Computer Methods and Programs in Biomedicine 262 (2025) 108640

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

LMTTM-VMI: Linked Memory Token Turing Machine for 3D volumetric medical image classification

Hongkai Wei [a], Yang Yang [a], Shijie Sun [a,*], Mingtao Feng [b], Rong Wang [a], Xianfeng Han [c]

[a] School of Information Engineering, Chang'an University, Xi'an, 710064 Shaanxi, China
[b] School of Computer Science and Technology, Xidian University, Xi'an, 710126 Shaanxi, China
[c] College of Computer & Information Science, Southwest University, 400715 Chongqing, China

ARTICLE INFO

Keywords:
Neural Turing Machine
Token Turing Machine
Collaborative memory network
MedMNIST v2 dataset
3D volumetric medical image classification

ABSTRACT

Biomedical imaging is vital for the diagnosis and treatment of various medical conditions, yet the effective integration of deep learning technologies into this field presents challenges. Traditional methods often struggle to efficiently capture the spatial characteristics and intricate structural features of 3D volumetric medical images, limiting memory utilization and model adaptability. To address this, we introduce a Linked Memory Token Turing Machine (LMTTM), which utilizes external linked memory to efficiently process spatial dependencies and structural complexities within 3D volumetric medical images, aiding in accurate diagnoses. LMTTM can efficiently record the features of 3D volumetric medical images in an external linked memory module, enhancing complex image classification through improved feature accumulation and reasoning capabilities. Our experiments on six 3D volumetric medical image datasets from the MedMNIST v2 demonstrate that our proposed LMTTM model achieves average ACC of 82.4%, attaining state-of-the-art (SOTA) performance. Moreover, ablation studies confirmed that the Linked Memory outperforms its predecessor, TTM's original Memory, by up to 5.7%, highlighting LMTTM's effectiveness in 3D volumetric medical image classification and its potential to assist healthcare professionals in diagnosis and treatment planning. The code is released at https://github.com/hongkai-wei/LMTTM-VMI.

1. Introduction

Biomedical imaging is essential in contemporary medical diagnostics and disease management, with computed tomography (CT), radiography, and other modalities providing vital information for early detection of conditions such as tumors, clots, and fractures [1,2]. However, interpreting these images requires substantial medical expertise and poses challenges in time and resource management.

Despite advances in deep learning, which have led to significant strides in understanding and recognition of medical images, such as using 3DCNN to recognize Melanoma [3], using attention-based 3D multi-instance learning to detect COVID-19 [4], the integration of these deep learning technologies with biomedical imaging has still faced some obstacles.

Models such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Transformer [5] can be used to deal with spatial–temporal feature tasks, such as in 3D volumetric medical image task. However, there are also other types of neural networks, such as those that interact with external memory. Examples include the Neural Turing Machine (NTM) [6] and the Token Turing Machine (TTM) [7]. In these models, the memory module records the information the model needs to handle sequential data. This is very helpful for the model when making inferences. The memory module records the information of all the past moments, and the model extracts the valid information from the past information by interacting with the memory module. Thus, it has better generalization and can handle more complex spatial–temporal data. Furthermore, the collaborative memory network (CMN) [8] is a neural network designed to process spatial–temporal data. It captures and exploits spatial and structural features within sequences through a collaborative memory mechanism, which results in better generalization and more efficient model performance.

To address these challenges, we introduce the Linked Memory Token Turing Machine (LMTTM), a novel approach that takes advantage of both TTM [7] and CMN [8]. By incorporating the core concepts of TTM with the collaborative memory mechanisms of CMN, LMTTM is designed to efficiently process and record spatial and structural features of 3D volumetric medical images in an external linked memory module. This architecture not only facilitates enhanced feature accumulation but also improves the network's reasoning capabilities for complex

* Corresponding author.
E-mail address: shijieSun@chd.edu.cn (S. Sun).

https://doi.org/10.1016/j.cmpb.2025.108640
Received 14 August 2024; Received in revised form 4 January 2025; Accepted 1 February 2025
Available online 11 February 2025

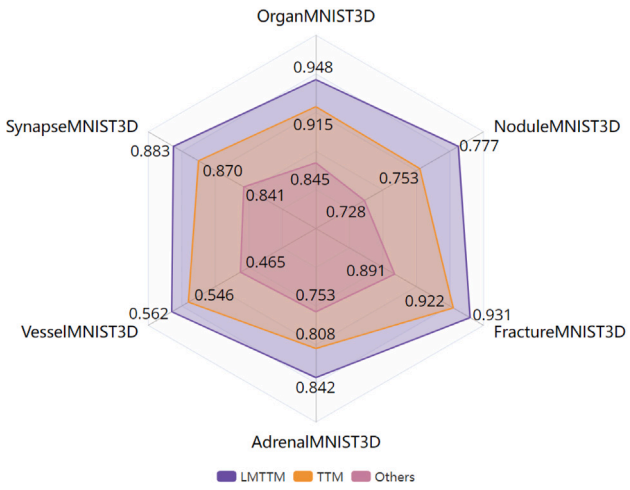

Fig. 1. LMTTM benefits TTM and other models on a broad range of datasets in 3D volumetric medical imaging tasks under the ACC metric.

image classifications. It with a significant performance boost on 3D volumetric medical imaging tasks, as shown in Fig. 1. The contributions of this paper can be summarized as follows:

- We introduce the Linked Memory Token Turing Machine (LMTTM), a novel recursive memory network that leverages a linked memory structure to efficiently manage the complexities of 3D volumetric medical image classification, surpassing the capabilities of earlier models.
- Our proposed model integrates the Tri-Temporal Memory Collaborative (T-TMC) and Memory-Augmented Feature Amplifier (MAFA), dynamically adjusting linked memory to enhance processing of spatial structural dependencies and boost feature accumulation and reasoning capabilities.
- Through comprehensive evaluations on the MedMNIST v2 datasets [9], we demonstrate that LMTTM outperforms both state-of-the-art (SOTA) models and its predecessor, TTM, in terms of accuracy and efficiency. This success lays the groundwork for future expansions of LMTTM into various AI domains, potentially leading to a unified memory approach for handling diverse data types across different fields.

2. Related works

Our review of related work in three parts highlights the evolution of deep learning and memory mechanisms in neural networks, from Recurrent Neural Networks (RNNs) to Neural Turing Machine (NTM) and Token Turing Machine (TTM), culminating in Linked Memory Token Turing Machine (LMTTM).

2.1. 3D volumetric medical image analysis

2D medical images offer clear, static views for diagnosis [10]. Ren et al. proposed LCGANT [11], which aids in early detection of lung cancer. Xu et al. introduced G2ViT [12], excelling in retinal vessel segmentation. However, 3D volumetric medical imaging provides richer spatial and structural information, significantly enhancing diagnostic and treatment capabilities, and is indispensable for addressing complex medical challenges.

Medical imaging techniques like Computed Tomography (CT) and Magnetic Resonance Angiography (MRA) offer comprehensive views for diagnosing complex conditions such as tumors and Medical imaging techniques like Computed Tomography (CT) and Magnetic Resonance Angiography (MRA) offer comprehensive views for diagnosing complex conditions such as tumors and cardiovascular diseases [13]. 3DCNN detects microhemorrhages in MRA [14], and geometric deep learning enables causal analysis and personalized prediction of 3D brain structures [15]. Additionally, Shaker et al. developed UNETR++ [16] for superior segmentation of 3D volumetric medical images. Yu et al. introduced DrasCLR [17], which excels in extracting nuanced features from 3D volumetric medical images to bolster downstream classification tasks. Wu et al. proposed VoCo [18], demonstrating its proficiency in a spectrum of 3D volumetric medical image classification tasks. Dritsas et al. presented a network [19] that stands out in classifying 3D volumetric medical images of ear canals, underscoring the impact of advanced deep learning in 3D volumetric medical imaging.

The integration of deep learning with 3D volumetric medical imaging has ushered in significant enhancements across diagnostic precision, therapeutic efficacy, and surgical precision [20]. This synergy enables the detailed mapping of a broad spectrum of pathologies and anatomical variations [21], leading to improved patient outcomes through accurate diagnoses and tailored treatment plans.

2.2. Neural network with memory

Since RNNs and Long Short-Term Memory (LSTM) [22–24] were applied to neural networks for medical image analysis, the memory mechanisms have seen a significant evolution. These models use cyclic structures to capture spatial–temporal dependencies, which are vital for tasks such as image segmentation and diagnosis. The Neural Turing Machine (NTM) [6] introduced a significant improvement by incorporating external memory. This concept has since been adopted by other models [25,26] to enhance their capabilities in the medical image analysis domain.

NTM is highly respected for its innovations in sequential data processing and storage. It especially excels in natural language processing (NLP). However, there are limitations in the application of NTM to computer vision (CV). The Token Turing Machine (TTM) [7] successfully overcomes the shortcomings of NTM by combining the external memory capability of NTM with the efficiency of the transformer architecture and applying the Token Summarization Mechanism [27,28] to efficiently read, write and process video frame information. There are some similar Token Summarization mechanisms [29–33]. TTM has proven effective for the classification of 3D volumetric medical images, such as CT and MRA, by capturing intricate spatial and structural features. However, its single-memory structure struggles with large datasets. Our study aims to address these limitations, based on the successes of TTM.

2.3. Collaborative memory networks

Collaborative Memory Network (CMN) [8] are adept at processing sequence data and memory tasks, capturing spatial–temporal dependencies through synergies of memory units [34]. CMNs dynamically capture user-project relationships through multiple hops [35], refining the collective neighborhood state with an attention mechanism [36]. This involves iterative memory coprocessing. In each step, the system queries both current and past memory to predict future states [37]. Additionally, memory modules are stacked together to form a deeper architecture, enabling complex feature recognition and iterative optimization of memory partitions.

We have developed Linked Memory Token Turing Machine (LMTTM) to overcome the limitations of traditional neural networks like TTM in complex sequence processing and to leverage CMN's collaborative memory enhancements. LMTTM's linked memory structure with efficient token summarization tackles large-scale dataset challenges and complex tasks, excelling in 3D volumetric medical image classification and demonstrating potential in various domains.

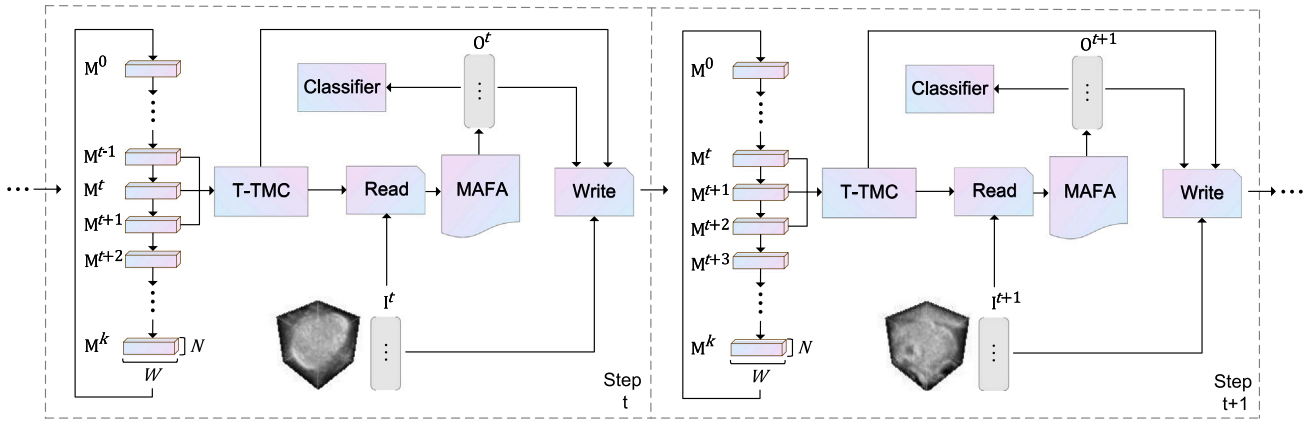

Fig. 2. Schematics of the Linked Memory Token Turing Machine (LMTTM). On the left, the network's processing at timestamp t is depicted, where the current memory block $\mathbf{M}^t$ and its adjacent blocks $\mathbf{M}^{t-1}$ and $\mathbf{M}^{t+1}$ are synergistically processed by the Tri-Temporal Memory Collaborative (T-TMC) module, which we will refer to as Simplified Memory. Simplified memory, along with preprocessed input tokens, is read through the Read module, and the extracted information is then refined into more efficient output tokens by the Memory-Augmented Feature Amplifier (MAFA) for image classification. The simplified Memory, input tokens, and output tokens are then written in the subsequent Memory block $\mathbf{M}^{t+1}$ via the Write module, completing the LMTTM cycle for the timestamp t and facilitating a linked cyclic update of the memory content. At timestamp $t+1$, the processing is shown on the right side of the schematics. The input tokens for this timestamp, $\mathbf{I}^{t+1}$, interact with the Linked Memory blocks. The T-TMC module dynamically adjusts these blocks to include the current memory block $\mathbf{M}^{t+1}$, the preceding memory block $\mathbf{M}^t$, and the future memory block $\mathbf{M}^{t+2}$. This dynamic adjustment aligns with the chain-like characteristic of the Linked Memory, catering to the present, past, and future memory at timestamp $t+1$. Consequently, the Read module can extract pertinent information, which is then refined by the MAFA for the classification task specific to timestamp $t+1$.

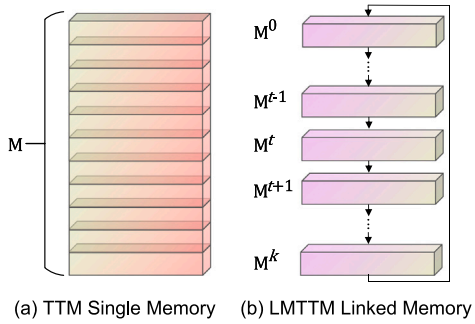

Fig. 3. Illustration of different memory. (a) is the memory structure of TTM, where the entire memory block is iteratively updated during the interaction. (b) is the memory structure of our LMTTM, which can be viewed as a chain of k memory units linked together, and is iteratively updated one memory unit at a time during the interaction.

3. Methodology

Linked Memory Token Turing Machine (LMTTM) aims to provide an innovative upgrade to the Token Turing Machine (TTM) architecture in order to realize a robust new approach to classify 3D volumetric medical images. As shown in Fig. 2, the core innovation of LMTTM is the transformation of otherwise sequential memory blocks into linked memory blocks. In timestamp t, the memory $\mathbf{M}^{t-1}, \mathbf{M}^t, \mathbf{M}^{t+1} \in \mathbb{R}^{m \times d}$ consists of multiple sets of m tokens of dimension d. In LMTTM, the interaction between the processing unit and the memory is performed entirely through read and write operations. Unlike TTM, memory read operations do not involve the entire memory block but rather blocks of memory for the past, current, and future moments. The result of the read operation is then passed to the Memory-Augmented Feature Amplifier (MAFA). The output of the MAFA is then written in future memory blocks.

In LMTTM, the input of the timestamp $\mathbf{I}^t \in \mathbb{R}^{n \times d}$ interacts with the memory $\mathbf{M}^{t-1}$ of the past time, the memory $\mathbf{M}^t$ of the current time and the memory $\mathbf{M}^{t+1}$ of the future time, to retrieve the relevant tokens. These tokens are then further processed to produce the output $\mathbf{O}^t \in \mathbb{R}^{r \times d}$. The output of the current step is then merged with the previous input, the memory of the past moments, the memory of the current moments, and the memory of the future moments, from which the relevant tokens are retrieved and written in the memory of the future moment $\mathbf{M}^{t+1}$, in preparation for the next timestamp $t+1$. To fulfill the need for prediction at each timestamp in many sequential decision-making tasks, LMTTM includes a linear output header at each timestamp.

3.1. Linked memory structure

As shown in Fig. 3(a), compared to the single contiguous memory structure used in TTM, LMTTM introduces an innovative memory chunking mechanism. As shown in Fig. 3(b), this mechanism further demonstrates this mechanism, in which the originally continuous memory space $\mathbf{M} \in \mathbb{R}^{k \times d}$ is divided into multiple blocks, each of which contains m tokens of dimension d. The blocks in this context are not standalone elements but rather interconnected sequentially to create a looped memory structure. This particular design enables the model to access and update tokens continuously and cyclically while processing information. As a result, the efficiency and flexibility of the model in handling spatial–temporal data are improved.

3.2. Tri-temporal memory collaborative

Tri-Temporal Memory Collaborative (T-TMC) in LM-TTM is crucial to extract memory blocks from linked memories. It combines the memory block $\mathbf{M}^{t-1}$ for the past moment, $\mathbf{M}^t$ for the current moment, and $\mathbf{M}^{t+1}$ for the future moment to form a coherent representation. This integration is critical for subsequent I/O operations, as it provides the necessary contextual information that allows the model to make more precise and in-depth decisions when processing spatial–temporal data. This process allows the model to simultaneously use cumulative knowledge of historical data, immediate information about current data, and predictions of future trends.

T-TMC enhances the model's grasp of time-series spatial–temporal dependencies and its adaptability to new data. During reads, the model uses this integrated memory to extract pertinent information, vital for accurate classification and prediction in 3D volumetric medical image analysis.

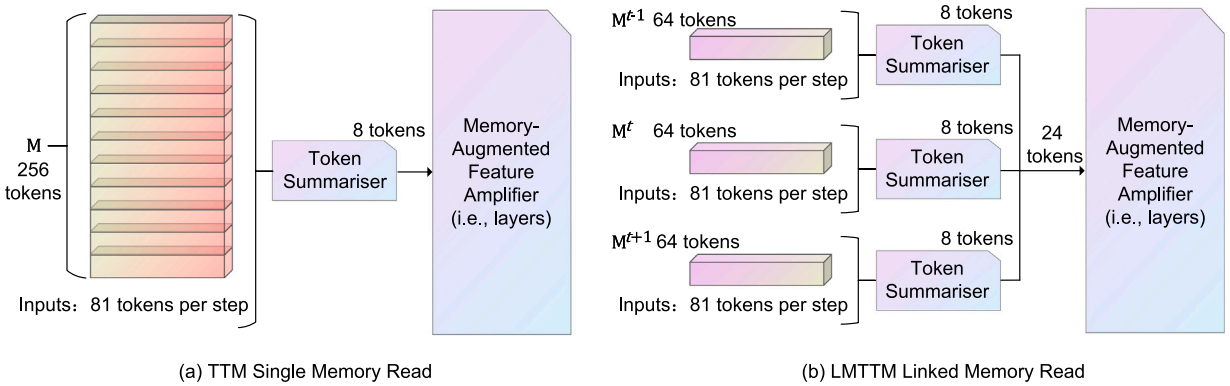

Fig. 4. Illustration of different Reads. (a) is the entire block of memory that will be read by the TTM. (b) is the coordinated read operation for the current moment t, the past moment $t-1$ and the future moment $t+1$.

3.3. Token summarization

Suppose that we are dealing with a sequence of tokens $\mathbf{V} \in \mathbb{R}^{n \times d}$, where n is the number of tokens and d is the dimensionality of the tokens. The core task of Token Summarization is to compress these tokens into a smaller set $\mathbf{Z} \in \mathbb{R}^{k \times d}$, where k is the number of tokens we wish to obtain, usually $k \ll n$. Our approach is similar to the Token Summarization mechanism in the TTM, but on top of the TTM, we add [29], a design choice motivated by the simplicity, full differentiability, and excellent performance achieved by the STTS approach in several domains. By incorporating STTS, our model is able to efficiently capture key information in both spatial and temporal dimensions when processing data, leading to efficient aggregation and summarization of tokens.

In the LMTTM token summarization module, one-dimensional convolutional techniques are utilized to dynamically fuse token information into linked memory. First, the input data $\mathbf{I}$ is normalized to obtain $\mathbf{I}'$. The attention weight matrix $\mathbf{A}$ is calculated by one-dimensional convolution and the GELU activation function, and then converted to the weight matrix $\mathbf{W}$. To ensure that the sum of the weights is 1, the Softmax function is applied to obtain the normalized weight matrix $\mathbf{W}'$. The input data $\mathbf{I}$ are passed through a token convolution layer to extract tokens and form a token matrix $\mathbf{F}$. Finally, the normalized weight tensor of attention $\mathbf{W}'$ is multiplied by the token matrix $\mathbf{F}$, and weighted aggregation is achieved by Einsum operation to obtain the labeled aggregate representation $\mathbf{Z}$. This formula realizes the summation over the index k through the Einsum operation, that is, as follows:

$$\mathbf{Z}_{ij} = \sum_{k=1}^{n} \mathbf{W}'_{ij} \mathbf{F}_{kj}, \tag{1}$$

where, n is the dimension of the tensor, which bounds the upper limit of the summation. In this way, we are able to combine the weights of the attention mechanism with the elements of the token matrix to obtain a new representation of each token.

This method merges attention weights with token elements to form a novel representation for each token. The aggregated result, denoted as $\mathbf{Z}$, enhances the model's generalization ability by using the dropout layer, while the LMTTM refines the token representation by effectively extracting crucial information through these processes. The algorithm of Token Summarization is shown in Algorithm 1.

3.4. Feature amplification and adaptive I/O

LMTTM efficiently extracts information from memory through the Memory-Augmented Feature Amplifier (MAFA), while the adaptive I/O mechanism ensures accurate read and write of information. This process not only optimizes the extraction and storage of information, but also significantly improves the accuracy and efficiency of the model in processing 3D volumetric medical image classification tasks by dynamically adjusting the I/O operations.

Algorithm 1: Token Summarization

Input: Input token sequence $\mathbf{I}$
Output: Compressed token sequence $\mathbf{Z}$
Initialization:
Normalize input tokens: $\mathbf{I}' = \text{Normalize}(\mathbf{I})$
Procedure: # Perform token summarization
1. Compute attention weights: $\mathbf{A} = \text{GELU}(\text{Conv1D}(\mathbf{I}'))$
Apply 1D convolution and GELU activation to obtain attention matrix $\mathbf{A}$
2. Normalize attention weights: $\mathbf{W}' = \text{Softmax}(\mathbf{A})$
Apply Softmax along the token dimension to ensure normalized attention weights $\mathbf{W}'$
3. Extract token features: $\mathbf{F} = \text{Conv1D}(\mathbf{I})$
Extract token features through a token convolution layer to form token matrix $\mathbf{F}$
4. Compute weighted aggregation: $\mathbf{Z} = \text{Einsum}(\mathbf{W}', \mathbf{F})$
Perform weighted aggregation of tokens using the Einsum operation:

$$\mathbf{Z}_{ij} = \sum_{k=1}^{n} \mathbf{W}'_{ij} \mathbf{F}_{kj}$$

Aggregate information from n input tokens into k compressed tokens
Output: Compressed token sequence $\mathbf{Z}$

3.4.1. Read from linked memory

TTM employs a cohesive approach to read memory inputs. It handles and merges inputs and outputs independently, providing it with unique benefits over neural Turing machines (NTMs) in managing continuous data. Similarly, our LMTTM adopts this unified memory input reading approach. However, in contrast to TTM, LMTTM integrates the idea of dynamic memory coprocessing. It segments a large memory block into smaller sections and concentrates on the past, present, and future instances of these smaller sections.

As shown in Fig. 4(a), TTM connects the memory $\mathbf{M}$ consisting of k tokens to the input stream $\mathbf{I}^t$ consisting of n tokens and summarizes these tokens into a smaller subset of r tokens. LMTTM, on the other hand, as shown in Fig. 4 (b), divides memory $\mathbf{M}$ into small blocks, each containing m tokens, and focuses on processing memory blocks $\mathbf{M}^{t-1}$ in past moments, $\mathbf{M}^t$ in current moments, and $\mathbf{M}^{t+1}$ in future moments. Thus, our read operator is defined as:

$$\mathbf{Z}^{t-1} = \text{Read}(\mathbf{M}^{t-1}, \mathbf{I}^t) = S_r([\mathbf{M}^{t-1} || \mathbf{X}^t]), \tag{2}$$

$$\mathbf{Z}^{t} = \text{Read}(\mathbf{M}^{t}, \mathbf{I}^t) = S_r([\mathbf{M}^{t} || \mathbf{X}^t]), \tag{3}$$

$$\mathbf{Z}^{t+1} = \text{Read}(\mathbf{M}^{t+1}, \mathbf{I}^t) = S_r([\mathbf{M}^{t+1} || \mathbf{X}^t]), \tag{4}$$

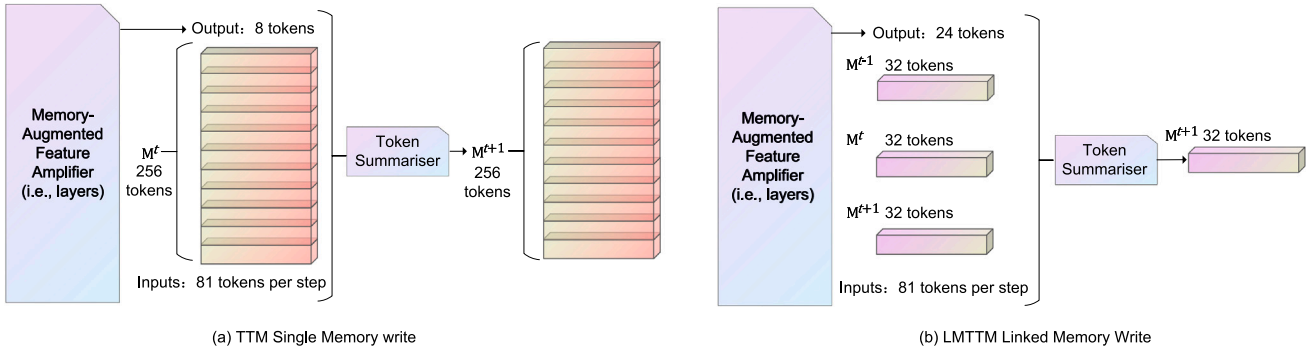

Fig. 5. Illustration of different Writes. (a) is the entire block of memories from the current moment in TTM being written to the next moment in the entire block. (b) is the entire block of tri-temporal memories from the current moment t, the past moment $t-1$, and the future moment $t+1$ being synergized and written to the future moment $t+1$ in LMTTM.

$$\mathbf{P}^t = [\mathbf{Z}^{t-1} \| \mathbf{Z}^t \| \mathbf{Z}^{t+1}], \tag{5}$$

where, $[\mathbf{M}^t || \mathbf{X}^t]$ denotes the concatenation of these two matrices. The $\mathbf{P}^t$ is obtained by performing read operations on the memory blocks of each of the three tensors and then concatenating the read results $\mathbf{Z}^{t-1}$, $\mathbf{Z}^t$ and $\mathbf{Z}^{t+1}$ to obtain $\mathbf{P}^t$. This process is essentially a mapping of $\mathbb{R}^{3(m+n)\times d} \rightarrow \mathbb{R}^{3r\times d}$. Thus, the read operator not only filters the input and the information in memory but also dynamically combines the information from the three important moments of the past, present, and future, which are subsequently passed on to the subsequent processing units. In particular, by reducing the number of tokens passed to the processing modules, we significantly reduce the computational cost of this phase.

Furthermore, we integrate a trainable localization embedding in each read module, enabling the token summarization module to perform content-based memory addressing, i.e., "content-based reading" in NTMs. This mechanism combines the location and content information of tokens, improving the accuracy of the model in recognizing information in memory without changing the original process.

3.4.2. Memory-augmented feature amplifier (MAFA)

In the LMTTM architecture, the processing unit is designed as $\mathbf{O}^t = \text{MAFA}(\mathbf{P}^t)$, which receives the tokens $3r$ obtained from the read operation $\mathbf{P}^t$ and processes them. The MAFA outputs a set of vectors $\mathbf{O}^t$ containing the $3r$ tokens, which are used in subsequent write operations and in generating predictions in the 3D volumetric medical image classification task. For tasks that require predictions at each timestamp, we introduce a fully connected layer after $\mathbf{O}^t$ to enhance the classification capability of the model, which is implemented as $\mathbf{Y}^t = \text{Classifier}(\mathbf{O}^t) = \mathbf{W}_o\mathbf{O}^t$, where $\mathbf{W}_o$ is the weight matrix of the fully connected layer used to map $\mathbf{O}^t$ to the input space of the classifier. This produces the final classification prediction.

Through our experiments, we determined that the network model achieved optimal performance with the standard Transformer serving as the Memory-Augmented Feature Amplifier (MAFA). We also investigated other architectures, including MLP and MLPMixer, as potential alternatives for the MAFA meta. Consequently, our model is capable of effectively learning the characteristics of 3D volumetric medical images and making highly accurate predictions in classification tasks.

3.4.3. Write to linked memory

Similar to our read operation, our write operation is designed as a token summarization process that is both simple and efficient. As illustrated in Fig. 5(a), the TTM write mechanism ensures that the information in memory $\mathbf{M}^t$ is preserved by enabling token reselection. The token updates in memory are performed by choosing tokens from $\mathbf{O}^t$ and $\mathbf{I}^t$ of the processing module. However, TTM memory operation is considered to be too time consuming and space consuming. Conversely, LMTTM (depicted in Fig. 5(b)) uses only the memory blocks of the past, present, and future tenses to update tokens during writing. It writes the updated tokens solely to the future memory block, making better use of future information. This method greatly improves efficiency and reduces time and space costs, surpassing TTM in accuracy.

Therefore, the process of our write operation can be expressed as the selection of m tokens, which represents the size of the future memory block. These tokens are chosen from a combination of tri-temporal memory blocks, input tokens, and output tokens using Token Summarization, which is denoted as:

$$\begin{aligned}\mathbf{M}^{t+1} &= \text{Write}(\mathbf{M}^{t-1}, \mathbf{M}^t, \mathbf{M}^{t+1}, \mathbf{I}^t, \mathbf{O}^t) \\ &= S_n([\mathbf{M}^{t-1} \| \mathbf{M}^t \| \mathbf{M}^{t+1} \| \mathbf{I}^t \| \mathbf{O}^t]),\end{aligned} \tag{6}$$

Like the read operation, write operation also incorporates a positional (+content) writing mechanism, which can be seen as a mapping function of $\mathbb{R}^{(3m+n+3r)\times d} \rightarrow \mathbb{R}^{m\times d}$. Subsequently, $\mathbf{M}^{t+1}$ is sent to the next round of memory in the LMTTM, forming a self-looping pattern. The algorithm of LMTTM Main Function is shown in Algorithm 2.

3.5. Loss function

In the LMTTM, after MAFA processes the tokens, the output $\mathbf{O}^t$ is utilized for generating predictions in the 3D volumetric medical image

Algorithm 2: LMTTM Main Function

Input: Input token $\mathbf{I}^t$, Memory $\mathbf{M}$
Output: Classifier Vector $\mathbf{Y}^t$, Updated memory $\mathbf{M}'$
Procedure:
1. $\mathbf{O}_{\text{list}} = [\ ]$ # Initialize an empty list
2. **for** $t = 1$ *to* T **do**
   - **Read Operation:**
   - $(\mathbf{M}^{t-1}, \mathbf{M}^t, \mathbf{M}^{t+1}) = \text{T-TMC}(\mathbf{M})$
   - $\mathbf{P}^t = \text{Read}(\mathbf{M}^{t-1}, \mathbf{M}^t, \mathbf{M}^{t+1}, \mathbf{I}^t)$
   - # Read past, current, and future memory blocks
   - **Processing:**
   - # Process memory blocks and input tokens
   - $\mathbf{O}^t = \text{MAFA}(\mathbf{P}^t)$
   - **Write Operation:**
   - $\mathbf{M}^{t+1} = \text{Write}(\mathbf{M}^{t-1}, \mathbf{M}^t, \mathbf{M}^{t+1}, \mathbf{I}^t, \mathbf{O}^t)$
   - $\mathbf{M}' \leftarrow (\cdots, \mathbf{M}^{t+1}, \cdots)$
   - # Write updated portion to linked memory
   - Append the output tokens: $\mathbf{O}_{\text{list}}$.append($\mathbf{O}^t$)
3. Combine outputs: $\mathbf{O}_{stack} = \text{stack}(\mathbf{O}_{list}, \text{dim} = 1)$
Stack outputs across time steps into a single tensor
4. Apply feature pooling:
$\mathbf{O}_{pooled} = \text{AdaptiveAvgPool1d}(1)(\mathbf{O}_{stack})$
Adaptive average pooling to aggregate features
5. Compute classification logits: $\mathbf{Y}^t = \text{Classifier}(\mathbf{O}_{pooled})$
Output: $(\mathbf{Y}^t, \mathbf{M}^{t+1})$

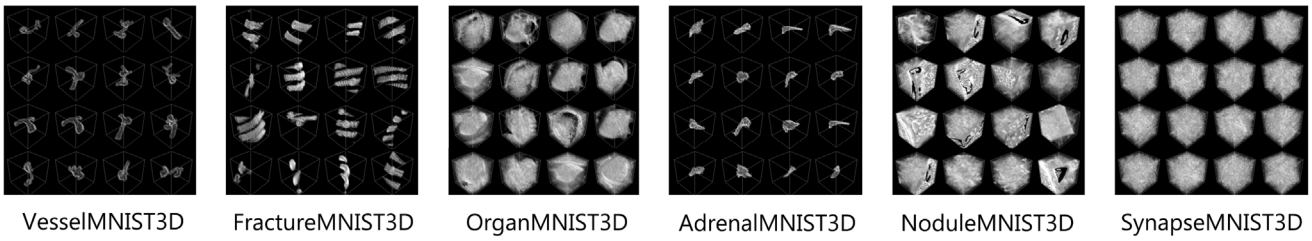

Fig. 6. Visualization of 3D datasets in MedMNIST v2. MedMNIST v2 contains a collection of six pre-processed 3D volumetric medical image datasets. It is designed to be educational, standardized, diverse, and lightweight and can be used as a general classification benchmark in 3D volumetric medical image analysis.

Table 1
MedMNIST v2 contains 6 biomedical 3D image datasets. (BC: binary classification task, MC: multiclassification task.).

Dataset name	Data modality	Task(#Classes)	#Samples	#Training/Validation/Test	#Resolution
OrganMNIST3D	Abdominal CT	MC(11)	1743	972/161/610	- → 28 × 28 × 28
NoduleMNIST3D	Chest CT	BC(2)	1633	1,158/165/310	1MM×1MM×1MM → 28 × 28 × 28
AdrenalMNIST3D	Shape from AbdominalCT	BC(2)	1584	1,188/98/298	64M×64M×64M → 28 × 28 × 28
FractureMNIST3D	Chest CT	MC(3)	1370	1,027/103/240	64M×64M×64M → 28 × 28 × 28
VesselMNIST3D	Shape from Brain MRA	BC(2)	1909	1 335/192/382	- → 28 × 28 × 28
SynapseMNIST3D	Electron Microscope	BC(2)	1759	1,230/177/352	1024 × 1024 × 1024 → 28 × 28 × 28

classification task. To quantify the difference between the predicted outputs and the true labels, we employ the Cross-Entropy Loss as our loss function, which is an effective choice for multi-class classification problems. It increases as the predicted probability diverges from the actual label, providing a clear indication of the model's accuracy. The loss $\mathcal{L}_{class}$ for a single sample is calculated as follows:

$$\mathcal{L}_{class} = \mathcal{L}_{CE} = -\sum_{c=1}^{C} y_{o^t,c} \log(p_{o^t,c}), \tag{7}$$

where, C is the number of classes, $y_{o^t,c}$ is a binary indicator of whether class c is the correct classification for observation o^t, and $p_{o,c}$ is the probability that observation o^t is of class c as predicted by the model.

Cross-Entropy Loss $\mathcal{L}_{CE}$ helps our LMTTM classify 3D volumetric medical images more accurately. It makes the model better by catching mistakes and improving its guesses, which is important for diagnosing diseases.

4. Experiments

4.1. Datasets

The MedMNIST v2 dataset [9] in Fig. 6 is a comprehensive collection of standardized biomedical images used by the Linked Memory Token Turing Machine(LMTTM). It consists of six temporally characterized 3D volumetric medical image datasets representing a variety of medical imaging modalities such as computed tomography (CT), magnetic resonance angiography (MRA), and electron microscope. These datasets are designed to be used for a variety of classification tasks, as shown in Table 1.

OrganMNIST3D: This is an 11-class task for the classification of human organs using 3D CT images from the Liver Tumor Segmentation Benchmark (LiTS) [38]. It uses bounding-box annotations from a separate study [39] to label the organs.

NoduleMNIST3D: This is a dataset based on LIDC-IDRI 32 [40], which belongs to the binary classification task in 3D images of chest CT scans, determining whether a sample is a positive or negative tumor class.

AdrenalMNIST3D: This is a dataset based on the binary classification task of the normal adrenal gland or the adrenal mass, which is based on 3D CT images.

FractureMNIST3D: This is derived from RibFrac [41], focusing on 3D CT images of rib fractures. It classifies rib fractures into three distinct categories.

VesselMNIST3D: This is based on the open access 3D intracranial aneurysm dataset [42], belonging to the MRA images of blood vessels in the brain in 3D, which is used to identify and categorize cerebrovascular structures.

SynapseMNIST3D: This comprises electron microsco-py images for distinguishing between excitatory and inhibitory brain synapses, utilizing a subvolume from MitoEM dataset [43] for dense 3D mitochondrial instance segmentation.

4.2. Evaluation metric

Accuracy (ACC) and the area under the ROC curve (AUC) are utilized for assessing the model performance on the MedMNIST v2 dataset. ACC indicates the percentage of accurately predicted samples compared to the total number of samples during inference. Conversely, AUC offers a more holistic evaluation of the model's predictive prowess. Normally, AUC values fall between 0.5 and 1, where scores below 0.5 suggest subpar inference performance, while values approaching 1 signify higher predictive ability.

The AUC and ACC metrics for various models, such as ResNets, autosklearn, AutoKeras, and Token Turing Machine (TTM), were evaluated. These metrics were then compared to the final performance metrics of our LMTTM model across six 3D datasets obtained from MedMNIST v2.

4.3. Implementation details

We trained and evaluated our LMTTM model on six 3D volumetric medical image datasets provided by MedMNISTv2. An official evaluation tool was employed to benchmark the model's performance against current state-of-the-art (SOTA) models. To improve robustness, various noises were introduced into the Memory module during the training phase.

Initially, we evaluate the model's training inference performance on six datasets. Subsequently, we conduct ablation studies to examine the impact of introducing various types of noise into the memory module and altering the memory capacities on the inference outcomes. The detailed experimental results will be discussed in the following section.

The experiments utilized six NVIDIA TITAN Xp GPUs, employing a learning rate of 0.0001 and the Adam optimizer. Features were derived from 3D volumetric medical images and processed in a sequence of 28 spatial–temporal frames. Read and write operations integrated current, past, and future memories for each timestamp, and a classifier produced the final fused outputs and predictions. The memory module was dynamically updated during the read and write operations, with the detailed procedure illustrated in Fig. 2.

Table 2
Comparison of the performance of LMTTM and other models on six 3D datasets under the MedMNIST v2 dataset.

Methods	# of params	OrganMNIST3D		NoduleMNIST3D		FractureMNIST3D		AdrenalMNIST3D		VesselMNIST3D		SynapseMNIST3D	
		AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18+2.5D[44]	11M	0.977	0.788	0.838	0.835	0.587	0.451	0.718	0.772	0.748	0.846	0.634	0.696
ResNet-18+3D[44]	33M	0.996	0.907	0.863	0.844	0.712	0.508	0.827	0.721	0.874	0.877	0.820	0.745
ResNet-18+ACS[44]	11M	0.994	0.900	0.873	0.847	0.714	0.497	0.839	0.754	**0.930**	0.928	0.705	0.722
ResNet-50+2.5D[44]	15M	0.974	0.769	0.835	0.848	0.552	0.397	0.732	0.763	0.751	0.877	0.669	0.735
ResNet-50+3D[44]	44M	0.994	0.883	0.875	0.847	0.725	0.494	0.828	0.745	0.907	0.918	**0.851**	**0.795**
ResNet-50+ACS[44]	15M	0.994	0.889	0.886	0.841	**0.750**	0.517	0.828	0.758	0.912	0.858	0.719	0.709
ACS-Conv [45]	11M	0.992	0.897	0.790	0.819	0.581	0.438	0.791	0.801	0.843	0.910	0.615	0.711
auto-sklearn[46]	–	0.977	0.814	**0.914**	0.874	0.628	0.453	0.828	0.802	0.910	0.915	0.631	0.730
AutoKeras[47]	–	0.979	0.804	0.844	0.834	0.642	0.458	0.804	0.705	0.773	0.894	0.538	0.724
FPVT [48]	–	0.923	0.800	0.814	0.822	0.640	0.438	0.801	0.704	0.770	0.888	0.530	0.712
TTM[7]	21M	0.995	0.915	0.873	0.870	0.662	0.546	0.816	0.808	0.855	0.922	0.709	0.753
LMTTM(Ours)	21M	**0.997**	**0.948**	0.891	**0.883**	0.691	**0.562**	**0.847**	**0.842**	0.917	**0.931**	0.737	0.777

Table 3
Average experiment results on six MedMNIST 3D datasets.

Methods	Average	
	AUC	ACC
ResNet-18 + 2.5D[44]	0.750	0.731
ResNet-18 +3D[44]	0.849	0.767
ResNet-18 + ACS[44]	0.843	0.775
ResNet-50 + 2.5D[44]	0.752	0.732
ResNet-50 +3D[44]	**0.863**	0.780
ResNet-50 + ACS[44]	0.848	0.762
ACS-Conv [45]	0.769	0.654
auto-sklearn[46]	0.815	0.765
AutoKeras[47]	0.763	0.737
FPVT [48]	0.746	0.623
TTM[7]	0.818	0.802
LMTTM(Ours)	0.847	**0.824**

Table 4
Impact of epochs on model performance.

Epoch	VesselMNIST3D		AdrenalMNIST3D	
	AUC	ACC	AUC	ACC
50	0.823	0.842	0.714	0.732
100	**0.917**	**0.931**	**0.847**	**0.842**
200	0.912	0.928	0.823	0.816

4.4. Results

4.4.1. Results on each dataset

The comparative results of our proposed LMTTM method against the current state-of-the-art (SOTA) methods on the AUC and ACC evaluation metrics for the MedMNIST3D dataset are presented in Table 2. Compared to existing SOTA methods and the previous TTM, LMTTM demonstrates superior learning performance on various evaluation metrics. Specifically, for the AdrenalMNIST3D dataset, we note a 4.0% increase in ACC and a 0.8% increase in AUC; for the OrganMNIST3D dataset, a 3.3% increase in ACC and a 0.1% increase in AUC; for the FractureMNIST3D dataset, a 1.6% increase in ACC; for the NoduleMNIST3D dataset, the ACC increased by 0.9%; and for the VesselMNIST3D dataset, the ACC increased by 0.3%.

Our experiments on the MedMNIST3D dataset show that LMTTM performs well in 3D volumetric medical image classification. Specifically, LMTTM excels in the AdrenalMNIST3D, OrganMNIST3D, FractureMNIST3D, NoduleMNIST3D, and VesselMNIST3D datasets. It surpasses the state-of-the-art in ACC and AUC for the AdrenalMNIST3D and OrganMNIST3D datasets. LMTTM also shows strong performance in ACC and AUC across the other datasets. Additionally, we will discuss the robustness and memory capacity in the following sections. These findings suggest that the proposed LMTTM design is not only effective but also demonstrates excellent generalization capabilities.

4.4.2. Average performance

As illustrated in Table 3, when comparing the average AUC and the average ACC of our method against other methods in all datasets, our LMTTM model attains an average AUC of 84.7% and an average ACC of 82.4% in six different datasets. LMTTM notably surpasses all baseline models, including ResNets, AutoML methods, and its predecessor TTM methods, in terms of average ACC, and also exceeds nearly all methods in terms of average AUC. The outstanding performance of LMTTM in both the average AUC and the average ACC across multiple datasets not only demonstrates its significant benefits in 3D volumetric medical image classification tasks but also introduces a novel token-based cyclic memory network architecture into the field.

Unlike other models, LMTTM's main benefits are its linked memory, dynamic Token Summarization method for reading and writing, and Transformer processing layer. This novel approach is expected to drive forward the development of 3D volumetric medical image classification technology and make a substantial contribution to medical diagnosis and related research.

4.5. Hyperparameter analysis

We also analyzed the impact of several hyperparameters on model performance, including the training epoch, learning rate, and batch size. A series of experiments were conducted to determine the optimal hyperparameters for the task. All hyperparameter experiments were based on a memory module size of 448 and a memory dimension of 448.

Training Epoch Selection. We evaluated the impact of epoch numbers on model performance, finding that excessive epochs lead to overfitting, while too few lead to undertraining. Testing epoch values of 50, 100, and 200, we identified 100 epochs as the optimal choice, achieving a balance between training and generalization, as shown in Table 4.

Learning Rate Sensitivity. The learning rate plays a crucial role in model convergence. A high learning rate (0.0010) caused significant loss fluctuations and failed to achieve optimal performance after 100 epochs. A moderate learning rate (0.0005) converged faster but resulted in lower accuracy. In contrast, a low learning rate (0.0001) achieved the best performance, balancing convergence stability and accuracy, as shown in Table 5.

Batch Size Dependency. The hyperparameter batch size also affects model performance. Larger batch sizes provide more stable gradient estimates, while smaller ones introduce noise. Experiments with batch sizes of 16, 32, and 64 showed that a batch size of 32 achieved the best performance, followed by 64, as shown in Table 6.

Through basic hyperparameter experiments, the optimal settings for subsequent experiments were determined: 100 epochs, a learning rate of 0.0001, and a batch size of 32.

Table 5
Learning rate effects on model performance.

LR	VesselMNIST3D		AdrenalMNIST3D	
	AUC	ACC	AUC	ACC
0.0010	0.621	0.648	0.582	0.589
0.0005	0.849	0.853	0.768	0.752
0.0001	**0.917**	**0.931**	**0.847**	**0.842**

Table 6
Batch size influence on model performance.

Batch	VesselMNIST3D		AdrenalMNIST3D	
	AUC	ACC	AUC	ACC
16	0.903	0.927	0.833	0.829
32	**0.917**	**0.931**	**0.847**	**0.842**
64	0.912	0.930	0.846	0.840

4.6. Ablation studies

4.6.1. Ablation study for linked memory robustness

To assess the robustness of the LMTTM memory module, we conducted experiments on six subsets of the MedMNIST v2 dataset. We introduced noise with various distributions into the memory module, such as uniform, Laplace, normal, exponential, gamma, and Poisson distributions, to mimic the variations that may occur in actual medical images, as illustrated in Table 7.

By evaluating the precision before and after the introduction of noise, we discovered that in certain instances, such as with the NoduleMNIST3D and AdrenalMNIST3D datasets, the inclusion of specific noise types (e.g., Laplace and Poisson distributions) enhanced model accuracy by 2.9% and 2.0%, respectively. For other datasets like VesselMNIST3D and FractureMNIST3D, Exponentially distributed and Poisson-distributed noises also led to accuracy increases of 1.7% and 2.1%. These findings indicate that LMTTM can effectively handle noise and even leverage it to improve learning in some scenarios.

Ensuring the stability of LMTTM in noisy settings is essential for medical applications, as medical image data often include noise. Our research not only verifies the potential of LMTTM for 3D volumetric medical image classification but also shows its effectiveness in handling complex data environments.

4.6.2. Ablation study for linked memory capacity

We conducted ablation studies on Linked Memory Capacity using the AdrenalMNIST3D and VesselMNIST3D datasets. The capacity size of the Linked Memory module within the LMTTM framework was adjusted and the model was re-trained to evaluate its importance. Furthermore, we carried out experiments on the TTM to showcase the Memory module's effectiveness. The results of these experiments are presented in Table 8.

The results indicate that the model with LMTTM as the backbone consistently outperforms the model with TTM as the backbone in terms of ACC metrics, given the same memory capacity. As illustrated in Table 9, on the VesselMNIST3D dataset, with a dimension of 352 and a memory size of 64, the ACC of LMTTM is about 1.60% higher than that of TTM. When the memory size is 160, the ACC of LMTTM is approximately 1.22% higher than TTM. Similarly, on the AdrenalMNIST3D dataset, with a dimension of 256 and a memory size of 64, the ACC of LMTTM exceeds that of TTM by around 1.15%. When the memory size is 256, the ACC of LMTTM is higher than TTM by about 1.51%.

The results indicate that, for both TTM and LMTTM, an increase in memory capacity leads to improved model performance on the dataset.

5. Discussion & conclusion

Discussion of Limitations. Although the LMTTM presents a robust framework for 3D volumetric medical image classification, it is not without limitations:

- **Linked Memory Flexibility:** The linked memory in LMTTM, though innovative, does not yet match the functional adaptability of the human brain, particularly when dealing with varying dataset sizes and complexities. This restricts the model's scalability and adaptability.
- **Domain Diversity:** LMTTM's development and validation have been largely confined to the medical imaging domain. Its adaptability to other domains and data types remains unexplored, which limits its versatility and broader applicability.
- **Broader Generalization:** With performance primarily assessed on the MedMNIST v2 dataset, LMTTM's ability to generalize across different datasets and real-world applications is yet to be fully established.

Challenges and Future Work. To overcome these limitations, our future research will focus on:

- **Human Brain-Inspired Memory Enhancements:** We aim to develop memory architectures that emulate the human brain's functional segmentation, allowing for more efficient and adaptive processing of large and complex datasets.
- **Cross-Domain Model Applicability:** We plan to extend LMTTM's application beyond medical imaging, targeting a unified memory capable of handling diverse data types across various domains, thus enhancing its versatility and applicability.
- **Comprehensive Performance Validation:** We will conduct extensive experiments on additional datasets and in different domains to provide a comprehensive evaluation of LMTTM's performance and to uncover any domain-specific challenges.

To address the limitations and align with our future work, we plan to upgrade the LMTTM's memory architecture by initially drawing inspiration from the differentiable neural computer [49] and then emulating the human brain's 52 distinct regions to enhance its adaptability. We aim to integrate this enhanced model within various medical imaging domains before expanding its application across different sectors, to demonstrate its versatility and effectiveness.

Conclusion. This study introduces the Linked Memory Token Turing Machine (LMTTM), a novel approach to 3D volumetric medical image classification that surpasses its predecessor, TTM, and current state-of-the-art (SOTA) models. Our experiments demonstrate that LMTTM achieves average ACC of 82.4%, highlighting its effectiveness in accurately classifying complex medical images. While LMTTM has shown impressive results, we recognize its limitations and are actively planning future enhancements to refine the model further and expand its applications in medical diagnostics and artificial intelligence.

CRediT authorship contribution statement

Hongkai Wei: Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Yang Yang:** Writing – review & editing, Writing – original draft, Methodology. **Shijie Sun:** Writing – review & editing, Resources, Project administration, Data curation, Conceptualization. **Mingtao Feng:** Writing – review & editing, Software, Investigation. **Rong Wang:** Writing – original draft, Software, Resources, Data curation. **Xianfeng Han:** Writing – review & editing, Visualization, Resources, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 7
The result of mixing different types of noise in the memory module.

OrganMNIST3D		NoduleMNIST3D		FractureMNIST3D		AdrenalMNIST3D		VesselMNIST3D		SynapseMNIST3D	
Noise	ACC	Noise	ACC	Noise	ACC	Noise	ACC	Noise	ACC	Noise	ACC
None	0.926	None	0.818	None	0.469	None	0.790	None	0.901	None	**0.741**
Uniform	0.907	Uniform	0.803	Uniform	0.438	Uniform	0.795	Uniform	0.904	Uniform	0.725
Laplace	0.911	Laplace	**0.842**	Laplace	0.434	Laplace	0.767	Laplace	0.908	Laplace	0.713
Normal	0.917	Normal	0.823	Normal	0.473	Normal	0.787	Normal	0.897	Normal	0.726
Exp	0.920	Exp	0.818	Exp	**0.479**	Exp	0.789	Exp	**0.916**	Exp	0.713
Gamma	0.925	Gamma	0.830	Gamma	0.467	Gamma	0.783	Gamma	0.906	Gamma	0.714
Poisson	**0.931**	Poisson	0.828	Poisson	**0.479**	Poisson	**0.806**	Poisson	0.904	Poisson	0.726

Table 8
Results of ablation experiments on two datasets. In each column of the specified DIM, the TTM inference accuracy is shown on the left and our LMTTM is shown on the right.

AdrenalMNIST3D								VesselMNIST3D							
Memory	Dim = 64		Dim = 160		Dim = 256		Dim = 352		Memory	Dim = 64		Dim = 160		Dim = 256		Dim = 352	
	TTM	Ours	TTM	Ours	TTM	Ours	TTM	Ours		TTM	Ours	TTM	Ours	TTM	Ours	TTM	Ours
64	0.771	**0.775**	0.779	**0.783**	0.788	**0.796**	0.792	**0.800**	64	0.875	**0.886**	0.878	**0.889**	0.881	**0.894**	0.900	**0.906**
160	0.775	**0.779**	0.783	**0.792**	0.792	**0.804**	0.796	**0.808**	160	0.883	**0.889**	0.886	**0.892**	0.889	**0.900**	0.903	**0.917**
256	0.783	**0.792**	0.788	**0.796**	0.796	**0.808**	0.800	**0.812**	256	0.894	**0.897**	0.897	**0.900**	0.903	**0.903**	0.906	**0.919**
352	0.788	**0.800**	0.792	**0.804**	0.804	**0.812**	0.808	**0.817**	352	0.903	**0.917**	0.908	**0.919**	0.917	**0.922**	0.919	**0.925**

Table 9
Variation of ACC with memory when DIM is specified on different datasets. Left side is dim specified as 352 on VesselMNIST3D, right side is dim specified as 256 on AdrenalMNIST3D.

VesselMNIST3D (Dim=352)			AdrenalMNIST3D (Dim=256)		
Memory	TTM	LMTTM(Ours)	Memory	TTM	LMTTM(Ours)
64	0.903	**0.917**(1.60% ↑)	64	0.783	**0.792**(1.15% ↑)
160	0.908	**0.919**(1.22% ↑)	160	0.788	**0.796**(1.02% ↑)
256	0.917	**0.922**(0.55% ↑)	256	0.796	**0.808**(1.51% ↑)

Acknowledgments

This study was supported by the National Key R&D Program of China (Grant No. 2023YFB4301800).

References

[1] Woowon Lee, Amir Ostadi Moghaddam, Zixi Lin, Barbara L. McFarlin, Amy J. Wagoner Johnson, Kimani C. Toussaint, Quantitative classification of 3D collagen fiber organization from volumetric images, IEEE Trans. Med. Imaging 39 (12) (2020) 4425–4435.
[2] Nan Wu, Phang, et al., Deep neural networks improve radiologists' performance in breast cancer screening, IEEE Trans. Med. Imaging (2020) 1184–1194.
[3] Qian Wang, Li Sun, Yan Wang, Mei Zhou, Menghan Hu, Jiangang Chen, Ying Wen, Qingli Li, Identification of melanoma from hyperspectral pathology image using 3D convolutional networks, IEEE Trans. Med. Imaging 40 (1) (2021) 218–227.
[4] Zhongyi Han, Benzheng Wei, Yanfei Hong, Tianyang Li, Jinyu Cong, Xue Zhu, Haifeng Wei, Wei Zhang, Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning, IEEE Trans. Med. Imaging 39 (8) (2020) 2584–2594.
[5] Vaswani, et al., Attention is all you need, Neural Inf. Process. Syst. (2017).
[6] Alex Graves, Greg Wayne, Ivo Danihelka, Neural turing machines, 2014, ArXiv, arXiv:1410.5401.
[7] Michael S. Ryoo, Keerthana Gopalakrishnan, et al., Token turing machines, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 19070–19081.
[8] Dewen Seng, Guangsen Chen, Qiyan Zhang, Item-based collaborative memory networks for recommendation, IEEE Access 8 (2020) 213027–213037.
[9] Jiancheng Yang, Shi, et al., MedMNIST v2-a large-scale lightweight benchmark for 2D and 3D biomedical image classification, Sci. Data 10 (1) (2023) 41.
[10] Florin C. Ghesu, Edward Krubasik, Bogdan Georgescu, Vivek Singh, Yefeng Zheng, Joachim Hornegger, Dorin Comaniciu, Marginal space deep learning: Efficient architecture for volumetric image parsing, IEEE Trans. Med. Imaging 35 (5) (2016) 1217–1228.
[11] Zeyu Ren, Yudong Zhang, Shuihua Wang, A hybrid framework for lung cancer classification, Electronics 11 (10) (2022) 1614.
[12] Hao Xu, Yun Wu, G2ViT: Graph neural network-guided vision transformer enhanced network for retinal vessel and coronary angiograph segmentation, Neural Netw. 176 (2024) 106356.
[13] Zeyu Ren, Shuihua Wang, Yudong Zhang, Weakly supervised machine learning, CAAI Trans. Intell. Technol. 8 (3) (2023) 549–580.
[14] Qi Dou, Hao Chen, Lequan Yu, Lei Zhao, Jing Qin, Defeng Wang, Vincent C.T. Mok, Lin Shi, Pheng-Ann Heng, Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks, IEEE Trans. Med. Imaging 35 (5) (2016) 1182–1195.
[15] Rajat Rasal, Daniel C. Castro, Nick Pawlowski, Ben Glocker, Deep structural causal shape models, in: Computer Vision - ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, Springer-Verlag, Berlin, Heidelberg, 2023, pp. 400–432.
[16] Abdelrahman M. Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, Fahad Shahbaz Khan, UNETR++: delving into efficient and accurate 3D medical image segmentation, IEEE Trans. Med. Imaging (2024).
[17] Ke Yu, Li Sun, Junxiang Chen, Maxwell Reynolds, Tigmanshu Chaudhary, Kayhan Batmanghelich, DrasCLR: A self-supervised framework of learning disease-related and anatomy-specific representation for 3D lung CT images, Med. Image Anal. 92 (2024) 103062.
[18] Linshan Wu, Jiaxin Zhuang, Hao Chen, Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22873–22882.
[19] Stylianos Dritsas, Kenneth Wei De Chua, Zhi Hwee Goh, Robert E. Simpson, Classification, registration and segmentation of ear canal impressions using convolutional neural networks, Med. Image Anal. 94 (2024) 103152.
[20] Anindo Saha, Tushar, et al., Weakly supervised 3D classification of chest CT using aggregated multi-resolution deep segmentation features, in: Medical Imaging 2020: Computer-Aided Diagnosis, 2020.
[21] Zeyu Ren, Quan Lan, Yudong Zhang, Shuihua Wang, Exploring simple triplet representation learning, Comput. Struct. Biotechnol. J. 23 (2024) 1510–1521.
[22] Joyce Chelangat Bore, Peiyang Li, Lin Jiang, Walid M.A. Ayedh, Chunli Chen, Dennis Joe Harmah, Dezhong Yao, Zehong Cao, Peng Xu, A long short-term memory network for sparse spatiotemporal EEG source imaging, IEEE Trans. Med. Imaging 40 (12) (2021) 3787–3800.
[23] Xi Chen, Matthew R. Lowerison, Zhijie Dong, Nathiya Vaithiyalingam Chandra Sekaran, Daniel A. Llano, Pengfei Song, Localization free super-resolution microbubble velocimetry using a long short-term memory neural network, IEEE Trans. Med. Imaging 42 (8) (2023) 2374–2385.
[24] Ruipeng Zhang, Binjie Qin, Jun Zhao, Yueqi Zhu, Yisong Lv, Song Ding, Locating X-Ray coronary angiogram keyframes via long short-term spatiotemporal attention with image-to-patch contrastive learning, IEEE Trans. Med. Imaging 43 (1) (2024) 51–63.
[25] Yurong Chen, Hui Zhang, Yaonan Wang, Yimin Yang, Xianen Zhou, Q.M. Jonathan Wu, MAMA Net: Multi-scale attention memory autoencoder network for anomaly detection, IEEE Trans. Med. Imaging 40 (3) (2021) 1032–1041.

[26] Pengyu Wang, et al., MGIML: Cancer grading with incomplete radiology-pathology data via memory learning and gradient homogenization., IEEE Trans. Med. Imaging PP (2024).
[27] MichaelS. Ryoo, A.J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, Anelia Angelova, TokenLearner: Adaptive space-time tokenization for videos, Neural Inf. Process. Syst. (2021).
[28] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, Joao Carreira, Perceiver: General perception with iterative attention, 2021, Cornell University - arXiv.
[29] Junke Wang, Xitong Yang, Hengduo Li, Zuxuan Wu, Yu-Gang Jiang, Efficient video transformers with spatial-temporal token selection, in: European Conference on Computer Vision, 2021.
[30] Xuwei Xu, Changlin Li, Yudong Chen, Xiaojun Chang, Jiajun Liu, Sen Wang, No token left behind: Efficient vision transformer via dynamic token idling, in: Applied Informatics, 2023.
[31] Fanglei Xue, Qiangchang Wang, Zichang Tan, Zhongsong Ma, Guodong Guo, Vision transformer with attentive pooling for robust facial expression recognition, IEEE Trans. Affect. Comput. 14 (2022) 3244–3256.
[32] Xinjian Wu, Fanhu Zeng, et al., PPT: Token pruning and pooling for efficient vision transformers, 2023, ArXiv, arXiv:2310.01812.
[33] Yongming Rao, Zuyan Liu, Wenliang Zhao, Jie Zhou, Jiwen Lu, Dynamic spatial sparsification for efficient vision transformers and convolutional neural networks, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2022) 10883–10897.
[34] Hanyu Hu, et al., Horizontal and vertical crossover of sine cosine algorithm with quick moves for optimization and feature selection., J. Comput. Des. Eng. 9 (2022) 2524.
[35] Kyung Soo Kim, Doo Soo Chang, Yong Suk Choi, Boosting memory-based collaborative filtering using content-metadata, Symmetry 11 (4) (2019) 561.
[36] Yihao Zhang, Xiaoyang Liu, Learning attention embeddings based on memory networks for neural collaborative recommendation, Expert Syst. Appl. 183 (2021) 115439.
[37] Xunqiang Jiang, Binbin Hu, Yuan Fang, Chuan Shi, Multiplex memory network for collaborative filtering, in: Proceedings of the 2020 SIAM International Conference on Data Mining, SDM, 2020, pp. 91–99.
[38] Patrick Bilic, Christ, et al., The liver tumor segmentation benchmark (LiTS), Med. Image Anal. (2023) 102680.
[39] Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, Xiangzhi Bai, Efficient multiple organ localization in CT image using 3D region proposal network, IEEE Trans. Med. Imaging (2019) 1885–1898.
[40] Samuel G. Armato, McLennan, et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans, Med. Phys. (2011) 915–931.
[41] Liang Jin, Yang, et al., Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet, EBioMedicine (2020) 103106.
[42] Xi Yang, Ding Xia, Taichi Kin, Takeo Igarashi, IntrA: 3D intracranial aneurysm dataset for deep learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020.
[43] Donglai Wei, Zudi Lin, Daniel Franco-Barranco, Nils Wendt, Xingyu Liu, Wenjie Yin, Xin Huang, Aarush Gupta, Won-Dong Jang, Xueying Wang, Ignacio Arganda-Carreras, Jeff W. Lichtman, Hanspeter Pfister, MitoEM dataset: Large-scale 3D mitochondria instance segmentation from EM images, in: Medical Image Computing and Computer Assisted Intervention, MICCAI 2020, in: Lecture Notes in Computer Science, IEEE, 2020, pp. 66–76.
[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[45] Jiancheng Yang, Xiaoyang Huang, Yi He, Jingwei Xu, Canqian Yang, Guozheng Xu, Bingbing Ni, Reinventing 2d convolutions for 3d images, IEEE J. Biomed. Heal. Inform. 25 (8) (2021) 3009–3018.
[46] Matthias Feurer, Klein, et al., Efficient and robust automated machine learning, Neural Inf. Process. Syst. (2015).
[47] Haifeng Jiang, Qingquan Song, Xia Hu, Auto-Keras: An efficient neural architecture search system, 2018, Cornell University - arXiv.
[48] Jinwei Liu, Yan Li, Guitao Cao, Yong Liu, Wenming Cao, Feature pyramid vision transformer for medmnist classification decathlon, in: 2022 International Joint Conference on Neural Networks, IJCNN, 2022, pp. 1–8.
[49] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al., Hybrid computing using a neural network with dynamic external memory, Nature 538 (7626) (2016) 471–476.