In recent years, both non-neural and neural algorithms have been used in biosensors to process biomedical data to classify healthy and unhealthy users and quantify the biomarker levels. In recent years, there have been several studies of applying non-neural algorithms in the biosensing field, such as linear discriminant analysis (LDA), support vector machines (SVMs), naive Bayes (NB), k-nearest neighbours (kNN), decision tree (DT) and random forest (RF). Linear discriminant analysis (LDA) is a supervised ML algorithm LDA assumes linear decision boundaries and classifies the input features by probability

estimation. LDA can be applied to classify different biomarkers. [29]

Since there are no general rules for selecting a proper ML algorithm for biosensing data, recent studies often compare several algorithms with their performance. An eye-movement-controlled wheelchair system was developed based on Wavelet Transform-Support Vector Machine (WT-SVM) algorithm. The proposed prototype integrated the flexible biosensors in conductive hydrogel and flexible PDMS substrate to ensure comfortable usage by wheelchair users [30].

Wireless communications

Data collected by biosensors are transmitted to smartphones or other data processing platforms. In the real-world environment, wireless data transfer is often required to be integrated with wearable devices. However, due to the limitation of the power supply of wearables, low-energy wireless communication technology is preferable to be integrated with wearable devices. With power-saving wireless data transfer, streaming biosensing data could be sent to data storage and processing facilities. In addition, by applying wireless communication technologies, AI-assisted wearable biosensors can enhance flexibility and add new devices to build a wearable AI biosensor network since the fusion of sensing data from different biosensors can improve applications [31].

Recently, researchers began to use mobile processing units to process biosensing data collected by wearable biosensors. A smartphone-based surface plasmon resonance (SPR) biosensor has been developed to utilise a smartphone camera for readout and mobile processing unit (Dual Core ARM Mali-400 MP Processor) for image processing. And the system had comparable performance to the commercial SPR instrument with the advantages of low cost, small size and lightweight. Furthermore, a smartphone-based wearable electrochemical biosensor has been developed to monitor metabolite levels. Biosensing data collected by the wearable sweat sensing patch were transferred to smartphones via Bluetooth and processed using CPUs in smartphones. The smartphone-based biosensing system can be easily modified to adapt to a broad range of biomarkers. [32-33]

IoT in healthcare

IoT has efficiently revolutionized the world and created an influence on both living and working lifestyles. The most important IoT field is medical and health care, which is poised to alter the healthcare industry over the next decade because of its vast potential for various applications, ranging from hospital-centric to patient-centric [34]. Patients are getting more involved in their treatment due to the IoT technology, which would allow them to reach their doctors, organize appointments, and access their health records via a portal. The health workers and individuals could also monitor their health from home in real-time using the home monitoring systems, providing benefits for elderly people and patients with long-term diseases (Wickramasinghe and Bodendorf, 2020). The IoT in healthcare could improve patient care due to various smart devices like ultrasound, mobile X-ray machines, smart beds, glucose meters, blood-pressure analyzers, and thermometers units (Erin Lorelle Cook, 2020)

Recent advances in AI-assisted wearable biosensors

AI-assisted wearable biosensors coupled with smartphone-based sensing systems are efficient platforms for ongoing health and fitness monitoring. Researchers have started integrating energy harvesting techniques and wireless communication into AI-assisted wearable biosensors. The applications of the AI-assisted wearable biosensors provide a chance to develop a more personalised healthcare and telemedicine system.

Fitness and health status monitoring

A differential cardiopulmonary monitoring system was developed to monitor the status of COVID-19 patients, workers and athletes (Jeong et al., 2021). The device has a pair of inertial measurement units at opposite ends to measure the difference in the movement of the skin due to respiratory and cardiac activities. A tablet can provide a user interface, obtain the data wirelessly via Bluetooth, and transmit the sensing data to a cloud platform via the cellular data network. The data processing was performed on the cloud platform to extract physiological information such as heart rate, respiration, cough, body temperature, and activities.

Clinical wearable biosensor networks

The basic architecture of WAIBN consists of AI-assisted wearable biosensors, local personal devices, and cloud servers, where wearable biosensors can collect pathological and physiological information, which are uploaded via various wireless communication technologies, processed, and analysed by AI algorithms embedded in cloud servers, and eventually interactively displayed for non-medical background users in a developed user interface on personal devices. These early deployments of IoT technology imply a great potential for implementation in the medical field to improve the healthcare service qualities and the cost of healthcare service. However, due to the special nature of healthcare-related applications, WAIBN needs to have higher energy efficiency, longer lifetime, lower delay, and higher reliability. The final fully deployed medical-related WAIBN will provide users with accurate, real-time, continuous, and non-invasive monitoring of health-related biomarkers, interpretable summaries and visualisations of user information and timely warning of anomalies detected.

Challenges and Considerations

There is no doubt that integrating AI and machine learning into wearable medical devices comes with a slew of benefits. But this type of application is not without its challenges.

The World Health Organization (WHO) recently vocalized its concerns with the sharp increase of AI tool utilization in the healthcare field. They went as far as to say, “Precipitous adoption of untested systems could lead to errors by health-care workers, cause harm to patients, erode trust in AI, and thereby undermine the potential long-term benefits and uses of such technologies around the world.” Indeed, as with any advancements in technology, it is imperative that MedTech companies understand the challenges and considerations associated with AI use in wearables, including

privacy and security concerns, ethical implications, regulatory compliance, and others.

AI-driven wearable devices also raise ethical considerations, such as the responsible use of data, transparency of algorithms, and potential biases in decision-making. It is essential to ensure that AI models are developed and trained using diverse and representative datasets to avoid perpetuating biases or discrimination.

Conclusion

This article reviews the applications of AI in wearable healthcare biosensors and the feasibility of developing smartphone-based biosensing systems. The AI-assisted wearable biosensors can accurately and continuously monitor various biomarkers in biofluids for disease diagnosis and early intervention [36]. The incorporation of proper ML algorithms facilitates the streaming of high-dimensional big data processing to extract valuable health information from biosensing data. Since most commercially available smartphones have extra hardware components such as cameras, audio ports, NFC, 5G cellular networks and Bluetooth, the integration of smartphones with biosensing systems further enhances the accessibility of wearable biosensors and provides an opportunity for personalised healthcare. Recent studies have demonstrated that smartphones can receive, process and store biosensing data, display the results, wirelessly charge the biosensors and upload health data to the cloud server. The development of smartphone-based wearable biosensing has also triggered the applications of wearable AI biosensor networks in the medical field for more accurate remote diagnosis [37].

References

- [1] Ha, M., Lim, S., & Ko, H. (2018). Wearable and flexible sensors for user-interactive health-monitoring devices. *Journal of Materials Chemistry B*, 6(26), 4043–4064. <https://doi.org/10.1039/C8TB01192K>
- [2] Homayounfar, S.Z., & Andrew, T.L. (2020). Wearable sensors for monitoring human motion: A review on mechanisms, materials, and challenges. *SLAS Technology*, 25(1), 9–24. <https://doi.org/10.1177/2472630319873057>
- [3] Huifeng, W., Kadry, S.N., & Raj, E.D. (2020). Continuous health monitoring of sportsperson using IoT devices based wearable technology. *Computer Communications*, 160, 588–595. <https://doi.org/10.1016/j.comcom.2020.05.062>
- [4] Cheng, S., et al. (2021). Recent progress in intelligent wearable sensors for health monitoring and wound healing based on biofluids. *Frontiers in Bioengineering and Biotechnology*, 9, 765987. <https://doi.org/10.3389/fbioe.2021.765987>
- [5] Ates, et al. (2022). End-to-end design of wearable sensors. *Nature Reviews Materials*, 7(11), 887–907. <https://doi.org/10.1038/s41578-022-00424-3>
- [6] Baig, M.M., GholamHosseini, H., Moqem, A.A., Mirza, F., & Lindén, M. (2017). A systematic review of wearable patient monitoring systems—Current challenges and opportunities for clinical adoption. *Journal of Medical Systems*, 41(7), 115. <https://doi.org/10.1007/s10916-017-0760-1>
- [7] Sempionatto, J.R., Lasalde-Ramírez, J.A., Mahato, K., Wang, J., & Gao, W. (2022). Wearable chemical sensors for biomarker discovery in the omics era. *Nature Reviews Chemistry*, 6(11), 899–915. <https://doi.org/10.1038/s41570-022-00399-0>
- [8] Zhang, Z., Wang, L., & Lee, C. (2023). Recent advances in artificial intelligence sensors. *Advanced Sensors Research*, 2200072. <https://doi.org/10.1002/adrs.202200072>

- [9] Gao, S., He, T., Zhang, Z., Ao, H., Jiang, H., & Lee, C. (2021). A motion capturing and energy harvesting hybridized lower-limb system for rehabilitation and sports applications. *Advanced Science (Weinheim)*, 8(14), e2101834. <https://doi.org/10.1002/advs.202101834>
- [10] Cortisol Detection in Sweat. (2022). *Biosensors*, 12(10), 854. <https://doi.org/10.3390/bios12100854>
- [11] Yao, S., et al. (2023). A two-in-one annealing enables dopant free block copolymer based organic solar cells with over 16% efficiency. *Chinese Journal of Chemistry*, 41(6), 672–678. <https://doi.org/10.1002/cjoc.202200750>
- [12] Teymourian, H., et al. (2020). Wearable electrochemical sensors for the monitoring and screening of drugs. *ACS Sensors*, 5(9), 2679–2700. <https://doi.org/10.1021/acssensors.0c01258>
- [13] Teymourian, H., et al. (2020). Wearable electrochemical sensors for the monitoring and screening of drugs. *ACS Sensors*, 5(9), 2679–2700. <https://doi.org/10.1021/acssensors.0c01258>
- [14] Pérez, D., & Orozco, J. (2022). Wearable electrochemical biosensors to measure biomarkers with complex blood-to-sweat partition such as proteins and hormones. *Microchimica Acta*, 189(3), 127. <https://doi.org/10.1007/s00604-022-05122-4>
- [15] Min, J., et al. (2021). Wearable electrochemical biosensors in North America. *Biosensors and Bioelectronics*, 172, 112750. <https://doi.org/10.1016/j.bios.2020.112750>
- [16] Garg, M., et al. (2021). White graphene quantum dots as electrochemical sensing platform for ferritin. *Faraday Discussions*, 227(0), 204–212. <https://doi.org/10.1039/D0FD00126B>
- [17] Ha, S., et al. (2014). Low-power integrated circuit design for wearable biopotential sensing. In E. Sazonov & M.R. Neuman (Eds.), *Wearable Sensors* (pp. 323–352). Academic Press. <https://doi.org/10.1016/B978-0-12-418662-0.00014-7>
- [18] Jeong, H., et al. (2021). Differential cardiopulmonary monitoring system for artifact-canceled physiological tracking of athletes, workers, and COVID-19 patients. *Science Advances*, 7(14), eabf8412. <https://doi.org/10.1126/sciadv.abf8412>
- [19] Kim, J., Campbell, A.S., de Avila, B.E., & Wang, J. (2019). Wearable biosensors for healthcare monitoring. *Nature Biotechnology*, 37(4), 389–406. <https://doi.org/10.1038/s41587-019-0045-y>
- [20] Chae, H.Y., Lee, K., Jang, J., Park, K., & Kim, J.J. (2019). A wearable sEMG pattern-recognition integrated interface embedding analog pseudo-wavelet preprocessing. *IEEE Access*, 7, 151320–151328. <https://doi.org/10.1109/ACCESS.2019.2947683>
- [21] Bandodkar, A.J., & Wang, J. (2014). Non-invasive wearable electrochemical sensors: A review. *Trends in Biotechnology*, 32(7), 363–371. <https://doi.org/10.1016/j.tibtech.2014.04.005>
- [22] Jin, X., Liu, C., Xu, T., Su, L., & Zhang, X. (2020). Artificial intelligence biosensors: Challenges and prospects. *Biosensors and Bioelectronics*, 165, 112412. <https://doi.org/10.1016/j.bios.2020.112412>
- [23] Zhang, K., et al. (2021). Machine learning-reinforced noninvasive biosensors for healthcare. *Advanced Healthcare Materials*, 10(11), e2100734. <https://doi.org/10.1002/adhm.202100734>
- [24] Hoshyar, A.N., Al-Jumaily, A., & Hoshyar, A.N. (2014). The beneficial techniques in preprocessing step of skin cancer detection system comparing. *Procedia Computer Science*, 42, 25–31. <https://doi.org/10.1016/j.procs.2014.11.036>
- [25] Guo, J., et al. (2021). 5G-enabled ultra-sensitive fluorescence sensor for proactive prognosis of COVID-19. *Biosensors and Bioelectronics*, 181, 113160. <https://doi.org/10.1016/j.bios.2021.113160>
- [26] Chae, H.Y., Lee, K., Jang, J., Park, K., & Kim, J.J. (2019). A wearable sEMG pattern-recognition integrated interface embedding analog pseudo-wavelet preprocessing. *IEEE Access*, 7, 151320–151328. <https://doi.org/10.1109/ACCESS.2019.2947683>
- [27] Nikolenko, S.I. (2021). *Synthetic Data for Deep Learning* (Vol. 3). Springer. <https://doi.org/10.1007/978-3-030-75178-2>
- [28] Tucker, A., Wang, Z., Rotalinti, Y., & Myles, P. (2020). Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digital Medicine*, 3, 147. <https://doi.org/10.1038/s41746-020-00353-9>
- [29] Pandit, S., Banerjee, T., Srivastava, I., Nie, S., & Pan, D. (2019). Machine learning-assisted array-based biomolecular sensing using surface-functionalized carbon dots. *ACS Sensors*, 4(11), 2730–2737. <https://doi.org/10.1021/acssensors.9b01235>
- [30] Wang, X., Xiao, Y., Deng, F., Chen, Y., & Zhang, H. (2021). Eye-movement-controlled wheelchair based on flexible hydrogel biosensor and WT-SVM. *Biosensors*, 11(7), 198. <https://doi.org/10.3390/bios11070198>
- [31] Cui, F., Yue, Y., Zhang, Y., Zhang, Z., & Zhou, H.S. (2020). Advancing biosensors with machine learning. *ACS Sensors*, 5(11), 3346–3364. <https://doi.org/10.1021/acssensors.0c01611>
- [32] Liu, Y., et al. (2015). Surface plasmon resonance biosensor based on smartphone platforms. *Scientific Reports*, 5, 12864. <https://doi.org/10.1038/srep12864>
- [33] Wang, M., et al. (2022). A wearable electrochemical biosensor for the monitoring of metabolites and nutrients. *Nature Biomedical Engineering*. <https://doi.org/10.1038/s41551-022-00895-1>
- [34] Surantha, N., Atmaja, P., & Wicaksono, M.D. (2021). A review of wearable internet-of-things devices for healthcare. *Procedia Computer Science*, 179, 936–943. <https://doi.org/10.1016/j.procs.2021.12.083>
- [35] Nordrum, A. (2018). Italy launches new IoT network [News]. *IEEE Spectrum*, 55(9), 9–10. <https://doi.org/10.1109/MSPEC.2018.8510794>
- [36] van de Goor, R., van Hooren, M., Dingemans, A.M., Kremer, B., & Kross, K. (2018). Training and validating a portable electronic nose for lung cancer screening. *Journal of Thoracic Oncology*, 13(5), 676–681. <https://doi.org/10.1016/j.jtho.2018.01.017>
- [37] Pathinarupothi, R.K., Durga, P., & Rangan, E.S. (2019). IoT-based smart edge for global health: Remote monitoring with severity detection and alerts transmission. *IEEE Internet of Things Journal*, 6(2), 2449–2462. <https://doi.org/10.1109/JIOT.2019.2892073>

Treatment Group Assignment using Data from PerPain RCT

^{1st} Andrei Sirazitdinov¹

^{2nd} Jonas Tesarz²

^{3rd} Jürgen Hesser^{1,3,4,5}

andrei.sirazitdinov@medma.uni-heidelberg.de <https://orcid.org/0000-0003-3268-7778> <https://orcid.org/0000-0002-4001-1164>

Abstract—Randomized Controlled Trials (RCTs) are considered the gold standard for estimating causal effects. They come in several forms. In one of the approaches, despite the general randomness, treatment assignment for some patients can be done using a specific strategy. The goal of this paper is to test whether it would be possible to use causal inference methods to modify the assignment strategy so that it is based on the individual treatment effects estimated using causal inference methods, and thus potentially achieve better treatment outcomes. First, we modify several existing causal inference methods to work with uncertainty information. Then, we evaluate the performance of the methods on real-world RCT data. Finally, we assign new potential treatments based on the predicted outcomes. The results show that the proposed treatment assignment method predicts significant benefits for patients when using the proposed strategy compared to initial treatment assignment. These results can be used to design future studies to verify the findings.

Index Terms—causal inference, individual treatment effect estimation, graph neural networks

I. INTRODUCTION

Randomized controlled trials (RCTs) are widely considered state-of-the-art for establishing causal relationships between treatments and outcomes because of their rigorous design and ability to minimize bias through randomization. Due to their high cost and logistical requirements, RCTs typically involve fewer participants than observational studies. This limitation can lead to unbalanced data, which complicates analysis [1]. In addition, nonlinear relationships between patient characteristics and outcomes may exist in RCT data, requiring sophisticated analytical approaches [2].

RCTs take a variety of forms. In the most common form, treatment and placebo are randomly assigned. In other designs, such as in the PerPain project [3], patients with chronic pain are first randomized into groups receiving personalized

and non-personalized treatment, and then divided into three treatment groups based on patient characteristics. Participants receive either the treatment assigned to them by the clustering algorithm or a randomly selected alternative treatment. The main problem with such an approach is that the clustering algorithm used for treatment assignment may not be perfect, leading to insufficient treatment outcomes.

We propose to apply existing causal inference [4] methods for individual treatment effect (ITE) estimation to data from the PerPain RCT. Our hypothesis is that this approach will improve the estimation of potential outcomes, which are theoretical outcomes of all available treatments for the same individual. This methodology is based on the potential outcomes framework [5], which includes the concepts of actual, potential, and counterfactual outcomes.

In this paper, we present an algorithm designed to predict individual outcomes for each patient in the dataset coming from PerPain RCT studies and assign new treatments based on the best treatment effect. We first discuss the related work followed by a detailed description of our proposed method. Finally, we evaluate the effectiveness of our approach using data from the PerPain RCT. By using causal inference methods, we aim to contribute to more accurate and personalized treatment assignment strategies.

II. RELATED WORK

Numerous efforts have been made to estimate ITE. Among the most promising approaches are representation-based methods [6], and meta-learners [7]. Representation-based methods such as TARnet [8] transform the initial data set into hidden representations and then predict potential outcomes by running these representations through treatment-specific branches. Another approach within this category uses graph neural networks to incorporate information about connectivity between features [9] into the model. According to the authors, this can be effective when the training set size is small and the feature connectivity information is accurate. One approach from meta-learners trains a separate classifier for each treatment. This method is called T-Learner.

A major limitation of the discussed methods is that they do not account for uncertainties, since they only predict a simple mean of the outcome. Therefore, in this paper we extend the proposed methods to be able to generate the first and second moments of the outcomes distribution and thus account for the prediction uncertainty.

¹Mannheim Institute for Intelligent Systems in Medicine, 68167 Mannheim, Germany. ²Department of General Internal Medicine and Psychosomatics, Heidelberg University, Heidelberg, Germany. ³Interdisciplinary Center for Scientific Computing (IWR), 69120 Heidelberg, Germany. ⁴Heidelberg Institute for Theoretical Studies (HITS), 69120 Heidelberg, Germany. ⁵Medical School, Heidelberg University, 69117 Heidelberg, Germany. ⁶Central Institute for Computer Engineering (ZITI), 69120 Heidelberg, Germany. ⁷CZS Heidelberg Center for Model-Based AI, 69120 Heidelberg, Germany. The authors gratefully acknowledge the data storage service SDS@hd, supported by the Ministry of Science, Research, and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grants INST 35/1314-1 FUGG and INST 35/1503-1 FUGG. This study was funded by the Federal Ministry of Education and Research through projects LeMeDaRT (01ZZ2105A), PerPain (01EC1904B), Carl Zeiss Foundation Heidelberg Center for Model-based AI (P2021-02-001), BrainMEP (BW1_1276/04), and Q-HCC (01 KD 2214).

In a related study [10], the authors used image-based data from RCTs to assign new treatments based on the lowest uncertainty. They showed that personalized recommendations could effectively identify patients who would benefit most from treatment. In this paper, we present a method capable of making predictions with uncertainty using tabular data. We evaluate its performance on a real clinical dataset from the PerPain project [3]. After training the models, we determine the optimal treatment assignment based on the highest treatment effect, which is the difference between the pre-treatment and post-treatment outcomes. The contributions of this paper are as follows.

- 1) Extending existing methods to incorporate prediction uncertainties.
- 2) Applying and evaluating these methods on a real RCT dataset, leading to a novel potential treatment assignment strategy based on the best treatment effect for individual patients.

III. METHODS

In this paper we take the data from PerPain RCT. At the beginning of the study, patients were screened and completed baseline questionnaires. Then, after randomization into personalized and non-personalized treatment groups, patients were either assigned with a clustering algorithm to the appropriate treatment group or received a treatment opposite to their assigned treatment. At the end of treatment (EoT), patients completed the same questionnaires again. After each assessment, individual questionnaire scores or subscores were calculated, which involved summing or averaging responses to pre-defined individual questions. The difference between the end of treatment scores (Y_{EoT}) and the baseline scores (Y_{base}) can be attributed to the treatment effect.

A. Problem Formulation

Let us denote the vector of input features $X \in \mathbb{R}^M$ and the factual outcome $Y \in \mathbb{R}$ such that $Y = Y_{EoT}$. We formalize treatment allocation $T \in \{0, 1, 2\}$, reflecting the three possible treatments which are EMDR, EMI, and PERT. Using the potential outcomes framework [5] and a dataset $D = (x^i, y^i, t^i)_{i=0}^N$, our goal is to estimate the factual (for the treatment received) and potential (for the treatment not received) outcomes $\tilde{y}_0$, $\tilde{y}_1$, and $\tilde{y}_2$ for subjects assigned to treatments $t = 0$, $t = 1$, and $t = 2$, respectively. Using the estimated outcomes we are going to assign a new treatment $\tilde{t}$ based on the largest difference between the predicted outcome $\tilde{y}_t$ and the value at the baseline Y_{base} . Since the data are from an RCT, we can estimate the expected values of the potential outcomes $\tilde{\mu}_t = \mathbb{E}[\tilde{y}_t|x]$ using standard machine learning techniques and use the Mean Squared Error (MSE) defined by

$$\epsilon_{\text{error}} = \frac{1}{3} \sum_{t=1}^3 \frac{1}{N_t} (\tilde{\mu}_t - y_t)^2, \quad (1)$$

where y_t is the factual outcome of treatment t and N_t is the number of patients treated with treatment t to measure

the prediction quality. We additionally use a two-sample Kolmogorov-Smirnov test [11] to ensure that predicted outcomes come from the same distribution as the factual outcomes and that the new treatment assignment strategy is different from the original assignment by clustering.

B. Prediction Algorithms

Although we use data from RCTs, it is important to recognize that no RCT is without limitations. These include the presence of differences between treatment groups. In addition, standard methods such as linear regression estimate the treatment effect for the entire population, whereas our focus is on individual treatment assignment. Therefore, to estimate the potential outcomes, we use causal inference methods, namely TARnet of Shalit *et al.* [6], meta-learner T-Learner [7], and GNN-TARnet [9]. T-Learner is a model that takes covariates X as input and passes them directly through treatment specific fully connected layers. This can lead to higher bias if the treatment groups are unbalanced. Unlike T-Learner, TARnet first passes the covariates through the fully connected layers and then sends them to treatment-specific branches. This makes the covariates look like they come from the same distribution. In contrast to TARnet, GNN-TARnet assumes connectivity within features and uses multiple graph neural network (GNN) [12] layers with skip connections, while having the same outcome prediction branches. The algorithm takes as input a feature vector X and an adjacency matrix of the directed graph G describing the causal relationships between features.

C. Uncertainty Computation

Similar to the work of Joshua Durso-Finley *et al.* [10], we incorporate uncertainties into considered models. For that, we modify them to predict the first and second moments of the outcomes distribution. The Kolmogorov-Smirnov normality test showed that the factual outcomes come from a normal distribution. Thus, we estimate $\tilde{y}_t(x)$ as $\mathcal{N}(\tilde{\mu}_t(x), \tilde{\sigma}_t^2(x))$, where $\tilde{\sigma}_t(x)$ is a standard deviation of treatment t . To achieve this, we modify the output layers of the network to output $\tilde{\mu}_t(x)$ and $\tilde{\sigma}_t(x)$. The loss function for the used methods maximizes the log likelihood that the ground truth value comes from the predicted normal distribution. We additionally add the weighted with weight α MSE term ϵ_{error} like it was done in [13] as we noticed that it can improve the results.

D. Causal Graph Generation

GNN-TARnet requires a deterministic causal graph for best performance. Data from PerPain RCT is a compilation of multiple questionnaires that collect information about different aspects of the health of the patient. Each questionnaire has a final score. We propose to create a graph G based on the answers that make up the final score of the questionnaires. First, we concatenate the input features corresponding to the answers to baseline questionnaires with the baseline outcome scores of the same questionnaires. Baseline outcomes are

manually computed as summation or averaging of the questionnaire responses which can be error-prone. In order to avoid using manually computed outcomes, we replace the values of the baseline scores with zeros. Our assumption is that GNN-TARnet will be able to impute the values as it was shown in [9]. After that we create an adjacency matrix and store its elements as an array of tuples. The first element of the tuple is the index of an input feature and the second element is the index of the corresponding to it baseline score. The tuple corresponds to parent-child relationship in a causal graph.

IV. EXPERIMENTS AND RESULTS

A. Dataset

The dataset comes from the PerPain [3] randomized controlled trial. The aim of the RCT was to test the effectiveness of a personalized algorithm capable of assigning patients to therapies called *EMDR*, *EMI* and *PERT* based on their individual characteristics. As outcome variable we use the EoT score of the *Pain Severity* from MPI-D questionnaire [14], which describes the level of pain on a scale from zero to six. Higher scores indicate greater pain.

The input vector for GNN-TARnet, TARnet and T-Learner consists of the answers to the questions that make up the final scores. We propose to perform variable selection at the score level, i.e., we select all features that define the questionnaire score if the score influences the outcome. To find out which scores influenced the outcome most, we used the Pearson correlation coefficient [15]. After selecting the scores, we took unique responses to the questionnaires that contained the scores. The final dataset contains 87 patients with 21 selected features. Using different random seeds, we split the dataset into training and validation sets 200 times. We present the average MSE for the training and validation data splits. After combining the train and validation set into one set we train the best performing model on it and then evaluate our treatment assignment strategy on the previously unseen test set. In total 66 patients are used in the train and validation splits of which 21 belong to the validation set and 45 to the train set. The unseen test set consists of 21 patient.

B. Implementation Details

We train all our models with Stochastic Gradient Descent (SGD) [16]. All model parameters were found using the Random Search hyperparameters optimization algorithm from Keras tuner library [17], with a 0.2 validation split. For the computations, we used a system with the following specifications GeForce RTX 2060 SUPER, AMD Ryzen 7 3800X 8-core processor, and 32 GB of RAM.

C. Results

The results of the model performance on the validation and training sets can be seen in table I. It can be seen that GNN-TARnet showed the smallest error on the validation set. This difference is statistically significant with p-values equal to 0.01 for the tests between GNN-TARnet versus TARnet and

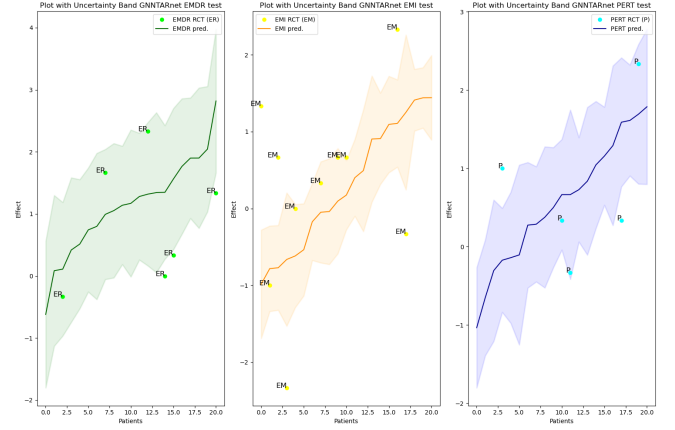


Fig. 1. Sorted effects for each treatment with confidence intervals on the test set. Higher values mean less pain after the treatment.

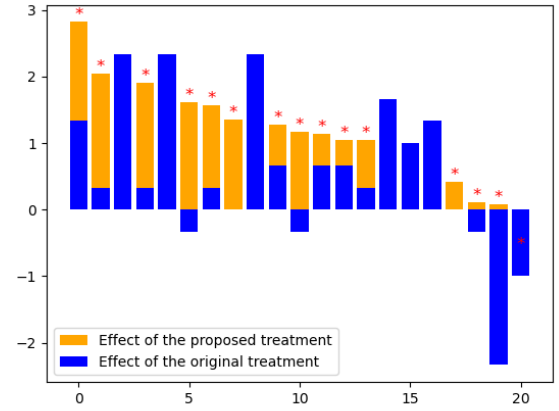


Fig. 2. Effect of the proposed treatment assignment strategy versus the effect of original treatments. Higher values mean less pain after the treatment. Values for which the estimated improvement exceeds 30% are marked with *.

T-Learner. The difference between the results of T-Learner and TARnet was not significant.

The results of treatment assignment using GNN-TARnet, the best performing on the validation set model, on the test set are presented in Figure 1. We can see that actual effects lie within or near the one σ area of the predicted effects. On the test set, we can also see that the model predicts EMDR to have better treatment effect than the other methods for most of the participants. Using two-sample Kolmogorov-Smirnov test we got that the predicted factual outcomes are not significantly different from the real outcomes with a p-value = 0.98. An average treatment effect for the original assignment was equal to 0.53 ± 1.11 . If we assign treatment using the proposed strategy, i.e., based on the highest effect, we get the estimated mean effect value equal to 1.17 ± 0.74 which is significantly better than the original assignment with a p-value = 0.01. The results of such assignment can be seen in Figure 2.

TABLE I
PERFORMANCE OF THE MODELS ON THE TRAIN AND VALIDATION SETS,
WHERE * INDICATES SIGNIFICANTLY BETTER RESULTS.

	ϵ_{train} error	ϵ_{val} error
T-Learner	0.99 ± 0.26	2.20 ± 0.28
TARnet	0.57 ± 0.03	2.25 ± 0.09
GNN-TARnet	0.68 ± 0.02	$2.09 \pm 0.07^*$

D. Discussion

In general, we can say that our algorithm recommends EMDR for most patients. When the other treatment outcomes are close to the predictions for EMDR, the differences are not significant. From the results, we can say that assigning EMDR to all patients seems to be a viable strategy, which needs to be verified by further studies. We believe that the new treatment assignment strategy using ITE estimation methods can potentially be applied to all types of data coming from RCTs with a similar design as the PerPain RCT. Applying this strategy to existing datasets can help determine if the original treatment assignment algorithm worked well or if there is a need to use an alternative approach. Our approach may not be applicable to pure RCTs without treatment assignment personalization.

The main drawback of the proposed method is that the new treatment assignment strategy needs to be verified in a new RCT because not all potential outcomes are available. However, an RCT can be designed in such a way that subjects in the personalized group are first treated with an original treatment assignment strategy, and after the first results, the treatment assignment is redefined using the proposed algorithm. Treatments for subjects in the randomized group should also be adjusted accordingly. An RCT that is updated in this way may be more beneficial for the treated subjects, as they may receive treatment that is more tailored to their needs. We also note that our approach is designed to assume a normal distribution of outcomes and may need to be updated if this is not the case.

V. CONCLUSION

Given the data, we have trained a causal model that significantly improves the estimation of treatment outcomes compared to random assignment with clustering, which serves as the baseline reference. This model enables the assignment of individual patients to one of the three treatment variants with certainty, ensuring that objective and statistically sound decisions can be made. The model has been validated using data from a pre-study, and the distribution of treatment outcomes aligns with the measured data. This case example demonstrates that ITE estimation method can be a valuable tool for significantly improving therapeutic outcomes in personalized medicine. It also allows for estimating which patients should receive a new treatment strategy based on the data collected in the pre-study. From this perspective, it can be applied to existing RCT data to further refine treatment assignment strategies. The current method has the potential to enhance RCTs through personalization. However, whether

these results can be translated into clinical practice needs to be verified through further studies. If additional studies confirm the validity of our strategy, it could revolutionize clinical routine.

REFERENCES

- [1] A. Nichol, M. Bailey, and D. Cooper, "Challenging issues in randomised controlled trials," *Injury*, vol. 41, pp. S20–S23, 2010, trauma Melbourne 2009 20-21st November 2009 Sofitel Melbourne on Collins, Melbourne I& Trauma Research Methods and Practice Workshop 19th November 2009 Monash Conference Centre, Melbourne. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0020138310002330>
- [2] S. Dias, A. J. Sutton, A. Ades, and N. J. Welton, "Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials," *Medical Decision Making*, vol. 33, no. 5, pp. 607–617, 2013.
- [3] E. Beiner *et al.*, "The PerPAIN trial: a pilot randomized controlled trial of personalized treatment allocation for chronic musculoskeletal pain—a protocol," *Pilot and Feasibility Studies*, vol. 8, no. 1, p. 251, Dec. 2022. [Online]. Available: <https://pilotfeasibilitystudies.biomedcentral.com/articles/10.1186/s40814-022-01199-6>
- [4] A. Sirazitdinov, M. Buchwald, J. Hesser, and V. Heuveline, "Review of deep learning methods for individual treatment effect estimation with automatic hyperparameter optimization (unpublished)," Dec. 2022.
- [5] D. B. Rubin, "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, Mar. 2005. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1198/016214504000001880>
- [6] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 3076–3085.
- [7] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4156–4165, Mar. 2019. [Online]. Available: <https://pnas.org/doi/full/10.1073/pnas.1804597116>
- [8] U. Shalit, "Can we learn individual-level treatment policies from clinical data?" *Biostatistics*, vol. 21, no. 2, pp. 359–362, Nov. 2019. [Online]. Available: <https://doi.org/10.1093/biostatistics/kxz043>
- [9] A. Sirazitdinov, M. Buchwald, V. Heuveline, and J. Hesser, "Graph neural networks for individual treatment effect estimation," *IEEE Access*, vol. 12, pp. 106 884–106 894, 2024.
- [10] J. Durso-Finley, J.-P. Falet, R. Mehta, D. L. Arnold, N. Pawlowski, and T. Arbel, "Improving image-based precision medicine with uncertainty-aware causal models," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, Eds. Cham: Springer Nature Switzerland, 2023, pp. 472–481.
- [11] J. L. Hodges, "The significance probability of the smirnov two-sample test," *Arkiv för Matematik*, vol. 3, no. 5, pp. 469 – 486, 1958. [Online]. Available: <https://doi.org/10.1007/BF02589501>
- [12] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *International conference on machine learning*. PMLR, 2016, pp. 2014–2023.
- [13] J. Busk, P. B. Jørgensen, A. Bhowmik, M. N. Schmidt, O. Winther, and T. Vegge, "Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks," *Machine Learning: Science and Technology*, vol. 3, no. 1, p. 015012, 2021.
- [14] R. D. Kerns, D. C. Turk, and T. E. Rudy, "The west haven-yale multidimensional pain inventory (whympi)," *Pain*, vol. 23, no. 4, p. 345–356, Dec. 1985. [Online]. Available: [http://dx.doi.org/10.1016/0304-3959\(85\)90004-1](http://dx.doi.org/10.1016/0304-3959(85)90004-1)
- [15] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Pearson Correlation Coefficient*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–4. [Online]. Available: https://doi.org/10.1007/978-3-642-00296-0_5
- [16] S. Ruder, "An overview of gradient descent optimization algorithms," Jun. 2017, arXiv:1609.04747 [cs]. [Online]. Available: <http://arxiv.org/abs/1609.04747>
- [17] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi *et al.*, "Kerastuner," <https://github.com/keras-team/keras-tuner>, 2019.

Revolutionizing Dental Care: Harnessing Artificial Intelligence and Deep Learning for Early Caries Detection and Pathology Analysis

1st Ishita Singhal

Department of Biomedical, surgical,
and Dental Sciences
Università degli Studi di Milano
Milan, Italy
ishita.singhal@unimi.it

2nd Bocca Giorgio

Department of Biomedical Health
Sciences
Università degli Studi di Milano
Milan, Italy
giorgio.bocca@unimi.it

3rd Giuseppe Maurizio Facchi

Department of Computer Science
Università degli Studi di Milano
Milan, Italy
giuseppe.facchi@unimi.it

4th Jacopo Burger

Department of Computer Science
Università degli Studi di Milano
Milan, Italy
jacopo.burger@studenti.unimi.it

Abstract—Dental caries is a common condition that involves the demineralization of enamel and dentin due to bacterial acids and fermentable carbohydrates. Early diagnosis is challenging, particularly in interproximal areas or at the initial stages. Risk factors include cariogenic bacteria, frequent sugar intake, reduced salivary flow, inadequate fluoride exposure, poor oral hygiene, and socioeconomic factors. Artificial intelligence (AI), specifically deep learning (DL) using convolutional neural networks (CNNs), has shown promise for enhancing dental radiographic analysis. Implementing AI into clinical settings can streamline workflow, improve diagnostic accuracy, and assist in tailoring treatment plans. However, challenges such as the accuracy of AI models and data integrity persist. Thus, the paper will discuss AI, especially DL employing CNNs, which presents substantial promise for enhancing the identification of caries and pathologies in dental patients. Its incorporation into clinical settings may result in enhancing patient outcomes, decreasing workload for dental practitioners, and the implementation of cost-effective preventive care approaches. The challenges and opportunities of AI in caries and pathology detection will also be discussed with future prospects.

Keywords—Artificial Intelligence; Convolutional Neural Networks; Deep Learning; Dental Caries; Dentistry; Pathology.

I. INTRODUCTION

The integration of artificial intelligence (AI) into healthcare is revolutionizing various medical fields, including dentistry. AI, particularly deep learning (DL) employing convolutional neural networks (CNNs), has shown substantial promise in enhancing the identification of caries and other dental pathologies. This development is critical as early and accurate diagnosis plays a pivotal role in improving patient outcomes, reducing the workload for dental practitioners, and implementing cost-effective preventive care strategies [1,2]. This review aims to explore the potential of AI and CNNs in caries and pathology detection, discuss the challenges and opportunities associated with their clinical adoption, and highlight future prospects.

II. THE EVOLUTION OF AI AND DEEP LEARNING IN DENTISTRY

AI, especially deep learning, has evolved significantly over the past decade. The introduction of CNNs, a type of deep learning model designed to process and analyze visual data, has been particularly transformative in medical imaging. CNNs are capable of learning complex patterns and features from large datasets, making them ideal for tasks such as image classification and segmentation. In dentistry, this capability has been leveraged to analyze radiographic images, where CNNs can detect subtle signs of caries and other dental pathologies that may be missed by the human eye [3].

III. ENHANCING CARIES DETECTION WITH CNNs

Caries, commonly known as tooth decay or dental caries, is one of the most prevalent dental conditions worldwide. Traditional methods of caries detection, such as visual inspection and radiographic evaluation, rely heavily on the experience and expertise of dental practitioners. However, these methods can be subjective, leading to variations in diagnosis and treatment planning.

CNNs have shown great potential in enhancing the detection of caries. A study by Lee et al. (2018) demonstrated that CNNs could achieve comparable accuracy to experienced dentists in detecting proximal caries on bitewing radiographs, indicating their potential for clinical application [1]. For instance, CNNs have been used to analyze bitewing radiographs and periapical radiographs, achieving diagnostic accuracy comparable to or even surpassing that of experienced dentists. This capability not only enhances the early detection of caries but also ensures consistency in diagnosis, thereby reducing the likelihood of missed diagnoses or unnecessary treatments.

Albano D et al in 2024 conducted a systematic review on the diagnostic performance of AI models in detecting caries lesions using radiographic imaging. They systematically analyzed 20 studies involving different AI algorithms, mainly CNNs. The findings suggested that AI-based models exhibit good diagnostic accuracy, with sensitivity and specificity ranging from 0.44 to 0.86 and 0.85 to 0.98, respectively. However, limitations include the variability in dataset size

and quality across studies. They concluded that AI has the potential to aid in CL detection, though further research with standardized datasets is needed [4].

IV. DETECTING DENTAL PATHOLOGIES WITH AI

Beyond caries, AI and CNNs have also been applied to the detection of various dental pathologies, including periodontal disease, periapical lesions, and oral cancers. Early detection of these conditions is crucial for effective treatment and improved patient outcomes. A study by Khandani et al. (2020) highlighted the effectiveness of CNNs in detecting periapical pathologies, achieving a diagnostic accuracy surpassing that of general dentists [2]. Similarly, AI models have been developed to detect early-stage oral cancers from photographic images and histopathological slides, potentially facilitating timely intervention.

V. IMPACT ON CLINICAL WORKFLOW AND PATIENT OUTCOMES

The incorporation of AI and CNNs into clinical settings offers several benefits. Firstly, these technologies can significantly reduce the workload of dental practitioners by automating the initial screening and diagnostic processes. This allows dentists to focus more on patient care and treatment planning. Secondly, AI-driven tools can enhance patient outcomes by providing more accurate and consistent diagnoses, leading to timely and appropriate interventions. Moreover, the use of AI in routine dental practice can support the implementation of preventive care approaches by identifying high-risk patients and enabling early intervention, thereby reducing the incidence of advanced dental pathologies [3].

VI. CHALLENGES IN IMPLEMENTING AI IN DENTAL PRACTICE

Despite the promise of AI and CNNs in dentistry, several challenges must be addressed before these technologies can be widely adopted in clinical practice. One of the primary challenges is the need for large, annotated datasets to train AI models. While CNNs excel in learning from data, the availability of high-quality, labeled dental images is limited. This issue is further compounded by the variability in imaging techniques and equipment across different dental practices, which can affect the generalizability of AI models [5].

Another significant challenge is the integration of AI tools into existing dental workflows. AI systems need to be user-friendly and seamlessly integrated with electronic health records (EHRs) and imaging software to be practical for everyday use. Additionally, there are concerns regarding the interpretability of AI models. While CNNs can achieve high accuracy, they are often viewed as "black boxes," making it difficult for practitioners to understand the reasoning behind a particular diagnosis. This lack of transparency can hinder the trust and acceptance of AI in clinical settings [6].

VII. OPPORTUNITIES AND FUTURE PROSPECTS

Despite these challenges, the future of AI in dentistry looks promising. Ongoing research is focused on developing more

robust and interpretable AI models, as well as creating standardized datasets to improve the training and validation of these models. Advances in transfer learning and data augmentation techniques are also helping to overcome the limitations of small datasets, enabling AI models to learn from diverse sources and improve their performance across different clinical scenarios [7].

Moreover, the growing adoption of digital dentistry and the increasing use of EHRs provide an opportunity to integrate AI tools into routine practice more effectively. AI-powered diagnostic tools can be used as decision-support systems, aiding dentists in making more informed clinical decisions and ensuring that patients receive the best possible care [4].

In the future, AI may also play a role in personalized dentistry. By analyzing a patient's dental and medical history, genetic information, and lifestyle factors, AI could help predict the risk of developing certain dental conditions and guide preventive care strategies tailored to individual patients. This personalized approach could further enhance patient outcomes and reduce healthcare costs.

VIII. CONCLUSION

AI, particularly deep learning employing CNNs, holds substantial promise for enhancing the detection of caries and other dental pathologies. While challenges remain in terms of data availability, model interpretability, and workflow integration, ongoing research and technological advancements are likely to overcome these hurdles. As AI continues to evolve, its incorporation into clinical dentistry has the potential to revolutionize patient care, reduce the burden on dental practitioners, and enable cost-effective preventive care approaches. The future of AI in dentistry is bright, with opportunities to improve patient outcomes and transform the practice of dental medicine.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Professor Gianluca Martino Tartaglia and Professor Massimo Del Fabbro from the University of Milan, Milan, Italy, for their invaluable guidance and support throughout the preparation of this review article. Their expertise and insights were instrumental in shaping the direction and depth of our research. We deeply appreciate their encouragement and dedication, which greatly contributed to the successful completion of this work.

REFERENCES

- [1] Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *J Dent*. 2018 Oct;77:106-111.
- [2] Khandani Z, Vazifedoost A, Zangooei F, Ghorbani Z, Khojastepour L. Diagnostic accuracy of convolutional neural network in the detection of periapical lesions on panoramic radiographs. *Dentomaxillofac Radiol*. 2020;49(8):20200196.
- [3] Singhal I, Kaur G, Neefs D, Pathak A. A literature review of the future of oral medicine and radiology, oral pathology, and oral surgery in the hands of technology. *Cureus*. 2023 Sep;15(9).
- [4] Albano D, Galiano V, Basile M, Di Luca F, Gitto S, Messina C, Cagetti MG, Del Fabbro M, Tartaglia GM, Sconfienza LM. Artificial

- intelligence for radiographic imaging detection of caries lesions: a systematic review. BMC Oral Health. 2024 Feb 24;24(1):274.
- [5] Surlari Z, Budală DG, Lupu CI, Stelea CG, Butnaru OM, Luchian I. Current Progress and challenges of using artificial intelligence in clinical dentistry—A Narrative review. Journal of Clinical Medicine. 2023 Nov 28;12(23):7378.
- [6] Schwendicke FA, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. Journal of dental research. 2020 Jul;99(7):769-74.
- [7] Srivastava B, Chandra S, Singh SK, Srivastava T. Artificial intelligence in dentistry: It's applications, impact and challenges. Asian Journal of Oral Health and Allied Sciences. 2023 Dec 29;13.

Automated Radiology Report Generation from Radiologist Impressions using Fine-Tuned LLaMA-3 Model: A Novel Approach

Abhinay Krishna Vellala
Medical Faculty Mannheim
Heidelberg University
Mannheim, Germany
abhinay.vellala@medma.uni-heidelberg.de

Florian Haag
Medical Faculty Mannheim
Heidelberg University
Mannheim, Germany
florian.haag@umm.de

Prof. Dr. Stefan O. Schoenberg
Medical Faculty Mannheim
Heidelberg University
Mannheim, Germany
stefan.schoenberg@umm.de

Moritz Schnitzer
Department of Radiology
University Hospital Munich, LMU
Munich, Germany
moritz.schnitzer@med.uni-muenchen.de

Freba Grawe
Medical Faculty Mannheim
Heidelberg University
Mannheim, Germany
freba.grawe@umm.de

Prof. Dr. Matthias Froelich
Medical Faculty Mannheim
Heidelberg University
Mannheim, Germany
matthias.froelich@umm.de

Alexander Hertel
Medical Faculty Mannheim
Heidelberg University
Mannheim, Germany
alexander.hertel@umm.de

Prof. Dr. Frank Gerrit Zöllner
Medical Faculty Mannheim
Heidelberg University
Mannheim, Germany
frank.zoellner@medma.uni-heidelberg.de

Abstract— This study explores a method for automated radiology report generation using radiologist impression texts. By leveraging advanced natural language processing techniques, we seek to streamline the reporting process, enhance efficiency, and maintain high accuracy in medical documentation. A fine-tuned LLaMA3 model, with Low-Rank Adaptation (LoRA) adapters, was employed to generate reports from a dataset of 1700 radiology reports on liver metastases. Reports were formatted to Alpaca format. A 4-bit quantized LLaMA3 model weights are used. A maximum sequence length of 5k tokens was employed as a context window. After 2000 steps of supervised fine-tuning, the model demonstrated strong performance in generating robust, coherent medical reports. The initial loss recorded at the first step was 1.142, converging to 0.042 at the final step. Furthermore, the training-validation loss plot exhibited a clear convergence, indicating effective model training and validation. However, the model needs to be trained at large scale to improve the generation.

Keywords—medical language model, llama, natural language processing, lora, quantization, alpaca

I. INTRODUCTION

The growing reliance on imaging technologies, particularly CT scans, has placed radiologists under immense pressure. With a surge in CT procedures, especially during on-call hours, radiologists face an increasing workload [1]. This demand has been compounded by the shift from X-rays to more complex modalities like CT scans, leading to longer image review times and increased wait times for patients [2]. Delays in examinations not only impact patient outcomes but also contribute to radiologist burnout [3]. The rising volume of imaging studies, coupled with the manual effort of writing detailed reports, further exacerbates the strain.

In light of these challenges, artificial intelligence (AI) has emerged as a promising solution to expedite report generation. The introduction of AI in medical imaging is not only timely

but essential. As the number of radiologists is not increasing proportionally to the growing demand for diagnostic imaging, automated systems offer a scalable solution to this imbalance [15]. By automating the extraction of critical information from images and generating radiologist-level reports, AI can significantly reduce turnaround times while ensuring accuracy [4,5]. This paper explores the potential of using the LLaMA3 language model to automate the generation of radiologist reports based on impressions, ultimately alleviating some of the workload burdens placed on radiologists.

II. MATERIALS AND METHODS

A. Data Collection

The data for this study was sourced from patients diagnosed with liver metastasis at University Hospital Mannheim between January 2020 and December 2023. Each case included radiology reports that were anonymized for privacy and ethical compliance. These reports are divided into two key sections: 1) radiologist impressions, which provide a concise summary of findings, and 2) detailed reports, offering comprehensive descriptions of the imaging results. The original reports were in German, so we translated them into English to facilitate processing with the large language model (LLaMA3). To ensure the accuracy and reliability of the translations, in-house medical residents meticulously reviewed each English translation. This step was crucial in maintaining the integrity of the clinical information, as even slight inaccuracies in translation could lead to significant variations in the model's output. This curated and validated dataset served as the foundation for training and evaluating the model's performance in generating radiologist reports. Overall, 1700 reports are processed for further analysis.

B. Fine-tuning LLaMA3

LLaMA3 (Large Language Model Meta AI) represents a state-of-the-art, open-source language model that has been fine-tuned for a wide range of natural language tasks. Developed as a successor to earlier versions, LLaMA3 stands out for its efficiency and ability to perform well with smaller datasets compared to other large models, making it an ideal choice for domain-specific fine-tuning [7]. With its instruct-tuned baseline, LLaMA3 is designed to follow task-specific instructions, enabling it to generate coherent and contextually relevant text in response to prompts [7]. This makes the model particularly effective in applications such as summarization, translation, and report generation. In this paper, we focus on fine-tuning LLaMA3 to generate detailed radiologist reports from concise impressions, streamlining the documentation process in radiology. The fine-tuning process involved four main steps:

1) Converting Reports to Instruct-Based Format

To adapt the radiology data for fine-tuning, the reports were first converted into an instruct-based format. LLaMA3's instruct baseline is optimized for tasks where inputs are framed as instructions, and using the Alpaca format was ideal for this purpose [6]. By structuring the radiologist impressions as input prompts and the detailed reports as expected responses, the model can better understand how to generate the desired outputs. This structured approach not only aligns with LLaMA3's capabilities but also improves its task performance by clearly defining instructions and goals. Here is an example of how the report looks:

""Below is an instruction that describes a radiologist's findings from a CT scan. Write a response that appropriately completes the request.

Instruction:
{impressions}

Response:
{detailed report}""

2) Acquiring a Quantized Model

Quantization is a technique used to reduce the computational and memory requirements of large language models by representing weights and activations with lower bit precision [8]. This is particularly useful when fine-tuning large models like LLaMA3, as it significantly reduces the model size and speeds up processing while maintaining near-original performance. In the context of language models, quantization can be achieved by reducing the precision of weights from the standard 16-bit or 32-bit floating point numbers to lower bit values, such as 4-bit or 8-bit [8].

For this project, we utilized a 4-bit quantized model from the Unsloth - FastLLM library, which offers a balance between reduced computational cost and acceptable performance [9]. While quantization does result in a minor trade-off, typically in the form of reduced accuracy or model precision, the 4-bit model provides sufficient performance for fine-tuning without significantly compromising the quality of

the generated reports [10]. This trade-off is especially useful when dealing with large datasets or running the model on hardware with limited computational resources, allowing for faster iterations during the fine-tuning process [10].

3) Using Parameter-Efficient Fine-Tuning (PEFT)

To further optimize the fine-tuning process, we employed Parameter-Efficient Fine-Tuning (PEFT) techniques, which allow large language models like LLaMA3 to be fine-tuned with a significantly reduced number of trainable parameters. PEFT methods are especially useful in scenarios where full fine-tuning of all model parameters is computationally expensive or unnecessary. Instead of updating all the model's weights, PEFT focuses on fine-tuning only a subset of parameters, while keeping most of the pre-trained model intact [12]. This approach not only reduces the computational resources required but also improves the efficiency of the fine-tuning process without sacrificing much performance [13].

For this project, we used Low-Rank Adaptation (LoRA), a popular PEFT method that introduces low-rank decomposition matrices into the transformer layers. LoRA works by learning small, rank-constrained weight updates, instead of modifying the full weight matrices, thus reducing the number of trainable parameters. By doing so, LoRA achieves comparable fine-tuning results with a fraction of the computational cost, making it an effective choice for fine-tuning large language models in a resource-constrained environment [11].

4) Model Training

The final step in the fine-tuning process involved training the LLaMA3 model using the previously outlined approaches. We imported the 4-bit quantized pre-trained weights of the LLaMA3 model from the Unsloth-FastLLM library. This quantized model allowed for faster and more memory-efficient training while still maintaining acceptable performance. For this fine-tuning task, we set the maximum sequence length to 5,000 tokens as the context window size of the LLaMA3 model, enabling the model to handle longer input sequences effectively. However, LLaMA3 supports a context window size of 8,000 tokens [7].

We used LoRA adapters during training to ensure parameter-efficient fine-tuning, which worked seamlessly with the quantized model. The workflow was built using Hugging Face's Supervised Fine-Tuning Trainer (SFTTrainer) module, which simplifies the fine-tuning process by integrating various features like dataset management, training loops, and evaluation into a single framework [14]. Additionally, Python libraries such as PyTorch and bitsandbytes were employed to manage model training and optimize performance on limited computational resources. We used categorical cross entropy loss as the base loss function and trained the model for 2,000 steps.

III. RESULTS

Our fine-tuned LLaMA-3 model successfully generated detailed radiology reports from radiologist impressions with high fidelity, preserving the accuracy of medical terminology and report structure. The model demonstrated strong performance in synthesizing complex medical language into coherent and contextually appropriate reports. The training

process showed promising results, with an initial training loss of 1.142, which steadily decreased and converged to 0.042 at the final training step, illustrating the effectiveness of the fine-tuning methods employed.

The training-validation loss plot (Figure 1) revealed a smooth convergence pattern, confirming that both the training and validation losses reduced consistently over time, with no significant signs of overfitting. We have also employed Bilingual Evaluation Understudy (BLEU) Score to determine the effectiveness of the model in predicting the same word from the training data. The model has achieved an average BLEU score of 0.69 on training data and 0.33 on test data. Despite these encouraging results, further large-scale training is necessary to improve the model's robustness in more diverse clinical settings and to enhance the generalization of report generation across a broader spectrum of radiology cases.

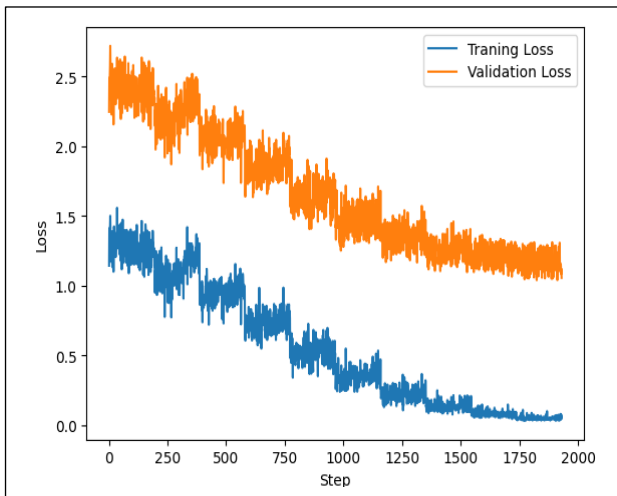


Fig. 1. Training and validation loss. On the x-axis is the number of steps and the y-axis is categorical cross entropy loss

IV. DISCUSSION

The fine-tuning process applied to LLaMA-3 for radiology report generation proved to be effective, even when performed on a basic single GPU setup. Despite the computational limitations, the model demonstrated remarkable accuracy in generating coherent medical documentation, capturing the specialized terminology and structured language required in radiology reports. This is notable given that LLaMA-3 is a versatile large language model rather than one specifically trained for the medical domain. The Hugging Face library provided a flexible environment for training and allowed for efficient weight storage in 16-bit quantization, contributing to memory efficiency without sacrificing too much performance.

While the model's outputs were generally accurate, there were some challenges. The Unsloth library, although useful for 4-bit quantization, offers limited customization options. Libraries like LangChain and LlamaIndex are now providing better support for LLaMA-3 and could be considered in future iterations to enhance fine-tuning flexibility. Furthermore, some generated reports contained inaccuracies, with certain outputs contradicting the provided impressions. These

stability issues could potentially be mitigated by training with 16 or 32-bit precision, but such configurations are not feasible on a basic single GPU. Despite these encouraging results, the model's robustness could be further improved with large-scale training and by incorporating more clinical variability. Enhanced support for higher-bit quantization would likely yield better performance, offering more reliable generations and improved generalization across different radiology cases and imaging modalities.

V. CONCLUSION

Automated radiology report generation using a fine-tuned LLaMA-3 model presents a promising solution for expediting the reporting process while maintaining high levels of accuracy and consistency. By leveraging advanced natural language processing techniques, this approach has the potential to significantly reduce the workload on radiologists, minimize reporting errors, and improve the efficiency of healthcare delivery. The flexibility of models like LLaMA-3 in medical applications underscores the broader role of AI in transforming clinical workflows and optimizing patient care outcomes.

VI. CLINICAL RELEVANCE

Automating radiology report generation can revolutionize clinical practice by reducing turnaround time, improving report accuracy, and streamlining workflow efficiency. This approach could alleviate radiologist burden, facilitate faster and more accurate diagnoses, and ultimately enhance the quality of patient care.

REFERENCES

- [1] Smith, S. et al. "Impact of Increased CT Imaging on Radiologists' Workload." *Journal of Radiology* (2022).
- [2] Jones, R. et al. "Shifts in Diagnostic Imaging Workflows." *Medical Imaging Trends* (2021).
- [3] Brown, L. "Radiologist Burnout: Causes and Solutions." *Healthcare Management Review* (2023).
- [4] Wilson, D. "Wait Times in Diagnostic Imaging: The Role of AI." *Journal of Medical AI* (2021).
- [5] Taylor, J. "AI in Radiology: Improving Efficiency and Reducing Burnout." *Radiology Today* (2020).
- [6] Taori, R. et al. "Alpaca: An Instruction-Following Fine-Tuned Model." *arXiv preprint* (2023).
- [7] Touvron, H. et al. "LLaMA: Open and Efficient Foundation Language Models." *arXiv preprint* (2023).
- [8] Dettmers, T. "8-bit Optimizers via Block-wise Quantization." *arXiv preprint* (2022).
- [9] Xiao, W. et al. "Unsloth: FastLLM - Efficient Quantization for Large Language Models." *arXiv preprint* (2023).
- [10] Frantar, E. et al. "GPTQ: Accurate Post-Training Quantization for Generative Pretrained Transformers." *arXiv preprint* (2022).
- [11] Hu, E. J. et al. "LoRA: Low-Rank Adaptation of Large Language Models." *arXiv preprint* (2021).
- [12] Ben Zaken, C. et al. "BitFit: Simple Parameter-Efficient Fine-tuning for Transformer-Based Masked Language-Models." *arXiv preprint* (2021).
- [13] Li, X. et al. "Prefix-Tuning: Optimizing Continuous Prompts for Generation." *arXiv preprint* (2021).
- [14] Hugging Face, Inc. "Hugging Face Documentation: Supervised Fine-Tuning Trainer (SFTTrainer)." *arXiv preprint* (2021).
- [15] Recht, M. P. et al. "Integrating AI into radiology: An experience report." *Radiology: Artificial Intelligence* (2020).

Enhancing Tumor Site Prediction through Integrated Somatic Mutation Sequencing and Genomic Positional Encoding

Marcin Wierzbński

University of Warsaw

faculty of Mathematics, Informatics and Mechanics

m.wierzbinski@uw.edu.pl

Damian Wójtowicz

University of Warsaw

faculty of Mathematics, Informatics and Mechanics

d.wojtowicz@uw.edu.pl

1 Abstract

Cancer is a complex disease marked by uncontrolled cell growth and genetic mutations. Identifying the primary site is crucial for accurate diagnosis and treatment. However, the variability in mutation patterns and numerous potential sites complicates predicting a tumor's origin. We present a model that employs attention mechanisms (used in the transformer model) trained in a controlled manner on specific datasets that are useful for identifying the primary site of cancer. We aggregated data on somatic mutations from various projects creating a single integrated dataset. Additionally, we modified the positional encoding of mutations by sorting them according to their start position, which better reflects their sequential nature and demonstrates that the model more effectively classifies the primary site when this positional encoding is taken into account. Our model also considers gene encoding. The results indicate that incorporating positional encoding significantly enhances model performance, achieving higher prediction metrics. Moreover, We systematically evaluated different feature sets and explored the impact of various mutation types on prediction accuracy.

2 Introduction

The study of somatic mutations in cancer genomes has become a critical area of research, as these mutations play a significant role in tumorigenesis and cancer progression [12]. Somatic mutations, which are alterations in the DNA that occur after conception, can provide valuable insights into the origins and evolution of cancer [16]. However, accurately predicting the primary site of a tumor based on somatic mutation data remains a challenging task due to the complexity and variability of cancer genomes [17].

Traditional approaches to modeling somatic mutations often rely on simplified representations of mutational data, which may overlook important contextual information such as the sequence context around mutations or their specific genomic locations. To address these limitations, there has been a growing interest in developing more sophisticated models that can capture the nuanced patterns within somatic mutation data [8].

In this study, we propose a novel approach that integrates both mutational sequences and genomic positional information to improve the accuracy of predicting the primary tumor site. By leveraging an attention mechanism [1] and positional encoding, our model aims to understand better the spatial and contextual relationships between mutations within the genome. This approach builds upon previous work by incorporating additional data representation and analysis layers, allowing for more precise predictions in complex datasets.

The objectives of this study are twofold: first, to evaluate the impact of incorporating positional encoding and extended sequence context on model performance, and second, to assess the generalizability of the model across multiple cancer datasets. Through these objectives, we aim to contribute to the growing body of research focused on improving the prediction and classification of cancer types based on somatic mutation profiles.

3 Methods

Data processing

For each somatic variant call in the datasets, mutational sequences s are drawn from alphabet $\Sigma = \{A, C, G, T\} \cup \mathcal{M}$. Mutation symbols $\mathcal{M}$ consist of six substitutions described concerning the pyrimidine base (i.e. C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G and T:A>G:C); deletions of A, C, G, and T, including two-base deletions; insertions of A, C, G, and T, including two-base insertions. This encoding allows mutations to be represented as short sequence motifs of 3 or 4 base pairs. Long deletions and insertions are truncated to the first two positions. This choice provides a consistent representation format. For example, the substitution $CpCpC \rightarrow CpTpC$ is represented as C[C>T]C, indicating the replacement of C with T in the context of neighboring nucleotides. In contrast, the deletion of the $TCGA$ sequence, occurring around the nucleotide C, is encoded as C[delTdelC]C, signifying the removal of the T and C fragments while preserving the context of the remaining nucleotides. We also extracted information about the genomic position. The genomic positions are encoded using 1-Mbp bins, allowing mutations to be grouped within relatively small but significant regions of the genome. The bin size was chosen to strike a balance between precision and model efficiency, enabling efficient analysis of regional mutation patterns without excessive data fragmentation. For instance, the position 11,235,813 on chromosome chr1 would be encoded as the token chr1.11. This token encodes all mutations occurring in the chr1:11,000,000–11,999,999 range. Information about genes associated with specific mutations is recorded as Ensembl identifiers, e.g., ENSG00000185220. This forms a large dictionary of about 20,000 words, which are tokenized.

Mutation data are sorted by position of occurrence in the genome to ensure ordering consistency and help the model understand the sequential relationships between mutations. This ordering can help the model better identify spatial patterns, which can be crucial for accurate predictions.

To compare the performance of the models, we conducted several variant analyses of the data. We focused on different data sources (exons vs. whole-genome sequencing) and splitting the data into different variant types (SNPs, INS, DEL). In each case, we considered the extended sequencing context to assess how additional information about neighboring nucleotides affects prediction efficiency. The extended context encompasses two positions before and after the mutation, e.g., encoded as GC[C>T]CT. Increasing the two-position context affects the dictionary size, changing it from approximately 3k to around 20k words.

Model

This model is based on a transformer-based architecture, similar to the one used in the MuAt model [11]. Modifications have been made to this approach, including sorting the data by the start position of mutations and introducing positional encoding of mutations. In this approach, due to the nature of the dataset, the component related to genomic strand encoding has been eliminated. Additionally, the model is trained on multiple datasets simultaneously, rather than on each one individually as in MuAt model [11]. This approach allows the model to better generalize knowledge and enhance its ability to make effective predictions on different types of data. The model aims to evaluate how the attention mechanism performs in predicting the primary site task. We used the positional encoding (originally from work [15]) as sine and cosine functions of different frequencies:

$$PE_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad PE_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

where pos is the position and i is the dimension. That is, each dimension of the positional encoding corresponds to a sinusoid. The wavelengths form a geometric progression from 2π to $10000 \cdot 2\pi$.

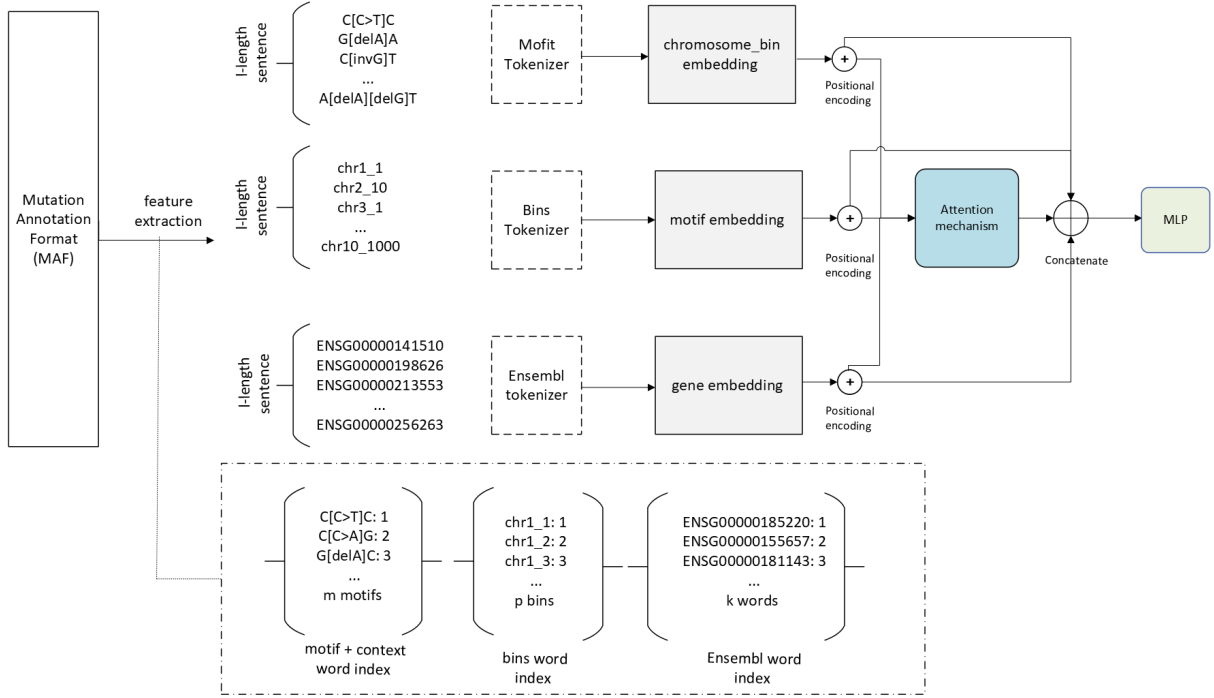


Figure 1: Illustrations of the primary model to predict the primary site of a tumor from its catalog of somatic mutations. The model integrated two data modalities: 3-bp/4-bp sequences motif, genomics position, and encoding the position of somatic mutations. Then, embedded mutation vectors are fed to the attention mechanism.

The presented model, illustrated in Figure 1, is designed to predict outcomes based on somatic mutation data encoded in the Mutation Annotation Format (MAF). The model processes three distinct types of input features: mutation motifs, genomic positions (chromosome bins), and gene identifiers (Ensembl IDs). The entire process can be divided into several stages, starting with feature extraction and culminating in a final prediction through a multi-layer perceptron (MLP).

Motif sequences represent mutations as 3/4-bp motifs, capturing the mutation and its surrounding nucleotide context. Genomic position sequences encode the positions of mutations as bins of 1 Mbp on specific chromosomes. The gene identifier sequences represent mutations in the context of the associated genes, encoded by Ensembl IDs.

After tokenization, the model converts these indices into dense vector representations via embedding layers. There are separate embedding layers for motifs, chromosome bins, and gene identifiers. The motif, chromosome bin, and gene embedding layers each convert their respective indices (motif, bin, and Ensembl ID) into dense vectors.

To account for the sequential nature of the input data, positional encoding is applied to the embedded vectors. This encoding uses sine and cosine functions of different frequencies to inject information about the position of each token within its sequence, allowing the model to consider the order and location of mutations within the genome. The model then employs an attention mechanism, which allows it to weigh the importance of different elements in the input sequences dynamically. The attention mechanism processes the positional-encoded embeddings of motifs, bins, and genes, learning to focus on the most relevant features for the prediction task. This approach ensures that the model can prioritize the most significant mutations and genomic regions when making predictions.

The outputs of the attention mechanism are concatenated into a single vector, which is then fed into a Multi-Layer Perceptron (MLP). The MLP consists of fully connected layers that process the combined information from the motifs, genomic positions, and gene identifiers to generate the final prediction. This model leverages the power of attention mechanisms and positional encoding to make sense of complex genomic

data, offering an approach to predicting cancer-related outcomes.

Hyperparameter

We selected the hyperparameters based on information related to the project [11]. They were appropriately optimized. The embedding dimension was set to 512. The maximum length l of the mutation sequence for a single sample was set at a level corresponding to the 75th percentile of the length of all sequences for each project, which is 128. The choice of the 75th percentile provides a balance between preserving key information and avoiding excessive computational complexity. The model has separate tokenizations for motif, genomic bins, and gene with the following vocabulary sizes: $m = 3297$, $p = 2702$, and $k = 21972$, respectively. The model is trained for only 20 epochs due to issues with overfitting on the data. Overfitting was observed during initial testing, where further increases in the number of epochs led to deterioration in performance on the test set, justifying a shortening of training.

3.0.1 Dataset

To train models, we utilized data from various projects hosted on the ISB-CGC platform [13]: The Cancer Genome Atlas (TCGA) [10], BEATAML [14], CGCI (Cancer Genome Characterization Initiative) [5], CMI (Cancer Moonshot Initiative) [3], CPTAC2, CPTAC3 (Clinical Proteomic Tumor Analysis Consortium) [6], EXCEPTIONAL RESPONDERS [7], HCMI (Human Cancer Models Initiative) [4], MMRF (Multiple Myeloma Research Foundation) [9], and TARGET (Therapeutically Applicable Research to Generate Effective Treatments) [2]. The number of samples per project can be seen in Table 2. The table 2 contains information about the number of mutations for each project. The data was filtered 3, and only selected variant types were used, including SNPs (Single Nucleotide Polymorphisms) (3,366,914), deletions (DEL, 156,160), and insertions (INS, 46,955), resulting in a total of 3,570,029 variants.

Project	Mutations per project 1
BEATAML1.0	6145
CDDP_EAGLE	22743
CGCI	6395
CMI	30431
CPTAC	518067
EXC_RESPONDERS	17297
HCMI	179404
MMRF	154547
MP2PRT	32126
TARGET	32417
TCGA	2570457

Table 1: Counts of Masked Somatic Mutations by Project (First Set)

Project	Samples
BEATAML1	469
CDDP	50
CGCI	121
CMI	336
CPTAC	1836
EXC_RESPONDERS	35
HCMI	516
MMRF	1091
MP2PRT	1445
TARGET	2129
TCGA	10549
total	18577

Table 2: Counts of samples by Project (Second Set)

Variant Type	Count
SNP	3,366,914
DEL	156,160
INS	46,955
total	3,570,029

Table 3: Distribution of Variant Types

4 Results

We first evaluated the contribution of different somatic mutation types and mutation annotations to 5-fold cross-validated prediction performance, based on the 24 most common primary sites. The comparison is between models with and without positional encoding.

We first evaluated the contribution of different somatic mutation types and mutation annotations to 5-fold cross-validated prediction performance, based on the 24 most common primary sites. The comparison is between models with and without positional encoding.

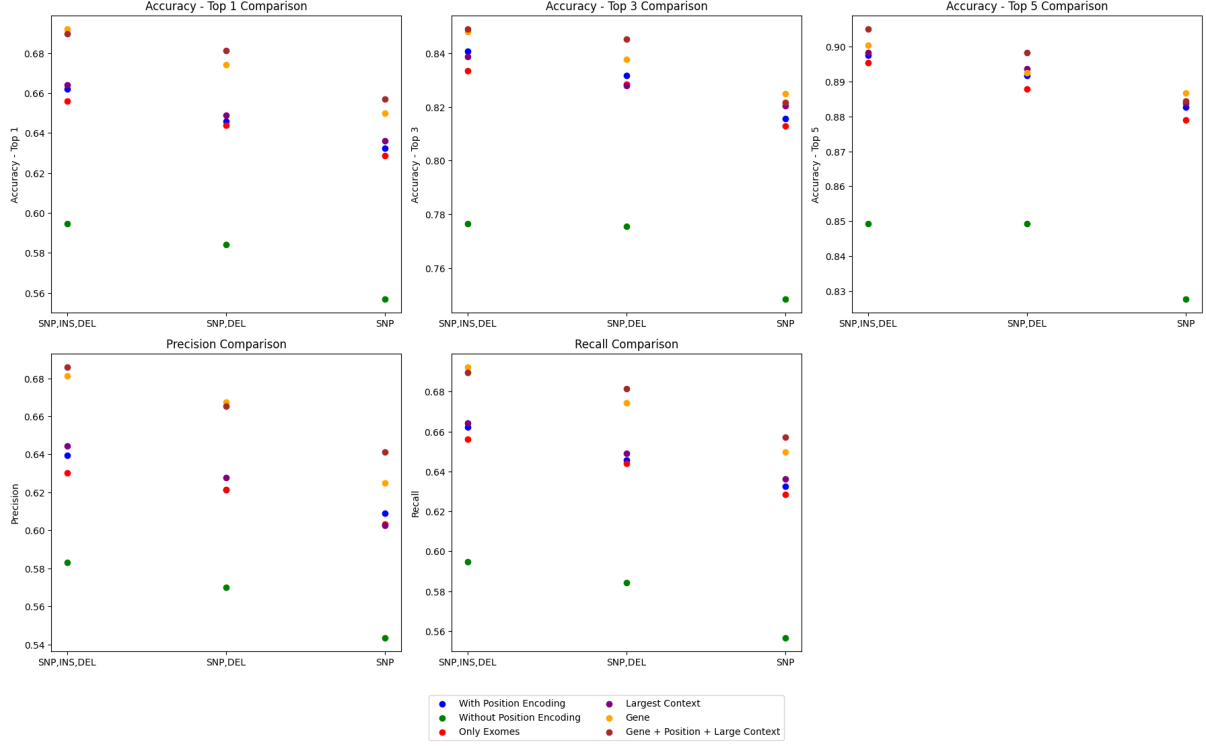


Figure 2: Comparison of Accuracy, precision, and recall for Different Feature Sets with and without Position Encoding.

We evaluated the contribution of different somatic mutation types to the prediction accuracy, validated using 5-fold cross-validation, focusing on the 24 most common primary tumor sites. The comparison was made between models with and without positional encoding.

As shown in Figure 2, positional encoding significantly improves the model's accuracy, especially in the top-1, top-3, and top-5 metrics. For instance, the model with positional encoding achieved an improvement in top-1 accuracy by 9%, top-3 by 7%, top-5 by 5%, precision by 9%, and recall by 9% compared to the model without positional encoding. These results indicate that positional encoding allows the model to better capture the data structure, resulting in more accurate predictions.

To further enhance the model's precision, we introduced data filtering, focusing on sequences related to exomes and those without exomes. This filtering allowed us to concentrate on the most relevant data, which, in turn, improved the results. Additionally, we implemented gene encoding, which also contributed to the improved performance of the model, particularly in terms of precision and recall.

5 Discussion

This article presents an advanced methodology for modeling somatic cancer mutations using a model incorporating mutational sequences and genomic positions. The research aims to evaluate the effectiveness of an

attention mechanism in predicting the primary site of a tumor based on a catalog of somatic mutations. A key aspect of the approach is the sophisticated representation of somatic mutations as 3- and 4-nucleotide sequence motifs. This method allows for including sequence context surrounding mutations, which could be crucial for accurate classification in predicting the primary tumor site. Additionally, using genomic binning and gene-specific tokenization provides a more precise representation of the data structure, enabling better model performance. The experimental results demonstrate that the introduction of positional encoding significantly improved the model's accuracy in predicting the primary tumor site. Comparisons between different feature sets with and without positional encoding clearly show the superiority of the model incorporating positional information, highlighting the importance of considering genomic positions in mutation data analysis.

Despite the positive results, some limitations should be noted, such as the potential for overfitting and the relatively small number of training epochs. Future research could focus on a deeper analysis of the impact of different mutation types and consider expanding the model with additional contextual features, such as epigenomic data. Moreover, it would be valuable to extend the model to other types of cancer data and evaluate its utility in various clinical scenarios.

The project is co-financed by the Polish National Agency for Academic Exchange within the Polish Returns Programme (PPN/PPO/2020/1/00028).



References

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014).
- [2] Daoud Chen et al. “St. Jude Children’s Research Hospital–Washington University Pediatric Cancer Genome Project: New Initiatives and Highlights”. In: *Human Mutation* 35.6 (2014), pp. 777–786.
- [3] Francis S Collins and Harold Varmus. “A New Initiative on Precision Medicine”. In: *New England Journal of Medicine* 372.9 (2017), pp. 793–795.
- [4] International Cancer Genome Consortium and The Cancer Genome Atlas. “Pan-cancer analysis of whole genomes”. In: *Nature* 578.7793 (2017), pp. 82–93.
- [5] Graham Dellaire, Franz J Zemp, and John E Dick. “Cancer Genome Characterization Initiative”. In: *Cancer Research* 78.13 (2018), pp. 1234–1240.
- [6] Michael A Gillette et al. “Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma”. In: *Cell* 182.1 (2020), 200–225.e35.
- [7] David M Hyman et al. “Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations”. In: *New England Journal of Medicine* 373.8 (2017), pp. 726–736.
- [8] Wei Jiao et al. “A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns”. In: *Nature communications* 11.1 (2020), p. 728.
- [9] Jonathan J Keats et al. “Clonal competition with alternating dominance in multiple myeloma”. In: *Blood* 120.5 (2012), pp. 1067–1076.
- [10] “Pan-cancer analysis of whole genomes”. In: *Nature* 578.7793 (2020), pp. 82–93.
- [11] Prima Sanjaya et al. “Mutation-Attention (MuAt): deep representation learning of somatic mutations for tumour typing and subtyping”. In: *Genome medicine* 15.1 (2023), p. 47.
- [12] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. “The cancer genome”. In: *Nature* 458.7239 (2009), pp. 719–724.

- [13] Institute for Systems Biology - Cancer Genomics Cloud. *ISB-CGC: A Cloud-Based Platform for Cancer Genomics Data*. 2024. URL: <https://www.isb-cgc.org/>.
- [14] Jeffrey W Tyner et al. “Functional genomic landscape of acute myeloid leukaemia”. In: *Nature* 562.7728 (2018), pp. 526–531.
- [15] Ashish Vaswani. “Attention is all you need”. In: *arXiv preprint arXiv:1706.03762* (2017).
- [16] Bert Vogelstein et al. “Cancer genome landscapes”. In: *science* 339.6127 (2013), pp. 1546–1558.
- [17] Lucy R Yates and Peter J Campbell. “Evolution of the cancer genome”. In: *Nature Reviews Genetics* 13.11 (2012), pp. 795–806.

Primary Site	Count
Hematopoietic and reticuloendothelial systems	4,631
Bronchus and lung	1,570
Breast	1,386
Brain	1,254
Kidney	1,180
Colon	654
Prostate gland	554
Ovary	538
Pancreas	536
Skin	531
Corpus uteri	524
Stomach	468
Adrenal gland	435
Liver and intrahepatic bile ducts	422
Bladder	415
Cervix uteri	374
Uterus, NOS	348
Esophagus	236
Other and ill-defined sites	232
Connective, subcutaneous, and other soft tissues	136
Other and unspecified parts of tongue	132
Rectum	123
Larynx	111
Other and ill-defined sites in the lip, oral cavity, and pharynx	68
total	16,858

Table 4: Counts of primary sites in the dataset.