A non-contrast multi-parametric MRI biomarker for assessment of MR-guided focused ultrasound thermal therapies

Sara Johnson, Blake Zimmerman, Henrik Odéen, Jill Shea, Nicole Winkler, Rachel Factor, Sarang Joshi, and Allison Payne

***Abstract*— *Objective:* We present the development of a non-contrast multi-parametric magnetic resonance (MPMR) imaging biomarker to assess treatment outcomes for magnetic resonance-guided focused ultrasound (MRgFUS) ablations of localized tumors. Images obtained immediately following MRgFUS ablation were inputs for voxel-wise supervised learning classifiers, trained using registered histology as a label for thermal necrosis. *Methods:* VX2 tumors in New Zealand white rabbits quadriceps were thermally ablated using an MRgFUS system under 3T MRI guidance. Animals were re-imaged three days post-ablation and euthanized. Histological necrosis labels were created by 3D registration between MR images and digitized H&E segmentations of thermal necrosis to enable voxel-wise classification of necrosis. Supervised MPMR classifier inputs included maximum temperature rise, cumulative thermal dose (CTD), post-FUS differences in T2-weighted images, and apparent diffusion coefficient, or ADC, maps. A logistic regression, support vector machine, and random forest classifier were trained in red a leave-one-out strategy in test data from four subjects. *Results:* In the validation dataset, the MPMR classifiers achieved higher recall and Dice than than a clinically adopted 240 cumulative equivalent minutes at 43°C (CEM_{43}) threshold (0.43) in all subjects.redThe average Dice scores of overlap with the registered histological label for the logistic regression (0.63) and support vector machine (0.63) MPMR classifiers were within 6% of the acute contrast-enhanced non-perfused volume (0.67). *Conclusions:* Voxel-wise registration of MPMR data to histological outcomes facilitated supervised learning of an accurate non-contrast MR biomarker for MRgFUS ablations in a rabbit VX2 tumor model.**

***Index Terms*— Focused ultrasound, thermal ablation, Supervised learning, multi-parametric MRI, MR-guided, histology registration**

Paper submitted for review on 01/24/2023. This study was funded by NIH grants P30CA042014, R37CA224141, R01CA259686, S10OD018482, and R03EB029204

S. Johnson (email: sara.l.johnson@utah.edu), H. Odéen, N. Winkler, and A. Payne are with the Department of Radiology and Imaging Sciences, University of Utah.

B. Zimmerman and S. Joshi are with the Biomedical Engineering Department, Scientific Computing and Imaging Institute, University of Utah.

J. Shea is with the Department of Surgery, University of Utah.

R. Factor is with the Department of Pathology, Duke University.

I. Introduction

MAGNETIC resonance (MR)-guided focused ultrasound (MRgFUS) ablation therapies are promising noninvasive alternatives to surgical tumor resection for many localized cancer indications. A critical challenge for noninvasive therapies is assessing treatment efficacy immediately after treatment. The gold-standard method for ensuring complete and clear margins is histology following surgical resection. Therefore, MRgFUS treatment biomarkers must accurately and precisely predict the thermal lesion as measured by histology to eliminate the need for surgical resection. Currently, there are two MR biomarkers used clinically for assessing the thermal ablation zone: the non-perfused volume (NPV) measured on contrast-enhanced T1-weighted imaging (CE-T1w) [1], [2] and the cumulative thermal dose (CTD) threshold of 240 cumulative effective minutes at 43°C (240 CEM_{43}) [3], calculated from MR temperature imaging (MRTI) [4], [5]. Despite their clinical adoption, both the NPV and CTD have demonstrated limitations in predicting the thermal lesion immediately following MRgFUS ablation.

Acute NPV is demonstrated to be a more accurate predictor of histological thermal necrosis than using CTD thresholds and is more widely used to predict the thermal lesion. However, the NPV immediately after ablation tends to overestimate the treated region [4], which can lead to mislabeling of viable tissues and potential tumor progression or recurrence. Transient effects such as edema and damaged, leaky blood vessels can conflate the interpretation of the NPV acquired immediately after thermal ablation [4]. NPV assessment also requires the administration of gadolinium contrast agent which introduces susceptibility artifacts that affect future MR temperature imaging and complicates treatment monitoring [6], [7]. Finally, continuing treatment after administering contrast agent can trap toxic contrast agents in the tissue [7].

The cumulative thermal dose is a non-contrast treatment assessment metric, typically computed using the proton resonance frequency method for MR temperature imaging (MRTI) and an assumed baseline body temperature [8], [9]. Thermometry images acquired throughout sequential sonications are integrated and thresholded at a predetermined lethal dose to predict the region of thermal necrosis [10], [11]. The clinically adopted 240 CEM_{43} threshold of CTD often un-

derestimates the true lesion size as assessed by standard histology techniques and NPV measurements [4], [12]–[14]. Underestimation may be a result of sub-ablative CTD levels causing delayed apoptosis due to irreversible thermal damage or temporary loss of perfusion [14], [15]. While CTD monitoring provides a real-time assessment metric, the threshold requires an *a priori* assumption of baseline tissue temperature and can be tissue and tumor-specific [16], [17]. For example, in clinical applications of MRgFUS in the brain, varying optimal CTD thresholds of 200 [18], [19] and 17 CEM_{43} [20] have been identified by comparing CTD to other MR biomarkers in patients. Overall, there is a need for alternative non-contrast MR biomarkers that can robustly predict the final treatment outcomes immediately following thermal ablation of localized tumors [10], [11].

There is currently no single non-contrast MR parameter that accurately predicts the thermally ablated lesion [4], [21]. We hypothesize that integrating thermal information with additional multi-parametric MR (MPMR) imaging that is sensitive to acute changes in tissue structure can improve acute thermal necrosis predictions. For example, T2 relaxation time immediately following thermal ablation therapies correlate with inflammation and edema formation [4], [22]. Additionally, changes in apparent diffusion coefficient (ADC) maps from diffusion-weighted imaging (DWI) are an indicator of coagulative necrosis and cytotoxic edema [23]. MPMR biomarkers have been previously investigated for acute MRgFUS treatment assessment. Hectors *et al.* utilized clustering algorithms to segment the MPMR feature parameter space into viable and nonviable groups, and demonstrated that a combination of T1 maps, T2 maps, and apparent diffusion coefficient (ADC) maps provides more accurate predictions than single metrics alone when comparing volume fractions with histology [6]. Although this type of analysis demonstrates a correlation between MPMR imaging and histology, it is not spatially specific; therefore, the precision, recall, and spatial similarity of the multi-parametric predictions remain unknown. Furthermore, studies for detecting prostate cancer from MPMR biomarkers show promising spatially specific results using manually-derived ground truth labels to train a supervised machine learning algorithm [24].

This study implements supervised machine learning classification to investigate a non-contrast MPMR biomarker for acute MRgFUS thermal treatment assessment in a multi-tissue VX2 rabbit tumor model. We leverage a previously validated MR-to-histology registration workflow [25] to generate the ground-truth labels for training from necrosis-labeled H&E volumes. Binary classification outcomes and spatial similarity are quantified on a voxel-wise level and compared to the NPV and 240 CEM_{43} clinical metrics. We demonstrate that combining MRTI and MPMR imaging in a supervised classifier can provide a more accurate acute non-contrast thermal lesion prediction than the 240 CEM_{43} metric. The primary benefits of an acute, non-contrast MR biomarker are the ability to continue ablation treatment after the initial treatment assessment, increased efficiency of MRgFUS ablations, and the reduction of contrast use.

II. Methods

The feasibility of supervised classification for developing MPMR biomarkers of the thermal lesion was tested in a rabbit tumor model to simulate tumor targeting and enable classifier learning on multiple tissue types. Figure 1 shows the anatomical geometry of the subjects during MRgFUS treatment. MPMR biomarkers, including MRTI, T2w imaging, and ADC maps, were collected before, during, and after MRgFUS ablation to generate the MPMR features for machine learning. Three supervised machine learning algorithms were trained and evaluated in a voxel-wise manner: a Logistic Regression classifier (LRC) and a Random Forest classifier (RFC), and a support vectors machine classifier (SVMC) [24], [26]–[28]. The predictive accuracy of the MPMR classifiers and standard clinical metrics were evaluated against the ground-truth histology label. Details for each of these steps are provided below.

A. MRI Acquisition and Clinical Biomarkers

All experiments were carried out in accordance with the approved Institutional Animal Care and Use Committee regulations at the University of Utah (Protocol 17-08012, approved 09/07/2017). Following intramuscular injection of VX2 tumors cells ($1x10^6$ cells in 50% media/Matrigel solution) into the quadriceps of four New Zealand white rabbits (2.5-3 kg), tumors were grown to approximately 2 cm in length, then ablated using an MRgFUS system (Image Guided Therapy, Inc.) with a 256-element phased-array transducer (Imasonic, Voray-sur-l'Ognon, France; 10-cm focal length, 14.4×9.8 cm aperture, f=940 kHz). Animals were intubated and anesthetized with isoflurane (2-3%) and monitored for vitals throughout the procedure. Ablation procedures were performed inside a 3T MRI scanner (Prisma^{FIT} Siemens, Erlangen, Germany) for tumor targeting, treatment monitoring, and post-treatment assessment. The MRgFUS system incorporated a single-loop MR receive coil around the targeted quadriceps to improve the MR image signal-to-noise ratio. All MR acquisition parameters are listed in Table I.

The rabbit was positioned on the treatment table as shown in Figure 1. High-resolution (0.5-mm isotropic) non-contrast T1w images of the entire treated leg (marked with ** in Table I) were acquired prior to and immediately after ablation for MR registration purposes. Coronal MPMR images of the lower leg were acquired immediately before (T2w, DWI), during (3D MRTI, PRF method), and 20 minutes after (T2w, DWI) the ablation procedure. The ablation procedure targeted 50% of the VX2 tumor and the surrounding muscle tissue. Sonication details for each subject are outlined in Table II. Approximately 40 minutes following the final sonication, CE-T1w MR images were acquired immediately following intravenous gadolinium injection (0.3 mL/kg ProHance) and animals were recovered. Due to the expected transient effects of ablation such as edema and latent apoptosis, which occurs for up to 72 hrs [4], [14], we defined the final treatment effect as the necrotic volume at 3-5 days after treatment. At this time-point post-ablation, the animal was re-positioned on the MRgFUS table and T1w and CE-T1w images were acquired for longitudinal MR and

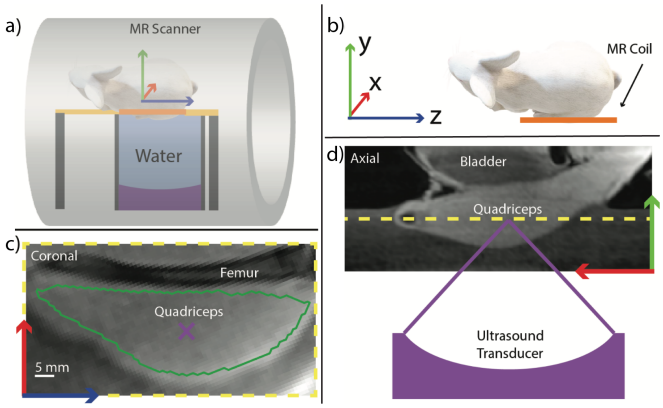

Fig. 1. a) Setup for the ablation and imaging procedures on the custom-built MRgFUS table and imaging coil. b) Subject positioning relative to the MR coordinate system. c) The coronal imaging plane (T1w) with the purple "X" indicating the location of the ultrasound geometric focus. The green contour is the quadriceps segmentation used for voxel-wise machine learning in this study. d) The location of the ultrasound transducer beneath the subject quadriceps in an Axial slice (T1w). The yellow-dotted line is the center of the multi-slice 2D or 3D coronal imaging plane.

histological registration purposes. Animals were then immediately euthanized for tissue excision and histology processing. A large region of the quadriceps which fully encompassed the tumor and MRgFUS ablation volume was excised and fixed in formalin for two weeks before sectioning and staining with H&E. Further details of gross tissue processing, sectioning, and staining with H&E are reported in the Appendix.

ADC maps were computed using a mono-exponential fit of the DWI signal at 2 b-values (b=20,500), averaged in three directions. Semi-automatic segmentation of the acute NPV was performed by thresholding the CE-T1w image intensity to encompass the non-enhancing region, followed by manual editing. All acute NPV segmentations were evaluated by an expert radiologist (NW) and scored using a Likert scale of 1-5. Expert-evaluated acute NPV segmentations were considered the clinical NPV prediction of tissue necrosis.

B. MRTI Prior Baseline Reconstruction

MRTI data was reconstructed and zero-fill interpolated to 1-mm isotropic resolution. A prior baseline approach similar to Bitton *et al.* [29] was implemented to account for localized temperature accumulation resulting from multiple sonications in the targeted tissue. This prior baseline approach is an improvement over the current clinical method for calculating MR thermal dose, which assumes a constant baseline temperature throughout the treatment and underestimates the total dose. In summary, the baseline phase from a prior sonication was used to calculate phase accrual at the start of subsequent sonications. A temporal criterion was applied to ensure valid baseline sonications were chosen to account for potential large B_0-field drift and subsequent phase-wrap errors. Therefore, the prior baseline sonication,(S_p), whose pre-heating baseline phase image is used as the baseline phase image for calculating the temperature change in the subsequent sonications, was reset to the current sonication when the time between sonications exceeded 10 minutes.Ten minutes is the simulated necessary cooling time for the muscle to return to a 1°C rise after a sonication reaching the maximum temperature elevation observed in this study. The details of the simulation implementation are described in [30], using a 0.5-mm iso-tropic T2w-anatomical segmented model of the rabbit hindlimb (Seg3D2, SCI Institute [31], University of Utah) and tissue properties from [32]). Based on the time between subsequent sonications, S_p was reset 1, 2, 1, and 5 times for animals 1-4, respectively. For each sonication, S_n, the local phase accumulation was calculated as:

$$\Delta\phi_{acc,n} = (\phi_{0,n} - \phi_{0,p}) - \Delta\phi_{BA,n}, \quad (1)$$

where $\phi_{0,n}$ is the baseline phase image for S_n, $\phi_{0,p}$ is the baseline phase image for S_p, and $\Delta\phi_{BA,n}$ is the bulk phase accrual resulting from B_0-field drift and inter-scan resonance frequency adjustments. $\Delta\phi_{BA,n}$ was calculated for each coronal slice of the MRTI volume by taking the mean value of $(\phi_{0,n} - \phi_{0,p})$ in a 15×15 voxel ROI in non-heated quadriceps muscle. For each sonication S_n, the phase change due to FUS heating and $\Delta\phi_{acc,n}$ were converted to temperature using the PRF method ($\alpha = -0.01$ ppm/°C) [8]. The absolute temperature profile for S_n, was calculated as the sum of these

TABLE I
MRI SEQUENCE PARAMETERS

Scan Type	Sequence	TR (ms)	TE (ms)	Flip Angle	Field of View (mm)	Pixel Bandwidth (Hz/Pixel)	Acquisition Resolution (mm)	Number Averages	Acquisition Time (mm:ss.ms)
MRTI	GRE-EPI (ETL=7)	25	11	14°	192×150×20	750	1.5×1.5×2.0	1	0:04.50
T1w*	VIBE	7.19	2.05	15°	256×192×52	250	1.0×1.0×1.0	1	1:03.00
T2w	SPACE	2000	300	120°	256×192×52	700	1.0×1.0×1.0	2	5:12.00
T1w**	VIBE	7.19	2.52	15°	256×192×56	250	0.5×0.5×1.0	3	6:19.00
Diffusion	SS-SE-EPI (ETL=92) (b=20,500)	7500	117	90°	160×116×20	1260	1.25×1.25×2.0	1	1:38.00

* T1w sequence with contrast for NPV segmentation.
** T1w sequence without contrast for image registration.
TR: Repetition Time, TE: Echo Time, EPI: Echo Planar Imaging, SPACE: Sampling Perfection with Application Optimized Contrasts Using Different Flip Angle Evolution, VIBE: Volumetric Interpolated Breath-Hold, GRE: Gradient Recalled Echo, MRTI: Magnetic Resonance Temperature Imaging, SS: Single Shot, SE: Spin Echo, and ETL: Echo Train Length.

TABLE II
MRGFUS TREATMENT PARAMETERS BY SUBJECT

Subj.	Tumor Vol.(mm^3)	Number of Sonications	Acoustic Power Mean $\pm$ 1 Std. (W)	Total Energy (kJ)	Maximum Temperature (°C)	Maximum CTD (CEM_{43})	Histology Vol.(mm^3)	Quadriceps Vol.(mm^3)
1	785.5	11	57 $\pm$ 17	23.14	55	$4.01e^{04}$	663	18,286
2	248.1	12	69 $\pm$ 18	26.00	78	$3.49e^{11}$	1,324	16,833
3	1,929.5	14	44 $\pm$ 9	18.59	72	$2.29e^{09}$	2,486	18,587
4	806.4	10	56 $\pm$ 9	18.55	78	$1.14e^{11}$	3,325	13,539

components and the baseline body temperature as measured by rectal fiber-optic probe monitoring animal body temperature at the time of the prior sonication $T_{FO,p}$:

$$T_{k,n} = \Delta T_{k,n} + \Delta T_{\phi_{acc,n}} + T_{FO,p}, \quad (2)$$

where T is temperature (°C), for k MRTI measurements in sonication n. To account for thermal dose accumulation between sonications, all time points between the previous and n$^{\text{th}}$ sonication, S_n, were padded with the value given by $\Delta T_{\phi_{acc,n}}$. Finally, the CTD in each voxel was calculated using the CEM_{43} model for each sonication and summing across all sonications [5].

C. Histological Registration

To create a co-registered data set for supervised classification training and validation, there were three registration steps: 1) motion during treatment 2) longitudinal changes in subject pose and position between ablation imaging and follow-up imaging [33], and 3) registration between histology and follow-up imaging. The registration methods used in this study have been previously developed [33] [25] but are briefly summarized here. During the MRgFUS procedure, the rabbit hindlimb may have slowly sunk deeper into the water bath, introducing minor motion errors over the course of the 3-hour treatment. To correct for this bulk motion in MRTI data, the final MRTI sonication magnitude image was registered to the post-treatment non-contrast T1w image using variance equalization and highly constrained (to limit deformation to bulk motion) elastic registration. Each of the prior MRTI sonication images was then registered to the deformed final MRTI sonication image using the same elastic registration method. The motion estimated by this registration process was also applied to the pre-treatment MR images and visual inspection was used to ensure the tissue-water boundary was aligned across the images. The bulk motion estimated during the registration process was less than 5 mm for all subjects.

Between treatment day imaging and 3-5 day follow-up imaging, there were unavoidable changes in the animal positioning in the MR scanner. A volume-conserving longitudinal registration algorithm was used to register follow-up images to treatment day images as tissue volume is preserved under normal physiological loading [33]. High-resolution T1w images (marked with ** in Table I) without contrast were used for registration since the features resulting from treatment are not visible. This registration step provided an invertible 3D diffeomorphism between the treatment day images and the follow-up images, which are registered to histology in the following step.

Finally, direct, voxel-wise comparison between ground truth histology and treatment day MR images required accurate histology registration. Histology necrosis annotations were registered to *in vivo* MPMR images to generate a volumetric label of histology necrosis. Briefly, a novel workflow was used to correct for deformation from every step of the tissue processing for histology. 3D surface registration was used to generate an invertible 3D diffeomorphism for each histology

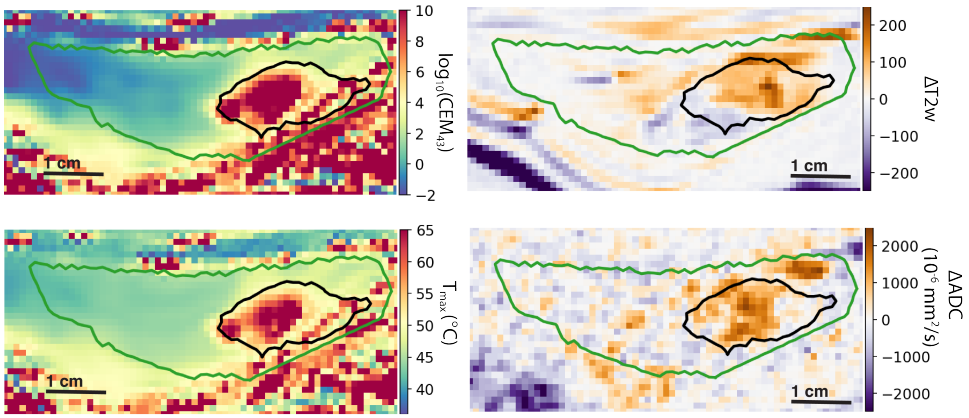

Fig. 2. Examples of multi-parametric images used as inputs to the classification model from a single coronal slice of subject 3, a) $\log_{10}$ of CTD (CEM_{43}), b) maximum absolute temperature achieved (°C), and the post- minus pre-ablation difference maps of c) T2w images, d) ADC maps ($\times \mathbf{10^{-6}}$ $\mathbf{mm^2}$/s). The ground-truth histology label is represented by the pink contour and the quadriceps segmentation is represented by the green contour.

slice in the *in vivo* MR images. Stacking the individual histology annotations of treated tissue provided a histology-derived volumetric label of the treated volume, co-registered with the follow-up images. See the Appendix for more details on 3D histology-to-MR registration. The diffeomorphisms from longitudinal registration and histology registration were composed, resulting in a voxel-wise alignment between the treatment day T2w images, ADC maps, MRTI data, and the histology necrosis ground-truth label.

D. MPMR Classification Models

To develop the supervised MPMR biomarkers, the dataset comprised all voxels in 3D anatomical segmentations of the quadriceps muscles (Seg3D2, SCI Institute [31], University of Utah) from all four subjects. The number of voxels per subject is given by "Quadriceps Vol." in Table II. The inclusion of the entire quadriceps for supervised learning resulted in 12.3% positive class voxels on average, as assigned by the 3D registered histology label ({C0: viable, C1: necrotic}). The number of positive class voxels (1-mm isotropic) per subject is given by "Histology Vol." in Table II. Four non-contrast MR image volumes were used for classifier inputs: 1) CTD, 2) maximum temperature projection in time (MTP), 3) T2w images, and 4) ADC maps (Figure 2). Overlapping in-plane 3×3 neighborhoods around each voxel were extracted from each of the four features, resulting in 36 total input features. For classifier training, input features were normalized to the mean and scaled to unit variance (scikit-learn, StandardScalar), with training and test sets normalized separately. First, optimal classifier hyper-parameters were identified with 5-fold cross-validation in the combined dataset from all subjects. Second, the hyper-parameter tuned classifiers were trained and validated in a leave-one-out (LOO) approach, where classifiers were trained on three subjects and validated on the fourth, for a total of n=4 folds. Details of classifier development are below.

Three binary classifiers were explored for the development of the voxel-wise MPMR biomarker: a logistic regression classifier (LRC; "liblinear" solver [34]), a support vectors machine classifier (SVMC; SVC implementation, "rbf" kernel [34]), and a random forest classifier (RFC; [34]). The LRC model was chosen as an option for a model with minimal complexity, while SVMC with radial basis functions and RFC were implemented in order to learn possible nonlinear relationships in the MPMR feature space. Random forests, with the capacity to create models with high complexity, are also prone to over-fitting data. The SVMC algorithm is also capable of higher complexity than logistic regression; however, is less susceptible to over-fitting in small datasets than RFCs.

For hyper-parameter tuning, the full dataset (N=67,235 voxels) was implemented for a 5-fold cross-validation grid search of classifier hyper-parameters. Folds were generated with a 70:30 random split, maintaining the original dataset imbalance (scikit-learn, StratifiedShuffleSplit [34]). To correct for class imbalance in the dataset, a balanced scorer was used for training the LRC and SVMC models, and balanced sampling was used for each decision tree batch for training the RFC model. Optimal hyper-parameters were selected to minimize the classifier complexity to reduce over-fitting while maximizing the 5-fold average Dice score (f1-score) in each classifier.For the LRC, an l1-penalty with C-parameter=0.05 was selected.For the RFC, the selected hyper-parameters were min_samples-split = 0.005, n_estimators = 50, and max_depth = 10. The selected hyper-parameters for the SVMC were C-parameter = 0.01 and gamma = 0.0001.

To test the accuracy and generalizability of the MPMR biomarkers, the classifiers with optimized hyper-parameters were re-trained in a LOO strategy. For the LOO analysis, the training data consisted of voxels from of three subjects and the validation dataset consisted of voxels from the fourth subject. Optimal thresholds for the output probabilities were selected to maximize the Dice score in the training data. Trained models and optimal thresholds were fit to the validation dataset to obtain probabilities for each voxel. LOO training was repeated for each validation subject for a total of n=4 folds. Similarly, the optimized thresholds for the clinical CTD biomarker were chosen to maximize the Dice score in the LOO training data sets. The binary NPV segmentation is dependent on a user-defined threshold between the hypo-intense NPV and the surrounding hyper-intense rim. For comparison to numerically continuous biomarkers, the NPV inter-user variability was simulated by applying a Gaussian blur to the expert-evaluated NPV segmentations (3-voxel kernel, $\sigma = 1$ voxel), generating a continuous 3-mm boundary varying from 0 to 1. For the clinical NPV and 240 CEM_{43} metrics, thresholds of 0.5 and 240 CEM_{43} were used respectively.

E. Biomarker Evaluation

Precision, recall, and Dice scores were compared for all MPMR and clinical biomarkers for each LOO fold. The Dice metric was chosen to score overall biomarker performance given the severe imbalance of the dataset (11% positive class). This metric provides a harmonic mean of the recall and precision scores. To assess prediction accuracy with more clinically relevant metrics, in each subject the percent difference in volume and the mean-distance-to-agreement (MDA) between the predicted thermal necrosis lesion and the histological necrosis label was calculated for each trained biomarker. See Appendix for MDA definition.

III. Results

A. Histological Registration

The acute NPV segmentations were representative of clinician segmentations, achieving an average score of 4.75 out of 5 in the four subjects by the expert radiologist. The volume-conserving longitudinal MR registration error was 1.26 ± 0.52 mm and the average volume change of tissue during registration was 0.24 ± 0.12 percent relative to the starting volume. The average total registration error between acute MR imaging and histology was 1.00 ± 0.13 mm.

To demonstrate MR-to-histology registration, the "delayed" NPV acquired 10 minutes prior to euthanasia 3 days post-ablation is compared to the registered H&E in Figure 3 in

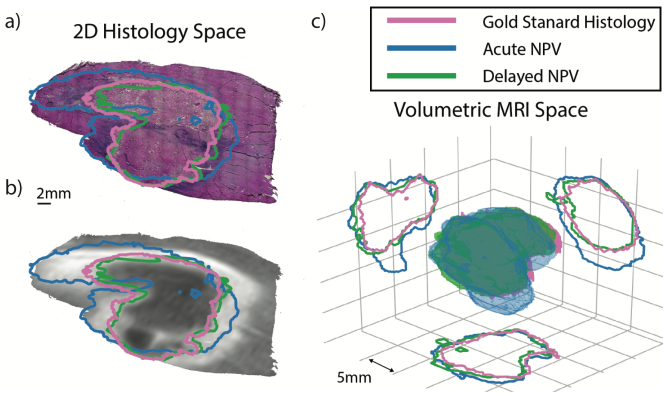

Fig. 3. a) A digital H&E image of a section acquired from subject 4, acquired at 2.5× magnification, and b) the registered resampled 3D MR CE-MRI image acquired during follow-up (delayed) imaging acquired 3 days following ablation. The ground-truth expert segmentation of histological necrosis is shown in pink, with the delayed NPV segmentation indicated in green, and the acute NPV segmentation evaluated in this study indicated in blue. c) The volumetric overlay of all three segmentations demonstrates good agreement between histological necrosis and the delayed NPV in all three views (grid lines = 5 mm).

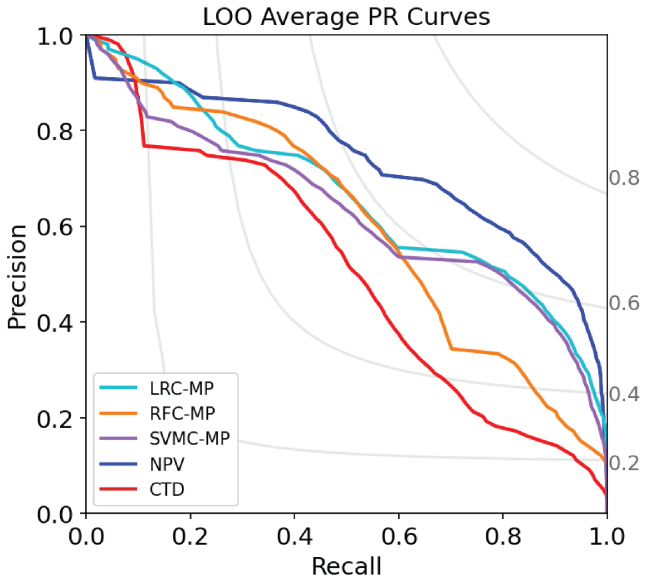

Fig. 4. Average test data Precision Recall-curves of MPMR classifiers and clinical NPV and CTD biomarkers (n=4). Dice iso-score contours are denoted by light-grey annotations.

subject 4. The delayed NPV segmentation on CE-MRI (Figure 3b) correlates to the region of coagulative necrosis in the registered H&E section (Figure 3a), which is characterized by pale eosin staining (pallor) of muscle and tumor tissue. A volumetric comparison of the longitudinally-registered 3D histology necrosis label (pink), delayed NPV (green), and acute NPV (blue) is shown in Figure 3c. In 2D cross-section profiles of the volumes in each dimension, the pink necrosis contour aligns well with the *in vivo* delayed NPV segmentation. As expected, the acute NPV, acquired 40 minutes post-ablation, is larger than the delayed NPV and necrosis label.

B. MPMR Leave-One-Out Classifier Training

MPMR classifiers with optimized hyper-parameters were trained in a LOO approach with four folds. All voxels from one subject served as the test dataset for each fold.The average precision-recall (PR) curves across all LOO folds for the MPMR biomarkers are shown in Figure 4 and compared to the clinical NPV (blue) and CTD metric (red) curves. All MPMR classifiers out-perform the CTD model, with the SVMC and LRC predictors surpassing a Dice score of 0.6.

In Figure 5, the training (dashed) and testing (solid) Dice scores are compared in each subject across MPMR classifiers. All MPMR classifiers achieved higher Dice scores than the optimized CTD in all subjects. Although RFC achieved the highest Dice score in the full dataset, the average RFC Dice score (0.58, Table III) was the lowest of all MPMR classifiers in the LOO training. Additionally, the test Dice score was on average 25.9% lower than the training scores for the RFC indicating potential data over-fitting. In contrast, testing scores were only 8.0% and 4.0% lower than the training scores for the LRC and SVMC models on average. Therefore, trained LRC and SVMC models generalized well to new subjects in each fold.

Precision, recall, and Dice scores for all LOO folds for all clinical and MPMR biomarkers are summarized in Figure 6 and Table III. Acute NPV, the current gold-standard MR biomarker of thermal ablation, achieved the highest average Dice score of 0.68 $\pm$ 0.10 (n=4), followed by the SVMC (0.63 $\pm$ 0.05) and LRC (0.63 $\pm$ 0.04), then RFC (0.58 $\pm$ 0.15). MPMR classifiers were slightly more precise than the acute NPV at the cost of reduced recall scores. Interestingly, the RFC achieved the highest average precision of all clinical and MPMR biomarkers of 0.62 $\pm$ 0.21.

In subjects 1 and 2, the LRC improved Dice over acute NPV by 2.2 and 8.9%, respectively; and in subject 2, SVMC improved Dice over acute NPV by 1.8%. However, acute NPV had the greatest Dice in subjects 3 and 4. In comparison to the optimized CTD, all three MPMR classifiers improved Dice scores in subjects 1-4 by 941.4%, 7.8%, 46.7%, and 13.7% on average across n=3 classifiers. Furthermore, the inter-subject variability in Dice scores, as measured by the standard deviation in Table III was the greatest for the 240 CEM_{43} and CTD models (0.27 and 0.26, respectively) and lowest for the SVMC and LRC models (0.05 and 0.04, respectively).

C. Optimized CTD Leave-one-Out Training

The optimal CTD thresholds calculated in each fold were 119.1, 120.0, 26.3, and 218.5 CEM_{43}, which are all lower than the accepted 240 CEM_{43} threshold. The 240 CEM_{43} and optimized CTD thresholds achieved similar average Dice scores of 0.43 $\pm$ 0.27 and 0.42 $\pm$ 0.26, respectively. LOO training for optimized CTD led to variable results, where Dice increased by 0.02 on average in 3/4 subjects and decreased by 0.09 in subject 3 (Table III). CTD optimization resulted in higher recall and lower precision as the CTD thresholds in the LOO training datasets were reduced. Inconsistent train-test score differences across folds (Figure 5) indicate that

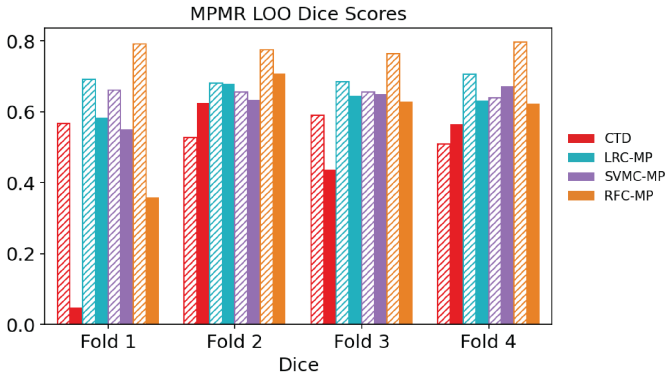

Fig. 5. Dice performance scores of the LOO training in 4 folds/subjects. Dice scores for the training data set (dashed) and test data set (solid) are shown for each MPMR classifier.

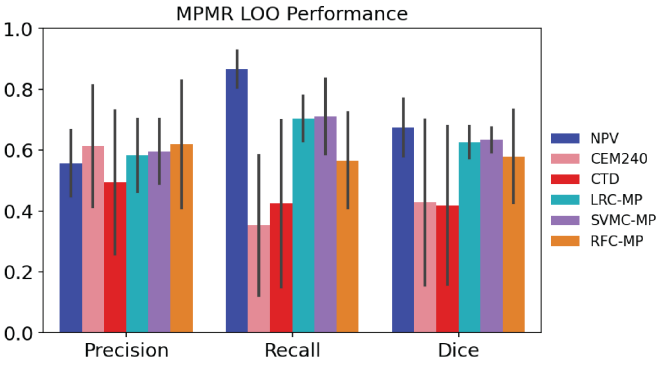

Fig. 6. Average Precision, Recall and Dice scores for clinical biomarkers and MPMR supervised classifiers. Error bars represent standard deviation across folds (n=4).

optimized thermal dose thresholds in training subjects do not generalize consistently to new "unseen" subjects.

D. Spatial Accuracy

Spatial performance was assessed in the test data of each subject in the corresponding LOO fold. Representative examples of clinical and MPMR model predictions in subject 4 are shown in Figure 7. The volume of intersection between the label and prediction is depicted by the opaque green volume, which is positively correlated to the Dice score. The average absolute value of the percent differences between predicted and labeled necrosis volumes were 59.5, 54.7, and 50.0% for the acute NPV, 240 CEM_{43} and CTD biomarkers, respectively. The absolute differences were 34.3, 30.4, and 28.8% for the LRC, SVMC, and RFC biomarkers, respectively (Table III).

High recall of the acute NPV is due to the consistent overestimation of the histological necrosis boundary (Figure 7a). Conversely, the 240 CEM_{43} and optimized CTD predictions did not cover the full extent of thermal damage (Figure 7b), or the optimized CTD threshold severely overestimated the necrotic volume in the LOO fold for subject 3. The LRC and SVMC prediction contours (Figure 7c-d) both align more closely with the optimized CTD contour than the RFC (Figure 7e), albeit with higher sensitivity than CTD. The RFC provides the most conservative prediction of the necrotic volume and the least variance in volume difference errors across folds ($\pm$ 4.5%).

The 5-mm grid in Figure 7 demonstrates the magnitude of contour errors for all predictions. These errors were quantified with the mean-distance-to-agreement (MDA) metric in the whole volume in each subject and are summarized in Table III. The 240 CEM_{43} and CTD metrics had the greatest MDA error across all subjects, ranging from 1.4-6.5 mm. Acute NPV MDA error was lowest overall with a range of 1.2-1.3 mm for all subjects. Of the MPMR classifiers, MDA was lowest for RFC (1.4-1.7 mm). The SVMC and LRC classifiers were similar with MDA errors ranging from 1.2-2.0 mm.

.

IV. Discussion

This study investigated using supervised machine learning with non-contrast MPMR biomarkers to provide a clinically useful and accurate acute prediction of the thermal lesion in an intra-muscular VX2 tumor model. Trained MPMR classifiers outperformed the clinical 240 CEM_{43} and optimized 120 CEM_{43} CTD prediction. Additionally, MPMR biomarkers achieved higher Dice scores than the acute NPV, the current clinical reference standard in 50% of subjects in a LOO training strategy. Although the contrast-enhanced NPV is an accurate biomarker of acute perfusion loss, perfusion changes can be temporary or transient. Acute perfusion also does not adequately measure delayed cellular apoptosis [14]. Finally, administering contrast agent to patients effectively ends the treatment, increasing patient costs if a second MRgFUS ablation treatment is required. Although MRTI provides an estimate of ablation volume during treatment, thermal metrics tend to underestimate the treatment effect. The voxel-wise supervised learning approach presented here is a promising proof-of-concept for a non-contrast, accurate MR biomarker for acute and intra-treatment assessments of MRgFUS in patients.

When considering oncological targets, a major concern is leaving a region of viable tumor untreated, leading to disease progression [6]. Such is the case for acute NPV, which overestimated the final histologically necrotic volume in this study. Although acute NPV achieved the highest mean recall, over-estimation poses a high risk of tumor recurrence. Knowledge of a consistent safety margin is valuable for clinician assessment of treatment success. In this study, the mean NPV MDA was 1.2 mm across all subjects with minimal variability ($\pm$ 0.0 mm), potentially serving as a safety margin for treatment assessment. Conversely, the thermal and MPMR metrics were highly specific but generally under-predicted necrosis. This bias could lead to unnecessary ablation of healthy tissue. However, a similar safety margin could be applied to an MPMR biomarker, since they each achieved low variability in MDA scores across subjects ($\pm$ 0.1-0.3 mm).

In contrast, the 240 CEM_{43} clinical biomarker underestimated histological necrosis in all subjects on average by 50%, with clinically significant inter-subject variability in MDA (1.3-6.5 mm). The 240 CEM_{43} metric has been previously reported to correlate well with thermal necrosis in rabbit

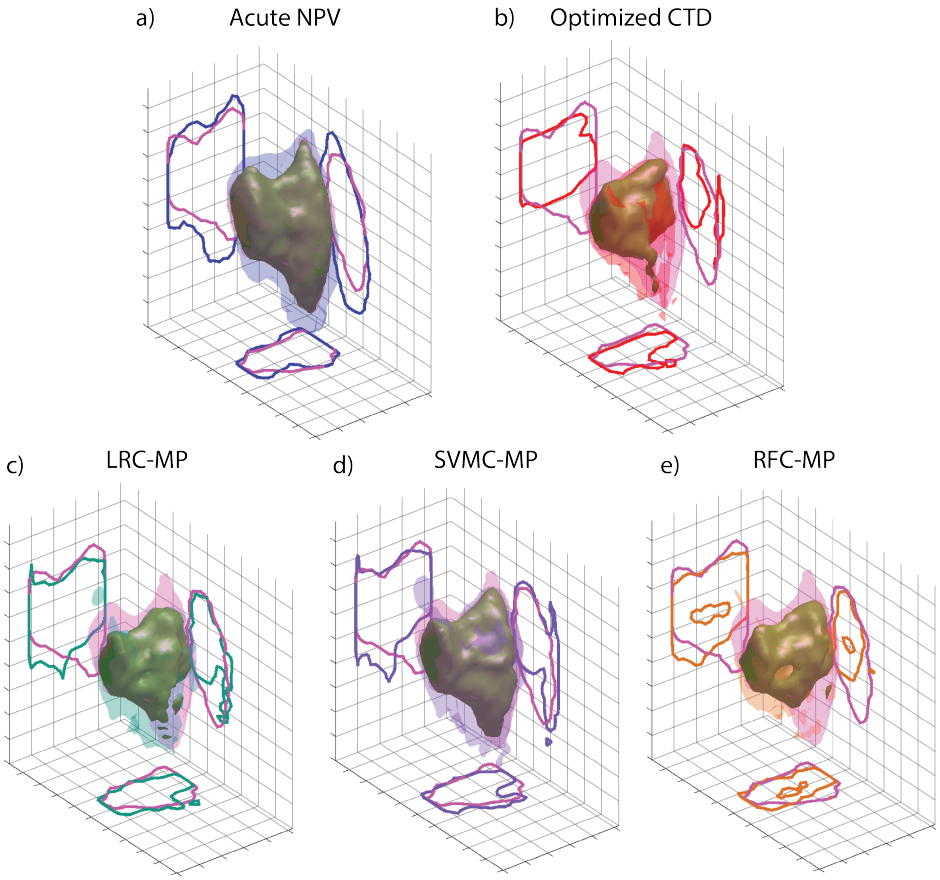

Fig. 7. 3D plots of the histological necrosis label (pink) and predictor segmentations in subject 4. The segmentations are visualized as overlapping volumes, where the volume of the intersection is represented by the opaque green volume. A 2D projection of the histological label and prediction segmentation is given in each dimension (grid-spacing = 5 mm).

muscle and VX2 tumor tissue in previous studies; however, these studies investigated non-overlapping FUS sonications and did not quantify thermal dose metrics with histological necrosis directly [11], [35]. A multi-point sonication ablation and histological correlation comparable to ours was performed by Wiljemans et al. [14], where the MR-derived 240 CEM_{43} area was compared to the necrosis area in a H&E cross-section at several time-points post-sonication. They found that the histological necrosis volume 3 days after ablation was up to 6.0 times larger than predicted by the 240 CEM_{43}. This finding corroborates our presented work, where a lower thermal dose threshold better predicted the final tissue lesion.

We found that the optimal threshold ranged from 26-218 CEM_{43} across training data sets. The lowest threshold correlates well with early studies in rabbit muscle by McDannold et al. which concluded that 31.2 CEM_{43} causes visible lesion damage as measured by contrast-enhanced T1w and T2w images acquired within 2 hours of sonications [36], [37]. Possible causes of discrepancy between optimal thresholds across studies are numerous, including the ground-truth metric for necrosis, the timing of post-sonication assessment, and volumetric versus cross-sectional correlation metrics. Although the "optimal" CTD for muscle tissue cannot be concluded by this study alone, our results are consistent with previous findings that 240 CEM_{43} is a conservative threshold of tissue necrosis, particularly in multi-point and large volume sonications [1], [38], [39].

The CEM_{43} metric is sensitive to errors in the underlying temperature data. Absolute temperature errors above 43°C due to inaccuracies and limitations of PRF MRTI are exponentially exacerbated in CEM_{43} calculations. Sources of error in MRTI data acquisition include unknown or inaccurate baseline temperatures, the tissue-specific water proton shift coefficient [8], and motion and respiration artifacts [40]. Bulk anatomical motion throughout hour-long ablation procedures can also cause misplacement of the thermal dose when it is calculated from serial MRTI acquisitions. This motion was corrected with MTRI magnitude image registration to the high-resolution post-treatment T1w image in the present study; however, current clinical workflows may not perform this type of correction. Additionally, indirect thermal damage in highly perfused tissues, such as the liver, may result from perfusion of higher temperature "ablated" blood to neighboring tissue [4]. The prior-baseline approach implemented in this study aimed to correct for progressive localized heat accumulation which is not currently measured in clinical implementations of the PRF method [29]. Increased recall from applying the prior baseline correction improved the Dice score for 240 CEM_{43} in subjects 1 and 3 by 20 % and 15%, respectively. Despite this correction, the clinical thermal dose prediction remained

TABLE III
PREDICTOR PERFORMANCE SCORES IN SUBJECTS 1-4

	Subject	NPV	CEM240	CTD	LRC-MP	SVMC-MP	RFC-MP
Dice	1	0.57	0.02	0.05	0.58	0.55	0.36
	2	0.62	0.60	0.63	0.68	0.63	0.71
	3	0.73	0.53	0.44	0.64	0.65	0.63
	4	0.78	0.56	0.57	0.63	0.67	0.62
	Average	**0.68** +/- 0.10	**0.43** +/- 0.27	**0.42** +/- 0.26	**0.63** +/- 0.04	**0.63** +/- 0.05	**0.58** +/- 0.15
Precision	1	0.43	0.32	0.25	0.46	0.42	0.31
	2	0.51	0.71	0.67	0.57	0.57	0.64
	3	0.64	0.70	0.34	0.65	0.66	0.76
	4	0.66	0.73	0.72	0.71	0.68	0.75
	Average	**0.56** +/- 0.11	**0.62** +/- 0.20	**0.50** +/- 0.23	**0.60** +/- 0.11	**0.58** +/- 0.12	**0.62** +/- 0.21
Recall	1	0.86	0.01	0.03	0.79	0.81	0.42
	2	0.81	0.52	0.59	0.84	0.71	0.79
	3	0.85	0.42	0.62	0.64	0.64	0.53
	4	0.95	0.46	0.46	0.57	0.66	0.53
	Average	**0.87** +/- 0.06	**0.35** +/- 0.23	**0.43** +/- 0.27	**0.71** +/- 0.13	**0.71** +/- 0.08	**0.57** +/- 0.16
MDA [mm]	1	1.3	5.6	6.5	1.2	1.3	1.6
	2	1.2	1.5	1.4	2.0	2.0	1.4
	3	1.2	4.1	2.5	1.7	1.5	1.7
	4	1.2	1.7	1.7	1.5	1.4	1.5
	Average	**1.2** +/- 0.0	**3.2** +/- 2.0	**3.0** +/- 2.4	**1.6** +/- 0.3	**1.6** +/- 0.3	**1.5** +/- 0.1
% Volume Difference	1	100.8	-89.4	-96.1	70.5	92.7	32.3
	2	58.9	-11.5	-26.9	47.1	23.6	22.0
	3	33.3	82.3	-39.8	-0.1	-2.1	-30.1
	4	45.1	-35.7	-37.2	-19.2	-3.0	-29.5
Average of Absolute % Volume Difference		59.5	54.7	50.0	34.3	30.4	28.5

% Volume Difference = $100 \times (predicted - label)/label$

an inadequate overall predictor of histological necrosis.

In comparison to CTD, changes in innate MR properties of damaged or thermally coagulated tissues do not require *a priori* knowledge of tissue properties nor rely on thermal history to compute. ADC maps and T2w images are sensitive to the presence or absence of fluid, which can be altered as a result of increased edema and immune infiltration or immediate coagulative necrosis after thermal ablation. Compared to thermal metrics alone, the inclusion of T2w and ADC information increased sensitivity and reduced false positives, which improved the localization of positively classified voxels to the treatment region. In Subject 1, CTD biomarkers severely underestimated the histological ablation volume. MPMR classifiers notably improved recall and Dice scores in this subject, from 0.05 to 0.58 for the LRC. In the remaining subjects, the MPMR biomarker outperformed the CTD-based predictions of thermal necrosis by 18% on average, as assessed by Dice scores.

The results demonstrate that a non-contrast, multi-parametric biomarker trained on multiple subjects may be less sensitive to errors in baseline temperature or MRTI imaging accrued during treatment. Pre- and post-treatment differences in innate MR contrasts are sensitive to transient tissue changes which correlate to final treatment volumes, such as acute perfusion loss or the presence of edema. These changes may indicate indirect thermal injuries which are not modeled by the CEM_{43} thermal dose equation. Multi-parametric imaging can provide a more robust characterization and prediction of thermal damage following FUS treatments than MRTI alone.

This study investigated three supervised machine learning classifiers for the non-contrast MPMR biomarker. Of these three models, the SVMC with radial basis function kernels and the LRC models achieved similar average precision, recall and Dice scores in the LOO analysis. These models also demonstrated low variability in Dice scores across LOO folds and small differences in test and train scores. Overall, these results indicate that LRC and SVMC models can generalize well to multiple test cases and were not over-fit during training. RFC performance was relatively diminished in the LOO analysis. Despite minimizing RFC complexity during hyper-parameter tuning, the average 26% decrease in test scores from the training score may indicate model over-fitting. The RFC model ultimately generated the highest precision of all clinical and MPMR biomarkers. As exemplified in Figure 7 and Supplementary Figure 1 (S1), the LRC and SVMC models were prone to misclassify the same negative-labeled voxels as the optimized CTD biomarker, indicating that these models were more susceptible to underlying errors in the MRTI data than the RFC model. Although the RFC Dice was the lowest of all MPMR classifiers, a more conservative and precise estimate of the lesion may be preferred clinically. An unbalanced scorer, such as the likelihood ratio, which can prioritize sensitivity or specificity, may also be used to tune classifiers to the desired clinical outcomes.

Overall, MPMR improvement over the clinical reference standard contrast-enhanced metric was not achieved; however, the MPMR sequences could be further optimized for the expected ranges of change in T2w and ADC maps. In DWI, optimizing the TE for the targeted tissue improves ADC map SNR. Acquiring multiple b-values can allow for quantitative separation of perfusion and diffusion-dominant effects [41]. Additionally, T2w images are qualitative, and although normalized for classifier training, the weighting could vary across subjects and scanners and over time. Quantitative T1 and T2 mapping are becoming increasingly fast and clinically applicable with techniques such as multi-tasking, fingerprinting [42], and multi-pathway multi-echo acquisitions [43]. This feasibility study demonstrates the potential for a more accurate and generalizable biomarker for post-ablation treatment assessment.

A. Limitations

The volumetric histological processing utilized in this study was costly and intensive, limiting the number of subjects and amount of data for this analysis. However, a quantified comparison of delayed NPV to registered histological outcomes may allow follow-up CE-MRI to serve as a surrogate label for supervised learning in the future. Due to the limited data acquired for this study, a voxel-wise approach to machine learning was implemented in this study. A 3×3 voxel neighborhood patch was used for each input feature to provide the models with contextual information and promote robust learning despite small registration errors ($<$3 mm). Although a 2D patch was used, the prediction from the classifiers was only for a single voxel. Voxel-wise predictions tend to be noisy and cannot fully leverage spatial connectivity and context to improve the prediction. For small sample sizes, predictions could be improved using connected components; however, this method requires knowing how many different locations have been treated, which may vary depending on near- and far-field heating. However, larger datasets from more subjects would allow for CNN or deep learning applications which could produce more connected 2D or 3D maps that are less susceptible to image noise. These complex algorithms may also learn post-treatment differences in multiple tissue types, increasing the likelihood of applying transfer learning to several applications.

Sources of error related to MR registration to histological outcomes are multi-fold. Registration errors of MR and histological registration were on the order of an MRTI voxel (1-mm isotropic); however, errors can vary across and within subjects and directly impact classifier scores. The longitudinal MR registration has been rigorously validated in Zimmerman *et al.* [33]. Although localized tissue swelling after ablation was not directly accounted for, anatomical landmarks near the region of the ablation, such as blood vessels, were utilized to determine target registration error. Single-voxel shifts in pre- and post-ablation registered ADC and T2w maps can introduce errors in difference map calculations. However, treatment effect classifiers often rely on pre- and post-treatment differences as inputs. Pre-processing with context-based radiomic filters or implementation of a CNN may reduce the impact of noisy or imperfectly registered MPMR data.

V. Conclusion

The motivation behind this work was to investigate non-contrast MPMR biomarkers which can predict histological treatment outcomes. The novel data set presented here has histology labels directly registered with the multi-parametric images with a spatial accuracy of approximately 1.0 mm. Histological registration facilitated supervised machine learning to fully leverage the information available in MPMR imaging for acute MRgFUS assessment. A voxel-wise logistic regression and support vector machine classifier using immediate post-treatment ADC, T2w and MR thermometry as input data performed similarly to the gold-standard contrast-enhanced non-perfused volume (NPV). MPMR assessment may provide more accurate predictions of thermal necrosis using non-contrast imaging methods. Future work includes collecting expansive quantitative multi-parametric data to optimize MR imaging protocols for the greatest predictive accuracy. Non-contrast MR biomarkers will allow more flexible non-invasive treatments, improving the clinical viability of MRgFUS treatments of localized tumors.

Acknowledgment

The authors acknowledge the direct financial support for the research reported in this publication provided by the Huntsman Cancer Foundation and the Experimental Therapeutics Program at Huntsman Cancer Institute; the authors also acknowledge support by the National Cancer Institute of the National Institutes of Health under Award Numbers P30CA042014, R37CA224141, R01CA259686, S10OD018482, and R03EB029204.

Appendix

A. 3D Histology-to-MR Registration Implementation

Tissue preparation for histopathology is divided into three main steps that introduce deformation into the tissue: tissue excision (D1), gross slicing (D2), and microtome sectioning (D3). Restoring the spatial relationship between MR and histopathology requires correcting deformations from each step of the destructive histopathology pipeline: block-face registration (R1), *ex vivo* registration (R2), and *in vivo* registration (R3). These steps are graphically outlined in Fig. 8. To complete the reconstructive pipeline, several imaging steps were required between tissue excision and sectioning. Briefly, the harvested quadriceps tissue was inked for MR orientation, fixed in formalin for one week, and embedded in an agar hydrogel (2.5%) for *ex vivo* MR imaging (T1-weighted VIBE, 0.5 x 0.5 x 1 mm resolution). The agar block was then grossly sliced along the head-foot axis of the ex vivo MR imaging with an industrial-grade meat slicer. The 3-mm thick gross slices ("blocks") were re-inked to maintain MR orientation, then embedded in whole-mount paraffin blocks for sectioning.

Each paraffin block was sectioned in 10 μm increments with a microtome, with digital block-face images acquired every 50

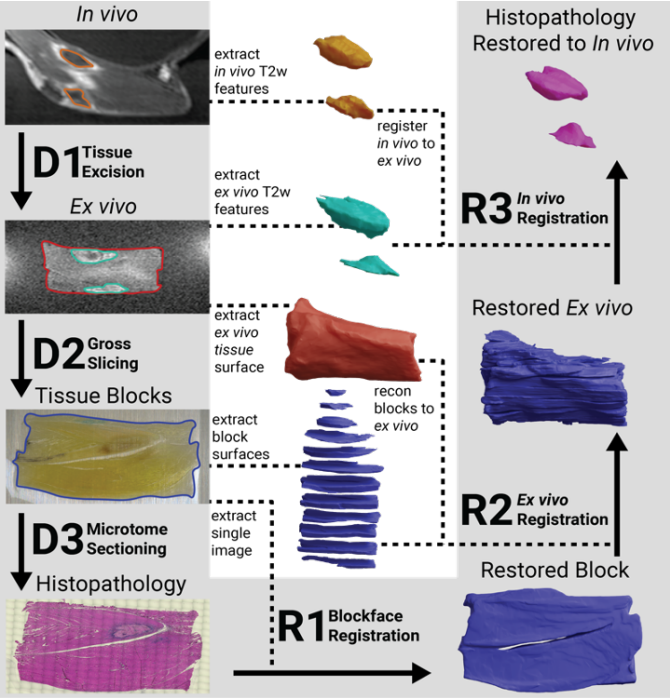

Fig. 8. Workflow for registering *in vivo* MR with volumetric histopathology. The three main steps for tissue processing are D1) tissue excision, D2) gross slicing, and D3) microtome sectioning. Dotted lines indicate information drawn from digital imaging during tissue processing that facilitates registration. The three registration steps to restore the relationship between histopathology and in vivo MR are R1) block-face registration, R2) *ex vivo* registration, and R3) *in vivo* registration.

μm, and sections acquired for H&E staining every 250 μm. H&E-stained sections were imaged with a digital bright-field microscope at 2.5 magnification (0.0076 mm resolution). For R1, each digital H&E slide was registered to the corresponding block-face image via intensity-based affine and subsequent multi-scale registration, resulting in a diffeomorphism between histology and block-face images. In R2, the tissue outline of each block-face image was segmented using a custom-trained V-Net neural network, and stacked to generate a 3D reconstruction of each tissue block. Reconstructed tissue blocks were sequentially registered together, starting with the center tissue block and working outwards, using surface-based registration [25], [44] to determine an affine and diffeomorphism for each block. The fully reconstructed block-face image surface was registered to the segmented tissue surface of the *ex vivo* MR image. In R3, anatomical and treatment-related features in corresponding *in vivo* and *ex vivo* T2-weighted images were used to generate 3D surfaces, which were registered together with surface-based registration [44]. The series of transformations and diffeomorphisms were composed to yield a single diffeomorphism between each histology section and *in vivo* MR, similar to [25].

B. Mean-distance-to-agreement (MDA)

The MDA is a measure of the mean spatial proximity of two segmentation contours, A and B. It is calculated by:

$$\frac{\Sigma_{i=1}^{n} d(A,B) + \Sigma_{i=1}^{m} d(B,A)}{m+n} \tag{3}$$

where $d(A,B)$ is the minimum Euclidean distance between any voxel m on the contour of segmentation A to any voxel n on the contour of segmentation B, and vice versa [45].

REFERENCES

[1] N. McDannold *et al.*, "Uterine leiomyomas: Mr imaging–based thermometry and thermal dosimetry during focused ultrasound thermal ablation," *Radiology*, vol. 240, no. 1, pp. 263–272, 2006.
[2] M. S. Breen *et al.*, "Mri-guided thermal ablation therapy: Model and parameter estimates to predict cell death from mr thermometry images," *Annals of biomedical engineering*, vol. 35, no. 8, pp. 1391–1403, 2007.
[3] G. C. Van Rhoon *et al.*, "Cem43° c thermal dose thresholds: a potential guide for magnetic resonance radiofrequency exposure levels?" *European radiology*, vol. 23, no. 8, pp. 2215–2227, 2013.
[4] S. J. Hectors *et al.*, "Mri methods for the evaluation of high intensity focused ultrasound tumor treatment: Current status and future needs," *Magnetic resonance in medicine*, vol. 75, no. 1, pp. 302–317, 2016.
[5] S. A. Sapareto and W. C. Dewey, "Thermal dose determination in cancer therapy," *International Journal of Radiation Oncology• Biology• Physics*, vol. 10, no. 6, pp. 787–800, 1984.
[6] S. J. Hectors *et al.*, "Multiparametric mri analysis for the identification of high intensity focused ultrasound-treated tumor tissue," *PloS one*, vol. 9, no. 6, p. e99936, 2014.
[7] N. M. Hijnen *et al.*, "The magnetic susceptibility effect of gadolinium-based contrast agents on prfs-based mr thermometry during thermal interventions," *Journal of therapeutic ultrasound*, vol. 1, no. 1, p. 8, 2013.
[8] V. Rieke and K. Butts Pauly, "Mr thermometry," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 2, pp. 376–390, 2008.
[9] R. T. Peters, R. S. Hinks, and R. M. Henkelman, "Ex vivo tissue-type independence in proton-resonance frequency shift mr thermometry," *Magnetic resonance in medicine*, vol. 40, no. 3, pp. 454–459, 1998.
[10] C. Diederich *et al.*, "Transurethral ultrasound applicators with directional heating patterns for prostate thermal therapy: in vivo evaluation using magnetic resonance thermometry," *Medical Physics*, vol. 31, no. 2, pp. 405–413, 2004.
[11] N. McDannold *et al.*, "Mri evaluation of thermal ablation of tumors with focused ultrasound," *Journal of magnetic resonance imaging*, vol. 8, no. 1, pp. 91–100, 1998.
[12] J. D. Hazle *et al.*, "Mri-guided thermal therapy of transplanted tumors in the canine prostate using a directional transurethral ultrasound applicator," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 15, no. 4, pp. 409–417, 2002.
[13] M. Kangasniemi *et al.*, "Multiplanar mr temperature-sensitive imaging of cerebral thermal treatment using interstitial ultrasound applicators in a canine model," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 16, no. 5, pp. 522–531, 2002.
[14] J. W. Wijlemans *et al.*, "Evolution of the ablation region after magnetic resonance–guided high-intensity focused ultrasound ablation in a vx2 tumor model," *Investigative radiology*, vol. 48, no. 6, pp. 381–386, 2013.
[15] B. Z. Fite *et al.*, "Magnetic resonance imaging assessment of effective ablated volume following high intensity focused ultrasound," *PloS one*, vol. 10, no. 3, p. e0120037, 2015.
[16] M. W. Dewhirst *et al.*, "Basic principles of thermal dosimetry and thermal thresholds for tissue damage from hyperthermia," *International journal of hyperthermia*, vol. 19, no. 3, pp. 267–294, 2003.
[17] P. S. Yarmolenko *et al.*, "Thresholds for thermal damage to normal tissues: an update," *International Journal of Hyperthermia*, vol. 27, no. 4, pp. 320–343, 2011.
[18] Y. Huang *et al.*, "Predicting lesion size by accumulated thermal dose in mr-guided focused ultrasound for essential tremor," *Medical physics*, vol. 45, no. 10, pp. 4704–4710, 2018.
[19] R. M. Jones *et al.*, "Accumulated thermal dose in mri-guided focused ultrasound for essential tremor: repeated sonications with low focal temperatures," *Journal of neurosurgery*, vol. 132, no. 6, pp. 1802–1809, 2019.
[20] N. McDannold, P. J. White, and G. R. Cosgrove, "Mri-based thermal dosimetry based on single-slice imaging during focused ultrasound thalamotomy," *Physics in Medicine & Biology*, vol. 65, no. 23, p. 235018, 2020.
[21] J. Pichat *et al.*, "A survey of methods for 3d histology reconstruction," *Medical image analysis*, vol. 46, pp. 73–105, 2018.

[22] V. M. Ferreira *et al.*, "Non-contrast t1-mapping detects acute myocardial edema with high diagnostic accuracy: a comparison to t2-weighted cardiovascular magnetic resonance," *Journal of cardiovascular magnetic resonance*, vol. 14, no. 1, p. 42, 2012.

[23] J. C. Plata *et al.*, "A feasibility study on monitoring the evolution of apparent diffusion coefficient decrease during thermal ablation," *Medical physics*, vol. 42, no. 9, pp. 5130–5137, 2015.

[24] D. L. Langer *et al.*, "Prostate cancer detection with multi-parametric mri: Logistic regression analysis of quantitative t2, diffusion-weighted imaging, and dynamic contrast-enhanced mri," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 30, no. 2, pp. 327–334, 2009.

[25] B. E. Zimmerman *et al.*, "Histology to 3d in vivo mr registration for volumetric evaluation of mrgfus treatment assessment biomarkers," 2020.

[26] C. Zhu *et al.*, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 4, pp. 550–560, 1997.

[27] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[29] R. R. Bitton *et al.*, "Improving thermal dose accuracy in magnetic resonance-guided focused ultrasound surgery: Long-term thermometry using a prior baseline as a reference," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 181–189, 2016.

[30] S. Johnson *et al.*, "Validation of hybrid angular spectrum acoustic and thermal modelling in phantoms," *International Journal of Hyperthermia*, vol. 35, no. 1, 2018.

[31] CIBC, 2016, seg3D: Volumetric Image Segmentation and Visualization. Scientific Computing and Imaging Institute (SCI), Download from: http://www.seg3d.org.

[32] P. Hasgall *et al.*, "IT'IS database for thermal and electromagnetic parameters of biological tissues, version 3.0, september 1st; 2015," 2015.

[33] B. Zimmerman *et al.*, "Learning multiparametric biomarkers for assessing mr-guided focused ultrasound treatments," *IEEE Transactions on Biomedical Engineering*, 2020.

[34] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] J. D. Hazle, R. J. Stafford, and R. E. Price, "Magnetic resonance imaging-guided focused ultrasound thermal therapy in experimental animal models: correlation of ablation volumes with pathology in rabbit muscle and vx2 tumors," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 15, no. 2, pp. 185–194, 2002.

[36] N. J. McDannold *et al.*, "Usefulness of mr imaging-derived thermometry and dosimetry in determining the threshold for tissue damage induced by thermal surgery in rabbits," *Radiology*, vol. 216, no. 2, pp. 517–523, 2000.

[37] N. McDannold, K. Hynynen, and F. Jolesz, "Mri monitoring of the thermal ablation of tissue: effects of long exposure times," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 13, no. 3, pp. 421–427, 2001.

[38] M. C. Pilatou *et al.*, "Mri-based thermal dosimetry and diffusion-weighted imaging of mri-guided focused ultrasound thermal ablation of uterine fibroids," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 29, no. 2, pp. 404–411, 2009.

[39] M. Voogt *et al.*, "Volumetric feedback ablation of uterine fibroids using magnetic resonance-guided high intensity focused ultrasound therapy," *European radiology*, vol. 22, no. 2, pp. 411–417, 2012.

[40] W. A. Grissom *et al.*, "Hybrid referenceless and multibaseline subtraction MR thermometry for monitoring thermal therapies in moving organs," *Medical Physics*, vol. 37, no. 9, pp. 5014–5026, 2010.

[41] L. Shanshan *et al.*, "Intravoxel Incoherent Motion Diffusion-weighted MR Imaging for Early Evaluation of the Effect of Radiofrequency Ablation in Rabbit Liver VX2 Tumors," *Academic Radiology*, vol. 25, no. 9, pp. 1128–1135, 2018.

[42] Y. Chen *et al.*, "Three-dimensional mr fingerprinting for quantitative breast imaging," *Radiology*, vol. 290, pp. 33–40, 2019.

[43] C. C. Cheng, F. Preiswerk, and B. Madore, "Multi-pathway multi-echo acquisition and neural contrast translation to generate a variety of quantitative and qualitative image contrasts," *Magnetic Resonance in Medicine*, pp. 1–12, 2019.

[44] J. Glaunes, A. Trouvé, and L. Younes, "Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.

[45] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, 2015.