CHAPTER 4

A Gradient boosted regression tree ensemble model using wavelet features for post-acquisition macromolecular baseline isolation from Brain MR spectra

Chapter Highlights

- ***Application of Gradient Boosted tree ensemble model***
- ***Macromolecular Baseline isolation***

1.1. Introduction:

^{1}H MR spectra of human brain manifest line-broadened signals of macromolecules (MM) and lipids (Lip) at short echo-time (TE) underlying the metabolite components of potential interests [1]. Metabolite quantitation from MRS spectra provides essential clinical inferences for diagnostics and therapy, as well as in research. Clinical strength MRS spectra are low SNR, baseline and lineshape distorted signals, and hence difficult to quantify. Primarily, other than the metabolite contents, all other components are considered nuisance parts referred as macromolecular baseline (MMBL) and are not included in the parameterization. But, from recent studies, it has been established that specific MMs undergo potential pathological alterations in brain disease conditions like tumor, multiple sclerosis, and stroke, and some relevant information might get ignored by discarding these details [2]. The cumulative information of MM and metabolite components variation during pathology would provide better diagnostic inference.MM signals in human brain find their origin from different amino acid protons, and have shorter T_1 and T_2 times compared to the metabolites. In frequency domain representation, these components overlap with the metabolite peaks complicating the parameterization of the metabolites as well as the macromolecular peaks. For ^{1}H MRS at

different magnetic field strengths, different MM peaks have been identified and reported by previous publications [3-5]. The standard nomenclature protocol for an MM peak is $M_{xx,}$ where subscript 'xx' implies the resonant frequency of a component in ppm. At short TEs, these peaks cover the whole frequency range underlying the metabolites. Also, MM components have different T_1 and T_2 relaxation times from metabolites [1].

To eliminate or isolate MM-peaks from metabolite spectra, various prospective as well as retrospective methods have been proposed in the past studies. Among prospective methods, are: (i) taking long-TE time acquisitions to suppress the contribution of MM in spectra [6, 7], and (ii) method based on T_1 and T_2 differences by using inversion recovery (IR) methods, either single IR or double IR, to obtain metabolite-nulled or MM-nulled spectra [8-10]. Both the approaches provide good suppression but has various limitations in terms of peak information loss, SNR reduction, weighting issues (T_2-weighting for long TE, and T_1-weighting for IR-methods) [1]. Among retrospective methods, different time- and frequency-domain fitting-based approaches have been used. Either using metabolite-nulled MM spectra for parameterization or mathematical modelling of MM spectra using a set of gaussian, Lorentzian or Voigt model functions has been the two main approaches [15-21]. Hankel-Lanczos singular value decomposition (HLSVD) based methods, followed by Advanced Method for Accurate, Robust and Efficient Spectral fitting (AMARES) were among the first methods used MRS domain [11-14, 17]. For MM parametrization, HLSVD does not take prior knowledge of MM components, whereas AMARES is highly user intensive in terms of creating a knowledge base of peaks. Machine learning and deep learning-based approaches for metabolite fitting and spectral mining have also been used in recent years for MRS metabolites mapping but discarding MMs with other baseline artifacts and noise. Most deep learning (DL) models use 2D convolutions to capture spatial features and require very large dataset for training, which is often a limitation with biomedical data. CNN models have recently been used in some

published works for MRS spectra metabolite isolation and extraction [22-24]. Recurrent neural networks and LSTMs have been introduced recently, which are capable of handling 1D sequential data/time series data (although, not for MM isolation), but they are highly complex, computation heavy, and require large dataset for training. Therefore, an independent study focussing on isolation and parametrization for macromolecular baseline is needed for total parameterization of MR spectra to improve overall quantitation and diagnosis.

In this study, a novel approach of combining wavelet based temporal feature extractor from 1D spectra with gradient boosted decision tree model (XGBoost) in a multioutput-regression framework was undertaken to separate overlapping MM baseline and metabolites from noise degraded MRS spectra acquired post-scan and parameterizing the individual components of isolated MM spectra. Through this study, a wavelet-machine learning (ML) combination model was developed to achieve highly competitive results over small and medium dataset compared to an equivalent DL model. An MR acquisition-model-agnostic proposition i.e., linear combination model of fitting has been taken to isolate macromolecular baseline from metabolite spectra. The individual metabolite basis-set were obtained from ISMRM fitting challenge, 2016 dataset [25]. Distinct baseline spectra were simulated by summing individual macromolecule peak of gaussian lineshape generated using macromolecular basis parameters given in Table 4.1. The metabolite peaks follow Lorentzian lineshape because of the exponential decay component in time-domain representation and macromolecular peaks follow gaussian peak profile. Because of this combination, *in-vivo* MRS spectra are considered to follow a near-Voigt lineshape as mathematically, a convolution between a Lorentzian and gaussian results in Voigt-shaped signals. To make this mathematical representation better imitative of *in-vivo* spectra, variation in amplitude, lineshape, frequency and phase shifts, and SNR of individual basis were also implemented. The final summed spectra were used for the fitting and MM isolation study.

Table 4.1: Parameters used to generate Macromolecular spectral baseline (expected at B_0 = 3T). (Small variations in values can be found in different literature) [26-28]

MM peak, M_{xx}	Chemical shift (in ppm)	Mean Amplitude (Normalized to peak amp at $M_{3.97}$)	Mean linewidth (in Hz)	Probable MM component
$M_{0.94}$	0.94	0.72	25.20	Leucine, Isoleucine, Valine
$M_{1.22}$	1.22	0.28	21.10	Threonine
$M_{1.43}$	1.43	0.38	15.90	Alanine
$M_{1.70}$	1.63	0.05	7.50	Lysine, arginine, leucine
	1.68	0.05	13.0	
	1.81	0.05	13.0	
$M_{2.05}$	1.99	0.45	29.03	Glutamate, glutamine
	2.04	0.36	20.53	
$M_{2.27}$	2.27	0.78	17.89	Glutamate, glutamine
$M_{2.54}$	2.57	0.04	5.30	β-methylene protons of aspartyl groups
$M_{3.00}$	3.00	0.30	14.02	Lysine
$M_{3.21}$	3.11	0.11	17.89	Valine-H_β, αCH protons of protein amino acids
	3.22	0.11	10.0	
	3.27	0.11	10.0	
$M_{3.71}$ + $M_{3.79}$	3.71	0.64	33.52	αCH protons of protein amino acids
	3.79	0.07	11.85	
$M_{3.97}$	3.97	1.0	37.48	

4.1.1. Theory

An acquired *in-vivo* time-domain MRS spectrum, also called free induction decay (FID), is a discrete, complex-valued signal. The real and imaginary part are called absorption and dispersion spectra. Mathematically, it can be represented as a parametric linear combination model given by:

$$S(t) = S_X(t) + S_M(t) + \eta \quad (4.1)$$

where,

$$S_X(t) = \sum_{k=1}^{K} A_k . B_k . \exp(j\varphi_k) \exp(-d_k t + 2\pi j f_k t) \quad (4.1a)$$

$$S_M(t) = \sum_{m=1}^{M} A_m . \exp(j\varphi_m) . F_g(f_c, t) \quad (4.1b)$$

Here, $S_X(t)$ is the summed metabolite spectra and $S_M(t)$ is the summed macromolecular contribution. A, B, φ, d, f are amplitude, metabolite basis profile, phase-shift, damping, and frequency shift of k^{th} metabolite and m^{th} macromolecular component. F_g is the gaussian basis for macromolecular components, and η denotes gaussian noise of mean zero and standard deviation σ, and other artifacts. t is the discrete time samples.

The isolation of macromolecular baseline from a noisy MR spectrum can be considered as an inverse problem of estimation where a loss minimization is performed between noisy data and known ground truth obtained from equation (4.1b). Rearranging equation (4.1) as $S(t) = S_M(t) + [S_X(t) + \eta]$, $\min(\|S_M(t) - S'_M(t)\|^2)$ can be performed to obtain a least-squared fit.

For an *in-vivo* MR spectrum, there are two major obstacles because of strong metabolite-macromolecule overlap and high noise content: (1) this inverse problem of fitting is ill-conditioned, and (2) efficient reconstruction of small peaks embedded in noise in frequency spectrum.

In the present study, a novel approach of gradient boosted wavelet-feature tree model in a multioutput-regression framework for MRS spectral fitting was adopted to address the ill-posed inverse problem, and the model is trained and validated over a simulated dataset to learn the inverse problem. The approach has been adopted to develop the model for a noise robust fitting and MM isolation with the bias-variance reduction.

4.1.2. Feature extraction with wavelet transform

It is a powerful multi-scale signal analysis tool which decomposes a signal into its components of varied resolution by scaling and shifting a localized-support mother wavelet function. The selection of a wavelet basis to decompose a signal in terms of wavelet and scaling function depends upon the nature of signal under investigation. Replacing the infinitely oscillating basis of Fourier transform (FT) with locally oscillating real wavelet basis for Discrete wavelet transform (DWT) provides optimal and sparser representation of signals containing singularities. But real wavelets suffer from oscillations, shift variance, aliasing, and directionality, noting that FT does not suffer from these shortcomings. Complex wavelet transform proposes Fourier like complex basis representation of scaling and wavelet function. DTCWT takes real and imaginary wavelet functions as individual orthonormal bases forming two filterbank trees (dual tree approach) and capable of overcoming the above issues for real as well as complex signals with only 2x redundancy for 1D signals. Since MRS signals are complex-valued signals, in the present study, DTCWT was chosen for the wavelet decomposition and optimal presentation of time-frequency response coefficients. The detailed information about DTCWT implementation can be obtained from [29]. The scaled coefficients capture local information and noise from different parts of spectra which may not be considered globally, and training over individual scale-coefficients may help in reducing variance. Also, a level thresholding over the coefficients reduces the noise and provide a sparse representation of data to reduce computational complexities as well as faithful reconstruction of denoised spectra.

4.1.3. Learning model for regression: XGBoost

Gradient boosting decision tree technique has been an important machine learning method in a wide area of application. Since the introduction of XGBoost, this technique has given state-of-the-art results and the model used as a standalone or in ensemble has been performing highly in a variety of competitions and challenges like Kaggle and KDDCup 2015. XGBoost provides

a highly scalable model, capable of capturing complex feature dependencies during training, over a wide range of application and different-sized dataset when domain specific feature analysis has been handled properly. eXtreme Gradient Boosting is a scalable ensemble gradient boosted decision tree machine learning model where a final decision is taken by aggregating the prediction from a set of individual weak learners thereby helping to reduce the bias of the model over the dataset. The detailed information about implementation and application was published and can be obtained from [30]. On small and medium-sized datasets, it is reported to perform better than neural nets. It also supports multioutput classification and regression modelling utilizing sci-kit learn pipeline.

For n spectra in a dataset, and i = {1, 2, …, n}, we denote a spectrum from training set as x_i, corresponding ground truth as g_i for network training. The final predicted output is denoted as $\boldsymbol{y_i}$ for the corresponding target spectrum y_i. An XGBoost tree model uses K additive weak learners as functions for output prediction using a regularised loss objective function:

$$L = \sum_{i}^{n} l(y_i, y_i) + \sum_{k}^{K} \Lambda(f_k) \tag{4.2}$$

Here, the first term l is a differentiable loss function to measure the difference between predicted and target spectrum. For fast optimization of the objective function, gradients and hessians of the loss function were used in the algorithm. The second term Λ is the penalty term for the regression tree functions, f_k. The penalty term is a regularization function which, in addition to controlling the complexity of the model, helps overcome the ill-conditioned inverse problem of estimation. L1- and L2-regularization schemes as penalty functions were used in the present study.

Initially, while making a choice for a training model, the performance of Random Forest regressor and SVM regressor was also evaluated along with XGBoost among the ML-based

methods. The XGBoost model performed better in both cases: i) when the whole spectrum was presented directly as features to the model, or ii) when wavelet coefficients of a spectrum were presented as features to the model.

4.2. Methods

The study performed was divided into three main steps: data augmentation using simulated basis set, feature-set generation, and selection, modelling a network architecture to perform fitting for MMBL isolation from a given spectrum.

1.2.1. Simulation of dataset

In this study, we have performed analysis over simulated as well as *in-vivo* dataset. Metabolite basis-set were obtained from [25] and for MM basis-set, parameters reported for individual MM components were obtained from [1, 18, 22, 23].

The standard steps followed to simulate realistic dataset are:

(i) The individual basis spectrum of 18 metabolites were considered: Ala, Asp, Cr, NAA, NAAG, GABA, Glc, Gln, Glu, Gly, GPC, GSH, Lac, mI, PC, PCr, PE, Tau. Each basis was normalized to its peak maximum.

(ii) The relative concentration range for each metabolite was obtained from literature [26-28]. These concentration ranges were divided evenly in 5 points for each metabolite. The mean concentration value for each metabolite were multiplied with respective basis and summed to generate the ground truth. Similarly, each individual metabolite basis multiplied with randomly selected concentration from its range and were summed used to generate a set of 10000 FIDs.

(iii) 17 individual peaks of Voigt lineshape were generated using amplitude, linewidth, and chemical shift parameters for individual MM basis (Table 4.1). Variation in individual MM components were generated, by selecting values randomly from within an amplitude range in (max. amplitude±15%), and linewidth within (characteristic linewidth±20%) range, and summing to generate combined MM baseline FIDs

(iv) the metabolite and MM FIDs were then added in the ratio of 1:0.75 within ±20% variation

(v) A line-broadening was performed on each summed FIDs in 5-15 Hz range and were frequency and phase shifted in the range of ±10 Hz and ±5º respectively. This step was performed to accommodate the effects of T1 and T2 values, phase effects of eddy currents, and B_o inhomogeneity seen during *in-vivo* acquisitions.

(vi) the resulting FIDs obtained were Fourier-transformed and were resized to 1024 data points between chemical shift range of 0-4.5 ppm.

(vii) To add noise, last 200 points of the acquired *in-vivo* spectra were used to calculate the standard deviation, $\sigma_{in-vivo}$ and SNR. By varying this value in ±25% range and keeping mean = 0, gaussian noise was added to obtain final Training dataset.

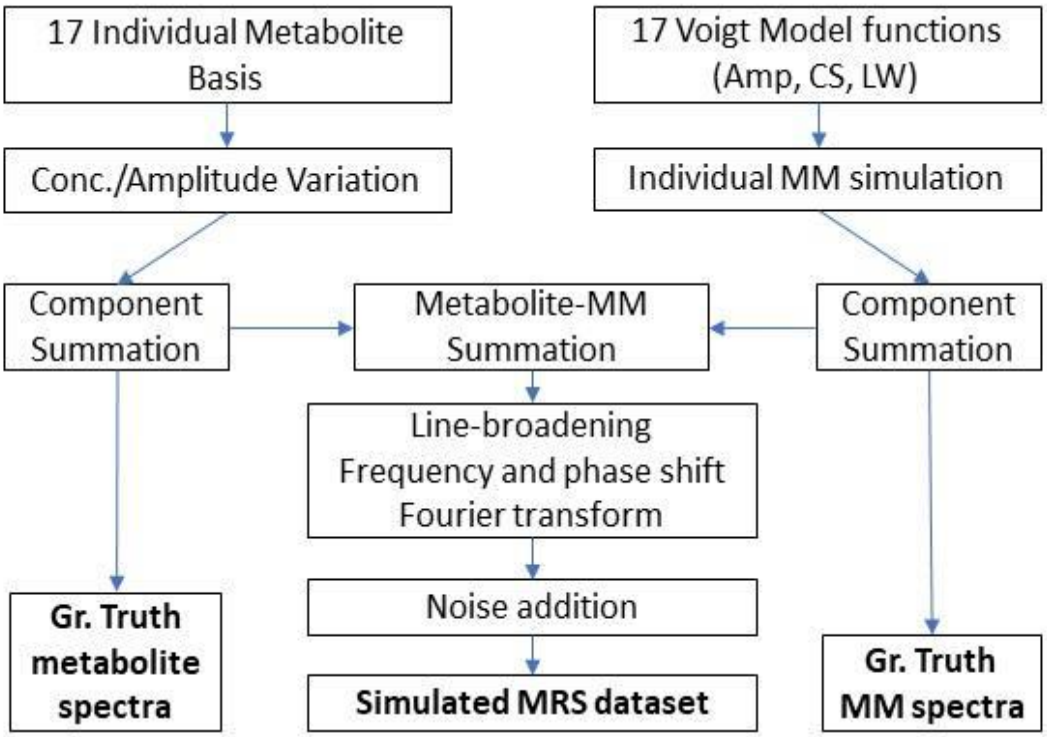


Figure 4.1: Schematic of the steps followed for simulation of brain MRS spectra.

The process of augmentation of simulated data from basis metabolites and macromolecular components were performed in Spyder-python package (version 3.7; Python Software Foundation). A set of 10000 complex-valued spectra was simulated by the abovementioned method (Figure 4.1). The set of 10000 spectra simulated were split into training, validation, and testing sets into the ratio of (70/15/15) %. During the simulation and augmentation, attention has been taken to keep the bias in the dataset as low as possible. Here, by bias, it is meant that the dataset should not have spectra closely mimicking either healthy or a diseased condition in large percentage to influence the learning towards a certain type of spectrum.

1.2.2. *In-vivo* dataset

Retrospective single voxel spectra of 5 patients with tumor of benign type were obtained after clearance from Institutional Ethical Committee. The acquisition protocols were: PRESS sequence at TE/TR = 30/2000 ms, spectral width = 2500 Hz, Number of data points = 1024,

from Siemens Magnetom 3T scanner. A set of 500 spectra was generated by varying overall linewidths and noise content of the spectra using method as described in section 4.2.1 steps (iii) and (iv). The standard deviation of last 200 points from an *in-vivo* FID was used for noise addition to simulated spectra. The set of 500 *in-vivo* spectra was used as a separate test set (test set 2) along with simulated test set (test set 1).

1.2.3. Fitting model architecture*:*

In the present study, we have followed two approaches of feature presentation to the XGBoost network. In the first approach, to capture the globally inter-related variations in the spectra, the training set of concatenated wavelet coefficients of spectra was presented as features to the fitting module. Secondly, to encapsulate local variabilities, we took inspiration from traditional window-based time-series methods like Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) [31] to create a regression model for spectral fitting (Figure 4.2). Window-based time-series methods in machine learning regression models help in capturing temporal patterns in the data by dividing the sequence into time windows. These windows can be of different variations e.g. fixed length, sliding, expanding, rolling type. These methods are especially useful for predicting continuous values in regression tasks. Each method can be tailored to the specific problem based on the nature of the temporal dependencies. We have designed our model using wavelet transforms in combination with windowed data for time-frequency analysis by transforming the time series into the frequency domain, capturing local patterns as features at different scales. The proposed model utilizes these features from noisy and ground truth data to perform the loss minimization and obtain a least squared fit for the macromolecular component isolation.

To generate features-set, following steps were taken:

1. DTCWT mother wavelet at different scales were used as a windowed filter to obtain detailed and approximation feature coefficients from each spectrum. In this implementation, 5 scales of wavelet decompositions were used for feature coefficient extraction.
2. A thresholding operation was performed on the wavelet coefficients for initial noise and artifacts reduction, using a method reported in our previous work [32] to obtain a sparse feature-set.
3. The coefficients from each scale were concatenated and flattened into a single vector for each individual spectrum, to be used as feature vector for training the regression model (method 1 in Figure 2).
4. To generate windowed instances of features, the coefficients of individual scale were concatenated for each spectrum to obtain windowed instances of original spectra for training the model (method 2 in figure 4.2).

XGBoost presents several hyperparameters of the objective function to tune for efficiently training a model over a specific dataset addressing issues of overfitting, underfitting, noise robustness. To tune the hyperparameters of the base regressor, 'HyperOpt' was used which is an open-source package in python and uses a Sequential Model-Based Optimization algorithm called Tree-based Parzen Estimators (TPE) to select the optimised hyperparameter values from a given range [33]. The hyperparameters, their functions, the range used for the present regression model tuning, and final values used during prediction are given in Table 4.2 as:

Table 4.2: Hyperparameters for the proposed model

Hyperparameter	**Range of values used (Lower: Step: Upper)**	**Final value**
Number of estimators (n_estimator)	4:4:32	16

Maximum tree depth (max_depth)	4:2:10	5
L1-regularization (reg_alpha)	0.01:0.05:0.1	0.025
L2-regularization (reg_lambda)	0.1:0.05:0.5	0.3
Learning rate (eta)	0.05:0.05:0.9	0.5
Pruning parameter (gamma)	0.00:0.05:0.5	0.05

After setting values of optimized parameters using HyperOpt, the model was trained individually for subsampling values = 0.60, 0.75, 0.80, 0.90, and no subsampling, and respective MSE and SSIM was obtained. At subsampling value of 0.75 along with other optimized hyperparameters, the model fitting response was found to perform better and more robust to external noise.

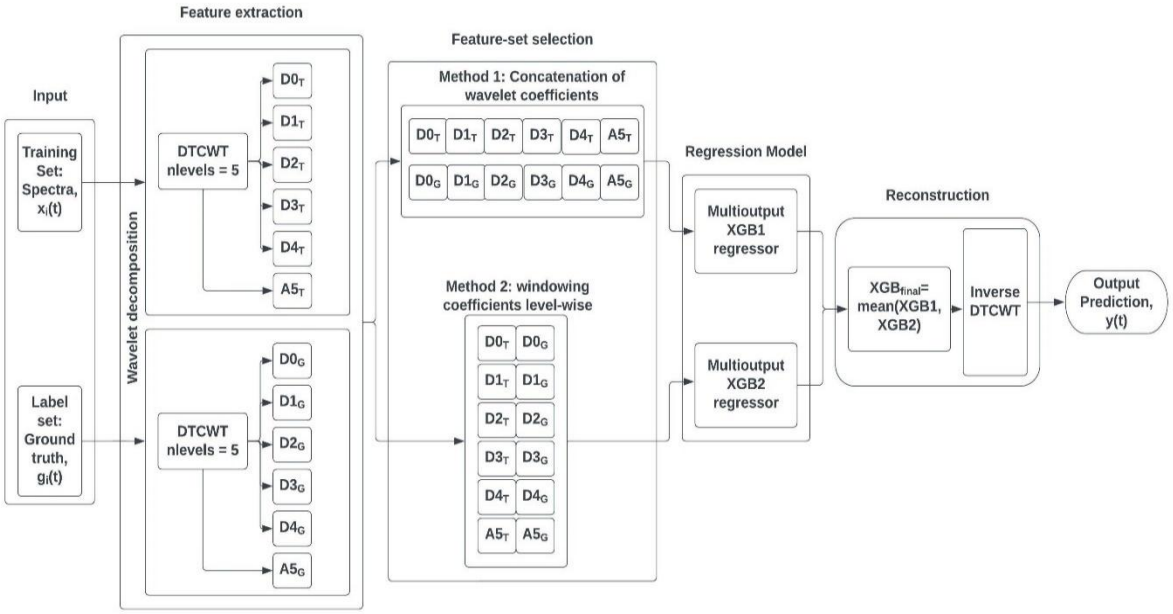


Fig 4.2: Architecture of the proposed fitting model

The outputs obtained from the regression modules XGB1 and XGB2 (by method 1 and 2) were averaged. Since, any frequency or phase shift during simulation was not performed, the spectra in our dataset were aligned before and after processing and hence, taking an average further

reduce any noise, if present. The mean value of prediction, XGB_{final} was then inverse DTCWT tranformed as final reconstruction step to obtain the final prediction, y(t).

Therefore,

$$XGB_{final} = mean(XGB1, XGB2) \qquad 4.3(a)$$

$$y(t) = IDTCWT\left(XGB_{final}\right) \qquad 4.3(b)$$

1.3. Results and Discussion:

To investigate the performance of the model algorithm, three propositions were considered. (i) comparison of different regression models on training dataset with proposed method for MM isolation from noisy dataset, (ii) parameterization of individual MMs from isolated MM spectra, (iii) the effect of noise variation on the proposed model.

4.3.1. MM isolation from noisy dataset: Comparison with different regression models

For comparison, two ML models and two DL models have been chosen. (a) Random forest (RF) regression model has been chosen as it follows a similar approach of ensemble decision tree as XGBoost. The hyperparameters for RF regressor has been obtained by HyperOpt tuning. (b) SVM as regressor is another ML model used for this comparison and radial basis function kernel '*rbf*' has been used during training of the model. (c) 1D-CNN has previously been used in MRS research for metabolite peak analysis and quantification. Therefore, a 4-layer 1D-CNN model was designed for MM isolation and comparison with the proposed model. Each layer consisted of 2 convolutions and a max-pooling layer, followed by flattening and 2 fully connected dense layers (output layer neuron = 1024). Adam [38] was used as optimizer and MSE, the objective loss function and metric, for training the network. The detailed architecture of CNN model is given in Appendix I. (d) An LSTM model, which is a special kind of recurrent

neural network, has been used in this comparative study as LSTM models have previously been used for time-series data, but not specifically MRS spectra. The architecture of LSTM model has an input shape: (1024, 1), followed by two LSTM modules with 64 and 32 units (neurons) in respective output state, and two dense layers of neurons 32 and 1 to obtain output of same shape as input.

Three statistical measures were used for model performance evaluation on the training, evaluation, and test dataset. (i) Root mean squared error (RMSE): Root mean-squared error is a measure to show the amount of deviation of residual error between original and predicted signal. It is calculated as: $RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \tilde{x})^2}{N}}$,where x_i is the original signal and $\tilde{x}$ is the predicted signal, (ii) Structural similarity index (SSIM): it is an objective measure commonly used in comparing structural quality of signals and images before and after processing. It is calculated as: $SSIM\ (X,Y) = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)}$, where predicted signal is represented as X and ground truth signal as reference signal Y; $\mu_X\ and\ \mu_Y$ are the mean and σ_X and σ_y are the standard deviation of X and Y respectively; σ_{XY} is covariance of X and Y; $c_1$and c_2 are stabilizing constant term, each for mean and standard deviation. (iii) Coefficient of determination (R^2-value): for a regression model, it provides a measure of variance in the dependent variable explained by the independent variable, i.e., how well the regression model fits and reproduce the data. It is calculated as: $R^2 = 1 - \frac{SS_{residue}}{SS_{total}}$, where $SS_{residue}$ is the sum of squares of the residual error and SS_{total} is the total sum of errors. The range of values for R^2 varies between 0 and 1, where 0 implies baseline model and 1 implies best fit regression model.

The performance evaluation of the proposed method along with the methods for comparison has been presented below in Table 4.3.

Table 4. 3: Performance metrics of the models for MM isolation from noisy spectra

Training dataset	**Training models**	**Validation set**			**Test set 1 (simulated)**			**Test set 2 (*in-vivo*)**
		RMSE	**SSIM**	**R^2-score**	**RMSE**	**SSIM**	**R^2-score**	**RMSE**
Noisy spectra as feature Input: **Group 1**	Random Forest	0.2942	0.8827	0.8137	0.3032	0.8715	0.114	0.3101
	SVR	0.3831	0.7431	0.6824	0.4202	0.7126	0.6203	0.4663
	1D-CNN	**0.2829**	0.8478	0.7327	0.2985	0.8322	0.7270	0.3143
	LSTM	0.3121	0.8265	0.7210	0.3376	0.8092	0.7077	0.3550
	XGBoost	0.2873	**0.8991**	**0.8533**	**0.2939**	**0.8817**	**0.8508**	**0.3047**
Wavelet features of noisy spectra as Input: **Group 2**	Random Forest	0.2663	0.9102	0.8613	0.2741	0.9021	0.8582	0.2903
	SVR	0.4031	0.7212	0.6043	0.4346	0.7088	0.6021	0.4692
	1D-CNN	0.2348	0.8503	0.7415	0.2655	0.8474	0.7311	0.3139
	LSTM	0.3397	0.8151	0.7177	0.3558	0.8023	0.705	0.3724
	Proposed	**0.2173**	**0.9378**	**0.8959**	**0.2358**	**0.9187**	**0.8702**	**0.2686**

From the results presented in Table 4.2, it can be inferred that the proposed regression model performs better for MM isolation task. Although, in terms of RMSE, CNN has a slight edge over other models, when spectra are presented as features, the SSIM and R^2-values clearly indicate that the proposed approach is better at maintaining the overall shape of MM spectrum, and reproducibility of fit. Figure 4.3 presents the fitting results of MM isolation using different methods.

To further investigate the importance of realistic MMBL simulation and efficiency of our model over previously unseen dataset, experimentally obtained metabolite-nulled MM spectra using double-IR semi-LASER sequence from 10 healthy adults at TE = 32, 44, 56, 62 ms (prefrontal and occipital brain region) were obtained from [4] and augmentation was performed

as earlier to generate a set of 1000 spectra and the fitting was performed over the trained model. RMSE and SSIM values obtained for this dataset was 0.2781 and 0.9015 respectively, which supports the model performance when using simulated spectra close to mimicking *in-vivo* spectra for training the model. The slight decrease in MSE and SSIM values for *in-vivo* dataset (Test set 2) points out the scope of further improvement in knowledge base of MM parameter during training and validation steps to reduce the bias favoured towards simulated signals.

4.3.2. Parameterization of individual MMs from isolated MM spectra:

The peak amplitude of individual metabolites/MM in an MR spectrum is directly proportional to its concentration. Peak heights/amplitudes and their ratios has been used as metrics for relative quantification of the metabolites as well as in classification and assessments in brain related diseases. Retrieving the peak information from individual component of MMs will be helpful in further quantitation and disease-specific analysis.

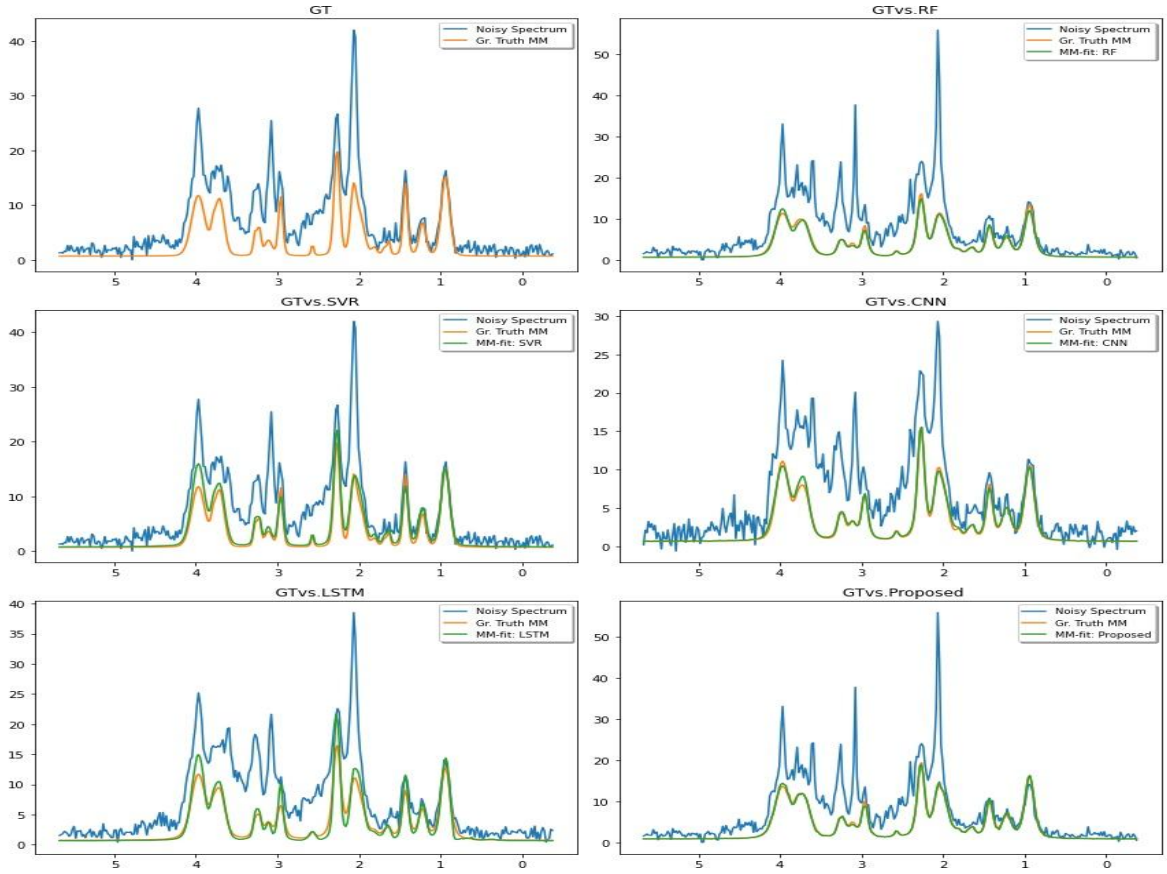


Fig 4.3: Comparison of MM spectra isolation. In the figure panel, Ground Truth-MM for given noisy spectrum is compared with fitted MM spectra of different models respectively. Top left: Noisy spectrum with GT-MM; Top right: RF model fitted MM; Middle Left: SVR model fitted MM; Middle right: CNN model fitted MM; Bottom left: LSTM model fitted MM; Bottom right: Proposed method fitted MM. [Colour code - Blue: Noisy spectrum, Orange: GT-MM, Green: Fitted MM for the model, respectively].

The peak height and linewidth of individual MM components were calculated from isolated MM spectra using the xgboost-regression model. Since, HLSVD and AMARES have previously been used to parameterize MM components from metabolites-nulled MM spectra, these methods have been implemented on the isolated spectra from the proposed model for parameterization and comparison. In the following Table 4.4, a comparison between ground truth and individual MM peak parameters for three different methods has been presented.

Table 4.4: MM peak parameterization: (mean over 500 simulated MMBL spectra)

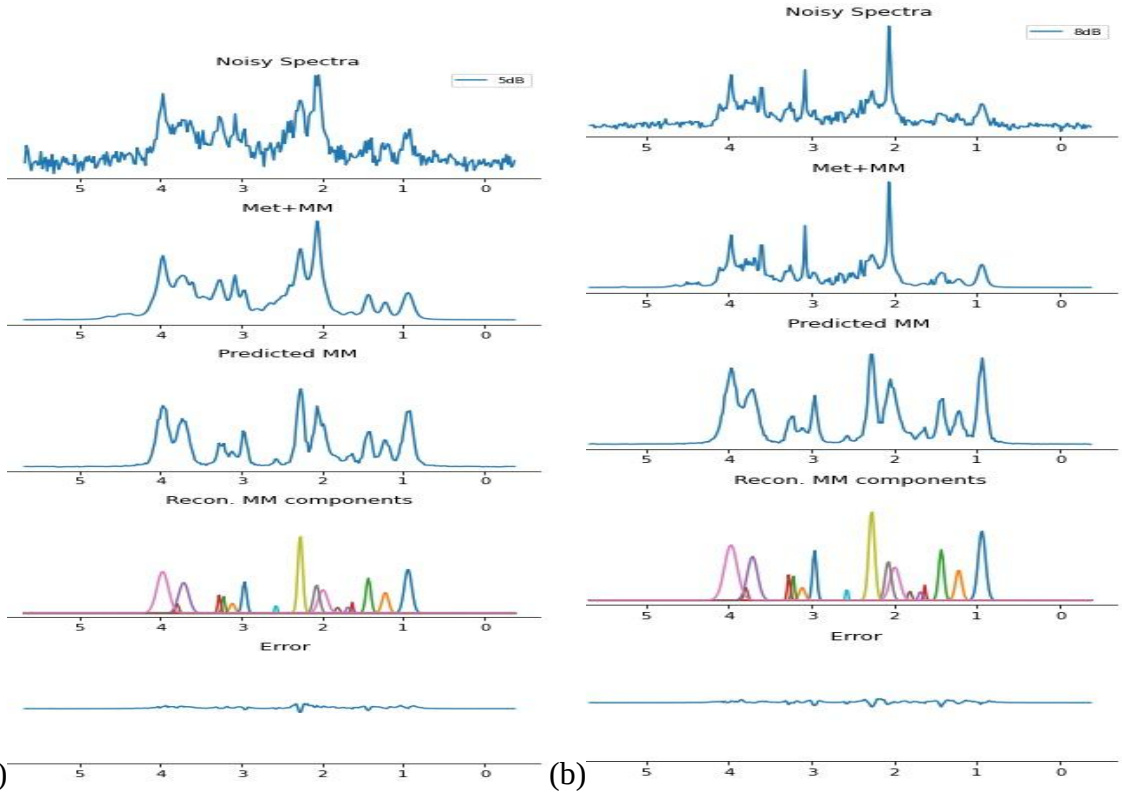
MM peaks	Ground Truth		HSVD		AMARES		Proposed	
	Amp (a.u.)	LW (Hz)	Amp (a.u.)	LW (Hz)	Amp (a.u.)	LW (Hz)	Amp (a.u.)	LW (Hz)
$M_{0.94}$	0.72	25.20	0.7034	26.11	0.7262	23.78	0.7236	24.82
$M_{1.22}$	0.28	21.10	0.2578	22.81	0.2833	21.07	0.2728	21.29
$M_{1.43}$	0.38	15.90	0.3677	16.45	0.3909	15.06	0.3813	15.57
$M_{1.70}$ (1.63+1.68, 1.81)	0.05 0.05	12.25 13.0	0.0213 0.0207	18.40 19.01	0.0487 0.0474	12.41 12.89	0.0494 0.0494	12.37 13.04
$M_{2.05}$ (1.99, 2.04)	0.45 0.36	29.03 20.53	0.4159 0.3403	29.55 21.06	0.4293 0.3654	30.36 20.72	0.4393 0.3612	29.53 20.63
$M_{2.27}$	0.78	17.89	0.7438	18.83	0.792	18.30	0.771	18.03
$M_{2.54}$	0.04	5.30	0.0177	7.331	0.0334	5.904	0.0368	5.455
$M_{3.00}$	0.30	14.02	0.2728	14.92	0.2923	15.27	0.2955	14.21
$M_{3.21}$ (3.11, 3.22+3.27)	0.11 0.11	17.89 10.0	0.0993 0.1049	18.22 10.83	0.1211 0.1026	18.25 9.481	0.1071 0.1083	18.02 10.19
$M_{3.71}$	0.64	33.52	0.6221	34.32	0.6285	33.17	0.6321	33.38
$M_{3.79}$	0.07	11.85	0.0589	12.99	0.0597	13.22	0.0677	11.78
$M_{3.97}$	1.0	37.48	0.9321	39.88	0.9747	38.47	1.043	37.22

Since, HLSVD algorithm used singular value decomposition approach and has largely been used for large water peak suppression in MR domain, the parameterized results were better for longer peak but the deviation for small peak values were prominent. AMARES, on the other hand, has been the one of the standard methods for fitting MR spectra and the fitting parameters

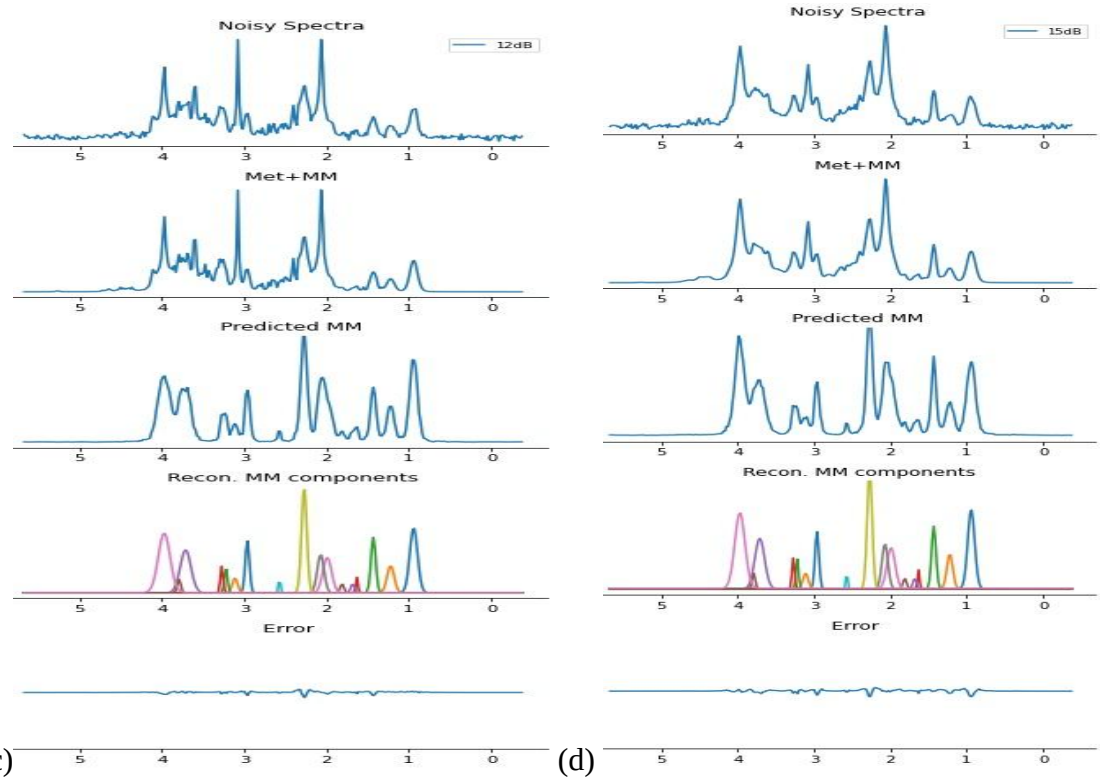
obtained were close to the ground truth parameters for large as well as small peaks. The peak parameters obtained by the proposed method were better especially for small peaks and competent with AMARES results for large peaks in MM spectra. Since, AMARES require user intervention for individual peak information during fitting, there is a clear advantage for our method, once trained, for further analysis.

4.3.3. Reconstructing Individual components from spectra of different noise level

The presence of noise in post-acquisition MR spectra presents an important constraint in parameterization and quantitation of spectral component. In this study, the present method was tested on spectra with variable SNR for MM isolation and faithful reconstruction of individual MM components after parameterization. In Figure 4, spectra with SNR=5, 8, 12, 15 dB were taken and reconstruction was performed. The low reconstruction error for MM isolated spectra suggests that the present method is quite robust to noise in the training spectra set for the parameterization of individual components.



(a) (b)



(c) (d)

Figure 4.4: Isolation of MM spectra and Individual MM components from Spectra of variable noise levels. (a) 5dB, (b) 8dB, (c) 12 dB, (d) 15dB. The y-scale (amplitude) in Predicted MM and Recon: MM components plot of every figure have been scaled up for better visualization of isolated MM spectra and reconstructed Individual MM components.

4.3.4. MM isolation in the presence of Residual Waterpeak

The presence of residual water peak in the MR spectra brings additional complications in terms of (i) residual water peak height of very high magnitude compared to metabolites, (ii) added baseline for metabolites present near 3.5-5.5 ppm range. A separate standalone study to address these factors and is being undertaken in a separate study. We have attempted MM isolation with our present method adding residual water peak to our training dataset. From our attempt, we have found that for residual waterpeak height <= 3xHighest peak of training spectra, our model is able to isolate MM spectra with little hyperparameter tuning of our model. For larger water residue, the results deteriorate, and therefore a separate study is ongoing to look at the

feature presentation and model modification to include residual waterpeak effects. The results of MM isolation with residual peak have been presented in Table 4.4 and Figure 4.4:

Table 4.5: Results for MM isolation from spectra with residual waterpeak

Noisy data with Res. Water content	Res. waterpeak height <= 3 x Highest peak of training spectra			Res. waterpeak height > 3 x Highest peak of training spectra		
	RMSE	SSIM	R^2-value	RMSE	SSIM	R^2-value
Proposed method	0.2988	0.8973	0.7836	>0.55	<0.74	<10

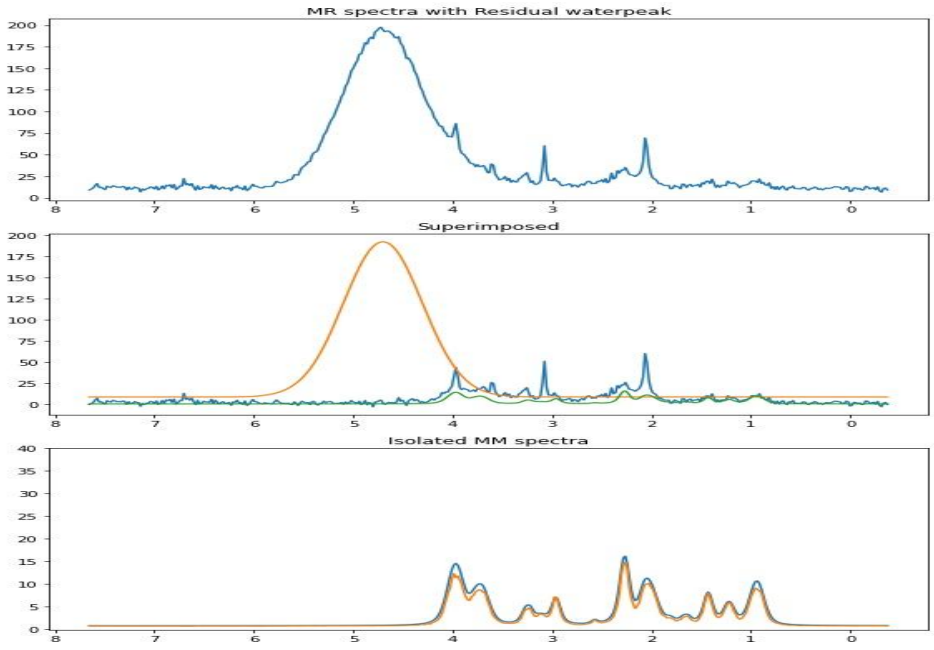


Figure 4.5: MM spectrum isolation from residual water peak included noisy MR spectra. Top: Noisy spectra with residual waterpeak; Middle: Superimposed individual components contribution of Metabolites, Macromolecular Baseline (GT) and Residual waterpeak, [Blue: Noisy spectrum, Green: GT-MM, Orange: Residual waterpeak]; Bottom: Isolated MM from noisy spectra with residual waterpeak (GT vs Proposed method) [Blue: GT-MM, Orange: Fitted MM from the proposed method]

Metabolite-MMBL isolation is a necessary step for MRS spectra analysis. It has direct implication on the quantification of chemical biomarkers present for diagnosis and differentiating/staging of diseases. The results in above sections shows that the proposed the wavelet-supervised machine learning approach presents a viable model-agnostic method for baseline fitting and MM isolation to improve quantification of metabolites as well as macromolecular components.

4.4. Conclusion:

In clinical acquisition settings, ^{1}H-MRS spectra are inherently corrupted because of noise, broad overlapping baseline over metabolite peaks, frequency and phase shifts, water residues and some arbitrary artifacts, like ghosts [33]. The quantitative assessment of metabolites and macromolecular biomarkers aids in disease evaluation and progressions. Acquisition times of MRS spectra require long duration scans for good quality output, making acquisitions for children and elderly particularly difficult. Prospective efforts of using IR methods [3-10, 36] have provided rich information, and using this information as a knowledge base, post-acquisition correction methods can effectively improve the overall quality of spectra for further quantitative assessments.

In the current study, the isolation of MMBL by using wavelet-feature based gradient boosted regression tree as our fitting model was highly efficient for simulated spectra dataset. The algorithm has been implemented on *in-vivo* test set 2 and results obtained are promising to use in an MRS spectroscopic quantitation framework. Further, approaches to reduce dependencies on annotated data during training are being investigated that will be helpful in improving the learning of ML model for real applications.

A machine learning approach was favoured over widely popular DL methods with the idea that most of the medical dataset are small or mid-sized in volume compared to the datasets normally

used for training different state-of-the-art DL networks. The hyperparameter tuning of the XGBoost-based networks requires less efforts, and are computationally less intensive in general, as compared to deep neural networks [34, 35]. Training on a dataset of 7000 spectra in the present study produced better results compared to an equivalent CNN trained on the same dataset.

Publication out of this study:

Applied Magnetic Resonance (2023) 54:637–655
https://doi.org/10.1007/s00723-023-01537-8

Applied
Magnetic Resonance

ORIGINAL PAPER

A Gradient Boosted Regression Tree Ensemble Model Using Wavelet Features for Post-acquisition Macromolecular Baseline Isolation from Brain MR Spectra

Chiranjeev Sagar[1] · Deepak Kumar Singh[2] · Neeraj Sharma[1]

Received: 18 October 2022 / Revised: 6 March 2023 / Accepted: 20 March 2023 /
Published online: 1 April 2023