frontiers | Frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 06 May 2022
doi: 10.3389/fnagi.2022.810873

A Tensorized Multitask Deep Learning Network for Progression Prediction of Alzheimer's Disease

Solale Tabarestani[1†], Mohammad Eslami[2†], Mercedes Cabrerizo[1], Rosie E. Curiel[3,4], Armando Barreto[1], Naphtali Rishe[1], David Vaillancourt[4,5,6], Steven T. DeKosky[4,5], David A. Loewenstein[3,4,7], Ranjan Duara[4,7] and Malek Adjouadi[1,4*]*

[1] Center for Advanced Technology and Education, Florida International University, Miami, FL, United States, [2] Harvard Ophthalmology AI Lab and Harvard Medical School, Schepens Eye Research Institute, Massachusetts Eye and Ear, Boston, MA, United States, [3] Center for Cognitive Neuroscience and Aging, Psychiatry and Behavioral Sciences, University of Miami School of Medicine, Miami, FL, United States, [4] Florida Alzheimer's Disease Research Center, University of Florida, Gainesville, FL, United States, [5] Department of Neurology, University of Florida, Gainesville, FL, United States, [6] Department of Applied Physiology and Kinesiology, University of Florida, Gainesville, FL, United States, [7] Wien Center for Alzheimer's Disease and Memory Disorders, Mount Sinai Medical Center, Miami Beach, FL, United States

OPEN ACCESS

Edited by:
Shenghong Ju,
Southeast University, China

Reviewed by:
Mingliang Wang,
Nanjing University of Information Science and Technology, China
Anees Abrol,
Georgia State University, United States

***Correspondence:**
Solale Tabarestani
staba006@fiu.edu
Malek Adjouadi
adjouadi@fiu.edu

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Alzheimer's Disease and Related Dementias, a section of the journal Frontiers in Aging Neuroscience

Received: *08 November 2021*
Accepted: *14 March 2022*
Published: *06 May 2022*

Citation:
Tabarestani S, Eslami M, Cabrerizo M, Curiel RE, Barreto A, Rishe N, Vaillancourt D, DeKosky ST, Loewenstein DA, Duara R and Adjouadi M (2022) A Tensorized Multitask Deep Learning Network for Progression Prediction of Alzheimer's Disease. Front. Aging Neurosci. 14:810873. doi: 10.3389/fnagi.2022.810873

With the advances in machine learning for the diagnosis of Alzheimer's disease (AD), most studies have focused on either identifying the subject's status through classification algorithms or on predicting their cognitive scores through regression methods, neglecting the potential association between these two tasks. Motivated by the need to enhance the prospects for early diagnosis along with the ability to predict future disease states, this study proposes a deep neural network based on modality fusion, kernelization, and tensorization that perform multiclass classification and longitudinal regression simultaneously within a unified multitask framework. This relationship between multiclass classification and longitudinal regression is found to boost the efficacy of the final model in dealing with both tasks. Different multimodality scenarios are investigated, and complementary aspects of the multimodal features are exploited to simultaneously delineate the subject's label and predict related cognitive scores at future timepoints using baseline data. The main intent in this multitask framework is to consolidate the highest accuracy possible in terms of precision, sensitivity, F1 score, and area under the curve (AUC) in the multiclass classification task while maintaining the highest similarity in the MMSE score as measured through the correlation coefficient and the RMSE for all time points under the prediction task, with both tasks, run simultaneously under the same set of hyperparameters. The overall accuracy for multiclass classification of the proposed KTMnet method is 66.85 ± 3.77. The prediction results show an average RMSE of 2.32 ± 0.52 and a correlation of 0.71 ± 5.98 for predicting MMSE throughout the time points. These results are compared to state-of-the-art techniques reported in the literature. A discovery from the multitasking of this consolidated machine learning framework is that a set of hyperparameters that optimize the prediction results may not necessarily be the same as those that would optimize the multiclass classification. In other words, there is a breakpoint beyond which enhancing further the results of one process could lead to the downgrading in accuracy for the other.

Keywords: Alzheimer's disease, multitask learning, prediction, longitudinal regression, progression, neural network

INTRODUCTION

Extensive research has focused lately on using different machine learning techniques for the diagnosis and prognosis of AD. However, a retrospective of previous studies on multimodal datasets reveals some inconsistencies in modeling the relationship between the many features captured from the different recording modalities. Although several linear methods have been previously reported in the literature with the ability to linearly fuse the information from different modalities (Perrin et al., 2009), several authors have also suggested different non-linear approaches to fuse the multimodal biomarkers (Wang et al., 2012, 2018c; Huang et al., 2016; Tong et al., 2017).

The relatively low accuracy of the classification and regression techniques in delineating converter from non-converter groups and Mild Cognitive Impairment (MCI) from Cognitively Normal (CN) draws our attention to the diversity and heterogeneity of the potential features that could be extracted from the multimodal and multiclass AD datasets (Pellegrini et al., 2018). For example, Wei et al. (2016) proposed a classification method to distinguish non-converter MCI (MCI-NC) from converter MCI (MCI-C) by using an SVM classifier over features that are a combination of FreeSurfer-derived MRI features and nodal features derived from the thickness network. In another recent study (Lin et al., 2020), the authors developed an extreme learning machine (ELM grading method to efficiently fuse multimodal data and predict MCI-to-AD conversion within a 3-year duration. In Huang et al. (2021), subjects are classified as healthy controls, subjective cognitive decline (SCD), or amnestic mild cognitive impairment (aMCI) based on SVM and features extracted from white matter. The ability to detect subtle changes that could lead to a more accurate classification of MCI stable from MCI converter remains extremely challenging. This is why most machine learning models opt for binary classification as an initial step for determining relevant indicators of the model how to best separate these two very difficult MCI subgroups (Tolonen et al., 2018; Gupta et al., 2020).

With the advent of deep learning and their multilayer structure at elucidating lingering abstract steps of machine learning, especially as it pertains to the extraction of relevant features in multimodal multiclass classification and regression processes, there is great interest in their application to brain research in general and complex neurodegenerative diseases like Alzheimer's disease (Liu et al., 2016; Sarraf and Tofighi, 2016; Zhang et al., 2017; Amoroso et al., 2018; Choi and Jin, 2018; Fisher et al., 2018; Lu et al., 2018; Wang et al., 2018a). In Jo et al. (2019), an extensive review for applying deep learning in neuroimaging data is provided, with a focus placed on the diagnosis and prognosis of AD and its prodromal stages. In Kang et al. (2020), a CNN-based classifier with a specific regularization technique is proposed to distinguish early MCI vs. CN subjects using structural MRI and diffusion tensor imaging (DTI) as input to their CNN-based model. Liu et al. (2018b) proposed a cascaded CNN that makes use of multimodal patch-based features from different regions of the brain. Using MRI and PET images, their deep 3D-CNN algorithm could achieve good binary accuracy in differentiating AD vs. CN, progressive MCI vs. CN, and stable MCI vs. CN.

Autoencoders have also been explored for their ability to extract high-level complex patterns embedded in the features to enhance classification accuracy (Suk and Shen, 2013; Liu et al., 2016). For example, in Jha and Kwon (2017), a sparse autoencoder is used for binary classification of AD from cognitively normal (CN) subjects. The use of Recurrent Neural Networks has been proposed by Wang et al. (2018b) to predict a future stage of the patient using historical clinical records. A related study (Liu et al., 2018b) proposed a combination of CNN and Recurrent Neural Network (RNN) for feature extraction and classification. Considering the large size of PET images, instead of using 3D CNN, they employed 2D CNN to extract features from 2D PET slices. The extracted features were then used through gated recurrent units (GRU) for the classification of AD and MCI subjects from the CN group.

With significant efforts made for predicting cognitive scores to track disease progression and for anticipating a diagnosis label at future timepoints to determine a future stage of the disease, the correlation between categorical and numerical variables brings the potentially open question of whether jointly learning based approaches could leverage the learning performance of both classification and regression tasks. Liu et al. (2018c) proposed the use of a CNN model for joint regression and classification tasks. Using their deep multitask multichannel learning (DM^2L) framework, they reached an accuracy of 51.8% in a four-class classification process. In another study by Zhu et al. (2016a), multimodal feature fusion has been explored through a sparse multitask learning process to predict ADAS-Cog, MMSE, and AD stages simultaneously. Another attempt by Shi et al. (2018) is made to perform both tasks of binary and multiclass classification, where a two-stage stacked deep polynomial network is used, obtaining an accuracy of 55.34% in multiclass classification with higher accuracies obtained for binary classification. The multitarget regression approach can also be categorized in this domain of application. In Zhen et al. (2018), the authors encoded the inter-target correlation and the relationship between the input and output space *via* low-rank learning. In a study by Zhang and Shen (2013), a multi-modal multi-task (M3T) learning framework is used for the prediction of multiple clinical variables of MMSE and ADAS-cog from a multimodal dataset. With similar objectives (Zhu et al., 2014), utilized a matrix-similarity-based loss function combined with group lasso to select the best features for both classification and regression tasks.

In this study, a novel neural network architecture, structured as a Kernelized and Tensorized Multitask network (KTMnet) is proposed for processing two joint tasks of classification and longitudinal prediction simultaneously. This network uses dense layers to first extract features from each modality separately, then uses Gaussian kernel layers and tensorization over the modality fused feature space to non-linearly map the data from a low-dimensional space to a high-dimensional space. Empirical results show enhanced performance in comparison to all related methods reviewed in this article, especially when delineating the

challenging group of MCI (converters and non-converters) from CN in a multiclass classification scenario.

MATERIALS AND METHODS

Subjects

The clinical data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). A total number of 1,117 individuals consisting of 632 males and 485 females were considered for this study. The average age is 73.84 with total average years of education of 16.04. The average MMSE score of the population is 27.44 at baseline and 27.06, 26.82, and 26.02 at the next 6 and 12, and 24 months, respectively. At each follow-up visit, participants were labeled as AD, MCI (Mild Cognitive Impairment), and CN, and those participants from the MCI stage that converted to AD are labeled as the MCI to the AD group. The demographics of the subjects are given in **Table 1**.

Problem Description

In longitudinal AD studies, disease progression can be gauged *via* screening the categorical or numerical labels of participants through time. The categorical labels in ADNI are AD, MCI (including the converter and non-converter groups), and CN. On the other hand, there are also numerical measurements needed to assess cognitive impairment, which augment the in-depth analysis of the data. Mini-Mental State Examination or MMSE is the best-known clinical AD predictor that is accepted and used worldwide. While predicting the diagnosis labels is accomplished through classification methods and predicting the numerical value of neuropsychological test scores is performed through regression models, the underlying features for both tasks are constructed from similar sets of measurements. This relationship between these two types of modeling methods motivated researchers to train these highly interrelated tasks of regression and classification through multitask learning.

To model the progression of AD, a time frame of 24 months has been considered to assess the conversion prospects of the MCI group into AD. Therefore, only those subjects that completed a baseline scan (M0) and showed up for a follow-up visit 6 months later (M6), 12 months later (M12), and 24 months later (M24) were considered. Studying longitudinal AD cohorts could improve our understanding of AD pathogenesis. While most patients that have been diagnosed as belonging to the intermediate stage of MCI have been known to progress toward the AD stage, there is some evidence that some of them might stabilize at the MCI stage. However, the different conversion slopes for the different individuals suggest that this stable group is converting into AD in a much longer time frame. **Figure 1** shows the number of subjects in each category of AD over the 24-month duration.

The average longitudinal changes of neuropsychological test scores for the 4 subgroups are shown in **Figure 2**. It is observed that for AD and MCI-C populations, the mean of the cognitive test score for these groups has decreased over time by 13 and 12.7%, respectively. This suggests a continuous decline in health status and thus the need for predicting cognitive decline as early as possible.

Problem Formulation

The proposed Kernelized and Tensorized Multitask network (KTMnet) shown in **Figure 3** is structured to estimate the progression of Alzheimer's disease by predicting the categorical and numerical labels simultaneously. Let y_r be the sets of longitudinal neuropsychological test scores (MMSE) for the regression task (Task 1) and y_c be the sets of categorical labels for the classification task (Task 2). The input space for both tasks is the multimodal features of $\{x_{m_1},\ x_{m_2},\ x_{m_3},\ x_{m_4},\ x_{m_5}\}$, in which the vector x_{m_i} comprises the extracted measurements from modality *i*. Note that these input features are extracted from MRI, PET, CSF, cognitive tasks, and the risk factors at baseline. Hence, vectors y_r andy_c for this study can be established as $y_r = [Score_{M0},\ Score_{M6},\ Score_{M12},\ Score_{M24}]'$ and $y_c = [AD,\ MCI-C,\ MCI-NC,\ CN]'$, where *MCI-C* and *MCI-NC* define the MCI converter and non-converter groups, with the prime symbol (') defining the transpose function. The risk factor parameters considered are age, years of education, sex, and APOE4. The overall objective function, in this case, could be modeled as an algorithm in which $y_r = E_r(x_{m_1},\ x_{m_2},\ x_{m_3},\ x_{m_4}, x_{m_5})$ and $y_c = E_c(x_{m_1},\ x_{m_2},\ x_{m_3},\ x_{m_4}, x_{m_5})$ with E_r and E_c being the corresponding estimators. The architecture of the proposed KTMnet method is shown in **Figure 3**.

The proposed network consists of a series of operations defined through Eqs. (1, 2). Feature representation, modality fusion, and tensorization have been incorporated in an end-to-end artificial neural network to harness the advantage of performing regression and classification tasks jointly in a unified framework. The multitask framework aims to make use of the features extracted from each modality through modality fusion and tensorization to secure optimal accuracy for both prediction and multiclass classification when such tasks are run simultaneously. First, the feature vectors of each modality would be extracted by $F_{m_i}$and then all the features from different modalities will be fused by function f. Next, a 3D tensorization (T) is applied to the fused feature vector to represent higher-order relations between features. Finally, tensor features will be extracted by F and fed to the regressor function f_r and classifier function f_c.

$$Task\ 1\ :\ \hat{y}_r = f_r(F\left(\mathrm{T}\left(f\left(F_{m_1}\left(x_{m_1}\right),\ F_{m_2}\left(x_{m_2}\right),\ F_{m_3}\left(x_{m_3}\right),\ F_{m_4}\left(x_{m_4}\right),\ F_{m_5}\left(x_{m_5}\right)\right)\right)\right)) \quad (1)$$

$$Task\ 2\ :\ \hat{y}_c = f_c(F\left(\mathrm{T}\left(f\left(F_{m_1}\left(x_{m_1}\right),\ F_{m_2}\left(x_{m_2}\right),\ F_{m_3}\left(x_{m_3}\right),\ F_{m_4}\left(x_{m_4}\right),\ F_{m_5}\left(x_{m_5}\right)\right)\right)\right)) \quad (2)$$

The loss function used to calibrate jointly the longitudinal regression and classification tasks is as follows:

$$Loss = \alpha \times MSE\ \left(y_r,\ \hat{y}_r\right) + \beta \times\ l\left(y_c,\ \hat{y}_c\right) \quad (3)$$

TABLE 1 | Demographic characteristics of subjects used in this study.

Parameter	Value	Total	Alzheimer	MCI-C	MCI-NC	Control
Subjects	Number	1,117	157	191	441	328
Gender	f/m	485/632	73/84	75/116	184/257	153/175
Age	Year (mean ± std)	73.84 ± 7.07	76.77 ± 6.99	73.86 ± 7.47	70.85 ± 7.19	75.01 ± 5.71
Education	Year (mean ± std)	16.04 ± 2.78	14.63 ± 3.15	16.09 ± 2.74	16.09 ± 2.63	16.36 ± 2.68
MMSE	Number (mean ± std)	27.43 ± 2.46	23.24 ± 1.96	27.23 ± 1.75	28.30 ± 1.59	29.15 ± 1.01
CDR	Number (mean ± std)	1.25 ± 1.36	3.98 ± 1.51	1.62 ± 0.92	1.24 ± 0.74	0.03 ± 0.13

Label f/m stands for the number of females in comparison to males. Age, years of education, MMSE, and CDR of subjects in each category are presented by mean ± standard variation of that variable.

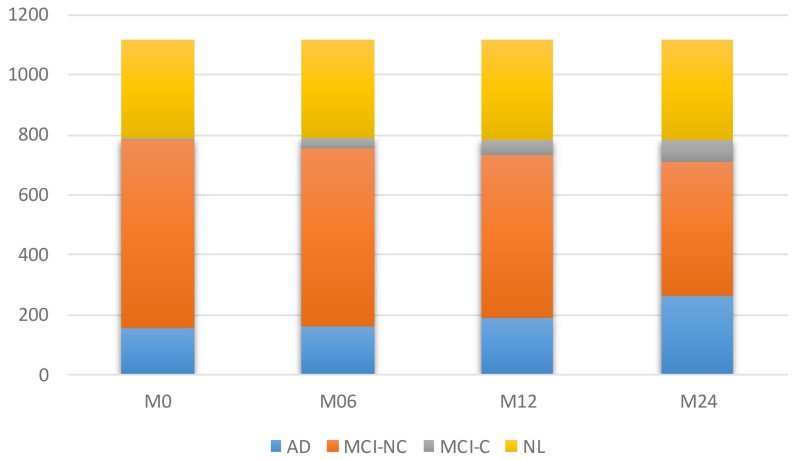

FIGURE 1 | Number of subjects in each of the four subgroups of AD at different time points.

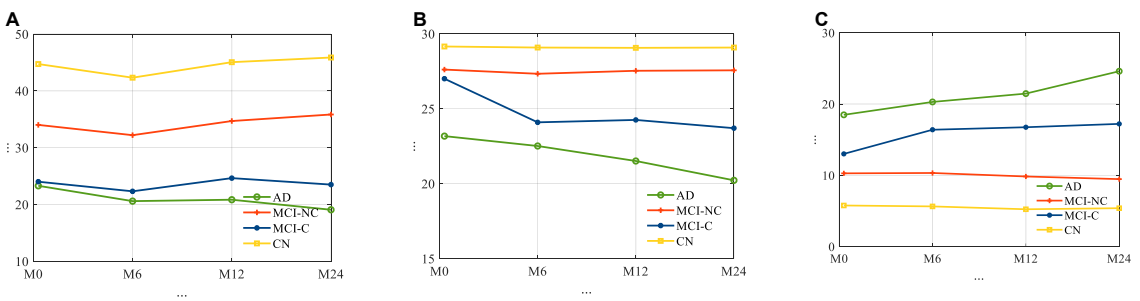

FIGURE 2 | The average trajectories of **(A)** RAVLT, **(B)** MMSE, and **(C)** ADAS11 score for subjects for four different classes of AD.

in which y is the target value and $\hat{y}$ is the value predicted by the network. The MSE is the mean square error for the regression task defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{ri} - \widehat{y}_{ri})^2 \quad (4)$$

And the categorical cross-entropy of $l\left(y_c,\ \hat{y}_c\right)$ is defined as:

$$l\left(y_c,\ \hat{y}_c\right) = \ -\ \frac{1}{N}\sum_{i=1}^{N}\left[y_{ci}\log\widehat{y_{ci}} + \left(1 -\ y_c\right)\log(1 - \widehat{y_c})\right] \quad (5)$$

where N is the number of observations and c is the number of categories assigned to the class label.

Network Architecture

This network architecture relies on convolutional neural layers to jointly perform the processes of tensorization and feature extraction. Given the schematic diagram of the network shown earlier in **Figure 3**, the main properties of the proposed network are as described in the following subsections.

Modality Fusion

The relational correlation of features within each modality and between the different modalities remains an important subject

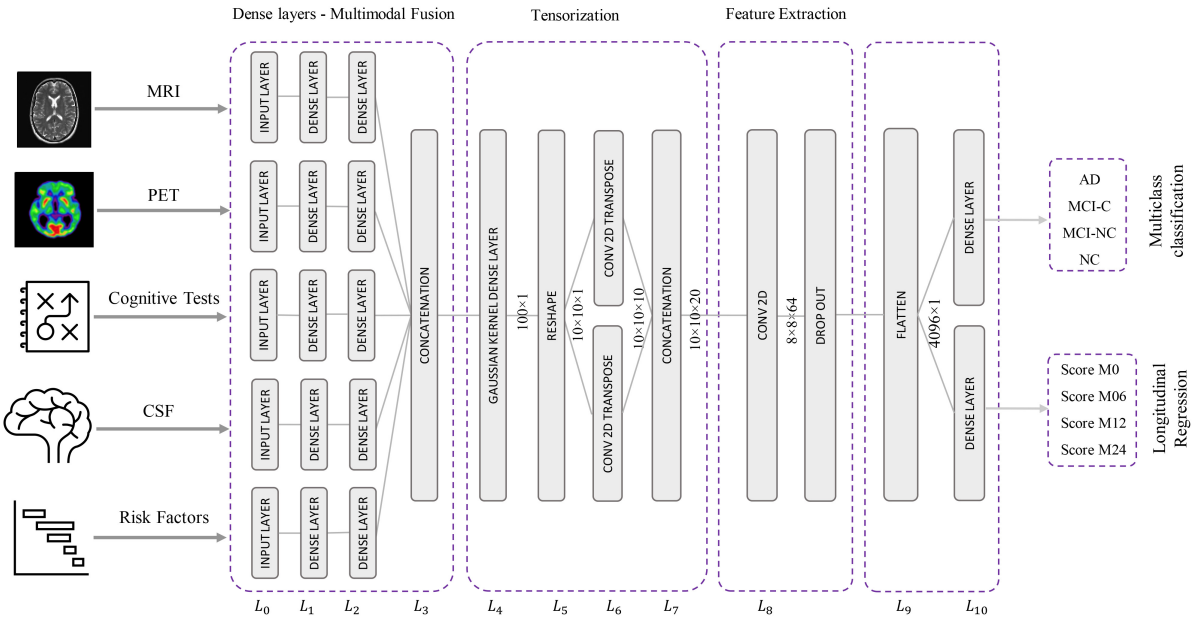

FIGURE 3 | Design architecture of the proposed KTM network.

in developing robust prediction algorithms. The importance of using and fusing relevant information from different modalities to improve classification is well documented in the literature, and some studies have shown significant improvement in comparison to relying on a single modality. For this reason, modality fusion has also been considered in the proposed network to incorporate the advantages of intra-modality and inter-modality feature representation. First, the network starts by transforming the raw features into a primary single modality representation space using fully connected layers. Two fully connected layers of L0 and L1 are then used to transform the extracted features from MRI, PET, CSF, neurocognitive measurements, and risk factor parameters into an initial intra-modality feature-space representation. Let n_mod be the length of the input feature vector of named modality mod, then L0 is the input layer for each modality with n_mod nodes. These single modality features are then processed *via* two fully connected layers of L1 and L2 with [2 × n]_mod and n_mod nodes followed by linear activation function layers. The intermodality feature space is then initiated by integrating the previous fully connected layers in L3, which concatenates the outputs of the L2 layer to create the new feature vector.

Tensorization

Complementary and shared information found in features from different modalities is an essential part of reliably modeling the progression of neurodegenerative diseases. However, concatenating the features from different modalities and processing them using a simple network will not consider the inhomogeneity of the multimodal dataset. Therefore, it is reasonable to transform the feature space into a higher dimensional receptive field to enable the network to find more meaningful relationships. A non-linear mapping function can map linearly inseparable data from a low-dimensional space into a high-dimensional space where it becomes possible to linearly separate the mapped data. The Gaussian kernel function is a representative function that is commonly used and is also adapted in neural networks (Srisuphab and Mitrpanont, 2009; Fei et al., 2016).

Tensorization is thus defined as transforming or mapping the lower-order data to higher-order data to improve the process of generalization afforded at this higher-order (Debals and De Lathauwer, 2015; Novikov et al., 2015). This means that when the data is not providing a satisfactory feature representation in a lower-dimensional space, transferring it to a higher dimensional space may improve the data analysis with the potential for retrieving hidden information in that same data. As an example, a vector can thus be reshaped into a 2D matrix or a 3D tensor of any arbitrary shape with width, height, and depth of W × H × D dimensions. Similarly, a matrix can also be reshaped into a higher-order tensor, by reshaping each column to a tensor of order K and stacking the results along the K + 1 dimension.

In this new architecture, a dense layer with a Gaussian kernel is used for kernelization and a convolutional neural network is used for tensorization, and both are used to extract higher-order features from fused multimodal features. In this way, a tensor with the size of 10 × 10 × 20 is generated using the following procedure:

- L4 uses Gaussian dense layer to assist tensorization.
- L5 reshapes the 100-node output vector of layer L4 to create a 2D 10 × 10 tensor.
- L6 performs 2D transpose convolutional filtering with a kernel size of 3 × 3, a stride of 1, padding type of "same," and linear activation function along with:

- 10 kernels with a dilation rate of 1.
- 10 kernels with a dilation rate of 2.
- Concatenation of the outputs from the two above dilation layers.

Feature Extraction

In this step, predictive features are extracted from the generated tensor. Since the feature extraction part is also based on 2D convolutional filtering with the network being trained in an end-to-end fashion, there is not a strong distinction for separating the network into the tensorization part and feature extraction part. The extracted features at the end of this stage make up the tensor. For this reason, 2D convolutional filtering is performed in L8 by using 64 filters with a kernel size of 4×4 and applying the ReLU activation function. A dropout rate of 10% is implemented to randomly deactivate the connection between the neurons during the training phase to overcome any potential for overfitting.

Classification and Longitudinal Regression

This last component of the network is dedicated to classification and regression. For this reason, L9 flattens the output of the L8 layer to build a vector with the size $4{,}096 \times 1$. The output of the L12 layer is connected *via* two fully connected networks with an L1 regularizer to the two output layers (i.e., y_r and y_c) in L10. Four nodes are assigned for the regression part, which has a ReLU activation function, and four nodes are assigned for the classification part with a Softmax activation function.

Optimizer Selection

In deep learning, choosing the right optimization method is key to tuning an accurate model. During the training, weights are iteratively updated until the network converges to a minimum cost function. Small learning rates will keep updating the weights with smaller steps, which could consequently lead to a minimal loss function. Updating the weights by taking large scales comes with the risk of skipping over the optimal weights. Still, some measure of caution should be taken when assuming smaller steps, as there is a risk of being trapped into some local minima.

For the proposed network, after testing several common optimization methods for training, the adaptive Adam algorithm has been selected as the optimization method. Adam, developed by Kingma and Ba (2014), is one of the most common and adaptive optimizers used in deep learning applications, which adaptively approximates lower-order moments to yield an efficient and easy-to-tune solution. The adaptive learning rate is estimated by retaining an exponentially decaying average of previously squared gradients along with keeping the exponentially decaying averages of past gradients. Using this optimization approach with a learning rate of 0.001 and with exponential decay rates for the moment estimates β_1 and β_2 of 0.9 and 0.999, respectively, resulted in a robust trained network that consolidates high precision, sensitivity, F1 score, and area under the curve (AUC) in the multiclass classification task with high similarity in predicted vs. actual MMSE scores at all-time points in the prediction task.

Regularization and dropouts were used to minimize the likelihood of overfitting in layers L4, L8, and L9. Feature dimensionality reduction is exploited to implicitly select and extract features between L1 and L2 and between L9 and L10. While all network layers from L1 to L10 are extracting features, the main part of the tensorization process is assumed to take place in layers L5 through L8 based on transposed and dilated convolutional filtering.

In summary, the proposed structure of the network accomplishes both classification and longitudinal regression tasks by enabling the network to utilize the complementary/shared information in the extracted features space. Integrating these two challenging tasks within a unified framework elevated the accuracy and robustness of the model by taking into consideration the inter-relatability between tasks in a multitask process. For training the network, an end-to-end learning process has been used to learn from both feature representation and modality fusion simultaneously to address both regression and classification tasks.

PREPROCESSING AND EXPERIMENTAL SETUP

Preprocessing

The procedure for predicting disease progression requires considering additional constraints. Subsequently, only the subjects that have a baseline scan and who showed up for a follow-up visit at 6, 12, and 24 months later, were considered in this longitudinal data collection.

The following preprocessing steps are performed in this analysis:

- Exclude all subjects whose cognitive score or diagnosis label has not been reported.
- Exclude the Aβ, P-tau, or Tau values, reported out of range (e.g., $> 1{,}300$ or < 80 for Tau).
- Remove the predictive biomarkers of ADAS13, MoCA, and CDR, which are found to be highly correlated with the status or label of the subjects. This was done so as not to bias favorably our longitudinal regression results which involve predicting future MMSE scores.
- Perform mean centering and normalization of training and test data using mean and variance of training data (z-score).

At the end of these preprocessing steps, a total number of 1,117 subjects, among them 328 CN, 191 MCI-C, 441 MCI-CN, and 157 AD subjects were considered for this study. **Table 2** provides an overview of the multimodal features used in this study.

Experimental Setup

Empirical evaluations were conducted on the Intel Xeon E7 with NVIDIA QUADRO M6000 GPU. The proposed network is implemented in Python with the Keras library (Chollet et al., 2015) using the TensorFlow backend (Abadi et al., 2016). For hyperparameter selection, a split of 15% of the data has been dedicated to threefold cross-validation trials, where the set

of hyperparameters (including α and β in equation 3) that achieved the minimum bias and variance has been selected. The hyperparameters are the number of kernels used in L9, L10, L11 in the range of {256, 128, 64, 32, 16, and 8} and β in the range of {10, 20, and 200} in which grid search has been performed. After hyperparameter selection, similar to the approach utilized in Suk and Shen (2013), Liu et al. (2018b), and Cao et al. (2018), 10-fold cross-validation trials were performed on the remaining 85% of data to avoid the occurrence of bias within a lucky partitioning. In each round of training set, 10% of data has been utilized as a validation set for monitoring the training process to prevent the network from overfitting. A batch size of 150 and the maximum number of epochs of 200 were set for this process and the training is stopped by monitoring the loss of validation with 30 patience epochs. We performed two sets of experiments to analyze the contribution of this work for each of the prediction tasks for evaluation purposes.

Task 1: Regression Task for Prediction of Disease Progression

In the following experiments, the first task of our KTMnet model is the longitudinal prediction of trajectories of the MMSE score. The neuroimaging modalities of MRI and PET, the cerebrospinal fluid (CSF) biomarkers, genetic information, and cognitive assessment tests have been used to create the multimodal data. Since the state-of-the-art algorithms used different performance metrics, to benchmark our method with other methods, network performance is measured by the following common metrics:

The Root Mean Square Error is defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2} \tag{6}$$

The R correlation coefficient with the formula given below:

$$R\left(Y, \hat{Y}\right) = \frac{\sum_{i=1}(\hat{Y}_i - \bar{Y})\,(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}(\hat{Y}_i - \bar{Y})^2}\sqrt{\sum_{i=1}(Y_i - \bar{Y})^2}} \tag{7}$$

With $\hat{Y}$ defining the predicted values, Y being the real values, N is the number of observations and $\bar{Y}$ is the average of the real values in Y. The RMSE metric measures the standard deviation of the residuals between the predicted and actual targets, while the correlation coefficient metric measures the weight of similarity between them. Low RMSE and high correlation coefficient are desirable, conveying how well the predictive model is approximating the targets.

Task 2: Classification Task for Prediction of Disease Status

For the classification task, the subjects were grouped according to the diagnosis label defined by ADNI as AD, EMCI, LMCI, and CN. The diagnosis label has also been tracked and labeled 24th months after their first visit and subjects are then labeled as MCI converter group (MCI-C) if they have been diagnosed as MCI at baseline and their diagnosis status has progressed into AD. The MCI Non-Converter group (MCI-NC) label is assigned to subjects whose diagnosis label did not change after 24 months. The network is trained to perform a 4-way classification (along with the longitudinal regression task) to predict the subjects' class labels after 24 months. In this second test using the features at baseline, the aim was to predict the probability of converting from MCI to AD, 24 months ahead of time.

TABLE 2 | Summary of multimodal features used for training and testing the KTMnet dataset.

Source	Features
MRI	Ventricular volume, Hippocampus volume, Whole Brain volume, Entorhinal Cortical thickness, Fusiform, Middle temporal gyrus, and intracranial volume (ICV)
PET	FDG, Pittsburgh Compound-B (PIB), AV45
Cognitive test	Rey Auditory Verbal Learning Test (RAVLT Immediate, RAVLT Learning, RAVLT Forgetting, RAVLT Perc Forgetting), Functional Activities Questionnaires (FAQ), Everyday Cognition (Ecog) scales: (EcogPtMem, EcogPtLang, EcogPtVisspat, EcogPtPlan, EcogPtOrgan, EcogPtDivatt, EcogPtTotal,EcogSPMem, EcogSPLang, EcogSPVisspat, EcogSPPlan, EcogSPOrgan, EcogSPDivatt, and EcogSPTotal)
CSF	Amyloid Beta (ABETA), Phosphorylated Tau Protein (PTAU), and Total Tau Protein (TAU)
Risk factors	Age, gender, years of education, and APOE4

RESULTS

Prediction Results

The prediction results for the MMSE test scores at baseline and at time points of 6 months, 12 months, and 24 months are summarized in **Table 3**. In this table, SVR is the conventional Support Vector Regression model. Since models reported in the literature and referenced in this table were using different numbers of features, preprocessing methods, and data modalities, we have taken a similar approach as Abrol et al. (2021) and compared our results with baseline models of SVR, Elastic Net, and Random Forest that were trained and tested using the same data that used to train and test the KTMnet model. The proposed model demonstrated an average RMSE of 2.32 ± 0.52 and a correlation of 0.71 ± 5.98 for predicting MMSE throughout the 24 months after baseline. **Figure 4** shows the scatter plots of predicted MMSE values vs. the actual target values at time points T0, T6, T12, and T24.

Multiclass Classification Results

In this experiment, the results of the multiclass classification considering the four groups of AD, MCI-C, MCI-NC, and CN are shown in **Table 4** with a comparison to other competing methods in the literature. In this multiclass classification process, it is important to investigate the classification performance of the network for each category of subjects. The total classification accuracy achieved by our proposed KTMnet method is 66.85 ± 3.77. In classifying the AD group from all other classes, the proposed network achieved a precision of

TABLE 3 | Comparison of longitudinal regression performance of the proposed network in contrast to other methods reported in the literature.

			T0		T06		T12		T24	
Study	**Data**	**Subjects**	**RMSE**	**Corr**	**RMSE**	**Corr**	**RMSE**	**Corr**	**RMSE**	**Corr**
Zhu et al. (2016a)	MRI + PET[a]	202	1.80 ± 0.13	0.57 ± 0.23	–	–	–	–	–	–
Liu et al. (2018c)	MRI + DEM[a]	1,984	2.37	0.57	–	–	–	–	–	–
Cao et al. (2018)	MRI[b]	755	2.37 ± 0.19	0.57 ± 0.05	–	–	–	–	–	–
Lei et al. (2017)	MRI[b]	445	1.75 ± 0.20	0.75 ± 0.08	2.31 ± 0.29	0.79 ± 0.10	2.48 ± 0.40	0.79 ± 0.12	3.00 ± 0.38	0.83 ± 0.06
Zhang and Shen (2013)	MRI + PET + CSF[b]	186	2.11 ± 0.35	0.65 ± 0.27	–	–	–	–	–	–
Elastic net	Multimodal*[b]	1,117	1.84 ± 0.35	0.71	2.58 ± 0.34	0.54	2.91 ± 0.53	0.51	3.64 ± 0.56	0.50
SVR	Multimodal*[b]	1,117	1.75 ± 0.44	0.42	2.02 ± 0.53	0.54	2.52 ± 31	0.54	3.12 ± 0.41	0.51
Random forest	Multimodal*[b]	1,117	1.74 ±	0.78	1.98 ± 0.45	0.67	2.36 ± 0.36	0.73	3.15 ± 0.28	0.70
Tabarestani et al. (2020)	Multimodal*[b]	1,620	1.62 ± 0.24	0.82	1.78 ± 0.22	0.86	2.24 ± 0.24	0.80	2.38 ± 0.21	0.81
KTMnet	Multimodal*[b]	1,117	1.79 ± 0.12	0.66 ± 0.81	2.10 ± 0.15	0.71 ± 0.92	2.42 ± 0.28	0.71 ± 0.41	2.97 ± 0.45	0.75 ± 3.10

**Multimodal here refers to using MRI, PET, DEM, CSF, and cognitive measurements without the inclusion of ADAS11, ADAS13, and CDRSB.*
[a]Imaging data.
[b]Tabular data.

70.49% ± 9.33, a sensitivity of 57.21 ± 9.41, an F1 score of 62.72 ± 10.11, and an AUC of 94%. In classifying the MCI-C group, the network reached a precision of 45.33 ± 7.22, a sensitivity of 50.79 ± 9.42, an F1 score of 47.72 ± 7.62, and an AUC of 83%. In classifying the MCI-NC group, the network reached a precision of 69.72 ± 8.63, a sensitivity of 67.57 ± 7.00, an F1 score of 68.16 ± 5.06, and an AUC of 84%. In classifying the CN group, the network reached a precision of 77.89 ± 6.62, a sensitivity of 79.78 ± 9.74, an F1 score of 78.10 ± 5.89, and an AUC of 94%.

Figure 5 illustrates the receiver operating characteristic (ROC) curves showing the capability of the network in discriminating between the four groups. This graph outlines the classification performance over all sets of possible thresholds. By varying the threshold, the observations are assigned to certain classes and the True Positive Rate on the y-axis is plotted against the False Positive Rate in the x-axis. **Figure 6** shows the confusion matrix for contrasting the correct and incorrect predictions. The CN population was the easiest population for the model to deal with, showing the lowest amount of false-positive in the MCI-NC and MCI-NC groups, and absolutely no miss-classification in the AD group. In contrast, the MCI-C represented the most challenging one, where the model confused several samples with the MCI-NC and AD groups. This raised the number of false-positive and false negatives in both the MCI-NC and AD groups and consequently degraded the precision and sensitivity in these two groups. There is currently no clear reason why some patients will stabilize in the MCI stage and others will transition into the AD stage.

Design Exploration

Three ablation experiments are conducted to evaluate the effectiveness of tensorization. In the first experiment, the tensorization and feature extraction modules (layers L_4 through L_9) have been removed and data were directly passed from layer L_3 to layer L_{10}. In the second experiment, the tensorization modules (layers L_4 through L_9) have been replaced by a dense layer which transforms the data from layer L_3 to a dimension of 100×1, and the output of this dense layer is passed onto the layer L_{10}. The third and last experiment is to keep the layer L_4 and to pass the results of this layer to layer L_{10}. This configuration keeps the Gaussian kernel dense layer but removes the next tensorization layers. **Table 5** summarizes the experimental results for the tasks of classification and regression. For each experiment, training is stopped by monitoring the loss value of the validation set with 30 patience epochs. Considering the results obtained in this study, the proposed KTMnet obtained the best results among different variations of the network structure. T-test has been performed between the prediction results of the proposed model and different model structures discussed in this subsection to measure the statistical significance of the results and the resulting p-values which were all less than 0.05. The most competing network (in terms of metrics) was the second configuration, where the network was taking advantage of a simple fully connected layer with the dimension of 100×1. This means that the Gaussian layer (in the third experiment) without tensorization and feature extraction modules (L_5 to L_9) becomes less useful. Need to mention that similar experiments to the second experiment have been conducted to explore the effectiveness of adding various hidden layers with different neuron sizes. In terms of RMSE and correlation coefficients metrics, all other configurations have resulted in almost the same performance. P-values between these sets of configurations were greater than 0.05 (showing no significant improvements between these results). Therefore, to keep the manuscript concise and easier to follow, only the results of adding a dense layer of size 100×1 have been reported in **Table 5**. Another interesting observation is that KTMnet converged faster and stopped with a smaller number of epochs.

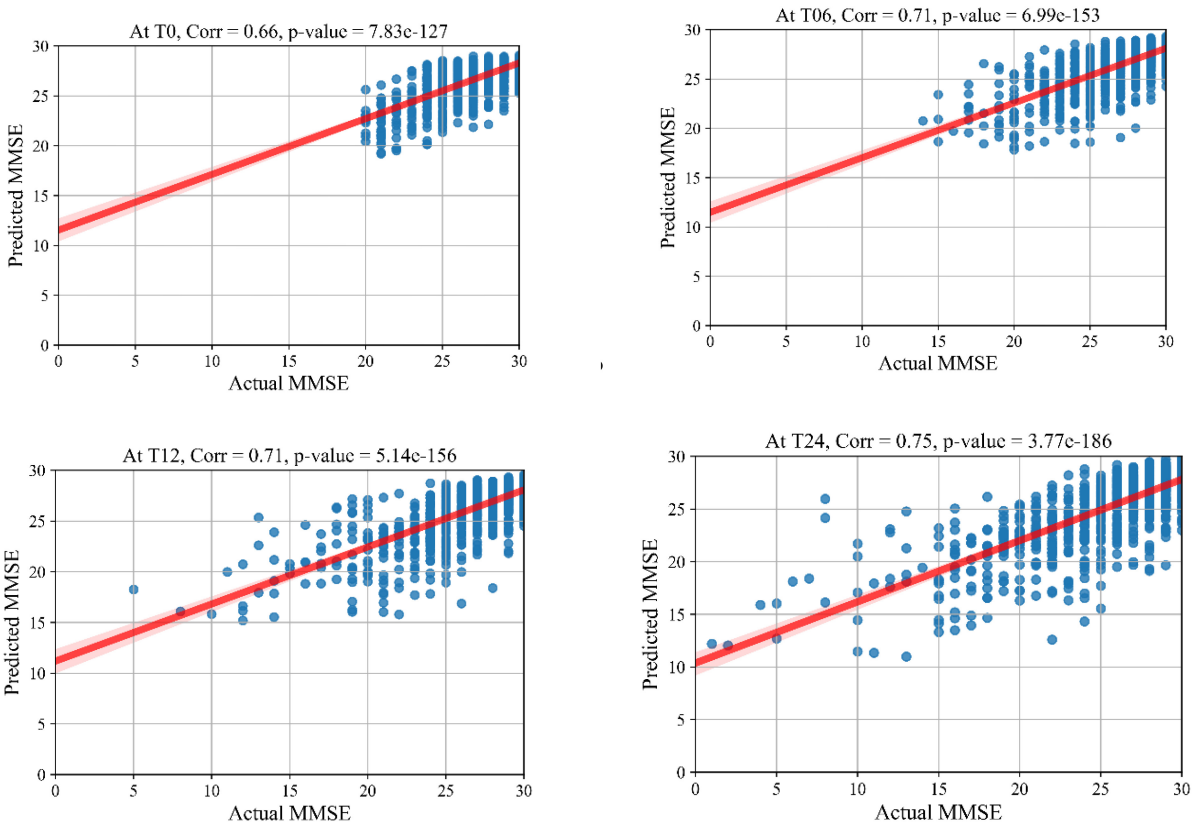

FIGURE 4 | Scatter plots of predicted MMSE values.

TABLE 4 | Comparison of 4-way multiclass classification performance of methodologies reported in the literature using ADNI dataset.

Study	Data	Subjects	Validation method	Accuracy
Liu et al. (2015)[a]	MRI	758	10-fold	46.30 ± 4.24
Liu et al. (2015)[a]	MRI + PET	331	10-fold	53.79 ± 4.76
Zhu et al. (2016a)[a]	MRI + PET	202	10-fold	0.619 ± 1.54
Liu et al. (2018c)	MRI + PET + DEM[a]	202	Independent test	51.80
Zhu et al. (2016b)	MRI + PET	202	10-fold	61.06 ± 1.40
Zhang and Shen (2013)	MRI + PET + CSF	805	10-fold	53.72 (max)
SVM	MRI + PET + CSF + COG + DEM	1,117	10-fold	58.49 ± 4.01
Random forest	MRI + PET + CSF + COG + DEM	1,117	10-fold	60.28 ± 2.83
KTMnet	MRI + PET + CSF + COG + DEM	1,117	10-fold	66.85 ± 3.77

[a]*DEM stands for Demographic information (Age, Gender, and Education).*

DISCUSSION

The deep learning network developed in this study, together with its unique architecture, is designed to perform both tasks of multiclass classification and regression simultaneously, predicts disease progression by tracking the MMSE test scores at four consecutive future time points in a time window spanning 24 months and assessing their categorical labels as (AD, MCI-C, MCI-NC, and CN). This objective has been accomplished through extracting and fusing the complex inter- and intra-modality features, extracting hidden features by using tensorization that projects the feature space into a higher-dimensional space, and eventually modeling the feature representation through non-linear transformations.

In the reported literature, binary classification of AD patients (including the converter and non-converter groups) has been taken into consideration (Moradi et al., 2015; Hojjati et al., 2017; Liu et al., 2017; Spasov et al., 2019). In these studies, attention was more focused on correctly classified subjects by measuring and reporting the metrics of sensitivity and specificity. However, the more challenging multiclass classification of AD cohorts using multimodal screening tests has not been fully

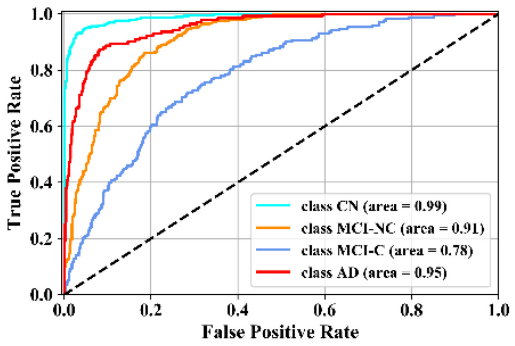

FIGURE 5 | Comparison of ROC curves of the KTMnet for AD vs. MCI-C vs. MCI-NC vs. CN.

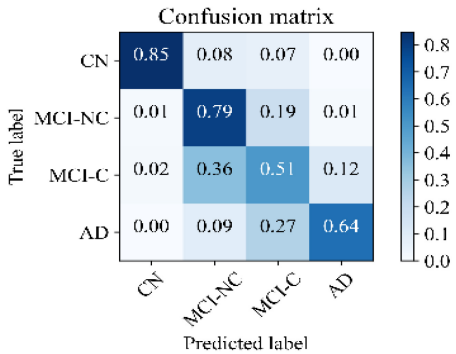

FIGURE 6 | Confusion matrix of the KTMnet model.

explored for the diagnosis and prognosis of AD. This topic becomes even more challenging when progression is assessed in a population of subjects without any preliminary information about their baseline disease category. In a multiclass classification scenario, where there is no auxiliary information to reduce the number of false-positive and false-negative samples, the probability of over and under diagnosis will be increased, making it more important to use additional metrics for performance evaluation purposes. **Table 6** summarizes specific studies that performed multiclass classification or longitudinal regression tasks for meaningful comparisons.

A noteworthy observation made on this model was the see-saw effect encountered during hyperparameter searching. Although we received better results in comparison to other methods reported in the literature, the classification and regression tasks were not in sync with each other. To be more specific, the regression task was falling from its optimum point when the parameters were tuned to increase classification accuracy, and the reverse was also true when the parameters were tuned for increasing prediction accuracy. This new study suggests that when adjusting the hyperparameters to maximize the results of a first given task (e.g., classification), may not necessarily yield a maximized accuracy in the second task (e.g., prediction), proving a breakpoint from which the same set of hyperparameters is to optimize both prediction and multiclass classification in a multitask framework. Other important issues that complicate this multitask process relate to (1) the imbalance in the number of subjects in each of the subgroups considered (NC, MCI-C, MCI-NC, AD), (2) the fact that the process involves multiclass classification involving the aforementioned 4 subgroups, and the prediction that is performed at all-time points of the longitudinal study.

Another set of experiments was conducted to test the full potential of the proposed network (this time as a single task model). By removing one of the two dense layers in L10, the network had its full degree of freedom to optimize its parameters for only one of the regression or classification tasks. When trained for regression task only, the results show an RMSE of 1.74 ± 0.13, 2.09 ± 0.14, 2.46 ± 0.19, and 3.10 ± 0.26 for T0, T6, T12, and T24, respectively, with an average mean RMSE score of 2.35 ± 0.53 for all four time points. The results are close to the results obtained from the network when it was trained and tested in a multitasking mode. To analyze the significance of the difference between the regression results of the single task and multitask model, the *p*-value between the predicted MMSE scores of the regression-only model with its counterpart predicted MMSE scores from the KTMnet model has been calculated. The *p*-values for all time points were bigger than 0.05, showing no significant difference when the model is optimized to only perform the regression task.

Similarly, another set of experiments has been repeated, aimed at the task of classification. This time, when the model has been set up as a single-task classification model, the model achieved an accuracy of 65.53 ± 3.75 which is again close to the multitask KTMnet accuracy, which is 66.85 ± 3.77. This demonstrates that, although the KTMnet model shows a see-saw effect when being optimized as a multitask model, seemingly unable to reach an ideal point in which both regression and classification tasks are each optimized to their full potential performance, the multitask learning approach is indeed helpful. The first supporting reason

TABLE 5 | Comparison of different configurations of the proposed model discussed as design exploration study.

	T0		T06		T12		T24		Acc
Experiment	**RMSE**	**Corr**	**RMSE**	**Corr**	**RMSE**	**Corr**	**RMSE**	**Corr**	
Design exploration 1	5.93 ± 1.29	0.52 ± 0.32	6.02 ± 1.17	0.50 ± 0.43	5.855 ± 1.30	0.51 ± 0.21	6.45 ± 1.08	0.52 ± 041	60.98 ± 3.07
Design exploration 2	1.84 ± 0.15	0.62 ± 0.27	2.46 ± 0.22	0.61 ± 0.18	2.50 ± 0.25	0.58 ± 0.25	3.17 ± 0.32	0.69 ± 0.38	64.42 ± 4.37
Design exploration 3	2.19 ± 0.20	0.56 ± 0.76	2.39 ± 0.35	0.62 ± 0.31	2.63 ± 0.29	0.62 ± 0.43	3.25 ± 0.32	0.70 ± 035	63.16 ± 5.13
KTMnet	1.79 ± 0.12	0.66 ± 0.81	2.10 ± 0.15	0.71 ± 0.92	2.42 ± 0.28	0.71 ± 0.41	2.97 ± 0.45	0.75 ± 0.31	66.85 ± 3.77

TABLE 6 | Summary of prediction tasks accomplished in the literature.

Method	Multitask	Classification type	Class name	Regression type	Modality	Subjects
Natarajan et al. (2014)	No	Multiclass	AD-MCI-CN	–	MRI	397
RELM (Natarajan et al., 2014)	No	Multiclass	AD-MCI-CN	–	MRI	214
Zhu et al. (2016b)	No	Multiclass	AD/MCI/CN and AD/MCI-C/MCI-NC/CN)	–	MRI—PET	202
JRMI (Zhu et al., 2016a)	Yes	Multiclass	AD/MCI/CN and AD/MCI-C/MCI-NC/CN	Single time point	MRI—PET	202
DM2L (Liu et al., 2018c)	Yes	Binary and multiclass	AD/MCI/CN and AD/pMCI/sMCI/CN	Single time point	MRI—Demographic	1,984
DW-S2MTL (Suk et al., 2016)	No	Binary and multiclass	AD/MCI/CN and AD/pMCI/sMCI/CN	–	MRI—PET—CSF	805
SMKMTL (Cao et al., 2018)	No	Binary	AD/MCI-C/MCI-NC/CN	Multiple cognitive scores	MRI	788
SAE (Liu et al., 2015)	No	Multiclass	AD/MCI-C/MCI-NC/CN	–	MRI and (MRI + PET)	758–331
SMTL (Lei et al., 2017)	No	–	AD/MCI/CN	4 time points	MRI	445
MSMT (Nie et al., 2017)	No	–	CN/MCI/AD	4 time points	Multimodal	818
CNN (Liu et al., 2018a)	No	Binary	AD/pMCI/sMCI/CN	–	MRI + PET	397
M3T (Zhang and Shen, 2013)	Yes	Binary	MCI-C/MCI-NC and AD/CN and MCI/CN	2y changes of MMSE	MRI + PET + CSF	186
MSJL (Zhu et al., 2014)	No	Binary	AD/CN, MCI/CN, MCI-C/MCI-NC	Single time point	MRI + PET + CSF	202

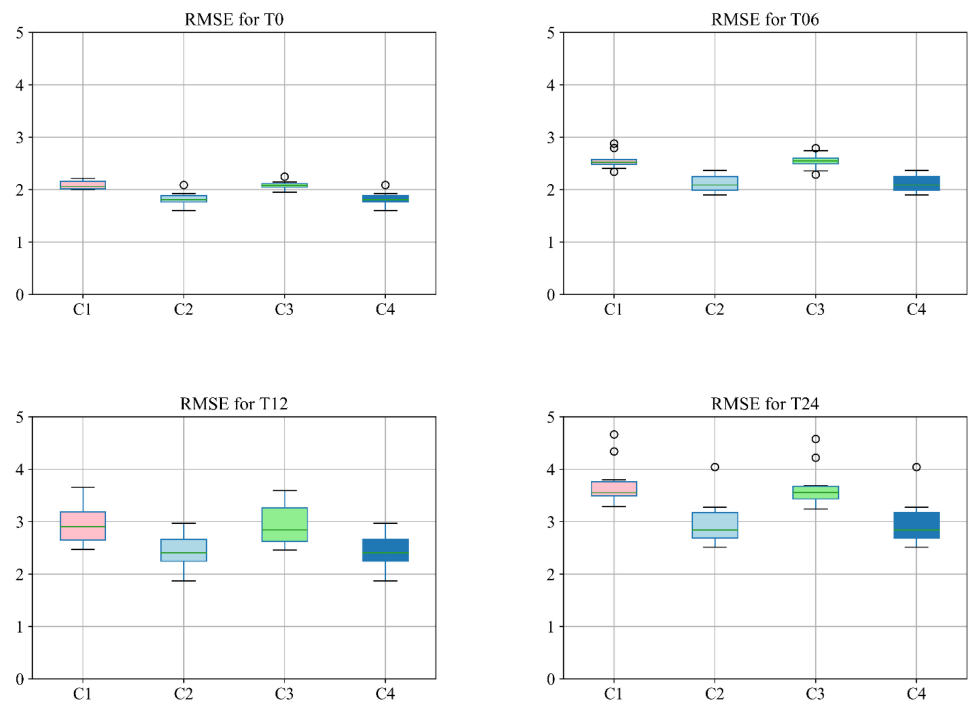

FIGURE 7 | Boxplot for RMSE of mixture category of subjects using different combinations of modalities. Here C1 stands for MRI + PET + RF, C2 stands for MRI + PET + RF + COG, C3 stands for MRI + PET + RF + CSF, C4 stands for MRI + PET + RF + COG + CSF.

for this last assertion is that when the model is designed to only do a single task of regression or classification, it is not able to pass the local optima and achieve better results than with a multitask model. Another supporting factor is the fact that to get classification and regression results from a single task model, two separate models need to be trained, which almost doubles

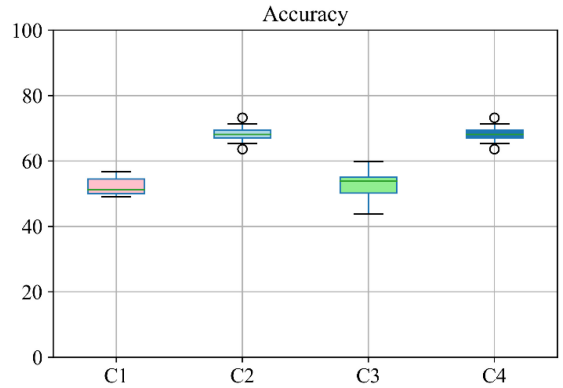

FIGURE 8 | Boxplot for accuracy of multiclass classification achieved through the proposed network based on a different combination of modalities. Here C1 stands for MRI + PET + RF, C2 stands for MRI + PET + RF + COG, C3 stands for MRI + PET + RF + CSF, C4 stands for MRI + PET.

the number of computations that are needed to perform both tasks separately.

The initial expectation from this experiment was that diagnostic labels and cognitive tests should be able to substitute for one another, i.e., they should be able to transform the feature space when being used as targets for a specific model. However, this was not the case in this study. While setting up the experiments, we also tested for the applicability of the model to predict other cognitive tests. Among all these three cognitive scores of (MMSE, RAVLT, and ADAS11) the best results were obtained with multitasking MMSE with diagnosis labels. Thus, we focused on reporting the results of this setup only.

Moreover, different combinations of modalities have been investigated to provide for more meaningful comparisons with other reported studies. Results provided in **Figure 7** demonstrate the influence of the different combinations of modalities in predicting the MMSE scores. Four different modality combinations have been considered, where RF signifies risk factor, with C1–C4 referring to the various combinations of the different modalities as indicated in the legend of **Figure 7**.

Moreover, the accuracy of the multiclass classification for predicting the progression of AD in a period of 24 months in terms of their categorical labels is shown in **Figure 8**. It should be noted that for the sake of uniformity, all the results reported in this study are generated using the same network shown in **Figure 3**. Therefore, the network that has been analyzed to yield the results shown in **Figures 7**, **8** used the hyperparameters (optimizer, dropout rate, decay rate, hidden layer size, and so on) that have been optimized exclusively with respect to the five modalities considered (MRI, PET, CSF, COG, and DEM).

CONCLUSION

In this study, a novel neural network structure with multitask learning, modality fusion, kernelization, and tensorization has been proposed to predict and classify the different stages of Alzheimer's disease in a multiclass population. Using the features collected at baseline, this newly developed network is shown to predict the cognitive status (through the MMSE scores) of the patients in a 24-month longitudinal study involving the AD/MCI-C/MCI-NC/CN groups [taking into consideration the converter (C) and non-converter groups (NC) in the MCI category]. Multitask learning has been explored to enhance prediction performance by incorporating the common relationship or interrelatedness between the regression and multiclass classification tasks. Furthermore, the power of modality fusion, kernelization, and tensorization have also been investigated to efficiently extract important features hidden in the lower-dimensional feature space without being distracted by those deemed irrelevant.

Empirical evaluations on the longitudinal multimodal ADNI dataset were conducted in this study to evaluate the model's performance. The results reveal that the proposed KTMnet framework not only predicts the cognitive scores with relatively high accuracy but can also enhance the multiclass classification accuracy for early stage diagnosis and prognosis of the MCI conversion group. It is emphasized here that although we are aware of the overlap that exists in the MMSE scores in between subject groups, making the prediction of MMSE scores difficult, we still removed from consideration in the training phase the predictive biomarkers of ADAS13, MoCA, and CDR, which are found to be highly correlated to MMSE. Their inclusion otherwise would have favored the proposed machine learning design and could have biased the accuracy for both prediction and multiclass classification.

In relation to **Figure 2**, for each cognitive test (RAVLT, MMSE, ADAS11) and each subgroup, this study shows that there may be a learning effect at 12 months, which continues to 24 months for CN and MCI-NC; however, for the AD and MCI-C groups, the learning effect seems to be overtaken by the disease effect beyond year 1.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

ST and ME designed and developed the model. ST analyzed the data, performed the statistical analysis, and interpreted the data. MA provided overall guidance on the design of the model. ST, ME, and MA wrote the manuscript. RD, SD, and DV provided clinical input on the multimodal framework of the model. RD provided input on the structure of the article. DL and RC provided input on the merits of the cognitive tests. AB, MC, and NR provided input on methodology and problem description. All authors revised the manuscript for important scientific content and final approval.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "TensorFlow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, (Berkeley, CA: Usenix), 265–283.

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., et al. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat. Commun.* 12:353. doi: 10.1038/s41467-020-20655-6

Amoroso, N., Diacono, D., Fanizzi, A., La Rocca, M., Monaco, A., Lombardi, A., et al. (2018). Deep learning reveals Alzheimer's disease onset in MCI subjects: results from an international challenge. *J. Neurosci. Methods* 302, 3–9. doi: 10.1016/j.jneumeth.2017.12.011

Cao, P., Liu, X., Yang, J., Zhao, D., Huang, M., and Zaiane, O. (2018). ℓ2,1-ℓ1regularized nonlinear multi-task representation learning based cognitive performance prediction of Alzheimer's disease. *Pattern Recognit.* 79, 195–215. doi: 10.1016/j.patcog.2018.01.028

Cao, X., Zhao, X., and Zhao, Q. (2018). "Tensorizing generative adversarial nets," in *Proceedings of the 2018 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, (JeJu, Korea: IEEE), 206–212.

Choi, H., and Jin, K. H. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behav. Brain Res.* 344, 103–109. doi: 10.1016/j.bbr.2018.02.017

Chollet, F., et al. (2015). *Keras*. San Francisco, CA: Github.

Debals, O., and De Lathauwer, L. (2015). "Stochastic and deterministic tensorization for blind signal separation," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, (Berlin: Springer).

Fei, J., Zhao, N., Shi, Y., Feng, Y., and Wang, Z. (2016). Compressor performance prediction using a novel feed-forward neural network based on Gaussian kernel function. *Adv. Mech. Eng.* 8:1687814016628396.

Fisher, C. K., Smith, A. M., and Walsh, J. R. (2018). Using deep learning for comprehensive, personalized forecasting of Alzheimer ' s disease progression. *Sci. Rep.* 9:13622.

Gupta, Y., Kim, J.-I., Kim, B. C., and Kwon, G.-R. (2020). Classification and graphical analysis of Alzheimer's disease and its prodromal stage using multimodal features from structural, diffusion, and functional neuroimaging data and the APOE genotype. *Front. Aging Neurosci.* 12:238. doi: 10.3389/fnagi.2020.00238

Hojjati, S. H., Ebrahimzadeh, A., Khazaee, A., and Babajani-Feremi, A. (2017). Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM. *J. Neurosci. Methods* 282, 69–80. doi: 10.1016/j.jneumeth.2017.03.006

Huang, L., Jin, Y., Gao, Y., Thung, K.-H., and Shen, D. (2016). Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiol. Aging* 46, 180–191. doi: 10.1016/j.neurobiolaging.2016.07.005

Huang, W., Li, X., Li, X., Kang, G., Han, Y., and Shu, N. (2021). Combined support vector machine classifier and brain structural network features for the individual classification of amnestic mild cognitive impairment and subjective cognitive decline patients. *Front. Aging Neurosci.* 13:687927. doi: 10.3389/fnagi.2021.687927

Jha, D., and Kwon, G.-R. (2017). Alzheimer's disease detection using sparse autoencoder, scale conjugate gradient and softmax output layer with fine tuning. *Int. J. Mach. Learn. Comput.* 7, 13–17. doi: 10.18178/ijmlc.2017.7.1.612

Jo, T., Nho, K., and Saykin, A. J. (2019). Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Front. Aging Neurosci.* 11:220. doi: 10.3389/fnagi.2019.00220

Kang, L., Jiang, J., Huang, J., and Zhang, T. (2020). Identifying early mild cognitive impairment by multi-modality mri-based deep learning. *Front. Aging Neurosci.* 12:206. doi: 10.3389/fnagi.2020.00206

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [preprint]*. Available online at: http://arxiv.org/abs/1412.6980 (Accessed January 14, 2019)

Lei, B., Jiang, F., Chen, S., Ni, D., and Wang, T. (2017). Longitudinal analysis for disease progression *via* simultaneous multi-relational temporal-fused learning. *Front. Aging Neurosci.* 9:6. doi: 10.3389/fnagi.2017.00006

Lin, W., Gao, Q., Yuan, J., Chen, Z., Feng, C., Chen, W., et al. (2020). Predicting Alzheimer's disease conversion from mild cognitive impairment using an extreme learning machine-based grading method with multimodal data. *Front. Aging Neurosci.* 12:77. doi: 10.3389/fnagi.2020.00077

Liu, J., Shang, S., Zheng, K., and Wen, J. R. (2016). Multi-view ensemble learning for dementia diagnosis from neuroimaging: an artificial neural network approach. *Neurocomputing* 195, 112–116. doi: 10.1016/j.neucom.2015.09.119

Liu, K., Chen, K., Yao, L., and Guo, X. (2017). Prediction of mild cognitive impairment conversion using a combination of independent component analysis and the cox model. *Front. Hum. Neurosci.* 11:33. doi: 10.3389/fnhum.2017.00033

Liu, M., Cheng, D., Wang, K., and Wang, Y. (2018a). Multi-Modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. *Neuroinformatics* 16, 295–308. doi: 10.1007/s12021-018-9370-9374

Liu, M., Cheng, D., Yan, W., Alzheimer's Disease and Neuroimaging Initiative (2018b). Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images. *Front. Neuroinformatics* 12:35. doi: 10.3389/fninf.2018.00035

Liu, M., Zhang, J., Adeli, E., and Shen, D. (2018c). Joint classification and regression *via* deep multi-task multi-channel learning for Alzheimer's disease diagnosis. *IEEE Trans. Biomed. Eng.* 66, 1195–1206. doi: 10.1109/TBME.2018.2869989

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., et al. (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Trans. Biomed. Eng.* 62, 1132–1140. doi: 10.1109/TBME.2014.2372011

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., and Beg, M. F. (2018). Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease. *Med. Image Anal.* 46, 26–34. doi: 10.1016/j.media.2018.02.002

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., and Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage* 104, 398–412. doi: 10.1016/j.neuroimage.2014.10.002

Natarajan, S., Saha, B., Joshi, S., Edwards, A., Khot, T., Davenport, E. M., et al. (2014). Relational learning helps in three-way classification of Alzheimer patients from structural magnetic resonance images of the brain. *Int. J. Mach. Learn. Cybern.* 5, 659–669. doi: 10.1007/s13042-013-0161-169

Nie, L., Zhang, L., Meng, L., Song, X., Chang, X., and Li, X. (2017). Modeling disease progression *via* multisource multitask learners: a case study with Alzheimer's disease. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 1508–1519. doi: 10.1109/TNNLS.2016.2520964

Novikov, A., Podoprikhin, D., Osokin, A., and Vetrov, D. (2015). Tensorizing neural networks. *arXiv [Preprint]* doi: 10.48550/arXiv.1509.06569

Pellegrini, E., Ballerini, L., Hernandez, M., del, C. V., Chappell, F. M., González-Castro, V., et al. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimers Dement (Amst).* 10, 519–535. doi: 10.1016/J.DADM.2018.07.004

Perrin, R. J., Fagan, A. M., and Holtzman, D. M. (2009). Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. *Nature* 461, 916–922. doi: 10.1038/nature08538

Sarraf, S., and Tofighi, G. (2016). Classification of Alzheimer's disease using fMRI data and deep learning convolutional neural networks. *arXiv [preprint]* doi: 10.48550/arXiv.1603.08631

Shi, J., Zheng, X., Li, Y., Zhang, Q., and Ying, S. (2018). Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. *IEEE J. Biomed. Health Inform.* 22, 173–183. doi: 10.1109/JBHI.2017.2655720

Spasov, S., Passamonti, L., Duggento, A., Liò, P., and Toschi, N. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage* 189, 276–287. doi: 10.1016/j.neuroimage.2019.01.031

Srisuphab, A., and Mitrpanont, J. L. (2009). Gaussian kernel approximation algorithm for feedforward neural network design. *Appl. Math. Comput.* 215, 2686–2693. doi: 10.1016/j.amc.2009.09.008

Suk, H. I., Lee, S. W., and Shen, D. (2016). Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. *Brain Struct. Funct.* 221, 2569–2587. doi: 10.1007/s00429-015-1059-y

Suk, H.-I., and Shen, D. (2013). "Deep learning-based feature representation for AD/MCI classification," in *Advanced Information Systems Engineering*, eds C.

Salinesi, M. C. Norrie, and Ó Pastor (Berlin: Springer), 583–590. doi: 10.1007/978-3-642-40763-5_72

Tabarestani, S., Aghili, M., Eslami, M., Cabrerizo, M., Barreto, A., Rishe, N., et al. (2020). A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. *NeuroImage* 206:116317. doi: 10.1016/j.neuroimage.2019.116317

Tolonen, A., Rhodius-Meester, H. F., Bruun, M., Koikkalainen, J., Barkhof, F., Lemstra, A. W., et al. (2018). Data-driven differential diagnosis of dementia using multiclass disease state index classifier. *Front. Aging Neurosci.* 10:111. doi: 10.3389/fnagi.2018.00111

Tong, T., Gray, K., Gao, Q., Chen, L., and Rueckert, D. (2017). Multi-modal classification of Alzheimer's disease using nonlinear graph fusion. *Pattern Recognit.* 63, 171–181. doi: 10.1016/j.patcog.2016.10.009

Wang, H., Nie, F., Huang, H., Yan, J., and Kim, S. (2012). "High-Order multi-task feature learning to identify longitudinal phenotypic markers for Alzheimer's disease progression prediction," in *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems*, (New York, NY: ACM).

Wang, S.-H., Phillips, P., Sui, Y., Liu, B., Yang, M., and Cheng, H. (2018a). Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *J. Med. Syst.* 42:85. doi: 10.1007/s10916-018-0932-7

Wang, T., Qiu, R. G., and Yu, M. (2018b). Predictive modeling of the progression of Alzheimer ' s disease with recurrent neural networks. *Sci. Rep.* 8:9161. doi: 10.1038/s41598-018-27337-w

Wang, X., Zhen, X., Li, Q., Shen, D., and Huang, H. (2018c). Cognitive assessment prediction in Alzheimer' s disease by multi-layer multi-target regression. *Neuroinformatics* 16, 285–294. doi: 10.1007/s12021-018-9381-1

Wei, R., Li, C., Fogelson, N., and Li, L. (2016). Prediction of conversion from mild cognitive impairment to Alzheimer's disease using MRI and structural network features. *Front. Aging Neurosci.* 8:76. doi: 10.3389/fnagi.2016.00076

Zhang, D., and Shen, D. (2013). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* 59, 895–907. doi: 10.1016/j.neuroimage.2011.09.069

Zhang, J., Li, Q., Caselli, R. J., Ye, J., and Wang, Y. (2017). Multi-task dictionary learning based convolutional neural network for computer aided diagnosis with longitudinal images. *arXiv [preprint]* doi: 10.48550/arXiv.1709.00042

Zhen, X., Yu, M., He, X., and Li, S. (2018). Multi-Target regression *via* robust low-rank learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 497–504. doi: 10.1109/TPAMI.2017.2688363

Zhu, X., Suk, H., Il, and Shen, D. (2014). A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. *NeuroImage* 100, 91–105. doi: 10.1016/j.neuroimage.2014.05.078

Zhu, X., Suk, H.-I., Lee, S.-W., and Shen, D. (2016a). Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis. *Brain Imaging Behav.* 10, 818–828. doi: 10.1007/s11682-015-9430-9434

Zhu, X., Suk, H. I. I., Lee, S. W., and Shen, D. (2016b). Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. *IEEE Trans. Biomed. Eng.* 63, 607–618. doi: 10.1109/TBME.2015.2466616

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.