cancers

MDPI

Article

Novel Human Artificial Intelligence Hybrid Framework Pinpoints Thyroid Nodule Malignancy and Identifies Overlooked Second-Order Ultrasonographic Features

Xiaohong Jia [1,†], Zehao Ma [2,3,†], Dexing Kong [2,3,4], Yamin Li [5], Hairong Hu [6], Ling Guan [7], Jiping Yan [8], Ruifang Zhang [9], Ying Gu [10], Xia Chen [10], Liying Shi [10], Xiaomao Luo [11], Qiaoying Li [12], Baoyan Bai [13], Xinhua Ye [14], Hong Zhai [15], Hua Zhang [16], Yijie Dong [1], Lei Xu [3,*], Jianqiao Zhou [1,*] and CAAU [‡]

[1] Department of Ultrasound, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200025, China
[2] School of Mathematical Sciences, Zhejiang University, Hangzhou 310013, China
[3] Zhejiang Qiushi Institute for Mathematical Medicine, Hangzhou 311121, China
[4] College of Mathematical Medicine, Zhejiang Normal University, Jinhua 321004, China
[5] Department of Ultrasound, Puyang People's Hospital, Puyang 457005, China
[6] Demetics Medical Technology, Hangzhou 310012, China
[7] Department of Ultrasound, Gansu Provincial Cancer Hospital, Lanzhou 730050, China
[8] Department of Ultrasound, Shanxi Provincial People's Hospital, Taiyuan 030012, China
[9] Department of Ultrasound, The First Affiliated Hospital, Zhengzhou University, Zhengzhou 450052, China
[10] Department of Ultrasound, Affiliated Hospital of Guizhou Medical University, Guiyang 550001, China
[11] Department of Ultrasound, The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming 650031, China
[12] Department of Ultrasound Diagnostics, Tangdu Hospital, Fourth Military Medical University, Xi'an 710038, China
[13] Department of Ultrasound, Affiliated Hospital of Yan'an University, School of Medicine, Yan'an University, Yan'an 716000, China
[14] Department of Ultrasound, The First Affiliated Hospital of Nanjing Medical University, Nanjing 210029, China
[15] Department of Ultrasound, Traditional Chinese Medical Hospital of Xinjiang, Urumqi 830000, China
[16] Department of Ultrasound, Anyang Tumor Hospital, The Fourth Affiliated Hospital of Henan University of Science and Technology, Anyang 455000, China
* Correspondence: xulei@qsmm.org.cn (L.X.); zjq11432@rjh.com.cn (J.Z.)
† These authors contributed equally to this work.
‡ The Chinese Artificial Intelligence Alliance for Thyroid and Breast Ultrasound (CAAU), Shanghai 200025, China.

Simple Summary: Deep learning-based computer-aided diagnosis has gained momentum in the radiology field thanks to the technological advances of convolutional neural networks (CNN). However, how to utilize the black-box predictions of these CNN models to the clinical routine still relies on radiologists' personal judgements. In addition, existing CNN models only improve radiologists' diagnosis when they outperform the radiologists, thereby limiting their added values for possible efficiency enhancement and improving mostly the diagnostic performances of junior radiologists.

Abstract: We present a Human Artificial Intelligence Hybrid (HAIbrid) integrating framework that reweights Thyroid Imaging Reporting and Data System (TIRADS) features and the malignancy score predicted by a convolutional neural network (CNN) for nodule malignancy stratification and diagnosis. We defined extra ultrasonographical features from color Doppler images to explore malignancy-relevant features. We proposed Gated Attentional Factorization Machine (GAFM) to identify second-order interacting features trained via a 10 fold distribution-balanced stratified cross-validation scheme on ultrasound images of 3002 nodules all finally characterized by postoperative pathology (1270 malignant ones), retrospectively collected from 131 hospitals. Our GAFM-HAIbrid model demonstrated significant improvements in Area Under the Curve (AUC) value (p-value $< 10^{-5}$), reaching about 0.92 over the standalone CNN (~0.87) and senior radiologists (~0.86), and identified a second-order vascularity localization and morphological pattern which

check for updates

Citation: Jia, X.; Ma, Z.; Kong, D.; Li, Y.; Hu, H.; Guan, L.; Yan, J.; Zhang, R.; Gu, Y.; Chen, X.; et al. Novel Human Artificial Intelligence Hybrid Framework Pinpoints Thyroid Nodule Malignancy and Identifies Overlooked Second-Order Ultrasonographic Features. *Cancers* **2022**, *14*, 4440. https://doi.org/10.3390/cancers14184440

Academic Editor: Emiel A.M. Janssen

Received: 15 July 2022
Accepted: 9 September 2022
Published: 13 September 2022

was overlooked if only first-order features were considered. We validated the advantages of the integration framework on an already-trained commercial CNN system and our findings using an extra set of ultrasound images of 500 nodules. Our HAIbrid framework allows natural integration to clinical workflow for thyroid nodule malignancy risk stratification and diagnosis, and the proposed GAFM-HAIbrid model may help identify novel diagnosis-relevant second-order features beyond ultrasonography.

Keywords: thyroid nodule; malignancy risk stratification; feature selection; second-order feature interaction; interpretable deep learning; radiology

1. Introduction

Ultrasound is the most widely used medical imaging tool to evaluate thyroid nodules, whose prevalence is found to be 30% to 68% [1,2], and the malignancy risks are evaluated based on the ultrasound nodular features according to Thyroid Imaging Reporting and Data System (TIRADS). While having many shared risk attributes, there exist TIRADS standards like ACR-TIRADS [3] proposed by American College of Radiology, Kwak-TIRADS [4] by Korean researchers, Eu-TIARDS [5] from the European Thyroid Association, K-TIRADS [6] from the Korean Society of Thyroid Radiology and the Korean Society of Radiology, as well as most recent C-TIRADS [7] proposed by the Chinese Medical Association and the Chinese Artificial Intelligence Alliance for Thyroid and Breast Ultrasound (CAAU). Despite being more standardized, no matter what TIRADS is taken for thyroid nodule screening, the lesion interpretations by radiologists are still subjective.

With the development of deep learning Convolutional Neural Networks (CNNs), radiologists start to use computer aided diagnosis (CADx) systems as a second opinion for thyroid nodule evaluation [8–10]. It consists of many layers of computations, taking the advantage of convolution computations over the images using kernels or filters of different sizes to extract features relevant to a classification task [11]. The outputs of each layer serves as the inputs for the next. The introduction of nonlinear activation functions in CNN models enables complex nonlinear mapping between their inputs and outputs. After many layers of computations, the activated spatial patterns of the perceptive field or the feature maps become incomprehensible to humans. Consequently, it is up to individual radiologists to make their final decisions. An immediate challenge facing the radiologists is how to objectively integrate such CADx systems into their clinical workflow such that the usage of these tools also becomes standardized. No less important is whether this integration can be made understandable or at least interpretable to humans. Otherwise, creating yet another black box on top of one does not address the concerns of radiologists as we would always prefer a conceptually meaningful explanation to an important decision. It is desirable to standardize the adoption of CADx systems in clinical practices by constructing a set of meaningful rules of criteria about how to use them in an objective way.

For medical image diagnosis, generating class activation heatmaps [12–15] localizes the deep representation of class-discriminating image regions of diseased tissues [9,16,17]. However, these techniques merely highlight the regions of importance for the class prediction with false color visualization (for instance, using typically red and blue respectively for positive and negative contribution), offering no further insight into what exact features from those outlined regions contribute to the model's predictions. An overlooked possibility is to take the diagnosis results of a trained CNN as an input feature, aggregate it to human-extracted ones for subsequent feature selection and weight optimization to refine the final prediction. In essence, the malignancy score outputted by a CNN is assigned with a weight representing its contribution to the final decision. Though how the CNN as a modular building block of the complete decision-making system, which we refer as Human Artificial Intelligence hybrid (HAIbrid) framework, comes about its decision is still opaque, that is no longer the case with regards to its contribution. Consequently, it enables us to generate a new set of TIRADS criteria that sum up contributions of both features noted by radiologists and a CNN model for malignancy risk stratification and diagnosis.

In this work, we chose Kwak-TIRADS as the basis for thyroid nodule evaluation as it has been widely adopted in China. In addition, we included extra vasculature-related ultrasound features [18] that describe both the localizations in relation to a nodule (i.e., being perinodular, close to the center, etc.) and their morphologies (being twisted or not twisted) to explore possibilities for establishing novel clinically-relevant TIRADS criteria. The complete set of candidate features is referred as TIRADS^{+} and listed in Appendix A Table A1. The interplay between cancerous tumors and vascularization is generally acknowledged in cancer research [19,20], and vascularity features observed in color Doppler images are an indispensable component in the T [21]. Some other malignancy-relevant features in ACR BIRADS, for instance posterior echoic features, heterogeneity of echo patterns, etc., are also included in the candidate feature set. In addition, the presence of hypoechoic or anechoic rim referred to as halo surrounding nodule introduced in C-TIRADS [7] and its thickness were also considered.

Hypothetically, co-occurrence of features or second-order feature interactions can constituent a unique malignancy-relevant feature. Just as an individual's inherent interdependent preferences over items of certain categories can be modeled by second or higher-order feature interactions using Factorization Machine (FM) [22] for improving the prediction of online users' click-through rate for advertising [23], there presumably can also exist interdependence between certain ultrasound features of malignant thyroid nodules. However, to our humble knowledge, this has never been explored in the context of the CADx field. Additionally, as different feature interactions may have unequal impacts on predictions, on top of an existing attention-FM variant (AFM) [24] that weighs the importance of each feature interaction, we introduce an additional gate mechanism that removes the interference of minimally relevant feature interactions, termed GAFM, to explore the usefulness of second-order interacting features in ultrasound-based thyroid nodule diagnosis.

As a proof of concept, we took a ResNet101 model [25] pretrained on ImageNet as a basis CNN and used the Area Under the Curve (AUC) of receiver operating characteristics as the main evaluation metric to compare the diagnosis results by standalone radiologists, CNN-based CADx, and by HAIbrid methods using feature selection modules of logistic regression (LR), conventional AFM model and our modified GAFM model. We compared two CNN systems integrated through our proposed GAFM-HAIbrid model, including an already-trained commercial system, to validate the usefulness of the model on an extra ultrasound image dataset of 500 nodules.

2. Materials and Methods

2.1. Clinical Samples

This study was based on the authorized clinical dataset with continuous data contribution from the consortium of CAAU [7] with 131 alliance hospitals led by Ruijin Hospital Affiliated to School of Medicine, Shanghai JiaoTong University. All registered data were from patients who had taken preoperative ultrasound examinations and had definitive postoperative pathological examination outcomes. The pathological results were taken as the gold standard for diagnosis evaluation by radiologists, CNN and the integrated approaches.

For each nodule, ultrasound scans of standard longitudinal and transverse planes were saved, including B-mode and color Doppler images. A retrospective cohort of 3002 thyroid nodules from 3002 patients continuously registered at CAAU alliance hospitals collected between January 2017 and August 2019 following the same protocol as in the previous publication was used for training and cross-validating our proposed model. Of these, 1732 nodules were determined to be benign, and 1270 nodules were malignant according to histopathological diagnosis. These samples were randomly partitioned into 10 folds for cross-validation experiments for which each fold was alternatingly taken as the validation set while the rest were taken as the training set. It is also important to note that dataset shift, which introduces substantial set to set variations, is a known deteriorating factor for accurate performance estimation [26]. To eliminate this effect, we adopted the distribution-balanced stratified cross-validation [27] algorithm to balance sample distributions in the feature space when partitioning a dataset according to radiologists' assigned features.

Data for an additional 500 nodules collected between February and April of 2021 were used to verify results obtained on the former cohort. Half of these nodules were benign and the other half were malignant, confirmed by pathological examinations.

2.2. Candidate Feature Set Construction

This part mainly describes the construction of candidate feature set from features extracted by radiologists and diagnosis results of a convolutional neural network for establishment of interpretable thyroid diagnosis criteria and malignancy risk stratification, as illustrated in Figure 1a.

The features extracted by radiologists are descriptions of ultrasound images as text data using feature dictionary defined in TIRADS. Each sample is represented by structured data in consistent order. One-Hot Encoding is used to convert text data to feature vectors, using N-bit status registers to encode N states, with each state having an independent register bit, and only one bit being valid at any time. For absent feature description, 0 was used to encode the feature to ensure consistent vector dimension.

In our study, the collected samples were labeled on the basis of Kwak-TIRADS and our defined extra features, most of which were already described in the published C-TIRADS [7] except features defined in color Doppler images, to explore possibilities to further improve the diagnosis. The complete candidate features are listed in Appendix A Table A1.

The diagnosis results from CNN-based CADx were included as an additional candidate feature for constructing candidate feature set together with human-extracted features. This is different from the conventional procedure where image features are extracted via a CNN model for later feature fusion and classifications.

2.3. Feature Selection Based on AFM and GAFM Model

The candidate features were thereafter used to construct weighted second-order feature interactions using either AFM or our proposed variant GAFM module according to the individual contribution of each feature to the diagnosis. The difference between AFM and GAFM is that an additional gate mechanism is introduced to only keep the top-ranked features according to their weights. The inclusion of the gate mechanism is because some insignificant second-order features, if not removed, can interfere the learned weights of other features and return sub-optimal combinations of features for the ultimate thyroid malignancy classifications. The architecture of our proposed GAFM model is illustrated in Figure 1b. The input of the model is the feature vector from the candidate feature set and we denote the feature vector as $\boldsymbol{X}$.

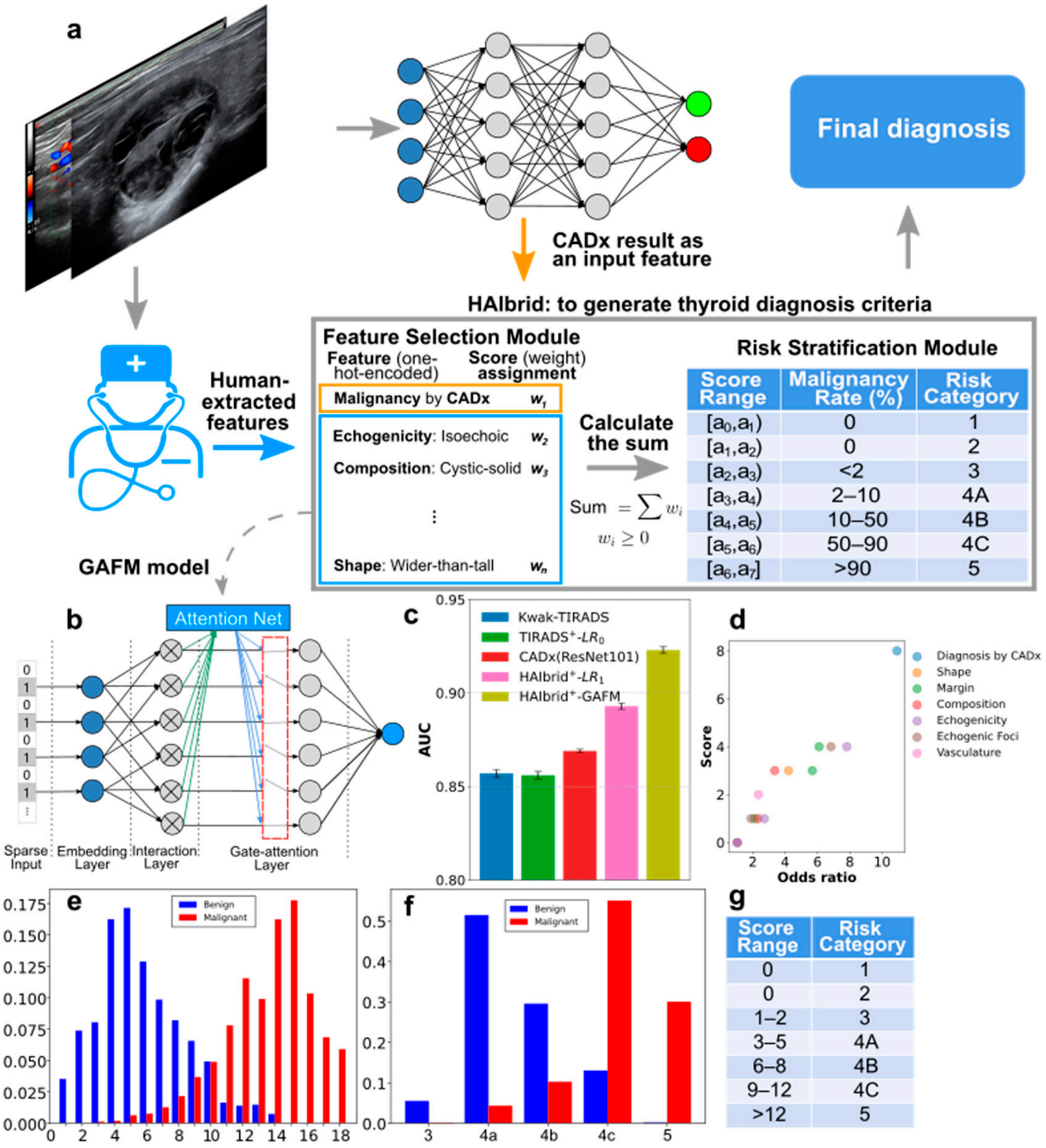
2.3.1. Embedding Layer

This layer adopts a sparse representation for input features and embeds each non-zero feature into a dense vector, which can be represented as

$$e_i = V_i x_i, i = 1, 2, \ldots, n, \tag{1}$$

where x_i is a non-zero feature in the feature vector $\boldsymbol{X}$ and n is the number of the non-zero features. The output of the embedding layer is then the concatenation of multiple embedding vectors as:

$$E = [e_1, e_2, \ldots, e_n], i = 1, 2, \ldots, n \tag{2}$$



Score Range	Malignancy Rate (%)	Risk Category
$[a_0,a_1)$	0	1
$[a_1,a_2)$	0	2
$[a_2,a_3)$	<2	3
$[a_3,a_4)$	2–10	4A
$[a_4,a_5)$	10–50	4B
$[a_5,a_6)$	50–90	4C
$[a_6,a_7]$	>90	5

Score Range	Risk Category
0	1
0	2
1–2	3
3–5	4A
6–8	4B
9–12	4C
>12	5

Figure 1. Our proposed HAIbrid method to establish CNN-boosted human-understandable thyroid diagnostic criteria and the corresponding results. (**a**) Scheme depicting the workflow: first, radiologists extract features from the ultrasound images and a convolutional neural network of arbitrary architectural design produces primary diagnosis results as a separate feature to construct a candidate feature set and then use a feature selection module that optimizes feature weights to generate the thyroid diagnostic criteria and finally return the diagnosis. The risk stratification module ensures that the malignancy rate for each risk category is within the corresponding specified range. (**b**) The neural network architecture of our proposed GAFM model consists of an input layer, embedding layer and interaction layer identical to an AFM, with an additional gate mechanism to filter out minimally relevant second-order interacting features. (**c**) Bar plot of the mean AUC values together with the corresponding standard deviations from 10-fold feature-stratified cross-validation experiments for diagnosing thyroid nodules using different methods. (**d**) The correlation between the assigned score of each selected feature by the GAFM model and the corresponding odds ratio. (**e**) The probability density functions of HAIbrid-TIRADS scores computed based on GAFM model for benign and malignant samples. (**f**) The distributions of samples according to our risk classifications. (**g**) Table of score ranges for risk classifications.

2.3.2. Interaction Layer

For pairwise interactions of n vectors, there are $n(n-1)/2$ interacting vectors, each of which is the element-wise product of two distinct vectors representing their interaction. This can be expressed as:

$$[\langle e_1, e_2 \rangle, \langle e_1, e_3 \rangle, \ldots, \langle e_{n-1}, e_n \rangle] \tag{3}$$

where e_i is the feature embedding and <□,□> represents the inner product of two vectors.

2.3.3. Attention Layer

The attention mechanism of the GAFM model is the same as AFM, which parameterizes the attention score with a multi-layer perceptron (MLP). The input of the gate-attention network is the pairwise interacting vectors, while the output is the attention scores, a vector whose dimension is consistent with the number of interacting vectors. The attention score of each feature interaction in the embedding space $a_{(i,j)}$ for the entire sample on average $a_{(i,j)}$ representing the feature importance is given by:

$$\alpha_{(i,j)} = \frac{1}{N}\sum_{N} a_{(i,j)} \tag{4}$$

The gate mechanism introduced in GAFM model sets a gate operation for each second-order interacting feature, and the number of gate operations are consistent with that of the interacting features. When the weight of the interacting feature is greater than the preset threshold, the gate state is open, and second-order interacting feature is retained, otherwise the opposite.

2.4. Model Training

We propose a two-step training procedure for our GAFM model. The first step is similar to the training process of an AFM model and there is no gate mechanism involved. After the model converges, we use the gate mechanism to remove minimally relevant second-order interacting features. The second step is to re-train the model after the gate operation with the remaining second-order interacting features and all first-order ones. The output of gate-attention layer can be expressed as:

$$L_{GAFM} = < W, X > + \sum_{(i,j)\in\Gamma} a_{(i,j)} < e_i, e_j > \tag{5}$$

where $\Gamma = \left\{ (i,j) \middle| \alpha_{(i,j)} > \varepsilon, i, j = 1, 2, \ldots, n \right\}$ is the subscript set of the remaining second-order interacting features after gate operation with a threshold of ε. For conventional AFM, no gate operation is applied, therefore all second-order interacting features are kept. The output of the GAFM model can be expressed as:

$$y_{GAFM} = sigmoid(L_{GAFM}) \tag{6}$$

The objective function is the cross-entropy loss function expressed as:

$$L(y, y_{GAFM}) = -y\log(y_{GAFM}) - (1-y)\log(1-y_{GAFM}) \tag{7}$$

After the model converges, we normalize the weights of all remaining features through the softmax function, then rank the coefficient of each feature in the model and select the top-ranked features. Following the design principle of Kwak-TIRADS, 15 top-ranked features were selected. The expression of the softmax function is as follows:

$$f(\mathbf{X})_j = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}} \tag{8}$$

where $\mathbf{X}$ is a K-dimensional vector, x_j is the *j-th* element and $f(\mathbf{X})_j$ is the softmax function value of feature x_j.

2.5. Risk Stratification

Once the feature selection and corresponding weight for each selected feature is determined after the model training, a 1–5 risk stratification system (including 4A, 4B and 4C) is constructed. This is performed by summing the weights for each sample *j*:

$$Score_j = \sum_{i \in O} W_i, \tag{9}$$

in which *O* represents the feature set for sample *j* that includes all assigned features by radiologists and the diagnosis by the CNN-based CADx system, while w_i represents the score for each feature *i* of sample *j*.

The malignancy rate for each summed score S (MR_{Score}) is computed as follows:

$$MR_{Score} = \frac{M_{Score}}{M_{Score} + B_{Score}}, \tag{10}$$

where M_{Score} represents the number of malignant samples with a summed score S while B_{Score} represents the number of benign samples with a summed score S. Thereafter, the score range $[a_i, a_{i+1})$ for each assigned risk category *i* according to the malignancy rate illustrated in the risk stratification module in Figure 1 is determined. Remember that during model training, all weights are normalized with their sum equals to 1. However, this is not user-friendly to radiologists. To mimic a conventional TIRADS score system, a multiplication factor to weight is manually optimized and each obtained weight is then rounded to an integer.

2.6. Control Experiments

To evaluate the effectiveness of our proposed diagnostic framework that takes advantages of both CNN-based CADx and radiologist-extracted features for thyroid nodules, we designed a series of control experiments as summarized in Table 1.

Table 1. Experiment ID, the corresponding method and short descriptions.

Exp. #	Method	Description
1	Kwak-TIRADS	Kwak-TIRADS criteria were followed by radiologists to extract features and to evaluate thyroid nodules.
2	Kwak-TIRADS + LR_0	Radiologists followed Kwak or candidate TIRADS$^+$ to extract features; LR worked as a classifier, denoted as LR_0.
3	TIRADS$^+$-LR_0	
4	CADx (ResNet101)	ResNet101 was used as an exemplar model for CADx.
5	Kwak-TIRADS/CADx	Return diagnosis of higher malignancy from either method.
6	HAIbrid-LR_0	Hybrid denotes combining features from TIRADS and diagnosis results from CNN-based CADx for classifications using LR merely as a classifier, denoted as LR_0.
7	HAIbrid$^+$-LR_1	Hybrid+ denotes the combination of features defined in TIRADS$^+$ with diagnosis results from CADx; LR_1 *, AFM or GAFM when each corresponding method was used as the feature selector and classifier simultaneously.
8	HAIbrid$^+$-AFM	
9	HAIbrid$^+$-GAFM	

* When using LR_1 as both a feature selector and classifier, the cross-entropy loss function is used as the cost function to learn the weights of each feature vector through stochastic gradient descent. After the model converges, the weight of each feature is ranked and the top-ranked features selected for later thyroid diagnosis.

Our proposed diagnostic framework to establish new understandable TIRADS criteria for thyroid nodules relies on using radiologist-extracted features according to TIRADS criteria and the diagnosis results from CNN-based CADx. As a simple ablation study to verify the significance of our HAIbrid approach, we included the diagnosis by radiologists according to Kwak-TIRADS criteria (method 1) or with a subsequent LR classifier (method 2), or with additionally defined candidate features (method 3), the diagnosis by conventional CADx (method 4), a simple integral approach that combine the diagnoses by radiologists and a CADx system by returning always the higher malignancy (method 5) or a plain integral approach that combines the features defined in TIRADS and the diagnosis results from CADx followed by using LR merely for classification (method 6) or as both a feature selector and classifier (method 7). To elucidate the significance of second-order feature interactions and the importance of further feature weight optimization, we included conventional AFM (method 8), our proposed GAFM variant (method 9) and a conventional method ignoring second-order feature interactions (method 7).

For verifying the results obtained from the cross-validation experiments, a CNN-based CADx system, AI-SONIC^{TM} Thyroid of software version 4.0 (2020) developed by Demetics Medical Technology (Hangzhou, China), was tested to check whether the proposed integration framework could also bring benefits to an independently trained CNN model.

2.7. Statistical Analysis

For cross-validation experiments, we used pairwise two-tailed t-test to evaluate whether the differences in mean AUC values between the GAFM model and control experiments were statistically significant. For comparison of two CNN models integrated through the GAFM-HAIbrid strategy, we subdivided the ultrasound images of 500 nodules to two groups, each of which were further partitioned to five subsets for both intra-group and cross-group statistical comparisons using the feature stratified partition method mentioned above. For intra-group comparisons of different diagnostic methods, as the dataset partitions were kept identical, pairwise two-tail t-tests were used for statistical evaluations. For cross-group comparisons, standard two-tail t-tests were used instead.

3. Results

We first performed a feature-stratified 10-fold cross-validation experiment on retrospectively collected 3002 nodules (1270 malignant ones) from our nationwide database contributed by 131 member hospitals. From the results, it can be seen that combining the CADx with TIRADS, denoted as HAIbrid+LR$_0$ (where LR was used merely as a classifier without feature exclusion) outperformed each separate method, i.e., Kwak-TIRADS and CADx (Supplementary Figure S1). Interestingly, the inclusion of extra primary candidate features made no difference when using LR as a classifier. In contrast, enabling feature selection capability of LR while providing extra candidate features with HAIbrid$^+$-LR$_1$ improved the diagnosis compared to the case of HAIbrid-LR$_0$ (0.893 ± 0.0017 vs 0.881± 0.0025). Furthermore, AFM-based methods including our proposed variant GAFM achieved significantly better performances (Figure 1c and Supplementary Figure S1) than the HAIbrid$^+$-LR$_1$. In addition, incorporating a gate mechanism to AFM by removing minimally relevant second-order features using GAFM significantly improved the diagnosis (p-value = 1.4×10^{-5} in Supplementary Figure S1).

The resulting TIRADS criteria, referred as HAIbrid-TIRADS which takes diagnosis from CNN-based CADx as an additional feature reweighted together with ultrasonographical features noted by radiologists, are listed in Supplementary Table S1 with the corresponding scores returned by the GAFM model for primary features and second-order feature interactions. The distribution of the summed scores and that of the corresponding risk stratifications for the benign and malignant samples confirmed by pathological examinations are shown in Figures 1e and 1f, respectively. The optimal threshold chosen by maximizing F1 score for thyroid nodule diagnosis based on HAIbrid-TIRADS scores was found to be 9 (for which 10 and above were considered suspicious to be malignant, Supplementary Figure S2); at this threshold, the precision and sensitivity were found to be 0.921 and 0.912 with the false positive and false negative rate being 0.102 and 0.088, respectively.

It is worth-noting that our GAFM model identified the second-order feature interaction concerning vasculature localization and morphology (noted by radiologists from Doppler ultrasound images in Supplementary Figure S3), i.e., being mainly perinodularly distributed and untwisted (OR = 2.349) to be important for thyroid nodule classifications. However, these vasculature-related features defined by radiologists in this study played no significant roles when fed directly to conventional classification methods such as LR, in which second-order feature interactions were ignored (Figure 1c and Supplementary Figure S1). On the ultrasound image test set of 500 extra thyroid nodule, the OR of the vasculature-related second-order feature interaction was 2.41, consistent to the OR calculated from the retrospectively collected 10-fold cross validation set.

Based on the 10-fold cross-validation experiments, another identified candidate feature in this study was punctate echogenic foci of undetermined significance, defined according to C-TIRADS [7], was found to have an OR of 1.891 (Supplementary Table S1). However, on the test set of 500 extra thyroid nodules, the OR was 1.27 instead.

Furthermore, we applied the GAFM-HAIbrid method on two different CNN models to assess the effectiveness of this approach on the test cohort by dividing the ultrasound images of 500 thyroid nodules to two groups of equal sizes in a feature-stratified manner, one for the trained ResNet101 model and the other for a CNN-based CADx system, AI-SONIC^{TM} Thyroid. It is clear that in both groups the HAIbrid results exceeded the diagnoses by the radiologists alone by large margins (Figure 2a) and were significantly better than the standalone CADx models (Figure 2a,c,d). The performances of the radiologists in these two groups were comparable (p-value = 0.065, Figure 2b). Interestingly, by combining radiologists' diagnoses through our HAIbrid approach, the performance gap between the under-performing ResNet and the AI-SONIC^{TM} Thyroid system (Figure 2a) was largely closed-off (from 0.8463 ± 0.0020 versus 0.9023 ± 0.0008 to 0.9167 ± 0.0019 versus 0.9263 ± 0.0027) despite a still measurable statistical difference (p-value ~ 2×10^{-4}).

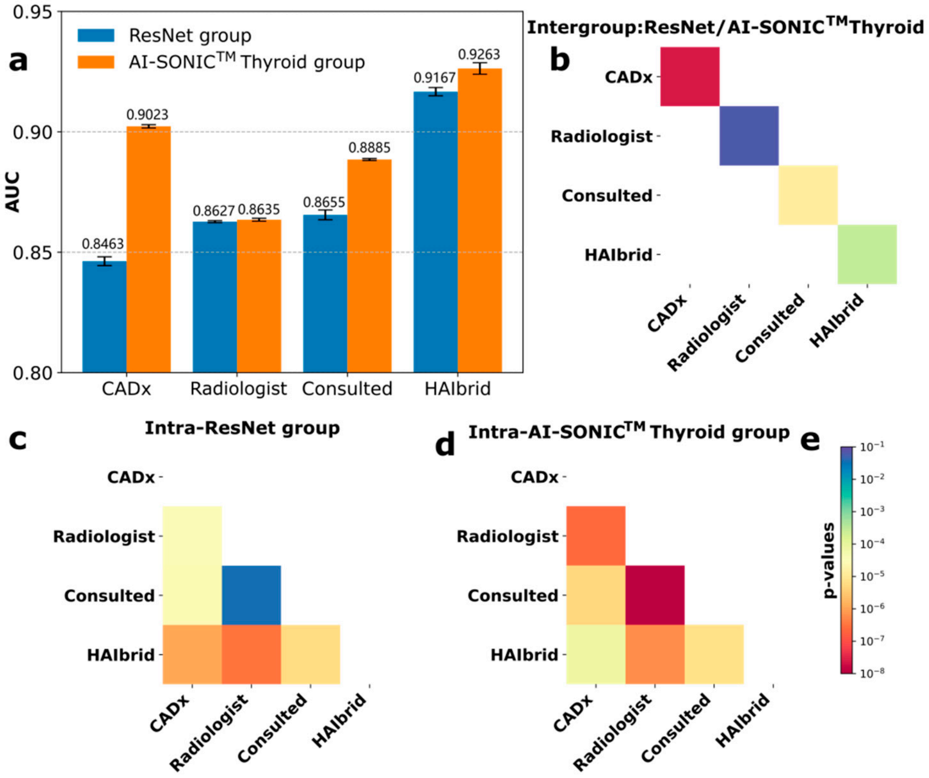


Figure 2. Comparison of two feature-stratified groups of equal sizes (250 cases each) to evaluate the importance of applying GAFM-based HAIbrid approach for improving thyroid nodule diagnosis performances of two CADx systems, the trained ResNet101 and commercial AI-SONIC^{TM} Thyroid. (**a**) The barplots of the diagnosis AUCs for CADx, radiologists, radiologists consulting the respective CADx and the HAIbrid approach in each group. (**b**) p-value heatmap for inter-group comparison. (**c**) p-value heatmap for the intra-group comparison for the ResNet group as well as for the AI-SONIC^{TM} Thyroid group (**d**). (**e**) The common p-value color-bar in logarithmic scale shared by (**b**–**d**).

4. Discussion

From the 10-fold cross validation experiments on retrospectively collected thyroid nodule ultrasound images, feature selection and weight reassignment in general can make an influential impact on the performance of the resulted computer-aided TIRADS criteria. For instance, enabling feature selection capability of conventional LR while providing extra candidate features improved the diagnosis compared to the case where LR was only used as a classifier, suggesting the relevance of feature selections on additional features for improving thyroid nodule diagnosis. Furthermore, the score of each selected feature correlated very well with their corresponding odds ratios (ORs), supporting that the learned scores reflect the relative malignancy risks of each feature. Implementing AFM-based feature selection modules especially our proposed GAFM variant within the HAIbrid framework that consider second-order interacting features further improved the diagnostic performances.

The GAFM-HAIbrid model identified a second-order vascular-related feature being mainly perinodularly distributed and untwisted. At first glance, it may appear contradictory to the notion that the thyroid nodule malignancy was more associated with the intranodular vascularity given by a study of 402 nodules, in which 23 were malignant [8]. However, according to a recent quantitative study of regional vascularity of 111 thyroid nodules, in which 27 were malignant, the mean vascularity index of peripheral region of the malignant nodules significantly higher than that of the benign nodules. In addition, the authors that suggested the contrary had only the binary definition of intra- and peri-nodular vascularity, different from our definition of five cases, and their identified association between regional vascularity and nodule malignancy may well be sample-dependent. In our study, 3002 nodules (including 1270 malignant ones) were continuously collected from 131 member hospitals of the CAAU, and should contain less sampling bias. Furthermore, though there was a claim that markedly chaotic central and peripheral vasculature was considered more suspicious for malignancy [28], to our humble knowledge, there has not been a quantitative study about what vasculature localization patterns and morphology concurrently correlate with thyroid nodule malignancy. Vasculature-related second-order feature from this test set was found to be 2.41 (not much different from 2.349), confirming the validity of it as a useful malignancy marker.

Based on the 10-fold cross-validation experiments, a first-order feature "punctate echogenic foci of undetermined significance" was also identified by the GAFM-HAIbrid method with an OR of 1.891. Its OR was different from the OR of 0.944 in the original C-TIRADS paper, where it was considered insignificant [7]. This is most likely due to differences in samples distribution though substantial overlapping cases could be expected as the data used in this study came from the same database maintained by CAAU. In the test cohort of the 500 extra nodules, its OR was 1.27 instead, suggesting the contribution of this feature for malignancy prediction was unstable. One explanatory factor to the observed high degree of variability is that all sample data were periodically uploaded to the CAAU database while different contributing hospitals could have different operational cycles. Nevertheless, this complicating factor had a negligible effect on our identified vasculature-related second-order feature.

It is also interesting to note that our test dataset results show that combining radiologists' diagnoses through our HAIbrid approach had a greater impact on the underperforming ResNet such that the standalone performance gap compared to AI-SONIC$^{\text{TM}}$ Thyroid system was largely closed-off. This suggests that standardizing the incorporation of radiologists' expertise to CNN-based CADx systems may outweigh the improvement of standalone CADx systems. In contrast, by inspecting the developments of high-profile deep learning algorithms from a historical perspective, it can be seen that they typically outperformed the previous state of the art algorithms by 1–3% [25,29–32]. It should also be noted that the better performing standalone EfficientNet-based AI-SONIC$^{\text{TM}}$ Thyroid system than the ResNet101 model may have benefited more from much larger available training samples than we had in this study. Nevertheless, this indicates that from a practical perspective, pushing the performance of a standalone CADx system to exceed senior radiologists' diagnostic performance for the real-world clinical workflow may no longer be an urgent need.

At the same time, the results show that our HAIbrid approach did not require the CNN-based CADx to be superior to radiologists for diagnosis performance improvement. This is different from other publications where the improvement of radiologists' diagnosis by consulting the results from CADx systems demands these systems to be more accurate than the radiologists [10,33]. Our approach instead can be perceived as an ensemble learning model [10,34] that is based on two constituent weaker learners, although one of them is radiologists. This methodology puts a softer requirement on the constituent raters that the predictions by CNN-based CADx systems are independent of the radiologists. Or in other words, there exists valuable complementary information that can be utilized to improve the final diagnosis. This is easier to fulfill because these systems are usually trained according to results from pathological examinations rather than radiologists' ratings in reference to a set of TIRADS criteria [35].

It is however desirable to relieve the burden of radiologists as much as possible. It is preferable to obtain the TIRADS features from another CNN system as done in reference [35] and to allow human intervention if needed. With this minimally human-intervened variant of HAIbrid approach, favorable image feature interpretability, diagnosis standardizability, and user-friendliness are simultaneously achieved. This is certainly of interest, but requires a large dataset with tremendous annotation efforts to train a CNN system to achieve accurate multi-class classification results. Other limitations of this work include that we have not tested the improvement of diagnostic accuracy for junior radiologists but focused on senior radiologists. Our consideration for this omission is mainly due to the fact that we focus on identifying novel malignancy-relevant features. In addition, in the clinical routine, junior radiologists' diagnoses need to be double-checked by a senior radiologist before vital decisions are made. We also have good reasons to believe that our framework will improve their diagnoses significantly also. It has been shown in numerous studies that improving senior radiologists' diagnostic accuracy is more difficult than junior radiologists by using a CNN system [10,34]. For single features, the highest weight was assigned to the prediction by the CNN system in our GAFM-HAIbrid model (Supplementary Table S1). As the two CNN systems evaluated in this work have equivalent diagnostic performances to senior radiologists or even better, it is expected that having a high weight on the CADx system when integrating with the junior radiologists' diagnoses will certainly be helpful as well for them. A third perceivable limitation is that an international multicenter study would be preferred to validate whether our findings in this work are universally applicable. However, this is for future investigation.

5. Conclusions

To conclude, our HAIbrid approach allows integration of CNNs of arbitrary architectural designs, combines thyroid nodule diagnosis results from CNN-based CADx system and radiologist-defined features, permits the establishment of novel human interpretable TIRADS criteria that outperform the original diagnostic methods and requires no extra efforts from radiologists to adapt and integrate to their clinical workflow. Our proposed GAFM feature selection method enables the identification of clinically relevant second-order feature interactions, which have been overlooked by radiologists and conventional feature selection methods. We anticipate that our integrated approach can be generally applicable for medical imaging-based diagnosis of other diseases beyond ultrasonography as long as radiologists may consult the second opinion from a CNN-based CADx system and our proposed GAFM model can be presumably adopted to identify potentially important second-order feature interactions that are overlooked in existing diagnostic criteria.

6. Patents

A patent application has been filed with publication number CN113889229A, and publication date 24 January 2022.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers14184440/s1. Figure S1. Bar plot of the mean AUC values in diagnosing thyroid nodules from 10-fold cross-validation experiments on 3002 nodules and the statistical comparisons; Table S1. Our HAIbrid-TIRADS criteria returned by the proposed GAFM model for thyroid nodule malignancy risk stratification from the cross-validation cohort; Figure S2. The identification of the optimal threshold to separate malignant from benign samples based on F1 score computed using our HAIbrid-GAFM model; Figure S3. Representative images for additionally defined vasculature-related features.

Author Contributions: Conceptualization, L.X., D.K. and J.Z.; methodology, Z.M. and L.X.; software, Z.M.; validation, X.J., Z.M., L.X. and J.Z.; formal analysis, Z.M. and L.X.; investigation, X.J. and Z.M.; resources, D.K., J.Z. and H.H.; data curation, all but L.X., D.K. and J.Z.; writing—original draft preparation, Z.M. and L.X.; writing—review and editing, L.X. and J.Z.; visualization, L.X.; supervision, L.X.; project administration, L.X., D.K. and J.Z.; funding acquisition, D.K. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 12090020, 12090025 and 82071928 and in part by the Natural Science Foundation of Zhejiang Province under Grant LSD19H180005.

Institutional Review Board Statement: Ethical approvals were waived by the local Ethics Committee of participating hospitals in view of the retrospective nature of the study and all involved ultrasound image reviews being performed were part of the clinical routine.

Informed Consent Statement: Written informed consent has been obtained from the patient representative to publish this paper.

Data Availability Statement: The data that support the findings of this study are available on reasonable request from corresponding author J.Z. after formal approval by the concerned Chinese regulating authorities.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Candidate Ultrasonographical Features

Table A1. Our defined candidate features (TIRADS+) of thyroid nodules in ultrasound images to establish a clinically relevant HAIbrid-TIRADS. The features shown in light gray are extra features defined with the aim to identify new potentially useful features for thyroid diagnosis.

Feature	Definition	Feature	Definition
Orientation (Shape)	Parallel (Wider-than-tall)	*Homogeneity of echo intensity*	Homogeneous
	Vertical (Taller-than-wide)		Inhomogeneous
Margin	Circumscribed	*Vasculature localization*	Avascular
	Ill-defined		Perinodular
	Irregular		Mixed
	Extra-thyroidal extension/micro-lobulated		Mainly intranodular
Composition	Solid		Mainly perinodular
	Predominantly solid	*Vascular morphology*	Not twisted
	Mixed cystic and solid		Twisted
Macro-calcification	Present	*Halo thickness*	Thin
	Absent		Thick

Table A1. *Cont.*

Feature	Definition	Feature	Definition
Echogenicity	Hyperechoic	*Posterior echo features*	Absent
	Isoechoic		Enhanced
	Hypoechoic		Shadowing
	Marked hypoechoic		Mixed
Echogenic foci	Micro-calcifications	Note: (1) Micro-calcifications correspond to punctate echogenic foci with or without shadowing; (2) Punctate echogenic foci of undetermined significance correspond to punctate echogenic foci without shadowing or comet-tail artifacts, therefore it is difficult to tell whether it is micro-calcification or colloid [7].	
	Comet-tail artifacts		
	Peripheral calcifications		
	No punctate echogenic foci		
	Punctate echogenic foci of undetermined significance [7]		

References

1. Guth, S.; Theune, U.; Aberle, J.; Galach, A.; Bamberger, C.M. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. *Eur. J. Clin. Investig.* **2009**, *39*, 699–706. [CrossRef] [PubMed]
2. Moon, J.H.; Hyun, M.K.; Lee, J.Y.; Im Shim, J.; Kim, T.H.; Choi, H.S.; Ahn, H.Y.; Kim, K.W.; Park, D.J.; Park, Y.J.; et al. Prevalence of thyroid nodules and their associated clinical parameters: A large-scale, multicenter-based health checkup study. *Korean J. Intern. Med.* **2018**, *33*, 753–762. [CrossRef] [PubMed]
3. Tessler, F.N.; Middleton, W.D.; Grant, E.G.; Hoang, J.K.; Berland, L.L.; Teefey, S.A.; Cronan, J.J.; Beland, M.D.; Desser, T.S.; Frates, M.C.; et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *J. Am. Coll. Radiol.* **2017**, *14*, 587–595. [CrossRef]
4. Kwak, J.Y.; Han, K.H.; Yoon, J.H.; Moon, H.J.; Son, E.J.; Park, S.H.; Jung, H.K.; Choi, J.S.; Kim, B.M.; Kim, E.-K. Thyroid imaging reporting and data system for US features of nodules: A step in establishing better stratification of cancer risk. *Radiology* **2011**, *260*, 892–899. [CrossRef] [PubMed]
5. Russ, G.; Bonnema, S.J.; Erdogan, M.F.; Durante, C.; Ngu, R.; Leenhardt, L. European Thyroid Association Guidelines for Ultrasound Malignancy Risk Stratification of Thyroid Nodules in Adults: The EU TIRADS. *Eur. Thyroid* **2017**, *6*, 225–237. [CrossRef]
6. Shin, J.H.; Baek, J.H.; Chung, J.; Ha, E.J.; Kim, J.H.; Lee, Y.H.; Lim, H.K.; Moon, W.; Na, D.G.; Park, J.S.; et al. Ultrasonography Diagnosis and Imaging-Based Management of Thyroid Nodules: Revised Korean Society of Thyroid Radiology Consensus Statement and Recommendations. *Korean J. Radiol.* **2016**, *17*, 370–395. [CrossRef]
7. Zhou, J.; Song, Y.; Zhan, W.; Wei, X.; Zhang, S.; Zhang, R.; Gu, Y.; Chen, X.; Shi, L.; Luo, X.; et al. Thyroid imaging reporting and data system (TIRADS) for ultrasound features of nodules: Multicentric retrospective study in China. *Endocrine* **2020**, *70*, 256–279. [CrossRef]
8. Ma, J.; Wu, F.; Zhu, J.; Xu, D.; Kong, D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. *Ultrasonics* **2017**, *73*, 221–230. [CrossRef]
9. Wang, L.; Zhang, L.; Zhu, M.; Qi, X.; Yi, Z. Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. *Med. Image Anal.* **2020**, *61*, 101665. [CrossRef]
10. Peng, S.; Liu, Y.; Lv, W.; Liu, L.; Zhou, Q.; Yang, H.; Ren, J.; Liu, G.; Wang, X.; Zhang, X.; et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: A multicentre diagnostic study. *Lancet Digit. Health* **2021**, *3*, e250–e259. [CrossRef]
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
12. Ramprasaath, R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv* **2016**, arXiv:1610.02391.
13. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. *arXiv* **2018**, arXiv:1710.11063.
14. Saurabh, D.; Harish, G.R. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. In Proceedings of the 2020 Workshop on Applications of Computer Vision, Snowmass Village, CO, USA, 1 March 2020; pp. 983–991.
15. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. *arXiv* **2020**, arXiv:1910.01279.

16. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225v3.
17. Chan, H.; Hadjiiski, L.M.; Samala, R.K. Computer-aided diagnosis in the era of deep learning. *Med. Phys.* **2020**, *47*, e218–e227. [CrossRef]
18. Papini, E.; Guglielmi, R.; Bianchini, A.; Crescenzi, A.; Taccogna, S.; Nardi, F.; Panunzi, C.; Rinaldi, R.; Toscano, V.; Pacella, C.M.; et al. Risk of malignancy in nonpalpable thyroid nodules: Predictive value of ultrasound and color-Doppler features. *J. Clin. Endocrinol. Metab.* **2002**, *87*, 1941–1946. [CrossRef]
19. Weis, S.; Cheresh, D. Tumor angiogenesis: Molecular pathways and therapeutic targets. *Nat. Med.* **2011**, *17*, 1359–1370. [CrossRef]
20. Kuczynski, E.A.; Vermeulen, P.B.; Pezzella, F.; Kerbel, R.S.; Reynolds, A.R. Vessel co-option in cancer. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 469–493. [CrossRef]
21. D'Orsi, C.J.; Sickles, E.A.; Mendelson, E.B.; Morris, E.A. *ACR BI-RADS®Atlas, Breast Imaging Reporting and Data System*; American College of Radiology: Reston, VA, USA, 2013; ISBN 155903016X.
22. Rendle, S. Factorization Machines. In Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, NSW, Australia, 13 December 2010; pp. 995–1000.
23. Cheng, H.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M.; et al. Wide & Deep Learning for Recommender Systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, Boston, MA, USA, 15 September 2016.
24. Xiao, J.; Ye, H.; He, X.; Zhang, H.; Wu, F.; Chua, T.S. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks. In Proceedings of the 26th International Joint Conference on Artificial Intelligence Main Track, Melbourne, Australia, 19–25 August 2017; pp. 3119–3125.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Moreno-Torres, J.G.; Saez, J.A.; Herrera, F. Study on the Impact of Partition-Induced Dataset Shift on k-Fold Cross-Validation. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 1304–1312. [CrossRef]
27. Zeng, X.; Martinez, T. Distribution-balanced stratified cross-validation for accuracy estimation. *J. Exp. Theor. Artif. Intell.* **2000**, *12*, 1–12. [CrossRef]
28. Anil, G.; Hegde, A.; Chong, F.V. Thyroid nodules: Risk stratification for malignancy with ultrasound and guided biopsy. *Cancer Imaging* **2011**, *11*, 209–223. [PubMed]
29. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
30. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 6105–6114.
31. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
33. Mai, W.; Zhou, M.; Li, J.; Yi, W.; Li, S.; Hu, Y.; Ji, J.; Zeng, W.; Gao, B.; Liu, H. The value of the Demetics ultrasound-assisted diagnosis system in the differential diagnosis of benign from malignant thyroid nodules and analysis of the influencing factors. *Eur. Radiol.* **2021**, *31*, 7936–7944. [CrossRef] [PubMed]
34. Dietterich, T.G. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*; MCS 2000. Lecture Notes in Computer Science; Kittler, J., Roli, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2000; Volume 1857, pp. 1–15.
35. Chen, Y.; Gao, Z.; He, Y.; Mai, W.; Li, J.; Zhou, M.; Li, S.; Yi, W.; Wu, S.; Bai, T.; et al. An Artificial Intelligence Model Based on ACR TI-RADS Characteristics for US Diagnosis of Thyroid Nodules. *Radiology* **2022**, *303*, 613–619. [CrossRef]