All rights reserved. No reuse allowed without permission.

Multimodal Diverse Granularity Fusion Network based on US and CT Images for Lymph Node **Metastasis Prediction of Thyroid Carcinoma**

Guojun Li^{1, 2}, Jincao Yao³, Chanjuan Peng³, Yinjie Hu², Shanshan Zhao⁴, Xuhan Feng², Jianfeng Yang⁴, Dong Xu³, Xiaolin Li^{1, 2}, Chulin Sha², Min He²

¹Academy of Medical Engineering and Translational Medicine, Tianjin University ²Hangzhou Institute of Medicine, Chinese Academy of Sciences ³Zhejiang Cancer Hospital ⁴Shaoxing People's Hospital

Abstract

Accurately predicting the risk of cervical lymph node metastasis (LNM) is crucial for surgical decision-making in thyroid cancer patients, and the difficulty in it often leads to over-treatment. Ultrasound (US) and computed tomography (CT) are two primary non-invasive methods applied in clinical practice, but both contain limitations and provide unsatisfactory results. To address this, we developed a robust and explainable multimodal deep-learning model by integrating the above two examinations. Using 3522 US and 7649 CT images from 1138 patients with biopsy-confirmed LNM status, we showed that multimodal methods outperformed unimodal counterparts at both central and lateral cervical sites. By incorporating a diverse granularity fusion module, we further enhanced the area under the curve (AUC) to 0.875 and 0.859 at central and lateral cervical sites respectively. This performance was also validated in an external cohort. Additionally, we quantified the modality-specific contributions for each nodule and systematically evaluated the applicability across various clinical characteristics, aiding in identifying individuals who can benefit most from the multimodal method.

The global incidence of thyroid cancer has surged 1 over the past 30 years[1], reaching over 586,000 new 2 cases in 2020[2]. Despite its generally indolent na-3 ture, thyroid cancer leads to cervical lymph node 4 metastasis (LNM) in up to 50% of patients[3]. Can-5 cer cells typically initially metastasize to the central 6 lymph nodes and subsequently spread to the lateral cervical site, increasing the risk of recurrence and 8 poor prognosis[4]. Consequently, LNM status signif-9 icantly influences the surgical approach for thyroid 10 cancer patients. Therapeutic lymph node dissection 11 (LND) of central and lateral cervical compartments is 12 normally recommended for individuals with central 13 and/or lateral cervical LNM[5]. While for patients 14 without LNM, although central LND remains con-15 troversial, prophylactic lateral cervical LND is not 16 advised[5, 6]. However, the current non-invasive di-17 agnostic accuracy of LNM is insufficient to guide 18 surgical decisions. For the central site, the primary 19 imaging methods, including Ultrasound (US) and 20 computed tomography (CT), provide average sensi-21 tivities of only 0.28 and 0.39[7], respectively. This 22 leads to a prevalent tendency for overtreatment to 23 prevent missed LNM detection and results in poten-24 tial complications such as recurrent laryngeal nerve 25

28

53

In recent years, the introduction of artificial in-29 telligence methods has significantly improved the 30 performance of LNM prediction. Several studies uti-31 lizing US images have employed various machine 32 learning methods, such as gradient boosting, ran-33 dom forests, neural networks, etc., achieving AUCs 34 in the range of 0.700 to 0.772[8, 9, 10] for predicting 35 central site LNM. Other studies focusing on extract-36 ing high-dimensional radiomic features or employing 37 deep learning methods to predict LNM status have 38 achieved AUCs spanning from 0.78 to 0.90[11, 12, 13] 39 for the central site and 0.62[14] for the lateral cervical 40 site. Similarly, in the case of CT images, methods 41 based on radiomic features extracted from thyroid 42 nodules have demonstrated predictive capabilities 43 for central site LNM at AUCs ranging from 0.710 to 44 0.770[15, 16]. 45

However, it's crucial to acknowledge that both US 46 and CT modalities have limitations owing to their 47 examination techniques. Though US images provide 48 high-resolution visuals of thyroid nodules' interior 49 and boundary characteristics, their limited field of 50 view poses challenges in assessing the spatial rela-51 + Cornesten This prentints exertis new second that has not been certified the provided at a second glass at ulcal provided ing 52

shachulin@gmail.com, hemin@him.cas.cn

injury and hypoparathyroidism. Therefore, there is a pressing need to improve the accuracy of LNM risk 27 assessment to assist surgical management.

These authors contributed equally to this work.

tissues. Conversely, CT images offer essential relative

All rights reserved. No reuse allowed without permission.

position information about the thyroid, lymph nodes, 54 and surrounding tissues, albeit at a lower resolution 55 compared to US images. Relying solely on unimodal 56 methods restricts the predictive capabilities of the 57 model. Considering that US and CT provide com-58 plementary information and are widely utilized in 59 thyroid cancer diagnosis, there's a great potential to 60 improve the performance by integrating US and CT 61 62 images through a multimodal approach for predicting LNM status. For instance, Zhao et al. developed 63 a multivariate logistic regression multimodal model 64 for predicting central LNM status by incorporating 65 clinical factors, US-derived diagnostic features, and 66 CT measurements, achieving an AUC of 0.827[17]. 67 Nevertheless, the study did not directly compare the 68 multimodal method's performance with unimodal 69 methods, besides, it utilized a simplified model that 70 overlooked the interaction between the two modali-71 ties, leaving the full potential of multimodal fusion 72 approach unexplored. 73

Leveraging deep learning methods for integrating 74 multimodal medical data has emerged as a promi-75 nent approach to enhance our understanding of com-76 plex diseases[18, 19, 20], with promises in tailoring 77 personalized diagnosis, prognosis, treatment, and 78 care[21, 22, 23, 24]. The central premise of multi-79 modal data integration is that diverse data sources 80 complement each other, augmenting information be-81 yond any individual modality. However, significant 82 challenges persist, such as data scarcity, sparsity, and 83 inter-modality complexity, limiting the full exploita-84 85 tion of data integration benefits. Recent advancements in deep learning methods within this domain 86 primarily focus on representation learning and fusion 87 techniques[25, 26], which include extracting mean-88 ingful representations with unlabeled data[27, 28] 89 and employing attention-based approaches to allow 90 more sophisticated fusion of cross-modality represen-91 tations [29, 30, 31, 32]. While these strategies exhibit 92 improvements in model performance, their use in the 93 biomedical field still requires broader testing across 94 diverse scenarios and adaptation to specific tasks 95 through dedicated study designs[33]. 96

In this study, we aim to improve the predictive 97 accuracy of LNM status for thyroid cancer patients 98 by developing a multimodal method incorporating 99 US and CT images. We curated a paired multi-100 modal dataset consisting of 3522 US and 7649 CT 101 images from 1138 patients with biopsy-confirmed 102 LNM status at both central and lateral cervical sites 103 (Fig. 1). To comprehensively integrate the consis-104 tent and distinct information of both modalities, we 105 first employed a multi-task network scheme to en-106 hance modal-specific feature learning (Fig. 2), which 107 achieved superior performance compared to cur-108

rently commonly used methods on unimodal mod-109 els. Next, we demonstrated that, even with a basic 110 feature fusion strategy, multimodal models consis-111 tently outperform their unimodal counterparts at 112 both sites. Furthermore, we designed a diverse gran-113 ularity fusion module, which learns the attention at 114 three granular levels from fine to coarse: dimension 115 level, modality level, and nodule level (Fig. 2). With 116 the incorporation of this module, our multimodal 117 model achieved AUCs of 0.875 and 0.859 at the cen-118 tral and lateral cervical sites respectively. Compared 119 to unimodal methods of US and CT, the multimodal 120 AUC improved by 5.5% and 10.1% respectively, at 121 the central compartment, and by 7.4% and 8.1% re-122 spectively, at the lateral cervical site. When evaluated 123 on an external validation set, our proposed model 124 demonstrated an AUC of 0.903 at the central site, 125 which robustly confirmed the generalizability of the 126 multimodal model. In addition, we comprehensively 127 evaluate the applicability of each modality on nod-128 ules with various characteristics to identify patients 129 who can best benefit from the multimodal method 130 (Fig. 1), which could significantly improve the clin-131 ical utility of multimodal models. In summary, we 132 presented a promising approach to mitigate the issue 133 of overtreatment in thyroid cancer. Our multimodal 134 AI system exhibits strong performance, high gen-135 eralizability, and substantial clinical utility, offering 136 significant potential for enhancing the diagnosis and 137 treatment of thyroid cancer. 138

Results

139

140

Patient Cohort

This study incorporated two datasets: a main cohort 141 and an external cohort. The main cohort comprised 142 patients who underwent thyroid examinations at Zhe-143 jiang Cancer Hospital from August 2018 to February 144 2021. To reflect the real clinical diagnostic conditions, 145 only necessary data quality control was performed, 146 with specific details outlined in the supplementary 147 material. After quality control, the main cohort con-148 sists of 1138 patients with a total of 1285 thyroid nod-149 ules. The external cohort, obtained from Shaoxing 150 People's Hospital in Zhejiang Province, also under-151 went the same quality control process, comprising 152 60 patients with 60 thyroid nodules. Both cohorts 153 included samples with matched US and CT data, 154 featuring multiple images of thyroid nodules, along 155 with their corresponding LNM status at the central 156 and lateral cervical sites. 157

The models were evaluated under an eight-fold 158 cross-validation setting, and various metrics were 159 employed to assess their performance. These met-160

All rights reserved. No reuse allowed without permission.

Modalities	LNM location	Models	ACC	AUC	SENS	SPEC	PREC	F1 score
US	Central site	Yu et al [12]	0.739	0.792	0.787	0.689	0.719	0.750
		ResNet	0.760	0.812	0.823	0.699	0.736	0.774
		Proposed	0.782	0.820	0.836	0.728	0.754	0.792
	Lateral cervical site	Yu et al [12]	0.729	0.754	0.729	0.731	0.736	0.728
		ResNet	0.747	0.725	0.826	0.665	0.713	0.763
		Proposed	0.767	0.785	0.843	0.693	0.737	0.784
СТ	Central site	ResNet	0.726	0.759	0.743	0.707	0.720	0.730
		Proposed	0.745	0.774	0.737	0.755	0.751	0.743
	Lateral cervical site	ResNet	0.746	0.758	0.812	0.679	0.718	0.760
		Proposed	0.764	0.778	0.725	0.797	0.796	0.751

Table 1: Performance of LNM status prediction using unimodal networks

rics encompassed accuracy (ACC), the area under 161 the ROC curve (AUC), sensitivity (SENS), specificity 162 (SPEC), precision (PREC), and the F1 score. We re-163 ported the mean metrics calculated from the eight-164 fold cross-validation process for a comprehensive 165 evaluation. It is worth mentioning that samples from 166 different folds were divided based on the individual 167 thyroid nodules, and nodules from the same patient 168 were consistently present within the same fold. In 169 addition, an undersampling strategy was applied in 170 this study to maintain a balance between positive 171 and negative categories. 172

173 Enhance modal-specific feature learning

¹⁷⁴ by employing multi-task models of each

175 modality

We start from enhancing modal-specific feature ex-176 traction to make the best use of each modality and 177 evaluate the feature capability in predicting LNM 178 status of each modality. US provides clear visualiza-179 tion of thyroid nodule attributes such as boundary, 180 shape, and internal structure (composition, calcifica-181 tion, echo characteristics). Meanwhile, CT images 182 encompass both thyroid nodules and the surround-183 ing anatomical context, offering insights into their 184 relationships. Therefore, we employ a multi-task 185 learning approach for each modality (as illustrated 186 in Fig. 2). Specifically, besides the LNM predic-187 tion task, we introduce two auxiliary tasks for US: a 188 nodule mask segmentation task to guide the model 189 to focus on the internal structural features, and a 190 nodule boundary segmentation task to emphasize 191 the boundary and shape of nodules. Likewise, for 192 CT, we introduce a nodule mask task and a tissue 193 boundary segmentation task to guide the model to 194 distinguish nodules and surrounding tissue regions, 195 respectively. We chose ResNet[34] as the backbone 196 to build multi-task models for each modality, due to 197 its simple structure, high popularity, and excellent 198 performance (Methods). For each unimodal network, 199

we trained it to complete the auxiliary segmentation 200 tasks in the first 100 epochs and added the additional 201 LNM prediction task in the following 200 epochs. We 202 compared our multi-task models for each modality 203 with ResNet models directly predicting LNM, and for 204 the US unimodal model, we also re-implemented the 205 network developed by Yu et al[12]. It shows that our 206 multi-task models for both modalities consistently 207 outperform their counterparts at both central and 208 cervical lateral sites, with an obvious improvement 209 of ACC and AUC (Table 1). 210

When comparing the unimodal performance of US 211 and CT, we have some interesting observations. First, 212 at the central site, the US models generally outper-213 form CT models, whereas there is no consistent win-214 ner between US and CT models at the lateral cervical 215 site. Moreover, at both the central and lateral cervical 216 sites, US models consistently exhibit higher sensitiv-217 ity, meanwhile, CT models consistently demonstrate 218 higher specificity. These results suggest that US is 219 more sensitive but less specific, while CT is the op-220 posite, highlighting the complementary information 221 provided by these two modalities. 222

Basic multimodal fusion methods outperform either unimodal model

Based on the multi-task unimodal models, we fur-225 ther evaluate the efficacy of integrating US and CT 226 for predicting the risk of LNM. We first examine 227 the multimodal performance using three basic fu-228 sion methods: concatenation, element-wise sum, and 229 element-wise multiplication, to fuse the unimodal 230 features extracted from US and CT and re-train the 231 multimodal network in an end-to-end manner. The 232 results clearly show that even with basic fusion meth-233 ods, multimodal models significantly improve per-234 formance. 235

For the central site prediction, the average multimodal AUC improved by 2.8% and 7.4% compared to US and CT unimodal respectively. Likewise, 238

223

All rights reserved. No reuse allowed without permission.

LNM location	Modality	Multimodal fusion operation	ACC	AUC	SENS	SPEC	PREC	F1 score
	US		0.782	0.820	0.836	0.728	0.754	0.792
	CT		0.745	0.774	0.737	0.755	0.751	0.743
Control site	Multimodal	Concat	0.810	0.853	0.862	0.759	0.782	0.819
Central site		Sum	0.806	0.850	0.855	0.757	0.779	0.814
		Product	0.812	0.840	0.861	0.763	0.790	0.820
		Average results	0.809	0.848	0.859	0.760	0.784	0.818
	US		0.767	0.785	0.843	0.693	0.737	0.784
	CT		0.764	0.778	0.725	0.797	0.796	0.751
Lateral	Multimodal	Concat	0.804	0.825	0.860	0.744	0.779	0.815
cervical site		Sum	0.808	0.854	0.878	0.742	0.771	0.819
		Product	0.811	0.835	0.867	0.755	0.784	0.821
		Average results	0.808	0.838	0.868	0.747	0.778	0.818

 Table 2: Performance comparison of unimodal and multimodal approaches using basic fusion methods

for the lateral cervical site prediction, the average
multimodal AUC outperforms the US and CT unimodal models by 5.3% and 6.0% respectively (Table
2). These results affirm the hypothesis that US and
CT modalities comprise complementary information,

and their integration can improve the performanceof LNM status prediction.

Further improve multimodal performance

²⁴⁷ by incorporating a diverse granularity

248 feature fusion module

Multimodal fusion using basic methods can combine 249 US and CT information and improve LNM predic-250 tion but is not able to fully consider the interaction 251 between these two modalities. Recent progress based 252 on the attention mechanism has shown superiority in 253 multimodal fusion. In our study, we adopted the at-254 tention mechanism simultaneously on three granular 255 levels to fully incorporate the information useful for 256 LNM prediction, which are feature dimensions level 257 (minimum granularity), modalities level (medium 258 granularity), and nodules level (maximum granular-259 ity). In specific, these include dynamically adjusting 260 the attention weights of different feature dimensions 261 to balance the common and specific features of the 262 two modalities, adapting the modality-specific atten-263 tion to learn the respective advantages for different 264 nodules, plus flexibly aggerate the features of other 265 nodules based on nodule-level attention to refine the 266 prediction, considering nodules with the same LNM 267 status should exhibit greater feature similarity. We 268 refer to our modality fusion methods as the 'diverse 269 granularity fusion' network (DGFNet, as illustrated 270 in Fig. 2, detail see Methods). Equipped with the 271 DGF module, our model demonstrates exceptional 272 predictive capabilities with AUCs of 0.875 and 0.859 273 at the central and lateral cervical sites respectively 274

(Table 3), indicating its remarkable performance in 275 predicting the risk of LNM. Particularly, the multi-276 modal AUC exhibited a significant improvement of 277 5.5% and 10.1% compared to the US and CT uni-278 modal models at the central site, respectively. And 279 a substantial enhancement of 7.4% and 8.1% respec-280 tively at the lateral cervical site. These results further 281 underscore the efficacy of integrating US and CT in 282 predicting LNM in thyroid cancer. Furthermore, in 283 comparison to the basic fusion methods, our DGFNet 284 model achieves superior performance in nearly all 285 metrics, providing comprehensive evidence for the ef-286 fectiveness of fusing multimodal features at different 287 granularities. 288

DGFNet demonstrates exceptional generalization abilities

Recognizing the importance of the generalizability 291 of multimodal networks in clinical applications, we 292 evaluated the efficacy of our DGFNet model using 293 an external test dataset. The primary cohorts were 294 partitioned into training and validation sets, and the 295 model with the highest accuracy on the validation 296 set was selected to predict the LNM status of pa-297 tients on the external cohort. Owing to constraints 298 related to data availability, our external evaluation 299 is only performed at the central site, the results are 300 presented in Table 4. Overall, our DGFNet model per-301 formed well on the external dataset, with an accuracy 302 of 0.817 and an AUC of 0.903, showing similar per-303 formance compared to the internal accuracy of 0.844 304 and an AUC of 0.898. This consistency underscores 305 the strong robustness and external generalizability of 306 our model. 307

289

All rights reserved. No reuse allowed without permission.

LNM location	Multimodal fusion methods	ACC	AUC	SENS	SPEC	PREC	F1 score
Central site	Basic fusion	0.809	0.848	0.859	0.760	0.784	0.818
	Diverse granularity fusion	0.826	0.875	0.848	0.803	0.813	0.830
Lateral	Basic fusion	0.808	0.838	0.868	0.747	0.778	0.818
cervical site	Diverse granularity fusion	0.838	0.859	0.862	0.814	0.835	0.842

 Table 3: Performance comparison of multimodal methods using different fusion techniques

 Table 4: Performance of LNM status prediction on internal and external validation set

Dataset	ACC	AUC	SENS	SPEC	PREC	F1 score
Internal validation set	0.844	0.898	0.854	0.833	0.837	0.845
External validation set	0.817	0.903	0.909	0.763	0.690	0.784

DGFNet dynamically adjusts the

³⁰⁹ contribution of US and CT in predicting

the LNM status prediction at different

311 sites

We next seek to delineate the contribution of each 312 modality in the DGFNet model on every nodule. We 313 analyze by quantifying the contributions of US and 314 CT within the DGFNet model using the integrated 315 gradients[35] and comparing them to their unimodal 316 counterparts. The results are presented in Fig. 3, 317 where a larger feature attribution value corresponds 318 to a greater contribution to the correct prediction in 319 DGFNet model, and the red or green denotes correct 320 or wrong prediction respectively. 321

The result shows that, at both the central site (Fig. 322 3a) and lateral cervical site (Fig. 3b), there is a no-323 table number of cases where the DGFNet can change 324 the unimodality to make positive contributions even 325 when it fails to give correct prediction in unimodal 326 models (attribution greater than 0 but red color) and 327 lead to a correct prediction in this multimodal ap-328 proach. In addition, when looking at the central and 329 lateral cervical sites separately, we find that, across 330 all samples, 64.9% of nodules exhibit higher US attri-331 bution over CT attribution at the central site, while 332 55.2% of the nodules show higher US attribution over 333 CT attributions at the lateral cervical site. These find-334 ings agree with our prior observations on unimodal 335 LNM prediction performance, highlighting a more 336 prominent role for US at the central site, whereas 337 both US and CT show comparable importance at the 338 lateral cervical site. In addition, this underscores 339 that the DGFNet model can dynamically adjust the 340 weights of the two modalities based on nodule char-341 acteristics, effectively leveraging the strengths of both 342 modalities. 343

DGFNet enhances model attention on the nodular region in US and CT images

To further investigate how our DGFNet model im-346 proves the LNM prediction performance, we gener-347 ated saliency maps for both US and CT images in 348 the multimodal network and compared them with 349 their unimodal counterparts. The results clearly show 350 that, for both US and CT images, the DGFNet model 351 significantly increases the attention towards the re-352 gion of interest compared to the unimodal models. 353 Specifically, within US images, the multimodal model 354 focuses more intensely on the nodules' peripheral 355 and inner hypoechoic region (Fig. 4a), whereas in CT 356 images, it narrows its focus to the nodules and their 357 immediate surrounding tissues (Fig. 4b), all of which 358 represent crucial regions providing key information 359 for LNM prediction. This directly proves the supe-360 riority of DGFNet in grasping meaningful medical 361 information over unimodal methods. 362

Identify patients who can best benefit from multimodal integration

The multimodal approach can effectively improve the 365 LNM prediction, however, it is often unfeasible to ex-366 amine all patients by both modalities in real clinical 367 settings. Hence, to make our DGFNet more appli-368 cable and useful for clinicians, we further seek to 369 identify patients who can best benefit from the mul-370 timodal approach. Given that the US examination 371 is cheaper and more commonly used, we analyzed 372 by identifying cases for whom adding CT as a sup-373 plementary modality would be advantageous. We 374 evaluated four well-established sonographic charac-375 teristics of the thyroid nodule during US diagnosis 376 including maximum diameter, margin characteristics, 377 aspect ratio, plus location in the thyroid for central 378 cite nodules, and categorized the nodules based on 379 the measurements. We then compared the prediction 380 performance in each category between the DGFNet 381

344

345

363

All rights reserved. No reuse allowed without permission.

³⁸² and the unimodal model of US and CT respectively.

(Results on more characteristics are illustrated in Sup plementary Material)

The analysis shows that the DGFNet model 385 achieves particularly high performance in specific 386 circumstances. This includes nodules with maxi-387 mal diameters between 20mm and 36mm, as well as 388 more extreme cases less than 12mm or larger than 389 60mm(Fig. 5a). Additionally, DGFNet excels in cases 390 exhibiting non-smooth borders (Fig. 5b), aspect ratios 391 surpassing 1 (Fig. 5c), and nodules situated within 392 the thyroid isthmus (Fig. 5d). Similar findings are 393 observed in the analysis conducted at the lateral cer-394 vical site (Supplementary Material). Therefore, the 395 DGFNet model is potentially particularly beneficial 396 and practical for patients with the above nodule char-397 acteristics. 398

399

Discussion

Patients often undergo multiple types of examina-400 tions in the diagnostic process, and the effective in-401 tegration of multimodal information can greatly im-402 prove diagnosis accuracy. Recent advancements in 403 artificial intelligence techniques have facilitated the 404 progress of deep-learning-based multimodal integra-405 tion methods, which have emerged as a trend in 406 cancer diagnosis in recent years. In this study, we 407 pioneered the development of a multimodal deep 408 learning approach that effectively integrates US and 409 CT modalities to successfully enhance the accuracy 410 of LNM prediction and further demonstrate its gen-411 eralizability in an external dataset. Moreover, by 412 conducting a series of comprehensive interpretability 413 analyses, we quantified the modality-specific con-414 tribution across nodules in various situations, and 415 investigated the attention heatmap of US and CT im-416 ages within the model, which not only shed light 417 on the reasons for the improved performance of the 418 multimodal model, but also improve the model's 419 applicability in clinical settings, and opens a new 420 avenue for mitigating the problem of overtreating 421 thyroid cancer. 422

The effective integration of multimodal data often 423 relies on a deep understanding of the domain knowl-424 edge involved with specific medical tasks. In our 425 study, a close collaboration between AI scientists and 426 clinicians allowed us to leverage our collective exper-427 tise in deep learning models, thyroid cancer, US and 428 CT images. This enabled us to strategically employ 429 multi-task learning techniques, facilitating the identi-430 fication of critical regions and extraction of essential 431 LNM-related features from both US and CT images. 432 Moreover, we introduced a novel diverse granularity 433 fusion network (DGFNet) that learns the attention 434

from three different levels, which excels in not only 435 effectively integrating shared and specific features 436 from multimodal data but also dynamically adjust-437 ing the weights of each modality's data for different 438 nodules. This approach demonstrates the potential 439 to optimize the utility of both US and CT images and 440 aggregate information from similar nodules, thereby 441 enhancing the model's overall performance and ro-442 bustness. 443

Besides the excellent performance of our developed 444 DGFNet model, our study has yielded valuable clini-445 cal insights through the multimodal approach. First, 446 it shows that unimodal methods based on US appear 447 to be more sensitive but less specific, while CT-based 448 unimodal methods are the other way around. Second, 449 it shows that the US modality generally plays a more 450 significant role than CT at the central site, whereas 451 there is no obvious difference between US and CT at 452 the lateral cervical site. Furthermore, by quantifying 453 the performance of the unimodal and multimodal 454 models for nodules within different diagnosis charac-455 teristics categories, we could pinpoint patients with 456 certain nodule characteristics who can potentially 457 best benefit from the multimodal approach. These 458 analyses offer valuable insights for accurately iden-459 tifying the appropriate patient population for mul-460 timodal diagnostic approaches in clinical practice 461 and guiding patients in selecting the most suitable 462 examination method. 463

In conclusion, through a close collaboration be-464 tween AI scientists and clinicians, this study suc-465 cessfully develops a multimodal approach aimed at 466 improving the LNM prediction for thyroid patients. 467 It paves the way for addressing the issue of overtreat-468 ment in thyroid cancer and provides new insights 469 in the integration of multimodal data for precise di-470 agnosis, representing an excellent scientific research 471 example originating from clinical practice and di-472 rectly addressing clinical necessities. 473

Methods

Patient Cohort

There are two cohorts included in this study. The 476 main cohort was obtained from Zhejiang Cancer Hos-477 pital in Zhejiang Province, China, consisting of 1360 478 patients. After the data screening process, a total of 479 1138 patients with 1285 nodules were retained for 480 analysis. The main cohort was utilized for the model 481 establishment and internal performance evaluation. 482 The second cohort, referred to as the external co-483 hort, was sourced from Shaoxing People's Hospital 484 in Zhejiang Province, China. Initially, this cohort in-485 cluded 126 patients, and after the data screening pro-486

All rights reserved. No reuse allowed without permission.

cess, 60 patients with complete data were included
for evaluation of model generalization (The patient
enrollment process is illustrated in Supplementary
Material). Ethical approval for the study was obtained from the ethics committees of both hospitals
and verbal informed consent was obtained from all
participating patients.

The inclusion criteria for this study encompassed 494 the following: (1) patients with thyroid nodules, (2) 495 patients who underwent cervical US and CT exami-496 nations, and (3) patients with confirmed pathological 497 status of cervical LNM. Exclusion criteria consisted 498 of the following: (1) missing US or CT data, (2) the 499 presence of measurement lines in the US images, and 500 (3) patients with multiple malignant thyroid nod-501 ules and metastatic cervical lymph nodes. As all the 502 thyroid nodules had the potential to metastasize, it 503 was impossible to determine which specific nodule 504 had metastasized to the cervical lymph nodes. After 505 the data screening process, the number of nodules 506 with and without metastasis in the central and lat-507 eral cervical sites for both cohorts is presented in the 508 Supplementary Material. 509

Multiple US images, including transverse and lon-510 gitudinal sections, as well as multiple CT images 511 from different slices, were available for most nodules 512 in the dataset. During each epoch of the training 513 process, one random US image and one random CT 514 image were paired together to form an image pair. 515 During the evaluation process, the US and CT images 516 with the largest nodal area were selected from the 517 multiple available images to form an image pair for 518 analysis. 519

520 Data pre-processing

Region of Interest Extraction. The methods used 521 for extracting the region of interest in both US and 522 CT images are similar and described as follows: 1) 523 We first performed a dilation operation on the mask 524 of thyroid nodules annotated by clinicians, using a 525 3x3 dilation kernel. The iteration steps were set to 526 40 and 25 for US and CT images, respectively. 2) We 527 determined the horizontal bounding rectangle of the 528 dilated region, with the height, width, and center 529 coordinates of the rectangle denoted as h, w, and 530 $(x_{\text{center}}, y_{\text{center}})$, respectively. 3) Using $(x_{\text{center}}, y_{\text{center}})$ 531 as the center and the larger value of *h* and *w* as the 532 side length, we obtained the external square of the 533 thyroid nodule. 4) The original US and CT images 534 were then cropped to reserve the region within this 535 square. If the square area exceeds the image bound-536 ary, the images are padded with zeros to fill the 537 exceeding part. 538

Image Augmentation. The cropped US and CT
 images were resized to 288x288 pixels and 96x96 pix-

els, respectively. To enhance the diversity of the data, 541 we applied additional data augmentation techniques 542 to both modalities. These techniques included rota-543 tion, horizontal flip, cropping and scaling, brightness-544 contrast transformation, and elastic transformation. 545 For rotation, the angle of rotation ranged from -15° 546 to 15°. Random cropping occurred with the cropped 547 area set to be between 90% and 100% of the original 548 size. The probability of applying these transforma-549 tions was set to 0.5, ensuring a balanced augmenta-550 tion effect. 551

Convolutional neural network architecture 552

The architecture of the proposed model is depicted 553 in Fig. 2. The model is composed of three distinct 554 branches: the US branch, the CT branch, and the Mul-555 timodal branch. Both the US and CT branches share 556 an identical structure, each comprising an encoder 557 and two decoders. The encoder adopts a pre-trained 558 ResNet[34] architecture, with ResNet34 and ResNet18 559 selected for the central and lateral cervical sites, re-560 spectively. Regarding the US branch, the decoders 561 are trained to delineate the mask and boundary of 562 thyroid nodules, directing the model's attention to-563 wards the internal and marginal regions of the nod-564 ules, correspondingly. Conversely, the CT branch's 565 decoders focus on segmenting the mask of thyroid 566 nodules and the boundary of surrounding tissue, fa-567 cilitating the model in comprehensively capturing 568 information about both the thyroid nodule and its 569 adjacent surroundings. 570

All the aforementioned decoders share the same 571 structure. Each decoder is constructed from 5 upsam-572 ple blocks, with every block encompassing 2 layers. 573 In the initial layer of each block, the input feature 574 is upsampled using bilinear interpolation. Subse-575 quently, the second layer comprises a convolutional 576 block, incorporating a convolutional layer featuring a 577 kernel size of 3×3, followed by batch normalization, 578 relu activation, and a dropout layer. Notably, to en-579 hance segmentation performance, short connections 580 interconnect the encoder and decoder components. 581

The US and CT encoders produce 512-dimensional 582 vectors through global average pooling. In unimodal 583 models, these vectors directly enter the classifier 584 for LNM prediction. In the multimodal model, the 585 unimodal vectors integrate within the multimodal 586 branch and then proceed to the classifier with the 587 same structure—a two-layer fully connected neural 588 network with 512 input nodes and a single output 589 node. 590

All rights reserved. No reuse allowed without permission.

⁵⁹¹ Diverse granularity fusion module

The diverse granularity fusion module comprises three branches, as depicted in the supplementary material. All branches are constructed using the attention mechanism.

⁵⁹⁶ In the dimensional correlation branch, the US and

⁵⁹⁷ CT features undergo a preliminary transformation as ⁵⁹⁸ outlined below:

$$Q_{\rm US}^{\rm D} = f_{\rm US} \times W_{\rm Q-US}^{\rm D} \tag{1}$$

$$V_{\rm US}^{\rm D} = f_{\rm US} \times W_{V-\rm US}^{\rm D} \tag{2}$$

$$Q_{\rm CT}^{\rm D} = f_{\rm CT} \times W_{Q-\rm CT}^{\rm D} \tag{3}$$

$$V_{\rm CT}^{\rm D} = f_{\rm CT} \times W_{V-\rm CT}^{\rm D} \tag{4}$$

⁵⁹⁹ Here, $f_{\rm US}$ and $f_{\rm CT}$ are the unimodal features of US and CT, respectively, and $W^{\rm D}_{Q-\rm US}$, $W^{\rm D}_{V-\rm US}$, $W^{\rm D}_{K-\rm CT}$, ⁶⁰¹ and $W^{\rm D}_{V-\rm CT}$ are trainable parameters. The product ⁶⁰² of $Q^{\rm D}_{\rm US}$ and $K^{\rm D}_{\rm CT}$, followed by the application of the ⁶⁰³ softmax function, results in the attention matrix $A^{\rm D}$, ⁶⁰⁴ which captures the interplay between various feature ⁶⁰⁵ dimensions of the US and CT modalities:

$$A^{\rm D} = \operatorname{softmax}\left(\frac{Q_{\rm US}^{\rm D} \times K_{\rm CT}^{\rm D^{\rm T}}}{\sqrt{d_k}}\right)$$
(5)

Here, d_k is the dimension of K_{CT}^D . The derived attention matrix is then utilized for the enhanced multimodal features:

$$f_{\rm US}^{\rm D} = V_{\rm US}^{\rm D} + A^{\rm D} \times V_{\rm US}^{\rm D} \tag{6}$$

$$f_{\rm CT}^{\rm D} = V_{\rm CT}^{\rm D} + A^{\rm D^{\rm T}} \times V_{\rm CT}^{\rm D} \tag{7}$$

⁶⁰⁹ Ultimately, the enriched features are amalgamated
 ⁶¹⁰ through concatenation along the dimension axis,
 ⁶¹¹ yielding the fused features:

$$f^{\rm D} = \operatorname{concat}\left(f^{\rm D}_{\rm US'} f^{\rm D}_{\rm CT}\right)$$
 (8)

In the modal weights branch, the US and CT features are first concatenated along the modal axis:

$$f_{\rm US-CT}^{\rm N} = \operatorname{concat}\left(f_{\rm US}, f_{\rm CT}\right) \tag{9}$$

Then the Q^{M} , K^{M} , and V^{M} are generated respectively:

$$Q^{\rm M} = f^{\rm M}_{\rm US-CT} \times W^{\rm M}_Q \tag{10}$$

$$K^{\rm M} = f^{\rm M}_{\rm US-CT} \times W^{\rm M}_K \tag{11}$$

$$V^{\rm M} = f^{\rm M}_{\rm US-CT} \times W^{\rm M}_V \tag{12}$$

The W_Q^M , W_K^M , and W_V^M are trainable parameters. ⁶¹⁶ Through the multiplication of Q^M and K^M , an attention matrix emerges, encapsulating the priority of the two modalities within separate nodes. ⁶¹⁹

$$A^{\rm M} = \operatorname{softmax}\left(\frac{Q^{\rm M} \times K^{\rm M^{\rm T}}}{\sqrt{d_k}}\right) \tag{13}$$

Subsequently, this attention matrix is employed to discussion adjust the relative significance of the two modalities: 620

$$f^{\rm M} = V^{\rm M} + A^{\rm M} \times V^{\rm M} \tag{14}$$

In the nodal correlation branch, US and CT features 422 are first merged along the dimensional axis: 622

$$f_{\rm US-CT}^{\rm N} = \operatorname{concat}\left(f_{\rm US}, f_{\rm CT}\right) \tag{15}$$

Then Q^N , K^N , and V^N are obtained respectively: 624

$$Q^{\rm N} = f_{\rm US-CT}^{\rm N} \times W_Q^{\rm N} \tag{16}$$

$$K^{\rm N} = f_{\rm US-CT}^{\rm N} \times W_K^{\rm N} \tag{17}$$

$$V^{\rm N} = f^{\rm N}_{\rm US-CT} \times W^{\rm N}_V \tag{18}$$

The W_Q^N , W_K^N , and W_V^N are trainable parameters. ⁶²⁵ The attention matrix is obtained and employed to ⁶²⁶ delineate the interrelation between distinct nodules: ⁶²⁷

$$A^{\rm N} = \operatorname{softmax}\left(\frac{Q^{\rm N} \times K^{\rm N^{\rm T}}}{\sqrt{d_k}}\right)$$
(19)

Refined features considering the similarity of different nodules emerge: 629

$$f^{\rm N} = V^{\rm N} + A^{\rm N} \times V^{\rm N} \tag{20}$$

The features from the three branches undergo 630 element-wise multiplication, resulting in the ultimate 631 fused features: 632

$$f_{\rm F} = f^{\rm D} \odot f^{\rm M} \odot f^{\rm N} \tag{21}$$

Nodule boundary extraction in US images 633

Firstly, the nodule boundary width (d) was deter-634 mined as a multiple (f) of the square region of in-635 terest's length. For our study, f was set to 0.08. Sec-636 ondly, the annotated thyroid nodule mask underwent 637 dilation and erosion operations to yield R_{dilation} and 638 R_{erosion} , respectively, with a kernel size of 3×3 and 639 iterations of 0.5*d*. Finally, the nodule's boundary 640 was obtained as the difference between $R_{dilation}$ and 641 $R_{\text{erosion}} (R_{\text{dilation}} - R_{\text{erosion}}).$ 642

All rights reserved. No reuse allowed without permission.

643 Boundaries of surrounding tissue

644 extraction in CT images

Firstly, bilateral filtering[36] was applied to preserve 645 the edges while reducing noise. The diameter of the 646 pixel field was set to 7, and the sigma values for both 647 the color space and coordinate space were set to 100. 648 Secondly, the Canny algorithm[37] was employed to 649 further extract the boundaries of the surrounding 650 tissues. The lower and upper threshold values were 651 set to -100 and 200, respectively. 652

Training Configuration

The base learning rate in our study was set to 654 1×10^{-4} , and we employed a cosine learning rate 655 schedule during the training process. The batch size 656 was set to 30, and we utilized the Adam optimizer 657 to optimize our model. A weight decay of 1×10^{-5} 658 was applied to mitigate overfitting. In this study, a 659 multi-task strategy was employed to address differ-660 ent tasks. For the classification task, specifically the 661 prediction of the LNM status, we utilized a binary 662 cross-entropy loss function. As for the segmenta-663 tion tasks, a combination of binary cross-entropy loss 664 and Intersection over Union (IOU) loss functions was 665 utilized. The model was initially trained for the seg-666 mentation tasks for the first 100 epochs, and then 667 the classification task was added and trained for the 668 remaining 200 epochs. 669

670 Interpretability Analysis Methods

We employed the integrated gradients[35] method to 671 enhance the interpretability of our model. Integrated 672 gradients is a feature attribution technique that cal-673 culates the integral of gradients along the path from 674 a chosen baseline to the input, resulting in an attri-675 bution value for each input feature. In our study, the 676 baseline is manually specified, and we select a base-677 line where the predicted probability of our trained 678 model is close to 0.5, indicating equal probabilities 679 for both LNM presence and absence. To determine 680 the contributions of US and CT images, we sum the 681 attributions of each pixel in the respective images. By 682 visualizing the attribution of each pixel, we generate 683 saliency maps for US and CT images. 684

685 Statistical Analysis

We assessed the performance of our model using several evaluation metrics, including accuracy, area under the curve (AUC), specificity, sensitivity (also known as recall), precision, and F1-score. To analyze the model's performance across different thresholds, we constructed receiver operating characteristic (ROC) curves, plotting sensitivity against specificity.

Hardware and Software

The computational resources utilized include an Intel 694 10900K CPU with a clock speed of 3.7GHz and 20 695 threads. The graphics card employed is a GEFORCE 696 RTX 3090, equipped with 10752 CUDA cores and 697 24GB of graphics memory. The programming lan-698 guage used for implementation is Python 3.9.7, and 699 the deep learning framework employed is PyTorch 700 1.10.0. 701

693

702

708

712

Data availability

Though this study was carried out with participant 703 consent, the dataset remains restricted in public access. For research inquiries, de-identified data can be obtained from the corresponding author upon 706 reasonable request. 707

Code availability

The code for model development and 709 interpretability analysis is accessible at 710 https://github.com/li10107/DGFNet. 711

References

- YuJiao Deng et al. "Global burden of thyroid 713 cancer from 1990 to 2017". In: *JAMA network 714 open* 3.6 (2020), e208759–e208759. 715
- [2] Hyuna Sung et al. "Global cancer statistics 716 2020: GLOBOCAN estimates of incidence and 717 mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 719 (2021), pp. 209–249. 720
- [3] Kyu Eun Lee et al. "Ipsilateral and contralateral rentral lymph node metastasis in papillary thy-roid cancer: patterns and predictive factors of nodal metastasis". In: *Head & neck* 35.5 (2013), pp. 672–676.
- [4] David T Hughes and Gerard M Doherty. "Central neck dissection for papillary thyroid cancer". In: *Cancer Control* 18.2 (2011), pp. 83–88.
- [5] Bryan R Haugen et al. "2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer". In: *Thyroid* 26.1 (2016), pp. 1–133.

All rights reserved. No reuse allowed without permission.

- [6] Claudio Gambardella et al. "The role of prophylactic central compartment lymph node dissection in elderly patients with differentiated thyroid cancer: a multicentric study". In: *BMC surgery* 18.1 (2019), pp. 1–8.
- [7] Mostafa Alabousi et al. "Diagnostic test accuracy of ultrasonography vs computed tomography for papillary thyroid cancer cervical lymph node metastasis: A systematic review and meta-analysis". In: *JAMA Otolaryngology– Head & Neck Surgery* 148.2 (2022), pp. 107–118.
- [8] Yinlong Yang et al. "Prediction of central compartment lymph node metastasis in papillary thyroid microcarcinoma". In: *Clinical endocrinol-* 0gy 81.2 (2014), pp. 282–288.
- [9] Jiang Zhu et al. "Application of machine learning algorithms to predict central lymph node metastasis in T1-T2, non-invasive, and clinically node negative papillary thyroid carcinoma". In: *Frontiers in medicine* 8 (2021), p. 635771.
- [10] Jong-Lyel Roh, Jin-Man Kim, and Chan Il Park.
 "Central lymph node metastasis of unilateral papillary thyroid carcinoma: patterns and factors predictive of nodal metastasis, morbidity, and recurrence". In: *Annals of surgical oncology* 18 (2011), pp. 2245–2250.
- [11] Meng Jiang et al. "Nomogram based on shear-wave elastography radiomics can improve pre-operative cervical lymph node staging for papillary thyroid carcinoma". In: *Thyroid* 30.6 (2020), pp. 885–897.
- [12] Jinhua Yu et al. "Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics". In: *Nature communications* 11.1 (2020), p. 4807.
- [13] Tongtong Liu et al. "Comparison of the application of B-mode and strain elastography ultrasound in the estimation of lymph node metastasis of papillary thyroid carcinoma based on a radiomics approach". In: *International journal of computer assisted radiology and surgery* 13 (2018), pp. 1617–1627.
- [14] Vivian Y Park et al. "Radiomics signature for prediction of lateral lymph node metastasis in conventional papillary thyroid carcinoma". In: *PLoS One* 15.1 (2020), e0227315.
- [15] Jingjing Li et al. "Computed tomographybased radiomics model to predict central cervical lymph node metastases in papillary thyroid carcinoma: a multicenter study". In: *Frontiers in Endocrinology* 12 (2021), p. 741698.

- Yun Peng et al. "Prediction of central lymph node metastasis in cN0 papillary thyroid carcinoma by CT radiomics". In: *Academic Radiology* 30.7 (2023), pp. 1400–1407.
- [17] Shanshan Zhao et al. "Combined Conventional Ultrasound and Contrast-Enhanced Computed Tomography for Cervical Lymph Node Metastasis Prediction in Papillary Thyroid Carcinoma". In: *Journal of Ultrasound in Medicine* 42.2 (2023), pp. 385–398.
- [18] Jana Lipkova et al. "Artificial intelligence for multimodal data integration in oncology". In: 798 *Cancer cell* 40.10 (2022), pp. 1095–1110. 799
- [19] Kevin M Boehm et al. "Harnessing multimodal data integration to advance precision oncology". In: *Nature Reviews Cancer* 22.2 (2022), pp. 114–126.
- [20] Julián N Acosta et al. "Multimodal biomedical AI". In: *Nature Medicine* 28.9 (2022), pp. 1773–1784.
- [21] Kevin M Boehm et al. "Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer". In: *Nature cancer* 3.6 (2022), pp. 723–733.
- [22] Xuejun Qian et al. "Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning". In: *Nature biomedical engineering* 5.6 (2021), pp. 522–532.
- [23] Richard J Chen et al. "Pan-cancer integrative histology-genomic analysis via multimodal deep learning". In: *Cancer Cell* 40.8 (2022), pp. 865–878.
- [24] Hong-Yu Zhou et al. "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics". In: *Nature Biomedical Engineering* (2023), pp. 1–13.
- [25] Ye Zhu et al. "Vision+ X: A Survey on Multimodal Learning in the Light of Data". In: arXiv preprint arXiv:2210.02884 (2022).
- [26] Chao Zhang et al. "Multimodal intelligence: Representation learning, information fusion, and applications". In: *IEEE Journal of Selected Topics in Signal Processing* 14.3 (2020), pp. 478–493.
- [27] Hassan Akbari et al. "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text". In: Advances in Neural Information Processing Systems 34 (2021), pp. 24206–24221.

All rights reserved. No reuse allowed without permission.

- Kaiming He et al. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.
- [29] Xiaohan Wang, Linchao Zhu, and Yi Yang.
 "T2vlad: global-local sequence alignment for text-video retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5079–5088.
- [30] Diping Song et al. "Deep relation transformer
 for diagnosing glaucoma with optical coherence tomography and visual field function". In: *IEEE Transactions on Medical Imaging* 40.9 (2021),
 pp. 2392–2402.
- [31] Shuai Zheng et al. "Multi-modal graph learning for disease prediction". In: *IEEE Transactions on Medical Imaging* 41.9 (2022), pp. 2207– 2216.
- [32] Richard J Chen et al. "Multimodal co-attention transformer for survival prediction in gigapixel
 whole slide images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 4015–4025.
- [33] Tsai Hor Chan et al. "Histopathology Whole
 Slide Image Analysis With Heterogeneous
 Graph Representation Learning". In: *Proceed- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15661–
 15670.
- [34] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [35] Mukund Sundararajan, Ankur Taly, and Qiqi
 Yan. "Axiomatic attribution for deep networks".
 In: *International conference on machine learning*.
 PMLR. 2017, pp. 3319–3328.
- [36] Carlo Tomasi and Roberto Manduchi. "Bilateral filtering for gray and color images". In: Sixth international conference on computer vision (IEEE Cat. No. 98CH36271). IEEE. 1998, pp. 839–846.
- [37] John Canny. "A computational approach to edge detection". In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), pp. 679–698.
 - Acknowledgments

883

This work was supported by the National Natural
Science Foundation of China (No. 82071946), the
Natural Science Foundation of Zhejiang Province
(No. LZY21F030001), the Pioneer and Leading Goose

R&D Program of Zhejiang (No. 2023C04039), the 888 National Key Research and Development Program 889 of China (2022YFF0608403), Youth Research Fund 890 Project of Shaoxing People's Hospital (Grant Num-891 ber 2022YB07), and the fund of Zhejiang Province 892 Medical and Health Science and Technology Project 893 (No. 2023KY581). We thank Y.G. for providing us 894 external validation set. 895

Author contributions

896

907

908

M.H., C.S., X.L., and D.X. conceived and planned 897 the study. C.S., G.L., and X.L. designed the research 898 framework. J.Y., C.P., S.Z., and J.Y. collected the 899 raw US and CT images, patients' clinical informa-900 tion, and image annotation. G.L., Y.H., and X.F. per-901 formed the data preprocessing and conducted the 902 performance analysis. G.L. designed the multimodal 903 fusion method and carried out model interpretation 904 analysis. G.L. and C.S. wrote the manuscript. All 905 authors commented on the manuscript. 906

Competing interests

The authors declare no competing interests.





Figure 1: Overall AI system for LNM risk prediction. The main cohort was employed for AI system development and evaluation, while the external cohort assessed the system's generalizability. After preprocessing, paired US and CT images are input into DGFNet, our deep learning model, to predict LNM status in central and lateral cervical regions. Post-AI system development, we conducted an extensive interpretability analysis comprising multimodal contribution assessment, saliency map visualization, and multimodal applicability evaluation.





Figure 2: DGFNet architecture. DGFNet consists of three branches: the US branch, CT branch, and multimodal branch. Each US and CT branch incorporates an encoder and two decoders. DGFNet concurrently performs five tasks: nodal mask and boundary segmentation in US images (guiding the model to focus on internal and marginal nodule features), boundary segmentation of nodules and surrounding tissues in CT images (guiding the model to focus on nodule and surrounding tissue features in CT images), and the final LNM prediction. The fusion of multimodal features in the latent space occurs within the diverse granularity fusion module, and the final results are generated by subsequent fully connected layers. The diverse granularity fusion module includes the dimensional correlation branch, modal weights branch, and nodal correlation branch, amalgamating characteristics from both modalities to provide a diverse granularity information integration. A detailed explanation of this module is available in the Methods section.



Figure 3: Attribution analysis of US and CT in predicting LNM status at central (a) and lateral cervical (b) sites. Each subfigure comprises four panels, with the shared horizontal axes indicating nodule indices. The central site includes 954 nodules, while the lateral cervical site includes 402 nodules. The values of the top two panels display attributions from US and CT images in the multimodal prediction, respectively. Column colors denote unimodal predictions, where green signifies accurate predictions and red indicates inaccuracies. The third panel illustrates the multimodal prediction results. Panel 4 represents the ground truth.



-1.00 -0.75 -0.50 -0.25 0.00 0.25 0.50 0.75 1.00

Figure 4: Examples of saliency map visualization results at central site. In these instances, both the US and CT unimodal models initially generated inaccurate predictions, whereas the multimodal models effectively rectified these to provide accurate predictions. The red curve delineates the nodule's boundary in the original US and CT images. The color red signifies an elevated likelihood of LNM development, whereas the color blue signifies the contrary.



Figure 5: Distribution of nodules with varied attributes and associated correct predictions ratio in central Site. Attributes encompass nodal maximum diameter (considering the larger of the maximum diameters from transverse and longitudinal US views) in US image(a), characteristics of margin (b), aspect ratio (calculated as the height divided by the width in transverse views) of nodules(c), and location in thyroid (d).