Multi-Modal Tensor Fusion for Alzheimer's Disease Recognition

Statistical Machine Learning II, Spring 2025

Tiffany Le

Dr. Wissam Ahmed

May 10, 2025

Abstract

Accurate and early diagnosis of Alzheimer's disease (AD) is vital for effective intervention and patient care. Traditional diagnostic approaches rely on single modalities such as clinical assessments, neuroimaging, or genetic markers, which may fail to capture the complex, multifaceted nature of AD. To address this challenge, this project proposes a multimodal tensor fusion network (MTFN) that integrates heterogeneous data sources, including visual imagery, demographics, and time-series data, to enhance AD recognition. Our approach leverages tensor representations to model intricate cross-modal interactions while preserving structural dependencies within each modality. Experimental results on publicly available AD datasets demonstrate that the proposed method outperforms the accuracy of the state-of-the-art deep learning classification. This work highlights the potential of tensor-based multimodal learning to advance precision medicine for neurodegenerative diseases.

I. Introduction

Modality fusion involves integrating and harmonizing diverse interrelated data sources to create unified representations [1], [2]. In healthcare, the abundance of multiple data modalities, such as Electronic Health Records (EHRs), medical imaging, genomic data, and data connected from sensors and wearable devices, makes integration essential. Seamlessly combining these sources provides a comprehensive view of a patient's health, enabling more accurate predictions and informed decision-making. This multimodal integration has emerged as a powerful tool for enhancing patient outcomes and driving advancements in healthcare practices [3].

II. Literature Review

Over the past decade, model fusion has gained increasing attention, leading to extensive research on integrating and analyzing multimodal healthcare data. The simplest way is to fuse features from different models by feature concatenation. i.e., stack multiple extracted features into a single vector. For example, Hung et al. [4] employed a multimodal approach using structured data from the EHR for the prediction of gastrointestinal bleeding events. Zhou et al. [5] implemented a multimodal large-scale strategy through the RadFusion model, which combines EHR data and high-resolution computed tomography (CT) scans to detect pulmonary embolisms. However, directly concatenating features from different modalities often faces challenges such as the curse of dimensionality. Approaches like multiple kernel learning [6], [7] have advanced this process by either fusing kernel matrices or encoding features separately before combining them for downstream tasks. Unfortunately, these methods often failed to capture interactions between modalities, limiting their ability to leverage complementary information. Some approaches address this by refining a shared latent representation after aggregating encoded features from each modality or by integrating predictions through weighted averaging. [8] These methods typically assume that all modalities share the same labels, which may not always be practical in real-world scenarios.

Advanced methods such as multi-attention modules [9], [10] have been introduced to enhance feature fusion by preserving shared information while maintaining the integrity of the original data. These modules dynamically assign attention weights to different modalities, effectively capturing complex dependencies and interactions between features. These approaches enhance the robustness of multimodal learning, mitigate information loss, and improve predictive performance. The main concern is that these approaches often emphasize shared features over modality-specific details. In addition, they tend to overlook patient similarities, an essential factor in improving disease prediction accuracy. Moreover, existing data fusion methods often involve an enormous number of parameters to characterize a given dataset. Therefore, it is important to develop efficient factorization techniques to represent the data while reducing the parameter space to a manageable size.

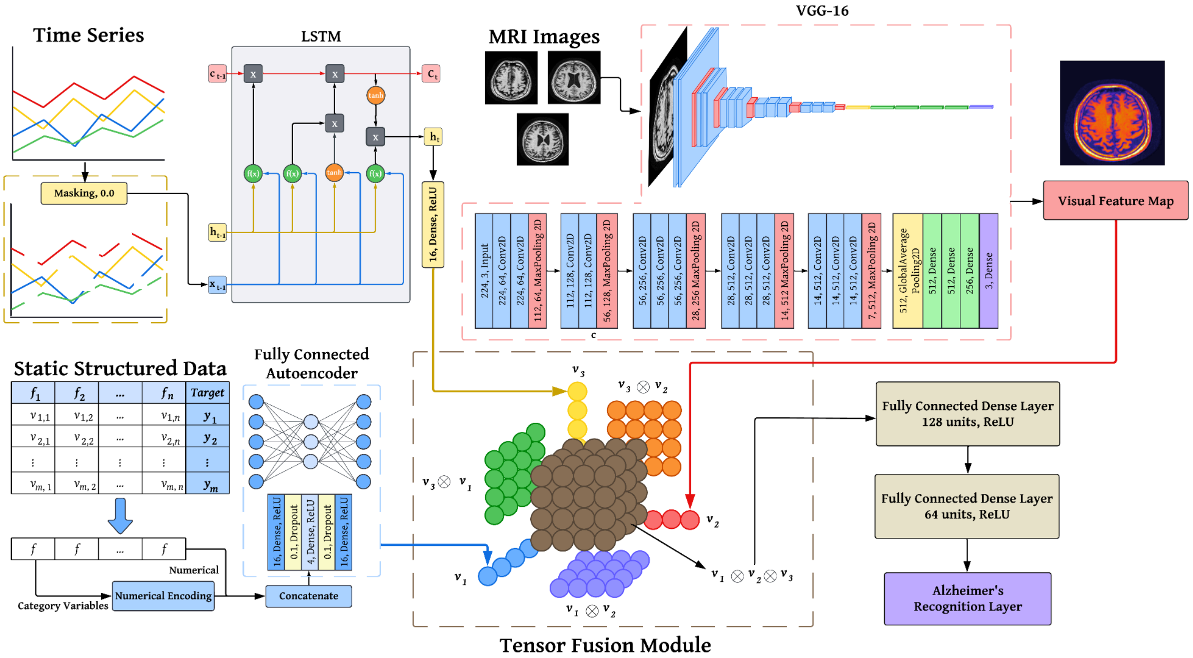
To tackle these challenges, this paper introduces a novel multimodal tensor fusion network (MTFN) that leverages tensor fusion to integrate and extract multimodal data, enhancing predictive modeling while effectively capturing both intra-modality and inter-modality dynamics. Specifically, this project proposes a modality fusion approach that constructs a tensor from diverse data modalities—including visual imagery, demographics, and time-series data—enabling the model to learn latent relationships between these modalities. This project applies this approach to Alzheimer’s disease (AD) recognition, leveraging multimodal data to improve diagnostic accuracy and disease progression modeling. Specifically, the proposed MTFN utilizes a fusion network that integrates embeddings from three input models: Magnetic Resonance Imaging (MRI) scans, patient demographic data, and time-sequential cognitive measurements. The image processing subnetwork employs a modified VGG16 to extract features from medical images. The demographics processing subnetwork processes structured attributes such as age, gender, marital status, and ethnicity through a multilayer perceptron (MLP) with fully connected and dropout layers. The time-series processing subnetwork incorporates a Long Short-Term Memory (LSTM) encoder, which learns compact representations by encoding and reconstructing input sequences, facilitating feature extraction and dimensionality reduction for sequential data. To integrate these modalities, the extracted features are fused using tensor fusion, which refines the representation and reduces dimensionality while preserving critical predictive information. The primary contributions of this work are as follows:

1) A novel tensor fusion approach that integrates convolutional layer weights and input feature data, improving dimensionality reduction and predictive accuracy.
2) A modality fusion framework that effectively captures intra-modality and inter-modality dynamics, enhancing feature representation.

The remainder of the paper is organized as follows. Section II presents the proposed methodology. Section III details the experimental setup and evaluates the effectiveness of the proposed approach through comparative analysis. Finally, Section IV concludes the paper and discusses potential directions for future research.

III. Methodology

The MTFN comprises three key components. First, three embedding subnetworks process unimodal features to generate embedding vectors that capture underlying representations of the original data. Next, a tensor fusion layer constructs a joint representation by modeling unimodal, bimodal, and trimodal interactions through a 3-way Cartesian product. Finally, an Alzheimer’s Disease Inference subnetwork utilizes the fused representation to perform the classification task. Fig. 1 demonstrates the overall framework of the proposed MTFN. The following subsections describe the MTFN in more detail.



f_1	f_2	...	f_n	*Target*
$v_{1,1}$	$v_{1,2}$	...	$v_{1,n}$	y_1
$v_{2,1}$	$v_{2,2}$	...	$v_{2,n}$	y_2
⋮	⋮	⋮	⋮	⋮
$v_{m,1}$	$v_{m,2}$	...	$v_{m,n}$	y_m

A. *Embedding Subnetworks*

1) **Visual Data Processing Subnetwork**: A modified VGG16-based deep convolutional neural network (CNN) architecture is developed for hierarchical feature extraction from medical imaging data, such as MRI scans. This project can leverage the VGG16 model's hierarchical convolutional architecture to capture spatial and structural patterns within imaging data effectively. The model's architecture consists of sequential convolutional layers that capture the spatial hierarchies of the features, moving from low-level edge detection to high-level abstract representations. By leveraging a pre-trained VGG16 model with weights optimized on ImageNet, the convolutional layers are utilized as feature extractors to ensure that the data's relevant spatial and structural patterns are effectively encoded and preserved. The operation of the convolution layer can be formalized as:

$$f_k^l = \delta\left(\sum_m W_{m,k}^l * f_m^{l-1} + b_k^l\right) \quad (1)$$

Where k indexes the output feature maps, l indexes the number of f_m^{l-1} layers, represents the *m*th feature map at layer l-1, δ(·) is the nonlinearity function, W indicates a weight tensor, and b is the bias term. To ensure the feature map is appropriate, average pooling is applied to standardize a fixed uniform shape and flatten it to a one-dimensional vector. Following this process, the vector is parsed through a series of fully connected, shared layers for embedding, consisting of dropout and fully connected layers to reduce overfitting. Furthermore, these layers reduce the high-dimensional feature map into a more compact representation for further efficient feature extraction.

2) **Static Structured Data Processing Subnetwork**: A Multilayer Perceptron (MLP) architecture is designed for meaningful, high-level feature and relationship extraction from static structured data. While there are alternative architectures, such as CNN with their spatial and sequential data processing capabilities [11], MLP's fully-connected neural network has been proven effective for extracting tabular features [12]. By transforming raw tabular data into higher-level embeddings, MLP can capture complex and linear inseparable relationships [13]. The static structured data processing subnetwork processes the input vector $x \in Rn$ through multiple fully connected layers, where n represents the number of input features. The MLP's fundamental computation is defined by:

$$y^{(l)} = f\left(\sum_i w_i^{(l)} x_i^{(l-1)} + b^{(l)}\right) \quad (2)$$

where $x_i^{(l}$ represents the input to the current layer, w(l) represents the weights for the current layer, b(l) represents the bias term, and f represents the activation function applied element-wise (such as ReLU, Sigmoid, or Tanh).

This transformation occurs at each neuron in the network, enabling the model to learn complex and non-linear patterns in each layer. This formulation can be utilized for dual purposes: compression, compressing input data into a latent representation, and reconstruction, reconstructing the input from the latent representation. This architecture focuses on the compression aspect for feature extraction, dimensionality reduction, and multimodal integrations.

3) **Temporal Data Processing Subnetwork**: Long Short-Term Memory (LSTM) networks are a subset of the recurrent neural networks (RNNs) class designed to model long-range dependencies in sequential data, such as a patient's vitals over time [14]. This LSTM-based autoencoder architecture encodes high-dimensional sequences into a compact latent representation while preserving the most informative temporal structure. This architecture employs an encoder LSTM to map the input sequence to a hidden state representation while maintaining sequential dependencies. Additionally, this encoder allows the compression of the high-dimensional hidden state into a lower-dimensional latent space.

In LSTM, different gates are responsible for determining what information is retained and forgotten over time. One of the gates is the input gate, which controls how much new information should be added to the cell state. This gate's equation is formalized as:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (3)$$

where it is the input gate activation at time step t and controls how much of the new input xt should be added to the cell state, σ represents the sigmoid activation function σ(x) = $\frac{1}{1+e^{-x}}$ which compresses values between 0 and 1, wi is the weight matrix associated with input gate, ht−1 represents the hidden state from the previous step t-1, xt is the input at the current time step t, [ht−1, xt] is the concatenation of the previous hidden state and current input, and bi is the bias term for the input gate. The forget gate determines what information should be removed or forgotten from the cell state. The gate's equation is formalized as

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \quad (4)$$

where σ is the activation of the forget gate at the time step t and indicates how much information is forgotten, wf is the weight matrix associated with the forget gate, and bf is the bias term for the forget gate. The value of ft typically ranges from 0 to 1, where ft ≈ 1 means that the past is fully retained and ft ≈ 0 implies that the gate is blocking all the information from the past. The final gate is the output gate that determines how much the updated cell state contributes to the next hidden state. The gate equation is formalized as

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \quad (5)$$

where ot refers to the output gate at time t and controls how much of the cell state should be seen as the hidden state, wo is the weight matrix associated, and bo is the bias term for the output gate. The architecture employs an LSTM decoder to reconstruct the original input sequence from the latent space and validate if the latent space appropriately preserves the most informative features of the data. Finally, an output layer resizes the decoder's output to the original input size for reconstruction. This architecture enables the capture of essential temporal patterns in the temporal data through the latent representations, which are essential for feature extraction. Through the sequence compression into a latent representation, the LSTM autoencoder can learn a structured embedding that captures the most informative, essential temporal patterns in physiological data. As a high-level representation, this embedding can later be utilized within the framework for downstream predictive modeling tasks.

B. *Tensor Fusion*

While base models usually rely on concatenation techniques to merge modalities, these techniques fail to capture the joint relationships between the different modalities. To model and capture insights from the joint relationships, this project employ Tensor Fusion for multimodal integration [15]. This approach constructs a structured representation of the data, including the individual modality information and interactions across all modalities. Tensor Fusion achieves this by computing the Kronecker product between the different feature vectors, preserving both modality-specific and modality interaction dependencies. Following with definitions from [16], in tensor structure, fibers are used to define the one-dimensional building blocks of the tensor, similar to rows or columns in a matrix, and span the horizontal, lateral, or frontal directions. Slices are two-dimensional sections of the tensor created by fixing a dimension of the tensor. They are the matrices spanning the horizontal, lateral, or frontal directions of the tensor, depending on whether the third, second, or first dimensions are fixed. Given outputs

from three subnetworks $v_1 \in R^{d_1}, v_2 \in R^{d_2}, v_3 \in R^{d_3}$ these three fibers are used as input feature vectors, each feature vector is first extended by adding a bias term

$\left\{(v_1, v_2, v_3) \mid v_1 \in \begin{bmatrix} v_1 \\ 1 \end{bmatrix}, v_2 \in \begin{bmatrix} v_2 \\ 1 \end{bmatrix}, v_3 \in \begin{bmatrix} v_3 \\ 1 \end{bmatrix}\right\}$. Next, the joint representation is calculated using the Kronecker product between each of the feature vectors, which is computed such that

$$T = \begin{bmatrix} v_1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} v_2 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} v_3 \\ 1 \end{bmatrix} \quad (6)$$

where $\otimes$ denotes the Kronecker product. The Kronecker product produces a higher-order tensor, specifically a third-order tensor T for three modalities, by computing all pairwise multiplications between the elements of the input vectors. Mathematically, the Kronecker product between two vectors $a \in R^m, b \in R^n$ is given by:

$$a \otimes b = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \dots & a_1 b_n \\ a_2 b_1 & a_2 b_2 & \dots & a_2 b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_m b_1 & a_m b_2 & \dots & a_m b_n \end{bmatrix} \in R^{m \times n} \quad (7)$$

The slices at the tensor's origin v1 $\otimes$ v2, v3 $\otimes$ v1, and v3 $\otimes$ v2 represent the relationship between two modalities. The full tensor T encodes all relationships across modalities, enabling the representation of both independent and joint feature interactions. The Tensor Fusion Module in Fig. 1 illustrates the tensor fusion process. By explicitly modeling feature interactions at different levels, including individual modality features, pairwise relationships, and higher-order dependencies, tensor fusion overcomes the limitations of concatenation techniques in capturing feature interactions. Additionally, using the Kronecker product preserves the structural relationships between multimodal dependencies. This fusion approach enhances multimodal learning by capturing complex interactions between modalities, leading to a richer feature representation.

IV. Results

A. Dataset Description

To evaluate the effectiveness of our proposed approach, this project conducts experiments using the Alzheimer's Disease Neuroimaging Initiative v1.0 (ADNI1) dataset. This project aims to conduct a three-way classification to predict patient categorization into the Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD) groups. The ADNI1 dataset is a collection of 800 total cohorts split across 200 CN, 400 MCI, and 200 AD cohorts. Multivariate clinical and neuropsychological assessments, magnetic resonance imaging (MRI), positron emission tomography (PET), and other biological markers are recorded in the ADNI1 dataset for a period from 2004 to 2010. This study utilizes the cohort's MRI imagery, demographic data, and neuropsychological assessment scores to evaluate the effectiveness of the proposed method. The utilized assessments are the Alzheimer's Disease Assessment Scale Cognitive (ADAS-Cog) versions 11, 13 and Q4. ADAS-Cog assessments are intended to evaluate a person's Alzheimer's disease progression, ranked on a numerical scale. Cohorts are recorded at 6-month intervals, at which point new imagery and assessments are recorded about the cohort. Although ADNI1 supports imagery from 3 Tesla MRI scanners, our experiments only consider 1.5 Tesla scans. To prepare ADNI cohorts for the MTFN model, a joint intersection of each cohort was first taken. Cohorts were separated and grouped by a unique cohort identification number. The cohorts' first MRI image was chosen to represent them over the entire course of the study. Furthermore, the cohort's time series variables, such as the ADAS assessments, were grouped to form a second-order tensor. Cohort data was imputed using feed-forward and feed-backward imputation. Cohorts with missing or bad data were dropped from the examination. After filtering, 382 cohorts remained in the dataset.

B. Experiments and Result Analysis

This project conducts three experiments to evaluate the proposed MTFN model:

(1) Determining the impact of input tensor size on the tensor fusion operation
(2) Examining the contribution of each modality to performance
(3) Comparing the proposed MTFN model with existing state-of-the-art models.

1) *Experiment 1*: In tensor fusion, the rank of input tensors determines the level of compression and the amount of information retained. A smaller size may lead to information loss, affecting the fidelity of the fused data, while a larger size can drastically increase computational complexity and lead to overfitting. Choosing an appropriate size is essential to balance data representation quality and model performance. To determine the optimal representation size for each submodality of data, this project performed a series of experiments.

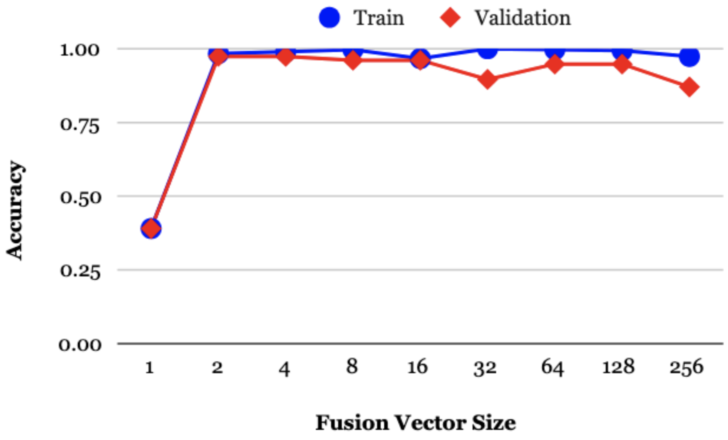
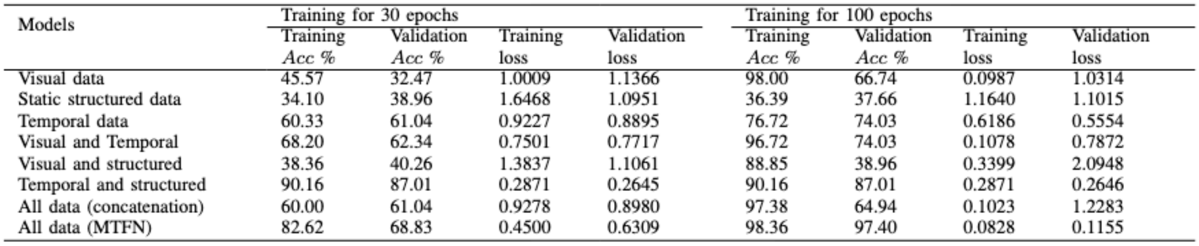


Fig. 2. Tensor Fusion Vector Size Comparison

Figure 2 presents the relationship between the fusion vector size and accuracy for both training and validation sets. As the fusion vector size increases from 1 to 2, both training and validation accuracy experience a sharp increase, suggesting that with a vector length of 2, the model is able to learn an accurate representation of the data. After a vector size of 2, training accuracy remains near perfect, indicating that the model has sufficient capacity to learn from the data. However, validation accuracy starts to slightly decrease after size 8, indicating potential overfitting as the fusion vector size increases. This pattern suggests that while higher fusion size improves training performance, it may lead to overfitting in validation accuracy.

TABLE I

THE ACCURACY AND LOSS COMPARISON OF OUR ABLATION STUDY.



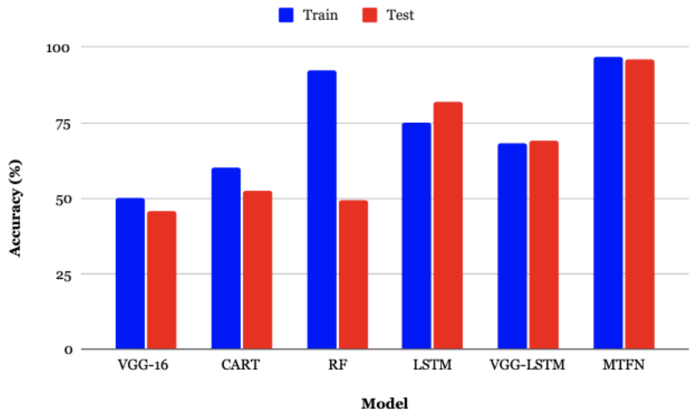
Models	Training for 30 epochs				Training for 100 epochs			
	Training *Acc %*	Validation *Acc %*	Training loss	Validation loss	Training *Acc %*	Validation *Acc %*	Training loss	Validation loss
Visual data	45.57	32.47	1.0009	1.1366	98.00	66.74	0.0987	1.0314
Static structured data	34.10	38.96	1.6468	1.0951	36.39	37.66	1.1640	1.1015
Temporal data	60.33	61.04	0.9227	0.8895	76.72	74.03	0.6186	0.5554
Visual and Temporal	68.20	62.34	0.7501	0.7717	96.72	74.03	0.1078	0.7872
Visual and structured	38.36	40.26	1.3837	1.1061	88.85	38.96	0.3399	2.0948
Temporal and structured	90.16	87.01	0.2871	0.2645	90.16	87.01	0.2871	0.2646
All data (concatenation)	60.00	61.04	0.9278	0.8980	97.38	64.94	0.1023	1.2283
All data (MTFN)	82.62	68.83	0.4500	0.6309	98.36	97.40	0.0828	0.1155

2) *Experiment 2*: This project investigates the significance of MTFN and analyzes the impact of each individual modality. Table I summarizes the classification performance of the model. This project also includes the performance results using feature concatenation for comparison. To facilitate this comparison, each modality's corresponding subnetwork is directly fed into a feed-forward network for Alzheimer's disease recognition.

From Table I, this project observes that all three unimodal models perform poorly. Among them, visual data yields the best performance, while static structured data achieves the lowest accuracy. Static structured data, which includes the cohort's age, educational background, gender, ethnicity, and marital status, appears to be a poor predictor when used in isolation. These types of data often lack the complexity needed for effective prediction.

For bimodal models, this project concatenates the embedding outputs of each unimodal model for intermediate fusion prediction. Generally, all bimodal models show improved performance compared to their unimodal counterparts, although a slight tendency to overfit remains. This behavior suggests that the bimodal models are learning some of the joint interactions between the two data modalities. However, this project allows us to believe that the bimodal representation alone is still insufficient to capture enough meaningful interactions.

In trimodal comparisons, this project evaluates the fusion version of our trimodal network against an intermediate fusion variant to assess the value gained through tensor fusion. The comparison reveals that the concatenation-based model suffers from significant overfitting, a problem also observed in the bimodal models. Without tensor fusion, even with three modalities, the model struggles to learn the unique interactions between the modalities, which hinders its ability to achieve robust performance and leads to overfitting. Therefore, tensor fusion plays a crucial role in capturing these interactions, significantly enhancing the model's classification performance on our dataset. Furthermore, this project also observes that despite the relatively small size of our dataset, the MTFN achieves convergence to optimal accuracy much faster than other models across fewer training epochs. This suggests that the fusion operation helps the model learn meaningful representations more efficiently, especially in environments where training data is limited.



3) Experiment 3: To evaluate the effectiveness of the proposed MTFN, this project further compares it with existing unimodal and multimodal models. The models selected as baselines for comparison are:

1) The Classification and Regression Tree (CART).
2) Random Forest (RF).
3) LSTM.
4) VGG-16
5) VGG-LSTM

TABLE II

BASELINE MODEL COMPARISON

Models	Accuracy (%)		Loss	
	Training	Validation	Training	Validation
VGG-16	50.1639	45.4545	1.3816	1.2756
CART	60.2459	52.4590	-	-
RF	92.2131	49.1803	-	-
LSTM	75.0000	81.9672	0.5668	0.4820
VGG-LSTM	68.1967	68.8312	0.7020	0.6795
MTFN	**98.3607**	**97.4026**	**0.0828**	**0.1155**

VGG-LSTM is a multimodal model consisting of VGG-16 and an LSTM. The two models are utilized for MRI and temporal feature extraction, respectively. The feature outputs are concatenated to obtain a prediction. Fig. 3 and Table II demonstrate the prediction performance.

V. Discussion

Most baseline models either fail to achieve significantly high classification accuracy on both the training and test datasets or suffer from substantial overfitting. I am able to surmise that due to our dataset's restrictive selection, many models struggle to achieve strong performance due to the limited sample of cohorts. However, the proposed MTFN approach captures complex joint features across multiple modalities, allowing the MTFN to perform strongly on this limited training set.

This fusion of features enhances the model's ability to learn meaningful interactions between data sources, resulting in the best performance. By leveraging these enriched representations, the MTFN is able to make more accurate classifications, even in the challenging environment of a limited dataset.

VI. Conclusion

In this paper, this project presents a tensor fusion approach that achieves high classification performance on the publicly available ADNI dataset. The proposed multimodal data fusion framework leverages three core embeddings to generate a third-order tensor that encapsulates unimodal, bimodal, and trimodal interactions across three distinct data modalities. Compared to existing data fusion models, our approach enables the model to learn multimodal interactions between input data, allowing for superior accuracy even with limited training data. For future work, I plan to explore several avenues to enhance and extend the proposed tensor fusion framework. One direction involves investigating the incorporation of additional modalities or features, such as genetic data or advanced imaging techniques, to further improve the model's predictive power. I also aim to develop methods for better handling missing or incomplete data in multimodal datasets, ensuring robustness in real-world applications.

Special Thanks to Dr. Yuxin Wen and Mason Li for their guidance and support on this project.

References

1] H.-Y. Zhou, Y. Yu, C. Wang, et al., "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics," Nature Biomedical Engineering, vol. 7, no. 6, pp. 743–755, 2023.

[2] J. Chen, Q. Li, F. Liu, and Y. Wen, "M3t-lm: A multi-modal multi-task learning model for jointly predicting patient length of stay and mortality," Computers in Biology and Medicine, vol. 183, p. 109 237, 2024.

[3] X. Xu, J. Li, Z. Zhu, et al., “A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis,” Bioengineering, vol. 11, no. 3, p. 219, 2024.

[4] C.-Y. Hung, C.-H. Lin, C.-S. Chang, J.-L. Li, and C.-C. Lee, “Predicting gastrointestinal bleeding events from multimodal in-hospital electronic health records using deep fusion networks,” in 2019 41st annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 2447–2450.

[5] Y. Zhou, S.-C. Huang, J. A. Fries, et al., “Radfusion: Benchmarking performance and fairness for multi-modal pulmonary embolism detection from ct and ehr,” arXiv preprint arXiv:2111.11665, 2021.

[6] H. Wen, Y. Liu, I. Rekik, et al., “Multi-modal multiple kernel learning for accurate identification of tourette syndrome children,” Pattern Recognition, vol. 63, pp. 601–611, 2017.

[7] F. Liu, L. Zhou, C. Shen, and J. Yin, “Multiple kernel learning in the primal for multimodal alzheimer’s disease classification,” IEEE journal of biomedical and health informatics, vol. 18, no. 3, pp. 984–990, 2013.

[8] J. Duan, J. Xiong, Y. Li, and W. Ding, “Deep learning based multimodal biomedical data fusion: An overview and comparative review,” Information Fusion, p. 102 536, 2024.

[9] Z. Ning, Q. Xiao, Q. Feng, W. Chen, and Y. Zhang, “Relation-induced multi-modal shared representation learning for alzheimer’s disease diagnosis,” IEEE Transactions on Medical Imaging, vol. 40, no. 6, pp. 1632–1645, 2021.

[10] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, “Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using mri,” IEEE Transactions on Neural Networks and Learning Systems, 2022.
[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[12] D. W. Ruck, S. K. Rogers, and M. Kabrisky, “Feature selection using a multilayer perceptron,” Journal of neural network computing, vol. 2, no. 2, pp. 40–48, 1990.

[13] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, “Multilayer perceptron and neural networks,” WSEAS Transactions on Circuits and Systems, vol. 8, no. 7, pp. 579–588, 2009.

[14] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” arXiv preprint arXiv:1707.07250, 2017.

[16] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” SIAM review, vol. 51, no. 3, pp. 455–500, 2009.