

ICHPRO: INTRACEREBRAL HEMORRHAGE PROGNOSIS CLASSIFICATION VIA JOINT-ATTENTION FUSION-BASED 3D CROSS-MODAL NETWORK

Xinlei Yu¹, Xinyang Li², Ruiquan Ge^{1,*}, Shibin Wu³, Ahmed Elazab⁴, Jichao Zhu⁵, Lingyan Zhang⁵, Gangyong Jia¹, Taosheng Xu⁶, Xiang Wan⁷, Changmiao Wang^{7,*}

¹Hangzhou Dianzi University, China ²The Chinese University of Hong Kong, Shenzhen, China

³Ping An Technology, China ⁴Shenzhen University, China ⁵Longgang Central Hospital of Shenzhen, China

⁶Hefei Institutes of Physical Science, Chinese Academy of Sciences, China

⁷Shenzhen Research Institute of Big Data, China

ABSTRACT

Intracerebral Hemorrhage (ICH) is the deadliest subtype of stroke, necessitating timely and accurate prognostic evaluation to reduce mortality and disability. However, the multifactorial nature and complexity of ICH make methods based solely on computed tomography (CT) image features inadequate. Despite the capacity of cross-modal networks to fuse additional information, the effective combination of different modal features remains a significant challenge. In this study, we propose a joint-attention fusion-based 3D cross-modal network termed ICHPro that simulates the ICH prognosis interpretation process utilized by neurosurgeons. ICHPro includes a joint-attention fusion module to fuse features from CT images with demographic and clinical textual data. We introduce a joint loss function to enhance the representation of cross-modal features. ICHPro facilitates the extraction of richer cross-modal features, thereby improving classification performance. Upon testing our method using a five-fold cross-validation, we achieved an accuracy of **89.11%**, an F1 score of **0.8767**, and an AUC value of **0.9429**. These results outperform those obtained from other advanced methods based on the test dataset, thereby demonstrating the superior efficacy of ICHPro. The code is available at our github ¹.

Index Terms— Joint-attention mechanism, Cross-modal fusion, Demographic and clinical text, ICH prognosis

1. INTRODUCTION

Intracerebral Hemorrhage (ICH) carries an extremely high mortality rate of more than 40%, with only 20% of survivors achieving functional independence [1]. Consequently, accurate prognosis prediction is of crucial importance for patients post-ICH in order to develop an appropriate treatment plan [2]. Experienced neurosurgeons predominantly rely on computed tomography (CT) scans, specifically the location, volume, and distinct texture features of the hemorrhage site, as

the primary determinants for judgment. Secondary indicators include the patient’s age, gender, and Glasgow Coma Scale (GCS) score [3] among others [4]. This process, however, is contingent on manual predictions by neurosurgeons, a labor-intensive task that may affect accuracy due to variability in doctors’ experience and subjective factors. To address these issues, early studies have employed machine learning techniques [5, 6], achieving certain levels of success, albeit with room for further improvement.

Despite the richer and more comprehensive information that can be obtained with cross-modal methods, their application in ICH prognosis remains limited and preliminary. Recently, there have been some advances, such as the fusion-based [7] and deep learning (DL)-based methods [8] that directly concatenated extracted image with clinical features. Also, GCS-ICHNet [9] improves performance by fusing images with domain knowledge using a self-attention mechanism. However, these methods lack an effective fusion mechanism, limiting the establishment of semantic connections and internal dependencies of features between modalities.

In response to these limitations, in this paper we propose a novel method boasting four key benefits: (1) The 3D structure provides more spatial texture features of hemorrhage locations. (2) The cross-modal structure incorporates more comprehensive demographic and clinical data, thereby enhancing the model’s understanding of the task. (3) The joint-attention mechanism directs the network to adjust regions of attention, facilitating the acquisition of richer and more effective fusion features. (4) The Vision-Text Modality Fusion (VTMF) loss, specifically designed for the cross-modal network, promotes better feature representations across the two modalities.

2. METHODOLOGY

As depicted in Fig.1, ICHPro comprises three components: the feature extraction module, the joint-attention fusion module, and the classification module. These modules represent the three consecutive stages of the entire process.

*Corresponding authors: Ruiquan Ge and Changmiao Wang.

¹Our source code is at: https://github.com/YU-deep/ICH_prognosis.git

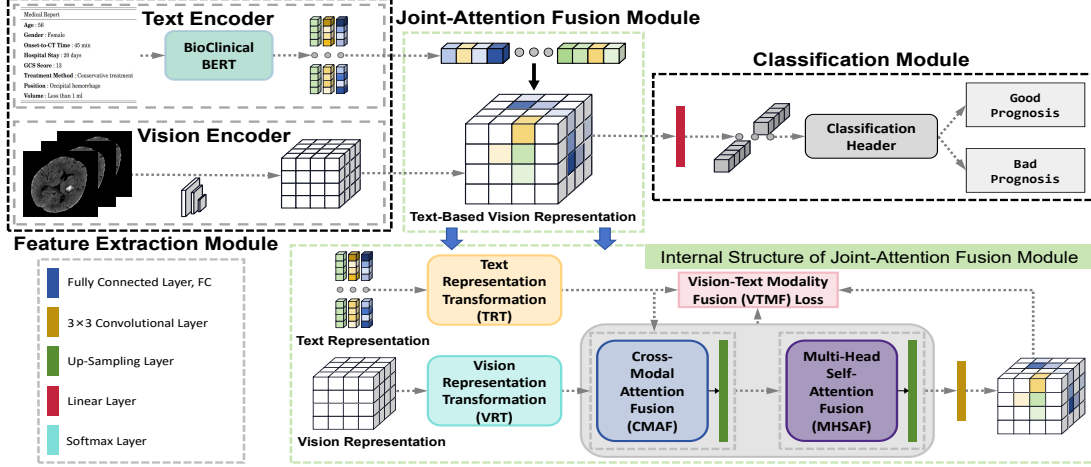


Fig. 1: The illustration delineated the architecture of ICHPro and the green-dashed box below represents the internal structure of the joint-attention fusion Module. The CMAF block is designed to facilitate the fusion of textual and visual modalities. Concurrently, the VTMF loss actively encourages the superior formation of representation of cross-modal features.

In the feature extraction module, we employ the pre-trained BioClinicalBERT [10] model as the text encoder to obtain text representation f^t , and the pre-trained 3D ResNet-50 [11] as the vision encoder to secure vision representation f^v . In the classification module, the pre-trained 1D DenseNet-121 is utilized as the classification header.

2.1. Joint-Attention Fusion Module

In this module, f^t is first fed into the text representation transformation (TRT) block and f^v into the vision representation transformation (VRT) block, respectively. This process yields a unified reconstructed text representation \tilde{f}^t and reconstructed vision representation \tilde{f}^v . These are subsequently processed through a cross-modal attention fusion (CMAF) block and a multi-head self-attention fusion (MHSAF) block, respectively, resulting in text-based vision representation f^{tbv} .

TRT and VRT Block. In these blocks, we transform f^t and f^v into similar structures, thereby fostering a stronger semantic connection between the two modalities. In the TRT block, f^t is multiplied by its transposition f^{tT} and then transformed through a fully connected (FC) layer and a reshaped layer, yielding \tilde{f}^t . In the VRT block, f^v is transformed through an FC layer followed by four up-sampling layers to obtain \tilde{f}^v .

CMAF Block. Partly inspired by the cross-modal fusion component in the CMAFGAN framework [12] which was originally designed for word-to-face synthesis tasks, we identified its potential for modal fusion and enhanced it to suit our task. Additionally, we incorporated a SoftPool layer [13] into the block to reduce computational overhead while preserving more information. Furthermore, the overall structure of the block was restructured.

As shown in Fig. 2, we initially diminish the size of inputs \tilde{f}^v and \tilde{f}^t through the FC layer, referring to these as x and y .

Following this, x and y are separately transformed into three feature spaces via 1×1 convolution layers, which are referred to as V_1, K_1, Q_1 and V_2, K_2, Q_2 , with w and superscripts denoting their corresponding weight matrices. Then, we can compute the matching degree as follows:

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^S \exp(s_{ij})}, \text{ where } s_{ij} = w^{Q1} x_i^T \times w^{K2} y_j, \quad (1)$$

$$\rho_{j,i} = \frac{\exp(t_{ij})}{\sum_{j=1}^S \exp(t_{ij})}, \text{ where } t_{ij} = w^{Q2} y_i^T \times w^{K1} x_j, \quad (2)$$

where β and ρ signify the matching degree in vision and text spaces, separately. We multiplied the matrices β and V_1, ρ and V_2 to get cross-modal attention feature map \mathbf{o}_x and \mathbf{o}_y .

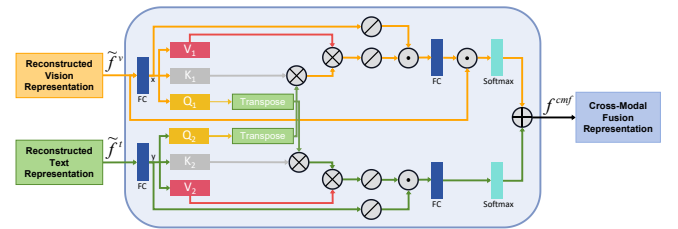


Fig. 2: Architecture of the proposed CMAF Block. \otimes denotes matrix multiplication, \circ signifies SoftPool, \oplus stands for matrix addition, and \oplus is representative of concatenation.

Subsequently, we apply SoftPool to the previously obtained $\mathbf{o}_x, \mathbf{o}_y, x$ and y to yield $\check{\mathbf{o}}_x, \check{\mathbf{o}}_y, \check{x}$ and \check{y} . Following this, we add the matrices $\check{\mathbf{o}}_x$ and $\check{x}, \check{\mathbf{o}}_y$ and \check{y} , and pass them through a linear layer to obtain \mathbf{o}_v and \mathbf{o}_w . Lastly, after applying a softmax layer to each, we can express f^{cmf} as follows:

$$f^{cmf} = \text{concat}(\gamma_1 * \mathbf{o}_v, \gamma_2 * \mathbf{o}_w). \quad (3)$$

MHSAF Block. We implemented a multi-head self-attention mechanism to map features to different subspaces via sev-

eral distinct linear transformations. Subsequently, we execute self-attention computations on each subspace to procure multiple output vectors, which are then concatenated.

2.2. Loss Function

In our study, we propose a joint loss function known as the VTMF loss. This loss is composed of three integral components. Firstly, the intra-modality and inter-modality alignment (IMIMA) loss is incorporated as a global loss. Its purpose is to map semantically similar samples from both intra-modalities and inter-modalities into a harmonious global space. Secondly, the similarity distribution matching (SDM) loss is employed to enhance semantic matching and to extract inherent dependencies between the two modalities. Finally, the function includes masked language modeling (MLM) loss, which serves to enrich semantic learning and augment textual comprehension.

IMIMA Loss. To accomplish alignment on both intra-modalities, such as Text-to-Text ($t2t$) and Vision-to-Vision ($v2v$), as well as inter-modalities, specifically Text-to-Vision ($t2v$) and Vision-to-Text ($v2t$), we map semantically related samples into related individual spaces and maintain the proximity of similar samples in the joint embedding space. We designate the negative sets for the sample as N . In intra-modalities is $\mathbf{N}_i^{intra} = \{y_j \mid \forall y_j \in N, j \neq i\}$ and in inter-modalities is $\mathbf{N}_i^{inter} = \{x_j \mid \forall x_j \in N, j \neq i\}$. Thus, the intra/inter loss can be expressed as follows:

$$\mathcal{L}_{intra/inter}^{A2B} = -\log \frac{\delta(f^A, f^B)}{\delta(f^A, f^B) + \sum_{f_k \in \mathbf{N}} \delta(f^A, f_k^B)}, \quad (4)$$

where $\delta(a, b) = \exp(a^T b)$. Therefore, IMIMA loss is:

$$\mathcal{L}_{IMIMA} = \mathcal{L}_{intra}^{t2t} + \mathcal{L}_{intra}^{v2v} + \mathcal{L}_{inter}^{t2v} + \mathcal{L}_{inter}^{v2t}. \quad (5)$$

SDM Loss. We employ SDM loss [14] to forge consistent semantic match, thus associating the representations across modalities. For each vision-text pair, we obtain a vision representation f_i^v and a text representation f_j^t . And we define $\{(f_i^v, f_j^t), l_{i,j}\}$, where $l_{i,j}$ is the matching label. When $l_{i,j} = 1$ means that (f_i^v, f_j^t) is a matched pair which denotes the two models from the same identity, while $l_{i,j} = 0$ indicates the unmatched pair. The true matching probability can be formulated as:

$$q_{i,j} = l_{i,j} / \sum_{k=1}^N l_{i,k} \quad (6)$$

Let $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T \mathbf{v} / \|\mathbf{u}\| \|\mathbf{v}\|$ denotes the dot product between L_2 normalized u and v (i.e. cosine similarity). The matching probability $p_{i,j}$ can be deemed as the proportion of the cosine similarity score between f_i^v and f_j^t to the sum of the cosine similarity score between f_i^v and $\{f_j^t\}_{j=1}^N$ [15].

Then the probability of matching pairs can be simply calculated with the following *softmax* function [14]:

$$p_{i,j} = \frac{\exp(\text{sim}(f_i^v, f_j^t) / \tau)}{\sum_{k=1}^N \exp(\text{sim}(f_i^v, f_k^t) / \tau)} \quad (7)$$

where τ the temperature hyperparameter to limit the probability distribution peaks.

The SDM loss of $v2t$ can be delineated as follows:

$$\mathcal{L}_{v2t} = KL(p_i \| q_i) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n p_{i,j} \log \left(\frac{p_{i,j}}{q_{i,j}} \right), \quad (8)$$

where q represents the true matching probability and p signifies the proportion of a specific cosine similarity score to the overall sum. The bi-directional SDM loss is the sum of the loss of $v2t$ and $t2v$.

MLM Loss. We adopt the design of the intrinsic loss function from BERT. The objective of MLM is to randomly obscure certain words in input texts. The model is then required to predict these hidden words, serving for assessment of loss.

Overall Objective. Based on the analysis above, the definition of VTMF loss can be calculated as follows:

$$\mathcal{L}_{VTMF} = \mathcal{L}_{IMIMA} + \alpha \mathcal{L}_{SDM} + \beta \mathcal{L}_{MLM}, \quad (9)$$

where α and β represent the weights of \mathcal{L}_{SDM} and \mathcal{L}_{MLM} , respectively, serving to dynamically balance the relative significance of these losses.

3. EXPERIMENT AND RESULTS

3.1. Experiment Setting

Dataset. In this study, we utilized a private ICH dataset obtained from our collaborative hospital, comprising a total of 294 cases with 149 indicating good and 145 bad prognoses. Each case included comprehensive CT imaging with demographic and clinical information including gender, age, onset-to-CT time, hospital stay, GCS score, and treatment method as well as hemorrhage position and volume. Each case was labeled with either a good or bad prognosis. The classification label is the prognosis outcomes of patients, which is determined by the Glasgow Outcome Scale (GOS) by neurologists. GOS is a rating scale that assesses patients' functional outcomes following brain injury and then according to it, neurologists can label each sample as good or bad. In terms of data preprocessing, we carried out several operations, there are the following steps: (1) Convert series 2D DICOM (Digital Imaging and Communications in Medicine) to 3D NIFTI (Neuroimaging Informatics Technology Initiative) format through dcm2niix. (2) Remove the skull and extract brain tissue with the Swiss Skull Stripper plugin in 3D slicers and the Numpy and Scipy package in Python. (3) Resample the images, constrained HU scales and perform Z-score standardization.

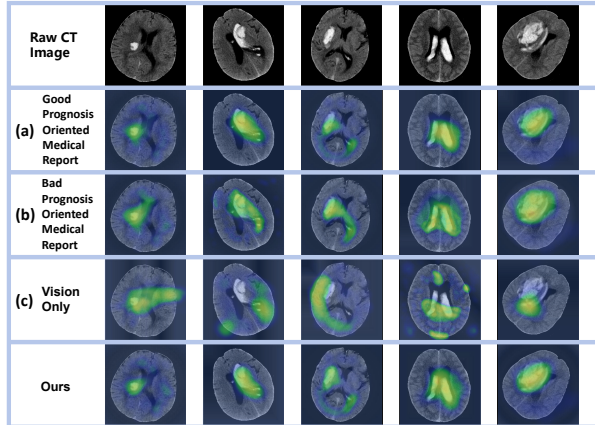


Fig. 3: This graph depicts the impact of different medical report texts on the regions of interest in the joint-attention mechanism.

Implementation Details. Experiments were conducted using two NVIDIA HGX A100 Tensor Core GPUs, employing the Adam optimizer. The training epoch, learning rate, and batch size were respectively set at 300, 0.0001, and 128. Finally, α and β in Eq 9 is learned as 0.84 and 0.45. All experiments were conducted through five-fold cross-validation.

3.2. Visualization Analysis

To better verify the interpretability of our work, we designed visual experiments. We conducted four comparative experiments, including (a) Good Prognosis Oriented Medical Report, (b) Bad Prognosis Oriented Medical Report, (c) Vision only (the same as the set in Sec.3.3) and our method. For (a) and (b), we wrote a good prognosis-oriented medical report and a bad prognosis-oriented medical report for the patients by adjusting the patient’s demographic and clinical information, with the help of neurosurgeons to simulate the impact of different prognosis-oriented medical reports on the network. The Score-CAM [16] method is applied to the last convolution layer of Vision Encoder, the last convolution layer of 3D ResNet-50. Especially, because it is a 2D method, we utilized it on the middle and lower 2D slices of visual features (selected the 25th slice out of the 64 slices).

As shown in Fig.3, for different oriented medical reports, our network can accurately locate the region of interest at the location of bleeding, in order to pay more attention to the areas with a higher correlation with prognosis results. For different reports, the region of interest will change to a certain extent with the change of text to match the text information. However, the main part of the region of interest is still determined by the CT image, which is consistent with the original intention of our network design and proves its rationality. .

3.3. Ablation Experiment

The Text-Only and Vision-Only models directly input the features extracted from their corresponding encoders into the MHSAF block, followed by a classification module. Compared to Vision-Only, ICHPro demonstrates a significant improvement, exceeding it by **9.79%** in accuracy and **0.0834** in AUC metrics. Our method learns modal fusion features that encompass richer demographic and clinical information. This enhances the extraction of more contextual information, thereby facilitating more accurate prognostic predictions.

Table 1: Results of modal ablation experiment.

Method	Acc(%)	Recall(%)	Prec(%)	F1 Score	AUC
Text-Only	69.15	65.10	71.11	0.6797	0.7534
Vision-Only	<u>79.32</u>	<u>77.18</u>	<u>82.72</u>	<u>0.7985</u>	<u>0.8595</u>
ICHPro	89.11	84.56	91.02	0.8767	0.9429

3.4. Attention Fusion Structure Experiment

We further conducted a comparative analysis of six methods, each comprising different permutations and combinations of CMAF and MHSAF blocks. We utilized the terms "Cross" and "Self" to individually denote these blocks. It is important to note that, the notation A-B implies that Block A is entered first, followed by Block B.

Table 2: Comparisons of attention fusion methods.

Structure	Acc(%)	Recall(%)	Prec(%)	F1 Score	AUC
Self Attention	82.71	78.52	83.17	0.8078	0.8671
Cross Attention	84.41	79.19	86.32	0.8260	0.8956
Self-Self Attention	76.94	71.81	80.95	0.7611	0.8025
Cross-Cross Attention	<u>87.11</u>	80.53	<u>89.40</u>	<u>0.8473</u>	<u>0.9108</u>
Self-Cross Attention	85.08	<u>81.21</u>	88.08	0.8451	0.8934
Cross-Self Attention	89.11	84.56	91.02	0.8767	0.9429

Methods incorporating cross-modal attention demonstrate superior performance compared to those lacking this addition, thereby confirming the effectiveness of the CMAF block. As indicated in Table 2, sequentially passing through the CMAF and MHSAF blocks yields optimal results. The former facilitates interaction between two modalities, establishing semantic connections and enriching feature expressions, while the latter captures the internal dependencies of fused features, thereby effectively capturing contextual relationships. This combination significantly amplifies the expressive power and generalization capabilities of cross-modal networks.

3.5. Loss Function Based Experiment

We employed three alternative loss functions for the comparative analysis of our model. These included two single cross-modal losses, \mathcal{L}_{blend} and \mathcal{L}_{cmpm} , and one joint cross-modal loss, \mathcal{L}_{CMFA} . Additionally, we conducted ablation experiments to demonstrate the effectiveness of each component.

As summarized in Table 3, joint losses, which amalgamate multiple optimization objectives, yield superior performance compared to single losses. As our global loss, \mathcal{L}_{IMIMA} outperforms the other four single losses due to its ability to align both intra and inter-modal. Although the

Table 3: Results of comparison and ablation experiment based on loss function.

Loss Function	Components			Acc(%)	Recall(%)	Prec(%)	F1 Score	AUC
	IMIMA	SDM	MLM					
\mathcal{L}_{blend} [17]				74.57	70.47	78.17	0.7412	0.7598
\mathcal{L}_{cmpm} [18]				73.56	69.13	76.78	0.7275	0.7852
\mathcal{L}_{CMFA} [19]				85.08	80.54	88.72	0.8442	0.8930
\mathcal{L}_{IMIMA}	✓			75.59	71.14	79.44	0.7506	0.7982
\mathcal{L}_{SDM}		✓		71.86	68.45	73.19	0.7074	0.7346
\mathcal{L}_{MLM}			✓	54.24	50.34	56.94	0.5344	0.5705
$\mathcal{L}_{IMIMA} + \alpha\mathcal{L}_{SDM}$	✓	✓		84.40	81.88	86.49	0.8412	0.8806
$\mathcal{L}_{IMIMA} + \beta\mathcal{L}_{MLM}$	✓		✓	76.27	73.83	79.02	0.7634	0.8194
\mathcal{L}_{VTMF}	✓	✓	✓	89.11	84.56	91.02	0.8767	0.9429

Table 4: Comparisons of ICHPro and other methods.

Method	Acc(%)	Recall(%)	Prec(%)	F1 Score	AUC
Image-Based Method (2D) [6]	74.23	67.11	75.98	0.7127	0.6933
GCS-ICHNet (2D) [9]	85.08	81.88	87.25	0.8448	0.8590
DL-Based Method (3D) [8]	81.02	78.52	83.31	0.8084	0.9141
Multi-Task Method (3D) [20]	85.42	79.86	89.80	0.8454	0.8998
UniMiSS (2D+3D) [21]	82.03	78.52	87.59	0.8281	0.8275
ICHPro	89.11	84.56	91.02	0.8767	0.9429

individual \mathcal{L}_{MLM} performs poorly when paired with losses bearing cross-modal capabilities, it can effectively enhance the contextual understanding of fused features. Compared to using \mathcal{L}_{IMIMA} independently, the addition of \mathcal{L}_{SDM} or \mathcal{L}_{MLM} increases accuracy by 8.81% and 0.68%, respectively. This demonstrates the efficacy of our additions. These findings suggest that the combination of all three losses can achieve optimal performance for each individual loss. Compared to another joint loss function \mathcal{L}_{CMFA} , \mathcal{L}_{VTMF} outperforms by 4.03% and 0.0499 in accuracy and AUC value, respectively. This analysis highlights the effectiveness of our loss function.

3.6. Comparative Experiments

We conducted a comparison of ICHPro with other advanced methods, using our dataset. The results are illustrated in Table 4 and Fig.4. The first four methods delineated in the table are specifically designed for the classification of ICH prognosis, while UniMiSS represents a universal network for medical image classification that utilizes a combination of 2D and 3D convolutional techniques.

Owing to the integration of domain knowledge, the accuracy of the 2D GCS-ICHNet essentially matches that of the 3D multi-task method, which relies solely on images. It also surpasses the universally applied 2D+3D UniMiSS method. Both ICHPro and the DL-based method incorporate comprehensive demographic and clinical information, rendering their AUC superior to all other methods. This highlights the enhanced robustness of networks that fuse information beyond images. Compared to the DL-based method, our performance is superior across all metrics, underscoring the effectiveness of our joint-attention fusion mechanism. When compared to the optimal indicators of other methods, ours improves accuracy by **3.69%** and the AUC value by **0.0288**. In addition, the ROC curve is closest to the upper left corner, indicative of its effectiveness in distinguishing between positive and negative samples. Our method demonstrates a comprehensive superiority over the comparison methods, and to our knowledge, it surpasses existing advanced methods in the task of ICH prog-

nosis classification on our dataset.

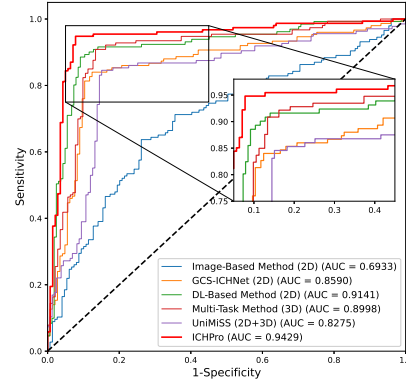


Fig. 4: ROC curves between ICHPro and other methods.

4. CONCLUSION

The absence of demographic and clinical information and the inefficient cross-modal fusion mechanism could hinder the effective extraction of cross-modal fusion features. To address this, in this paper, we proposed an ICHPro, a joint-attention fusion-based 3D cross-modal network, for ICH prognosis classification. Furthermore, we proposed a VTMF loss to enhance modal alignment and optimize networks. Our experimental results demonstrate the efficacy of our method. In the future, we aim to extend the network to an end-to-end model and augment classification tasks through segmentation, for improved outcomes. Additionally, our proposed method holds potential for application beyond ICH prognosis, extending to other medical cross-modal classification tasks.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was conducted by the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Longgang Central Hospital of Shenzhen (2023.10.26/No.2023ECPJ077).

6. ACKNOWLEDGMENTS

Support of the Zhejiang Provincial Natural Science Foundation of China (No.LY21F020017,2022C03043), Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (No.U20A20386), National Natural Science Foundation of China (No.61702146), Guangdong Basic and Applied Basic Research Foundation (No.2022A15110570) and Innovation Teams of Youth Innovation in Science and Technology of High Education Institutions of Shandong Province (No.2021KJ088) are gratefully acknowledged.

7. REFERENCES

- [1] Joseph P Broderick, James C Grotta, Andrew M Naidech, et al., “The story of intracerebral hemorrhage: from recalcitrant to treatable disease,” *Stroke*, vol. 52, no. 5, pp. 1905–1914, 2021.
- [2] Jonathan Rosand, “Preserving brain health after intracerebral haemorrhage,” *The Lancet Neurology*, vol. 20, no. 11, pp. 879–880, 2021.
- [3] Graham Teasdale, Andrew Maas, Fiona Lecky, et al., “The glasgow coma scale at 40 years: standing the test of time,” *The Lancet Neurology*, vol. 13, no. 8, pp. 844–854, 2014.
- [4] Zachary Troiani, Luis Ascanio, Christina P Rossitto, et al., “Prognostic utility of serum biomarkers in intracerebral hemorrhage: a systematic review,” *Neurorehabilitation and Neural Repair*, vol. 35, no. 11, pp. 946–959, 2021.
- [5] Lucas A Ramos, Manon Kappelhof, Hendrikus JA Van Os, et al., “Predicting poor outcome before endovascular treatment in patients with acute ischemic stroke,” *Frontiers in Neurology*, vol. 11, pp. 580957, 2020.
- [6] Jawed Nawabi, Helge Kniep, Sarah Elsayed, et al., “Imaging-based outcome prediction of acute intracerebral hemorrhage,” *Translational Stroke Research*, vol. 12, pp. 958–967, 2021.
- [7] Chen Wang, Xianbo Deng, Li Yu, et al., “Data fusion framework for the prediction of early hematoma expansion based on cnn,” in *International Symposium on Biomedical Imaging (ISBI)*. 2021, pp. 169–173, IEEE.
- [8] Amaia Perez del Barrio, Anna Salut Esteve Domínguez, Pablo Menéndez Fernández-Miranda, et al., “A deep learning model for prognosis prediction after intracranial hemorrhage,” *Journal of Neuroimaging*, vol. 33, no. 2, pp. 218–226, 2023.
- [9] Xuhao Shan, Xinyang Li, Ruiquan Ge, et al., “Gcs-ichnet: Assessment of intracerebral hemorrhage prognosis using self-attention with domain knowledge integration,” in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, pp. 2217–2222.
- [10] Emily Alsentzer, John Murphy, William Boag, et al., “Publicly available clinical BERT embeddings,” in *Proceedings of Clinical Natural Language Processing Workshop*, 2019, pp. 72–78.
- [11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.
- [12] Xiaodong Luo, Xiang Chen, Xiaohai He, et al., “Cmfagan: A cross-modal attention fusion based generative adversarial network for attribute word-to-face synthesis,” *Knowledge-Based Systems*, vol. 255, pp. 109750, 2022.
- [13] Alexandros Stergiou, Ronald Poppe, and Grigorios Kalliatakis, “Refining activation downsampling with softpool,” in *the International Conference on Computer Vision (ICCV)*, 2021, pp. 10357–10366.
- [14] Ding Jiang and Mang Ye, “Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2787–2797.
- [15] Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Arimitsu, “Semantic cosine similarity,” in *The 7th international student conference on advanced science and technology ICAST*, 2012, vol. 4, p. 1.
- [16] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu, “Score-cam: Score-weighted visual explanations for convolutional neural networks,” in *the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, 2020, pp. 24–25.
- [17] Weiyao Wang, Du Tran, and Matt Feiszli, “What makes training multi-modal classification networks hard?,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12695–12705.
- [18] Ying Zhang and Huchuan Lu, “Deep cross-modal projection learning for image-text matching,” in *the European Conference on Computer Vision (ECCV)*, 2018, pp. 686–701.
- [19] Ammarah Farooq, Muhammad Awais, Josef Kittler, et al., “Axm-net: Implicit cross-modal feature alignment for person re-identification,” in *the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 4477–4485.
- [20] Kai Gong, Qian Dai, Jiacheng Wang, et al., “Unified ich quantification and prognosis prediction in ncct images using a multi-task interpretable network,” *Frontiers in Neuroscience*, vol. 17, pp. 1118340, 2023.
- [21] Yutong Xie, Jianpeng Zhang, Yong Xia, et al., “Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier,” in *the European Conference on Computer Vision (ECCV)*, 2022, pp. 558–575.