

MINISTRY OF
EDUCATION
AND TRAINING

VIETNAM ACADEMY OF SCIENCE
AND TECHNOLOGY

GRADUATE UNIVERSITY OF SCIENCE AND TECHNOLOGY

.....***.....



Nguyen Thi Trang

**RESEARCH ON MULTIMODAL FEATURE FUSION
METHODS FOR SKIN LESION IMAGE
CLASSIFICATION**

**SUMMARY OF DOCTORAL DISSERTATION IN
COMPUTER SCIENCE**

Major: Computer Science

Code:9 48 01 01

Hanoi, 2026

**This dissertation was completed at the Graduate University of Science and Technology,
Vietnam Academy of Science and Technology**

Scientific supervisors:

1. Supervisor 1: Assoc. Prof. Dr. Nguyen Viet Anh, Institute of Information Technology, Vietnam Academy of Science and Technology
2. Supervisor 2: Assoc. Prof. Dr. Vu Van Hieu, Institute of Information Technology, Vietnam Academy of Science and Technology

Reviewer 1:

Reviewer 2:

The dissertation will be defended before the Graduate University-level Doctoral Dissertation Evaluation Committee convened at the Graduate University of Science and Technology, Vietnam Academy of Science and Technology, at on the day of,
.....

The dissertation is available at:

1. Library of the Graduate University of Science and Technology
2. National Library of Vietnam

INTRODUCTION

1. Urgency and research gaps

Classifying skin lesion images is meaningful in supporting early screening [1, 2, 3], but actual data vary widely in color, texture, shape, acquisition conditions and layer distribution. Danger classes often have few samples, many lesions from different classes are visually similar, while labeling requires experts. Craft features reflect color, texture and shape but depend on the design; Deep learning features have strong representation power but not all feature maps are useful. If the entire intermediate response is kept, the representation may contain noise and redundancy.

Image fusion, hand-crafted features, and clinical metadata help take advantage of complementary information [8, 9, 14], but direct fusion does not fully handle the differences in dimensionality, distribution, and confidence level of each source. Most models still process each image independently, without exploiting similarity relationships between samples. In semi-supervised learning, erroneous pseudo-labels can be reused and bias amplified, especially for rare classes.

The thesis identifies two consecutive gaps. First, there is a lack of a unified procedure for selecting feature maps according to appropriate entropy intervals, fusing them with handcrafted features, and balancing the representation. Second, there is a lack of a semi-supervised multimodal framework capable of constructing inter-sample relationships graphically and using only highly reliable pseudo-labels.

In terms of methodology, the thesis views each data source as a different observation angle of the same object. Photos provide visual morphological cues; Manually characterize domain knowledge encoding; metadata adds clinical context. Therefore, the problem is not just to increase the number of representation dimensions but to determine what information is worth keeping, how to normalize contributions between sources, and how to control errors when information is transferred between samples. This perspective governs the entire design from BlockNet to the semi-supervised graph model.

2. Objectives, objects and scope of research

The general goal is to develop multimodal feature fusion methods to improve the effectiveness of skin lesion image classification in conditions of large data variability, class imbalance, limited labels, and many sources of supporting information.

The first goal is to propose BlockNet, which selects feature maps according to the entropy range, supports balance in the feature space, and fuses deep learning features with handcrafted features. The second goal is to develop a graph-based semi-supervised multimodal learning framework, combining deep learning features, handcrafted features, and clinical metadata; exploit relationships between samples and unlabeled data through pseudo-labeling with confidence control.

The research object is the method of the organization, selection, fusion, and use of multimodal representations for classifying skin lesion images. The experimental scope includes VNCancer, HAM10000 and ISIC 2020. The thesis does not build a completely new backbone, does not conduct clinical trials and does not consider the model as an independent diagnostic tool.

The research method combines theoretical surveys, algorithmic modeling, computational experiments and quantitative analysis. Results were evaluated using Accuracy,

Precision, Recall, F1-score, AUC, Sensitivity, Specificity, confusion matrix and iterative validation suitable for imbalanced data.

The experiments were organized to answer three questions: does entropy selection eliminate less useful feedback; Does multimodal feature fusion improve discrimination compared to individual sources; and how securely unlabeled data can be exploited. When interpreting the results, the thesis prioritizes class balance indices instead of relying solely on Accuracy, and compares the confusion matrix and AUC to avoid biased conclusions by the majority class.

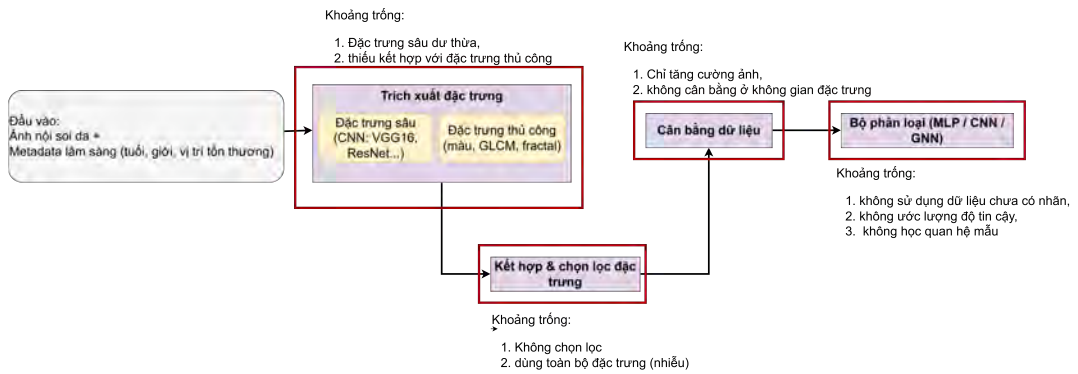
3. Contribution and structure of the thesis

The first contribution is BlockNet [CT3]. The model uses activation entropy to screen feature maps at many blocks of VGG16, convert the retained maps into vectors, fuse them with handcrafted features, and balance the data in the feature space. The contribution lies in the process of selective integration, fusion, and representation equalization, not in a new CNN backbone.

The second contribution is a multimodal semi-supervised learning framework on graphs [CT4]. Deep learning features, handcrafted features, and metadata are fused into a node representation; k nearest neighbor graph models inter-sample relationships; Predicted entropy is used to stratify confidence and select pseudo-labels before retraining.

Two contributions use entropy at two different levels: Chapter 2 calculates the entropy on the activation map to choose the representation; Chapter 3 calculates entropy on the predicted probability distribution to select unlabeled samples. Chapter 1 presents the theoretical basis; Chapter 2 presents BlockNet and experiments on VN-Cancer, HAM10000; Chapter 3 presents the multimodal model on graphs and experiments on ISIC 2020; The final section summarizes the results, limitations and future research directions.

Figure 1 summarizes the series of problems that the thesis addresses.



Hình 1: Research gap of skin lesion image classification system.

The research circuit therefore goes from intra-pattern representation to inter-pattern relations. Chapter 2 addresses the representational quality of each image before classification; Chapter 3 expands on data set structure, where a pattern is understood in relation to similar patterns. This division makes the two contributions clear and complementary: BlockNet reduces noise in the representation input, while the graph model and reliability mechanism reduce the risk when using relationships and predictive labels.

4. Research questions and overall approach

The thesis answers four linked questions: which feature map contains the appropriate amount of information; How feature modalities should be fused to exploit complementarity; Whether similarity relationships between samples can be modeled graphically; and what trust conditions must unlabeled data meet before reuse. These questions shift the focus from simply choosing a taxonomic architecture to organizing and controlling information throughout the entire learning process.

The approach is two-tiered. At the representation layer, activation entropy is calculated for each map, block-wise appropriate responses are retained and then fused with HSV, GLCM, and fractal features. At the data layer, the image representation combined with metadata forms nodes, k -NN graphs model inter-sample relationships, and GNNs transmit information. Predicted entropy is used in the semi-supervised phase to stratify confidence before creating the expanded training set.

The evaluation was carried out within the scope of methodological research. Comparisons between models must keep the same data split and training protocol. For imbalanced data, Macro F1, Macro AUC, Sensitivity and Specificity are considered along with Accuracy. Quantitative results indicate feasibility in the examined datasets, and do not replace clinical judgment. Constraints on metadata, domain bias, outliers, and label quality are used to determine the correct level of generalization of the conclusions.

The two contributions are connected by a common information-control principle. In Chapter 2, entropy operates before classification to retain feature maps whose activation distributions are informative for the target data. In Chapter 3, predictive entropy operates after classification to determine which unlabeled samples are sufficiently reliable for reuse. This distinction is important: activation entropy evaluates an intermediate representation, whereas predictive entropy evaluates uncertainty in the output distribution. The first mechanism controls what information enters the classifier; the second controls what information is allowed to return to the training set.

Accordingly, the thesis evaluates not only predictive performance but also the conditions under which each improvement is obtained. Feature-selection results are interpreted together with the retained representation and class-wise behavior. Graph-based results are interpreted together with graph construction, Trust-level distribution, and the number of pseudo-labeled samples actually used. This reporting perspective limits over-generalization and makes the two proposed pipelines reproducible within the stated experimental scope.

CHƯƠNG 1: OVERVIEW AND THEORETICAL BASIS

1.1. Multimodal Data and Feature Representations

Given the data set $D = \{(x_i, y_i)\}_{i=1}^N$, the model learns a mapping from the image or multimodal representation to one of C lesion classes. In case of metadata m_i , the pattern is represented by the pair (x_i, m_i) . Main difficulties include intraclass variability, interclass similarity, data imbalance, equipment differences, and lack of expert labels.

Hand-crafted features describe visually meaningful attributes [36, 37, 39, 41]: RGB/HSV histogram for color; GLCM, LBP, Gabor and wavelets for textures; HOG, boundary and fractal give shape or complexity. Deep learning features are extracted from convolutional layers [50, 51, 52], where each channel creates a feature map. The general deep representation can be written:

$$F_i^{(l)} = \{f_{i,1}^{(l)}, f_{i,2}^{(l)}, \dots, f_{i,K_l}^{(l)}\}, \quad (1.26)$$

Not every $F_k^{(l)}$ is useful. Some responses are nearly constant, some are too scattered or sensitive to noise. This is the basis for the thesis to evaluate the feature map before the fusion step instead of using the entire intermediate output.

In multimodal representation, an important requirement is to preserve supporting information but limit the dominance of high-dimensional sources. Normalization and projection to the common space are considered model-significant preprocessing steps, because they determine the distance between samples and directly affect the classifier or graph structure in later chapters. Therefore, the quality of feature fusion must be judged by both the final performance and the ability to explain the role of each modality.

1.2. Feature fusion and entropy

Multimodal fusion aims to create a common representation from complementary sources. With the deep learning feature $z_i^{(d)}$ and the handcrafted feature $z_i^{(h)}$, the concatenation takes the form:

$$x_i^{(dh)} = \mathcal{F}(x_i^{(d)}, x_i^{(h)}), \quad (1.28)$$

Given the prediction probability distribution $\mathbf{p} = (p_1, \dots, p_C)$, the prediction entropy is determined by:

$$H_{\text{pred}}(p_{\theta}(x)) = - \sum_{c=1}^C p_{\theta}(c|x) \ln p_{\theta}(c|x). \quad (1.33)$$

Low entropy indicates a more concentrated distribution, while high entropy indicates a more dispersed prediction. On the feature map, entropy is calculated from the normalized histogram of activation levels:

$$H_{\text{act}}(f) = - \sum_{r=1}^{K_f} p_r \log_2 p_r. \quad (1.34)$$

Two types of entropy are different in terms of object and purpose: feature map entropy used for representation screening in Chapter 2; Predicted entropy is used to control for erroneous pseudo-labels in Chapter 3. Entropy does not directly demonstrate class

discrimination ability, so any choice should be validated against classification performance.

Entropy in the thesis is not interpreted as an absolute measure of quality. A value that is too low may correspond to a nearly uniform response, whereas a value that is too high may reflect scattered or noisy activation. Therefore, the appropriate criterion is a block-dependent entropy interval instead of a global threshold. This approach retains maps with sufficient variability to characterize lesion structure, while reducing the number of redundant responses before merging.

1.3. Graph learning and semi-supervised learning

The graph $G = (V, E)$ represents each sample with a node [20, 21, 22], while edges describe similarity relationships. The adjacency matrix is determined from similarity and threshold or k nearest neighbor relationship:

$$A_{ij} = \text{sim}(x_i, x_j), \quad (1.36)$$

At each GNN layer, the node representation is updated from the node itself and its neighbors:

$$h_i^{(l+1)} = \sigma \left(\sum_{j \in \mathcal{N}(i)} W^{(l)} h_j^{(l)} \right), \quad (1.37)$$

Semi-supervised learning uses labeled and unlabeled sets simultaneously. The model is initially trained on labeled data, predicts unlabeled data, selects reliable predictions, and then retrain. The main risk is confirmation bias: false but confident erroneous pseudo-labels can skew the decision margin. Therefore, Chapter 3 only chooses the low-entropy group and does not consider pseudo-labels as reference labels.

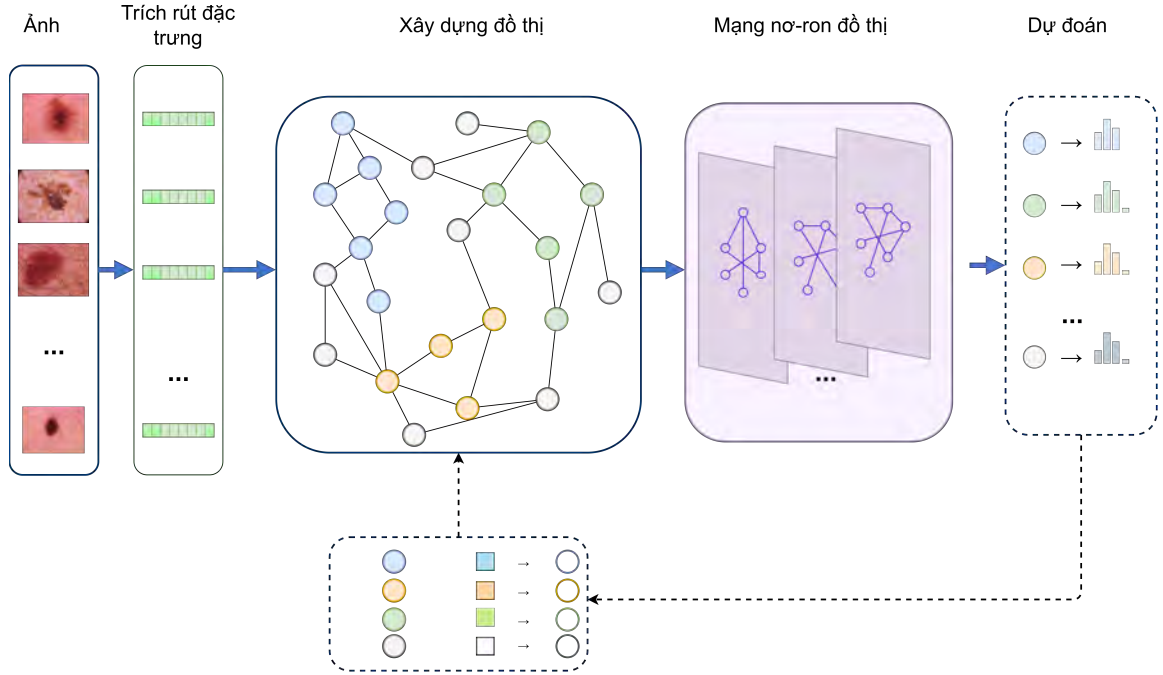
For graphs, the choice of node representation and the rules for creating edges are closely related. If the node representation does not accurately reflect clinical similarity, the k -NN graph may connect samples of different classes and propagate false information. Therefore, the thesis builds nodes after multimodal feature fusion, using cosine similarity and controlling the number of neighbors. Semi-supervised learning is performed only after the supervised model has generated a prediction distribution that is stable enough to assess reliability.

1.4. Basis for forming the methodology of the thesis

The overview reveals four alignment issues: input representation quality, heterogeneous source fusion, relationships between samples, and unlabeled data mining. Direct fusions do not guarantee noise removal or equalization of the contributions of each source. GNN can exploit inter-sample structure but depends on the quality of node and edge representation. Pseudolabeling expands the training data but must be accompanied by a reliability control mechanism.

From there, the thesis is organized in two directions. BlockNet processes the representation at the feature map level: extracting multiple blocks, calculating entropy, keeping responses in the right range, merging with handcrafted features, and equalizing the representation. The second direction processes data at the sample level: merging images with metadata, building a k -NN graph, learning GNN and selecting pseudo-labels according to predicted entropy.

Macro indicators are preferred because Accuracy can be dominated by the majority



Hình 1.7: Illustration of a learning pipeline on a graph for the problem of classifying skin lesion images. Image features and metadata are used to construct the graph in which each node represents a pattern and the edges reflect the degree of similarity between the patterns. GNN model propagates information between neighboring nodes to learn contextual representations, creating a basis for work exploit relationships between samples in semi-supervised learning settings.

class. Macro F1 and Macro AUC reflect balance between classes; Sensitivity and Specificity clarify the ability to identify malignant and benign samples. Results are interpreted within the scope of the experimental protocol and do not generalize to clinical effect.

Conclusion of Chapter 1

Chapter 1 establishes the foundation for the two contributions and clarifies the boundary between activation entropy and prediction entropy.

From the above bases, two roles of entropy are clearly distinguished. Activation entropy is a property of the intermediate map and serves feature selection; Predicted entropy is a property of the output probability distribution and serves to select the sample. Not identifying these two quantities helps avoid the inference that maps with high entropy necessarily produce uncertain predictions. They act at two different levels but implement the same principle of controlling information before moving on to the next learning step.

Analyzing the relationship between theoretical components

Deep learning features and handcrafted features are different in origin, size, and meaning. Deep learning features are optimized from data and are capable of describing complex structures, while HSV, GLCM, and fractals provide quantities associated with color, texture, and geometric complexity. If only direct coupling is used, high-dimensional branches can dominate the distance and gradient functions. Therefore, normalization, selection, and projection are not just technical operations but conditions for sources to contribute on a compatible scale.

Entropy was chosen because it can describe the dispersion of information without needing labels at the map level. However, entropy does not directly indicate whether the response belongs to the lesion structure or the image background. The thesis partly overcomes this by determining the appropriate range for each block and verifying the classification performance. This approach creates a bridge between the unsupervised benchmark at the representation layer and the supervised target at the output, while preserving the ability to trace the number of maps selected at each layer.

In graph models, the congruence assumption holds that samples close to each other in the representation space tend to share labels. The assumption is only reliable when the unified representation eliminates noise and encodes supporting information. Message passing can increase local consistency but also runs the risk of over-smoothing or propagating errors across edges to different layers. Therefore, the number of neighbors, GNN depth and pseudo-label selection mechanism must be evaluated together.

From there, the theoretical basis leads to three design principles. First, select the information before merging. Second, build inter-sample relationships after having a multi-modal representation. Third, only reuse predictions when the output distribution reaches a specified confidence level. These three principles form a logical flow throughout the two empirical contributions of the thesis.

Evaluation conditions and risk of bias

Deviations can arise from imaging equipment, lighting conditions, preprocessing, layer ratios, and patient segmentation. If images of the same patient are present in both the training set and the test set, the model can remember individual characteristics and generate optimistic results. Therefore, every step of choosing the entropy threshold, graph hyperparameters, and pseudo-labeling conditions is based solely on the training or validation set; The test set must be kept independent until the final evaluation step.

Class imbalance requires distinguishing overall performance from class-specific performance. Accuracy can increase when the model prioritizes the popular class while the recall of the dangerous class remains low. Macro F1 assigns equal weight to classes; AUC evaluates the ability to rank at multiple thresholds; Sensitivity and Specificity reflect the two sides of the decision. Confusion matrices help identify pairs of classes that are often confused and relate errors to the morphology or distribution of the data.

Reproducibility requires reporting the extraction and normalization settings, selected channels, random seeds, validation folds, and stopping criteria. Graph experiments must also specify edge construction, k , symmetry, and GNN depth; pseudo-labeling experiments must state the entropy thresholds and the number of samples retained at each Trust level. These controls prevent an observed improvement from being incorrectly attributed to a single component when several settings change simultaneously.

Statistical and practical significance must be distinguished through class-wise metrics and variation across runs. Attention weights indicate model contribution but not causality; low entropy indicates a concentrated distribution but not necessarily a correct prediction; and graph edges indicate similarity only in the selected representation. These interpretation boundaries define the methodological scope of Chapters 2 and 3.

CHƯƠNG 2: ENTROPY SELECTION-BASED MULTIMODAL FEATURE FUSION FOR SKIN LESION IMAGE CLASSIFICATION

2.1. Statement of problems and ideas

Given a set of labeled images, the goal is to learn a classifier from the unified representation but without including the entire feature map in the model. From the feature maps produced by CNN blocks and the handcrafted feature vectors, BlockNet selects informative maps according to entropy, converts them into vectors, performs feature fusion, and balances the data before training.

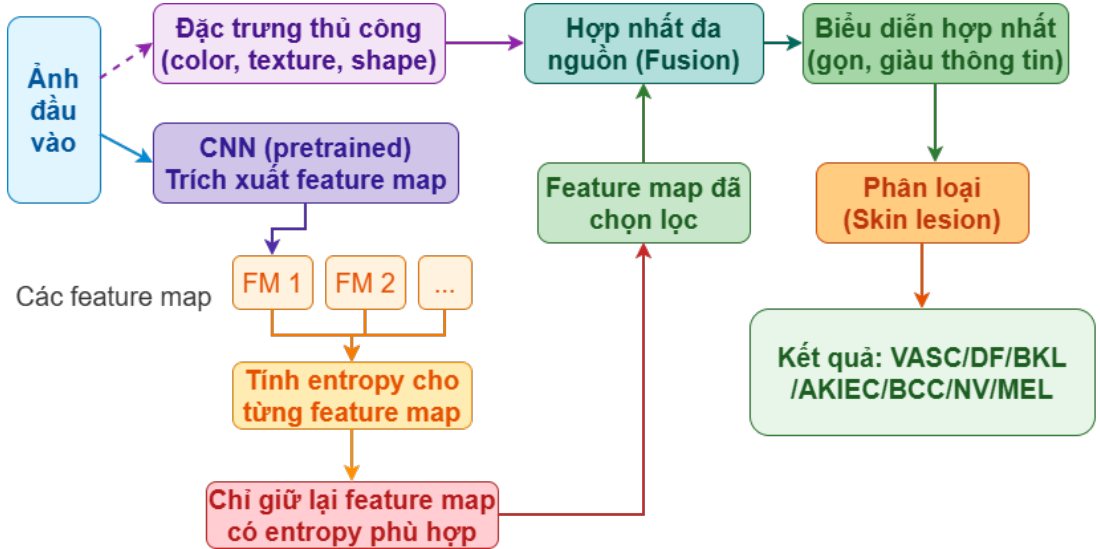
The entropy of a map F is calculated from the normalized histogram distribution:

$$H_{\text{act}}(F_m^{(b)}(x)) = - \sum_{r=1}^{K_m} p_r \log_2 p_r, \quad (2.16)$$

This value satisfies:

$$0 \leq H_{\text{act}}(F_m^{(b)}(x)) \leq \log_2 K_m \leq \log_2 N_m. \quad (2.17)$$

The entropy range is determined individually for each block. Maps with too little variability often contain poor responses; maps that are too scattered may contain noise. The method does not assume that the entropy is as high as possible, but rather chooses a range that is validated through classification.



Hình 2.1: Proposed method idea: feature map selection based on entropy and multi-modal feature fusion.

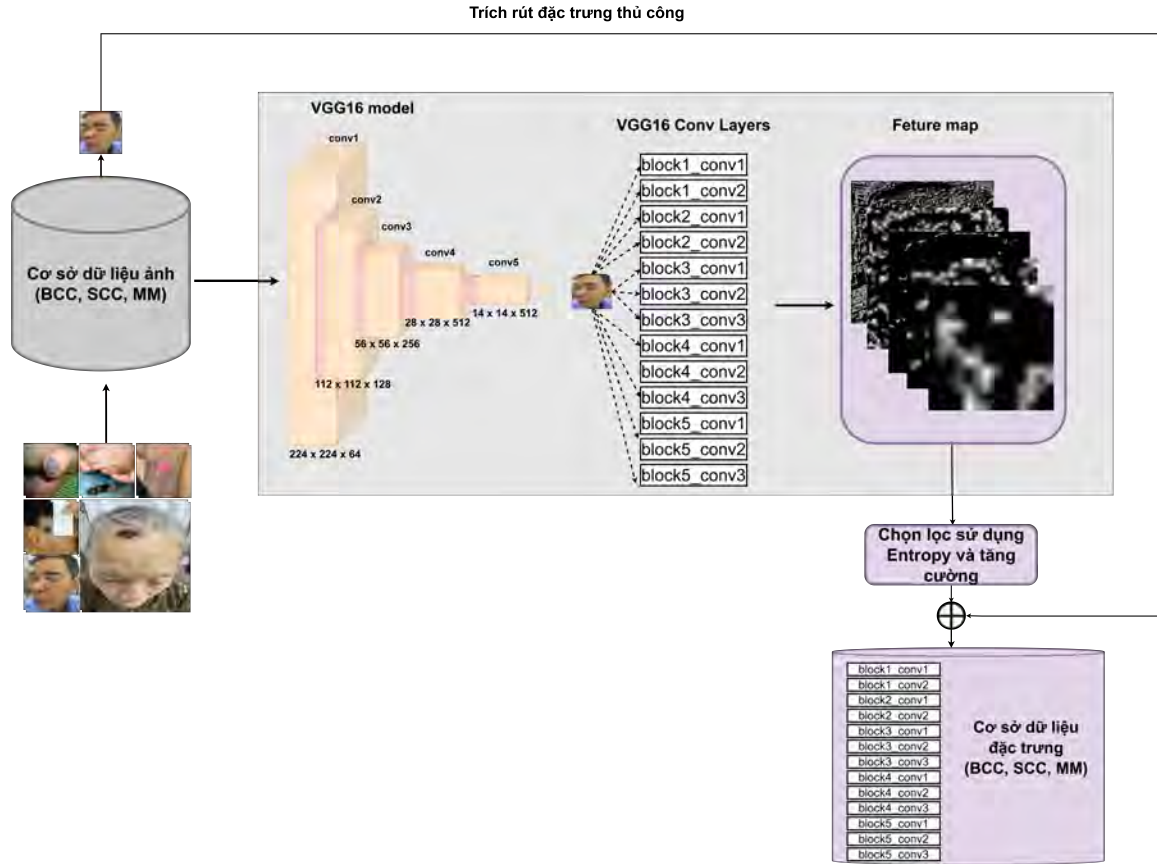
The BlockNet input combines preprocessed images with handcrafted features; Selection is performed separately at each block to match the level of abstraction of the response.

2.2. Overview architecture of BlockNet

BlockNet consists of five blocks: feature map extraction from VGG16; calculate entropy and select maps; deep vector generation; handcrafted feature extraction; feature

fusion and balancing in the feature space before the MLP classifier. The backbone is kept fixed in principle, thus contributing centrally to the organization of the performance.

Shallow blocks preserve local color, boundaries and texture; Deep blocks describe more semantic structure but may lose detail. Exploiting multiple blocks allows individual evaluation of each level of performance. Selected maps were vectorized and normalized before merging with the HSV, GLCM and fractal histograms.



Hình 2.3: Overview architecture of the BlockNet model.

The architecture is implemented in series: map extraction, entropy calculation, validation set determination, vectorization, concatenation with handcrafted features, representational equalization, and classification. Setting the post-fusion balancing ensures that the synthetic sample is created in the same space used by the classifier. This design also avoids direct imaging artifacts, which can produce inappropriate lesion structures and make it difficult to control for clinical significance.

In each block, the output has a different size and number of channels, so the vectorization step must preserve the processing order and create a representation of a certain size. The depth vectors after selection are normalized before merging, so that features with large value ranges do not automatically dominate the distance. The manual branch includes HSV color, GLCM texture, and fractal quantities, corresponding to three groups of signs commonly observed when characterizing lesions. The final unified representation is fed into the same classifier so that comparisons between profiles reflect changes in the selection step, rather than concurrent changes in the prediction architecture.

2.3. Determine the entropy range and select the map

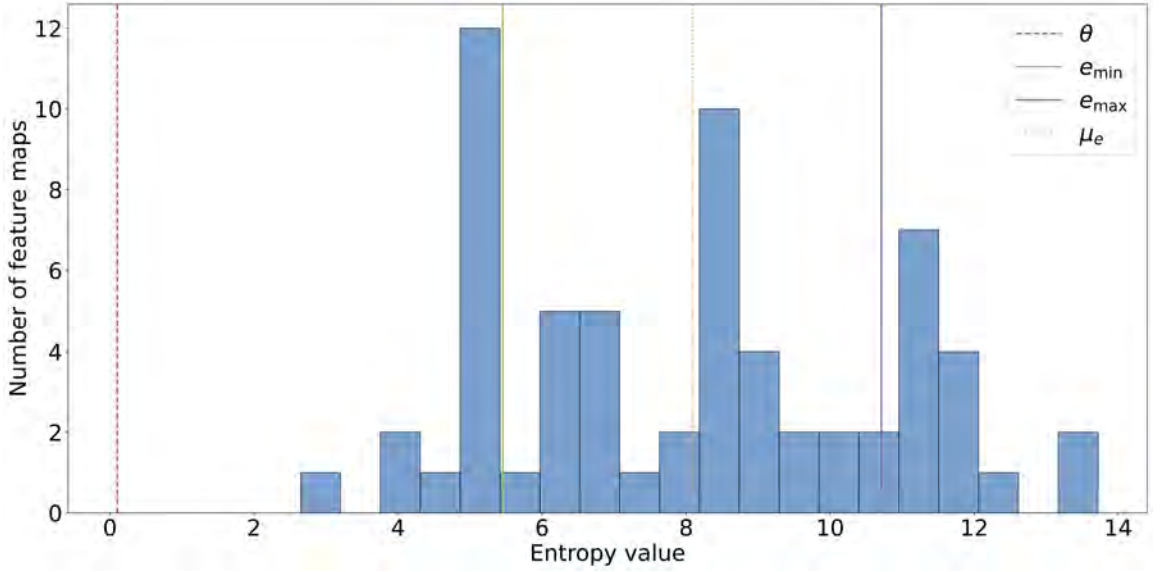
For block b , the valid feature-map set is determined by two thresholds $E_{\min}^{(b)}$ and $E_{\max}^{(b)}$:

$$E^{(b)} = \left\{ H(F_m^{(b)}(x_i)) \mid H(F_m^{(b)}(x_i)) > \theta^{(b)}, i = 1, 2, \dots, N, m = 1, 2, \dots, M_b \right\}, \quad (2.23)$$

The conditions for selecting a map are:

$$e_{\min}^{(b)} \leq H(F_m^{(b)}(x)) \leq e_{\max}^{(b)}. \quad (2.26)$$

Thresholds are not common to all blocks because the spatial size, number of channels, and activation distribution vary with depth. The entropy histogram provides quantitative observations to eliminate near-constant responses and over-dispersed responses.



Hình 2.8: Entropy distribution of feature maps at layer `block1_conv1` and thresholds used to determine the appropriate entropy range.

The entropy range is determined from the experimental distribution of each block. This is consistent with the fact that the number of channels, spatial size, and activation sparsity vary with network depth. A map is kept only when the entropy is within the selected domain; The resulting set may vary between blocks and between datasets. This mechanism relaxes the assumption that the same threshold can represent all layers and all image domains.

The histogram is normalized to a probability distribution before calculating entropy, so the measurement does not directly depend on the number of pixels in the map. The lower and upper thresholds simultaneously type near-constant responses and over-dispersed responses. Surveying by block also allows to identify which layer contributes the most valid maps. In the experiment, the threshold configuration is selected on the training data and kept constant during evaluation, ensuring that the test results do not participate in the feature selection decision.

2.4. Feature Fusion and Balancing in the Feature Space

After selection, the valid feature maps are aggregated into a deep representation. The final vector of sample x_i is the concatenation of the deep representation and the handcrafted features:

$$z_i^{(b)} = \psi\left(z_i^{(d,b)}, z_i^{(h)}\right), \quad (2.28)$$

To reduce the effect of class imbalance, the number of representations to be added to class k is determined by the largest class. The multiplier and the number of augmented samples are:

$$m_k^{(b)} = \left\lceil \frac{n_{\max}^{(b)}}{n_k^{(b)}} \right\rceil. \quad (2.29)$$

$$n_k'^{(b)} = n_k^{(b)} \cdot m_k^{(b)}. \quad (2.30)$$

Equalization is performed after fusion, in the feature space, to create additional representations for the minority class without changing the original image. Cross-entropy optimal classifier on the extended set. The thesis notes that the specific impact of balance needs to be further examined with class-weighted loss, focal loss [115] and SMOTE [116].

After concatenation, each sample retains a unique label and the entire balancing operation is performed in the training set. The test set is not used to determine the number of additional samples or parameter tuning. This is a necessary condition to avoid data leaks. Cross-entropy loss function optimal classifier; Options such as layer weights, focal loss or SMOTE are considered necessary controls when analyzing the contribution of the balancing step separately.

The coefficient $m_k^{(b)}$ determines the relative deficiency of class k compared to the largest class, while $n_k'^{(b)}$ represents the number of additional elements that need to be created. The separation of the two quantities helps distinguish the multiplicative ratio from the actual scale of addition. The synthetic samples only serve to adjust the training distribution; The number of real samples and the structure of the evaluation set are kept the same so that the final index reflects the original observation data.

2.5. Experimental data and setup

VNCancer [119] includes three layers BCC, MM and SCC, reflecting data acquired in the country with differences in equipment and shooting conditions. HAM10000 [3] consists of seven layers, with strong imbalance, in which the NV layer accounts for a large proportion. The two datasets allow evaluation of the method in a three-class and multi-class context.

Experiment using stratified five-fold cross-validation. The compared configurations use the same classifier and training protocol so that differences primarily reflect performance quality. Reporting metrics include Accuracy, Macro Precision, Macro Recall, Macro F1 and Macro AUC; Classwise analysis using ROC and confusion matrix.

VNCancer reflects acquisition conditions close to conventional images, while HAM10000 mainly consists of dermoscopic images and has seven layers with obvious imbalance. The two datasets provide a complementary test of image domain and number of classes.

Bảng 2.2: Skin lesion classes and quantity distribution in the VNCancer dataset

Symbol	English name	Vietnamese name	Quantity
BCC	Basal Cell Carcinoma	Basal Cell Carcinoma	2719
MM	Malignant Melanoma	Malignant Melanoma	314
SCC	Squamous Cell Carcinoma	Squamous Cell Carcinoma	833
Total number of images			3866

Bảng 2.3: Skin lesion classes and quantity distribution in the HAM10000 dataset

Symbol	English name	Vietnamese name	Quantity
MEL	Melanoma	Melanoma	1113
NV	Melanocytic nevus	Melanocytic mole	6705
BCC	Basal cell carcinoma	Basal cell carcinoma	514
AKIEC	Actinic keratoses and intraepithelial carcinoma	Actinic keratosis and intraepidermal carcinoma	327
BKL	Benign keratosis-like lesions	Benign keratosis lesions	1099
DF	Dermatofibroma	Dermatofibroma	115
VASC	Vascular lesions	Vascular lesions	142
Total number of images			10015

Note: Dermoscopy data, layers are sorted by decreasing number.

Preprocessing, batching, and hyperparameter selection must be performed consistently within each protocol; The indexes between the two sets are only used to identify trends and not directly compare absolute difficulty.

2.6. Effects of entropy selection

On VNCancer, the best configuration without using entropy reaches Accuracy 80.97% and F1-score 82.11%. When applying entropy selection, the block2_conv2 configuration achieves Accuracy 88.75%, F1-score 88.16% and Specificity 93.42%. The corresponding increase is 7.78 and 6.05 percentage points.

In deep blocks, the improvement is more obvious: block4_conv3 increases Accuracy from 49.73% to 86.85%. The results show that the post-screening response set is more suitable for the classifier in the considered configuration, but do not prove that individual entropy is a direct measure of usefulness.

Bảng 2.7: Results on the VNCancer dataset when using entropy selection

Block	Accuracy	Precision	F1-score	Specificity
midrule block1_conv2	88.41	88.02	87.58	93.12
block2_conv2	88.75	88.40	88.16	93.42
block3_conv1	88.24	87.95	87.57	93.05
block4_conv3	86.85	86.44	86.04	92.63
bottomrule				

VNCancer results show that map selection affects Accuracy, Precision and F1-score

simultaneously. The biggest improvement appears in the right configuration, while not every block increases equally. This observation favors distributional selection rather than keeping the entire response. However, the results must be understood as an impact of the whole process, since the hand-crafted feature, the balance, and the classifier jointly participate in generating the final prediction.

Five-fold cross-validation was used to reduce dependence on a single split of the data. The mean describes the overall trend, while the standard deviation shows the stability between folds. When two configurations have close means, dispersion and results on the minority class should be considered before selection. This way of reporting is more appropriate than just stating the configuration that reached the maximum score in one run.

2.7. Multimodal Feature Fusion Results

After selecting the map, BlockNet fuses the deep representation with handcrafted features. On VNCancer, the full configuration outperforms unimodal representations within the same protocol. The results confirm the complementarity between the CNN response and the domain description, and show that the organization of the representation is more important than simple concatenation.

On HAM10000, entropy selection improves the best F1-score from 63.78% to 65.00%; The increase is smaller than VNCancer due to the larger number of classes, strong imbalanced distribution, and high morphological similarity. ROC analysis noted that VASC AUC increased from 0.8681 to 0.9179, DF from 0.8294 to 0.8440, AKIEC from 0.7587 to 0.7727 and NV from 0.8409 to 0.8465.

Bảng 2.10: Performance comparison on the VNCancer dataset using 5-fold cross-validation. Results are presented as mean $\pm$ standard deviation (%).

Model	Accuracy	Precision	F1-score	Specificity
VGG16	92.19 $\pm$ 0.48	92.16 $\pm$ 0.49	92.15 $\pm$ 0.47	89.78 $\pm$ 1.54
ResNet101	93.59 $\pm$ 0.78	93.56 $\pm$ 0.71	93.53 $\pm$ 0.76	91.13 $\pm$ 0.44
EfficientNetB7	92.55 $\pm$ 0.54	92.65 $\pm$ 0.54	92.57 $\pm$ 0.54	91.15 $\pm$ 1.64
BlockNet–block1_conv1 bottomrule	95.58 $\pm$ 0.69	95.59 $\pm$ 0.69	95.59 $\pm$ 0.69	97.53 $\pm$ 0.37

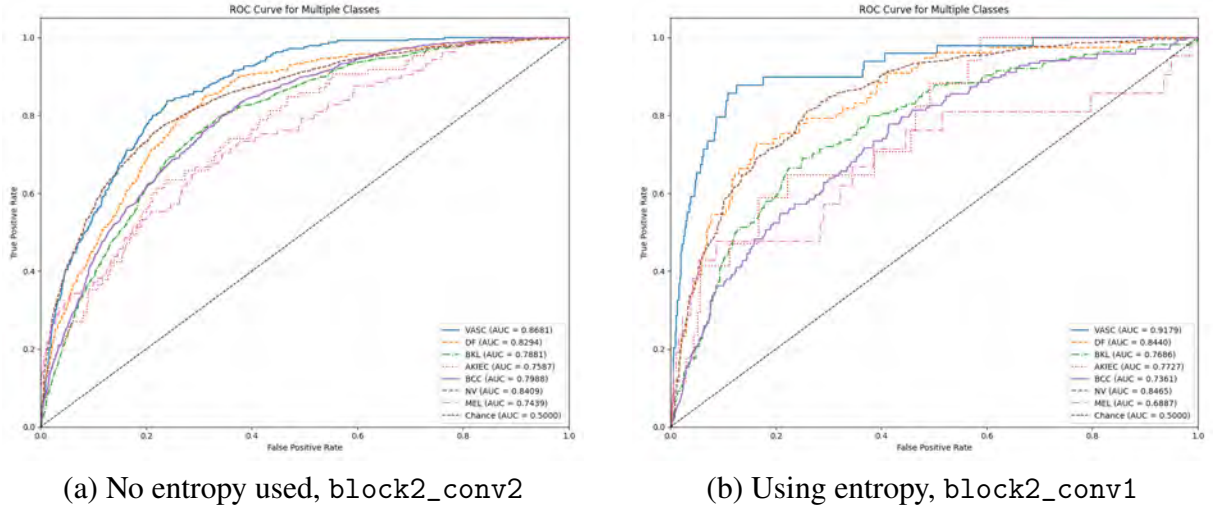
In HAM10000, the smaller difference reflects the multi-layer problem and the imbalance is more severe. AUC by class and confusion matrix show that rare classes are still difficult to identify despite high overall Accuracy. Therefore, BlockNet’s effectiveness should not be described by a single metric. A steady improvement across multiple measurements is more valuable than a large increase that only occurs in one configuration or class.

ROC curves before and after selection show changes in sample ranking ability at multiple decision thresholds. With multi-class data, AUC is considered class-wise and at the aggregate level; This helps detect cases where the common index obscures the rare class. Unfavorable results were retained in the analysis to identify method limitations and avoid reporting the best results.

2.8. Discussion of BlockNet results

First, entropy selection clearly changes the performance quality, especially at deep blocks. Second, the performance is not uniform between the two data sets, suggesting

that the amount of entropy must be tied to the block and data distribution. Third, hand-crafted features still complement CNN features in cases where color, texture, and shape have domain meaning. Fourth, balance in the feature space helps reduce the dominance of the large class but has not been completely separated from other components by exclusion experiments.



Hình 2.15: Comparison of ROC on the HAM10000 dataset before and after using entropy selection.

Conclusion of Chapter 2

BlockNet forms a representation selection, fusion, and equalization process. The results on VNCancer and HAM10000 provide evidence for the hypothesis that it is not necessary to use the entire feature map to achieve better representation.

The meta-analysis shows that entropy selection acts as an interpretable representational regulation mechanism. A reduction in the number of maps does not mean a loss of information if the eliminated responses are mostly homogeneous or less usefully dispersed. The remaining limitation is that the entropy range is experimental and has not been optimized with the loss function. This is the basis for studying the soft selection mechanism or threshold learning in the next models.

In terms of reproducibility, each configuration needs to store the backbone, blocks, entropy range, number of maps kept, normalization method, classifier, and random seed. Full reporting of these components allows distinguishing improvements due to selection from improvements due to changes in training settings. This is also a condition to expand BlockNet to another backbone or another image domain without defaulting to maintaining the threshold searched on the current data.

Additional analysis of BlockNet design and results

BlockNet is designed to be modular so that the impact of each block can be observed. For each image, VGG16 creates multiple maps at different depths. The histogram of activation values is normalized, then the entropy is calculated for each channel. The set of valid maps is not fixed in number but depends on the block's entropy distribution. This mechanism avoids the assumption that all backbone channels contribute equally to the skin damage problem.

After selection, vectorization converts the spatial response into a representation

suitable for feature fusion. HSV features support the description of color differences; GLCM reflects gray level and texture relationships; fractals add information about complexity. Groups were standardized before fusion to limit scale bias. Class balancing is performed on the fused representation of the training set, does not affect the test set, and does not create synthetic lesion images.

The results on VNCancer show that the appropriate selection configuration can simultaneously increase Accuracy and F1-score. This is notable because F1-score reflects the balance between Precision and Recall, being more suitable than Accuracy when the number of samples between classes varies. However, the improvement is not uniform in every block. Shallow layers easily retain color and edge responses, while deep layers carry semantic information but may contain specialized channels that are not suitable for the target data.

On HAM10000, the seven-layer problem clarifies the limits of the composite index. Dominant NV classes can have high Accuracy while rare classes still have low Recall. Therefore, the thesis further compares AUC by class, F1-score and confusion matrix. The pre- and post-selection ROC curves allow observing ranking likelihood changes at multiple thresholds, rather than depending on a single decision threshold.

The comparison needs to keep the same data split, preprocessing, classifier, and training schedule. If you simultaneously change the backbone or hyperparameters, the difference cannot be attributed to entropy. Therefore, the appropriate conclusion is that the entire BlockNet pipeline generates more efficient representations in the validated configurations; The individual contribution of each component needs to be further quantified by controlled ablation.

The scalability of the method lies in the fact that the entropy criterion does not depend on a specific disease class. However, the amount of entropy should not be transferred as is to the new backbone or image domain. Every change in map size, activation function, and acquisition condition can shift the distribution. The process of recalibration on the training set and external evaluation is a mandatory requirement before wider application.

A direct improvement is to build a subtraction experiment according to the component matrix: base backbone, entropy selection, handcrafted features, and representational equalization. Each configuration needs to be iterated with the same seed and folds, then report the difference along with the confidence interval. This design will identify the improvement due to each component, examine the interaction between selection and trade-off, and clarify the case where entropy is most beneficial for the rare class.

Implications for Representation Selection

Performance is not proportional to the number of retained maps: removing homogeneous, noisy, or redundant responses can produce a smaller yet more discriminative representation. Because shallow and deep blocks encode different information, their entropy ranges must be calibrated separately on the training set and validated without using the test set. Reports should include the retained-map count, representation dimension, and computation time so that performance gains can be interpreted together with their computational cost.

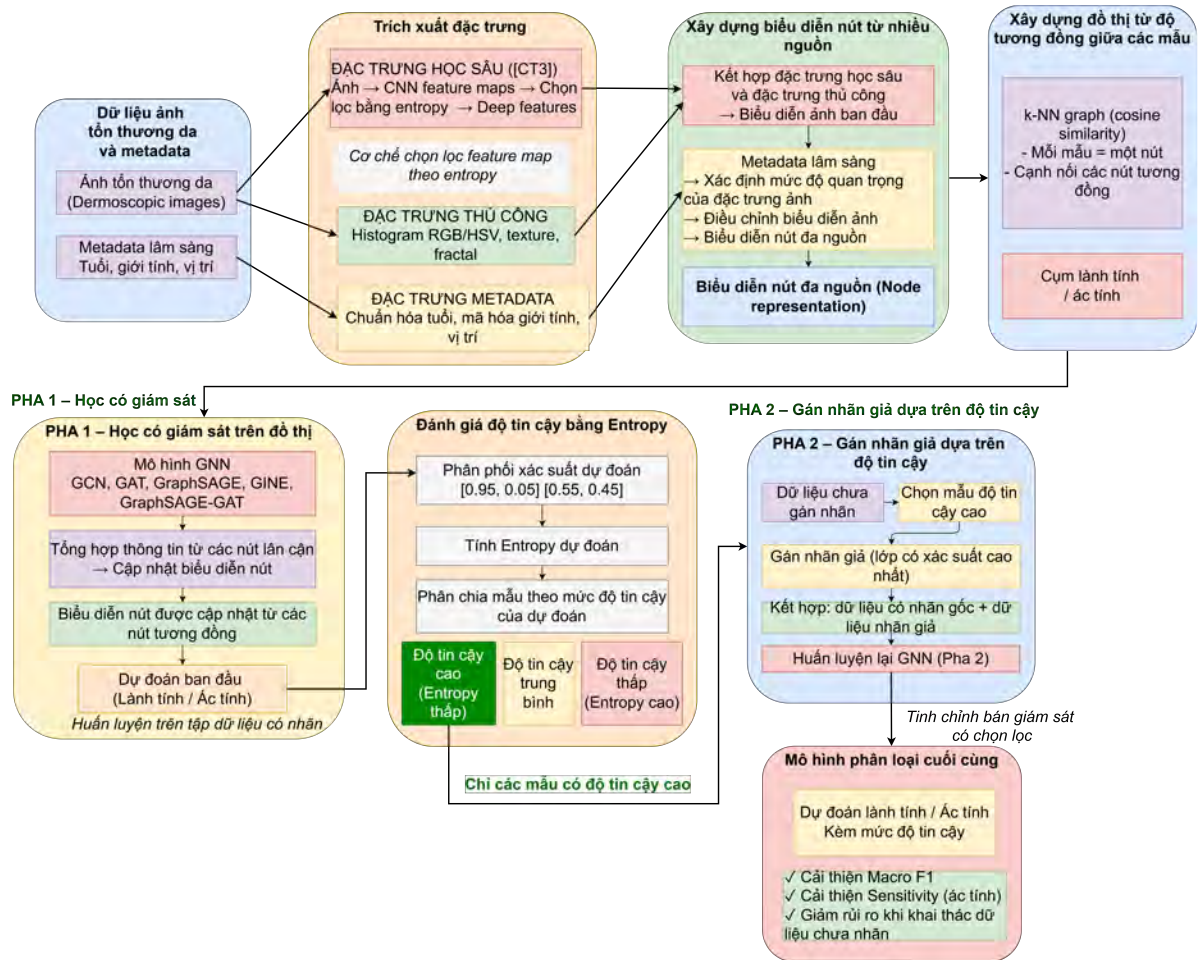
CHƯƠNG 3: GRAPH-BASED SEMI-SUPERVISED MULTIMODAL LEARNING

3.1. Overview problems and models

The labeled volume D_L and the unlabeled volume D_U contain images and clinical metadata. The goal is to learn a unified representation, construct an intersample graph, and controllably exploit D_U . The general representation of the pattern is:

$$\mathbf{x}_i^{img} = [\mathbf{x}_i^{deep}; \mathbf{x}_i^{hand}], \mathbf{h}_i = \Phi(\mathbf{x}_i^{img}, \mathbf{x}_i^{meta}). \quad (3.4)$$

The graph $G = (V, E)$ is constructed from the unified representation; Each node is a sample, edges connect samples that are close together according to cosine. The process includes phase 1 of supervised GNN training and phase 2 of calculating entropy, choosing highly reliable pseudo-labels and then retraining.



Hình 3.4: Overview of the proposed model.

Unlike Chapter 2, each sample in Chapter 3 is not classified independently but becomes a node in the graph. The node representation must combine image signatures, handcrafted features, and metadata before establishing neighbors. This design allows information from similar samples to support each other, but at the same time requires control of false edges and the propagation of unreliable predictions.

The process consists of two phases with clear boundaries. The first phase uses only the labeled set to learn the encoder, fusion module, and graph classifier. The best model of this phase is used for inference on the unlabeled set. The second phase calculates the entropy from the predicted distribution, stratifies the sample by confidence, and adds only the High Trust group to the training set. The test set is kept independent in both phases. This organization helps evaluate the improvement due to additional data without introducing test information into the pseudo-label selection process.

3.2. Attention-Based Fusion of Image Features and Metadata

Deep learning features are taken from the vector after pooling of VGG19, DenseNet121 and MobileNet [50, 52, 121]. Handcrafted features include RGB histogram, HSV histogram and fractal. The two visual groups are concatenated into z_i^{img} ; metadata including age, gender and lesion location are standardized and encrypted.

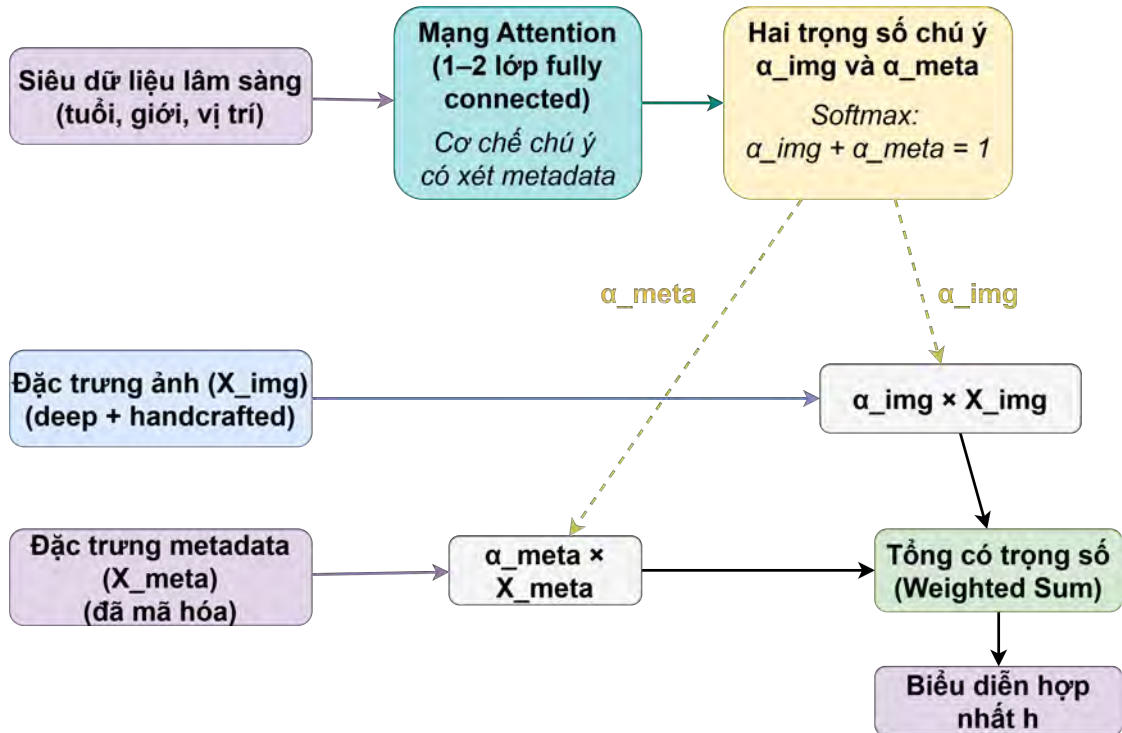
The two attention weights for image branch and metadata are calculated by:

$$\left[\alpha_i^{img}, \alpha_i^{meta} \right] = \text{softmax} \left(\mathbf{W}_a \mathbf{x}_i^{meta} + \mathbf{b}_a \right), \quad (3.14)$$

The representation that considers metadata is:

$$\mathbf{h}_i = \alpha_i^{img} \mathbf{x}_i^{img} + \alpha_i^{meta} \mathbf{x}_i^{meta}, \quad (3.16)$$

This mechanism does not generate three independent weights for deep, handcrafted, and metadata features; deep and handcrafted features are fused into the image branch before incorporating metadata.



Hình 3.6: Attention-based multimodal feature fusion incorporating metadata.

The two-branch attention mechanism learns relative weights for the image branch and the metadata branch. Metadata is not just concatenated at the end but participates in adjusting the unified representation. When data fields are missing or have low con-

confidence, the learned weights can reduce the contribution of the corresponding source. However, attention does not automatically equate to causal explanation; it only reflects the level of contribution in the model integration.

Vectors from multiple backbones are projected to compatible dimensions before feature fusion. The use of multiple architectures aims to capture different representational biases, rather than directly adding up all unnormalized outputs. Metadata is encoded by variable type and concatenated after the step of handling missing values. Thanks to this, attention weights are learned on a normalized representation and can vary from sample to sample instead of being fixed for the entire dataset.

3.3. Graph construction and GNN training

The neighbor set k -NN of node i is determined from cosine similarity:

$$\mathcal{N}_5(v_i) = \text{TopK}_{j \neq i}(s_{ij}, 5), \quad (3.18)$$

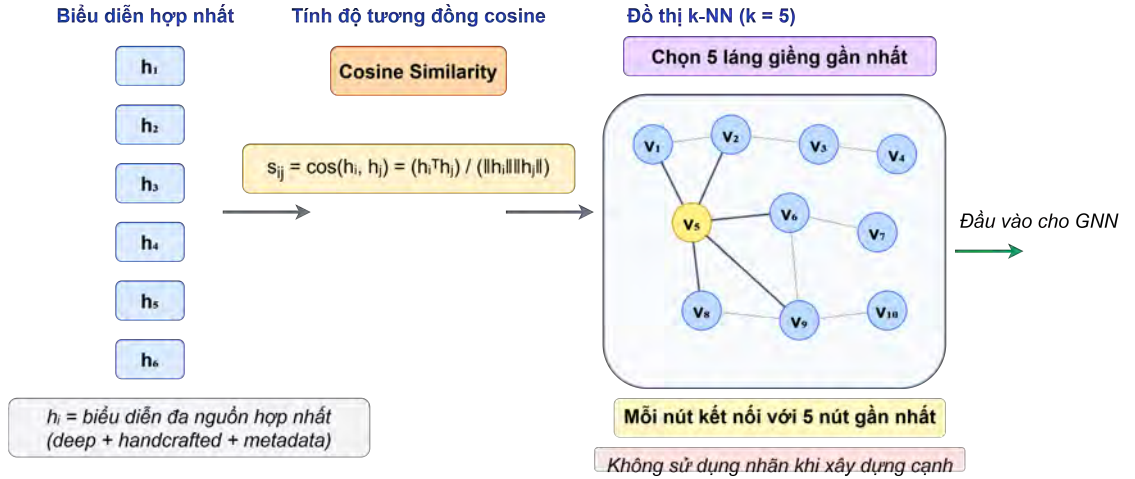
The edge set is created by:

$$E = \{(v_i, v_j) \mid v_j \in \mathcal{N}_5(v_i)\}, \quad (3.19)$$

At each layer, the node aggregates messages from neighbors and updates the representation:

$$\mathbf{h}_i^{(k+1)} = \text{UPDATE}^{(k)}\left(\mathbf{h}_i^{(k)}, \text{AGGREGATE}^{(k)}\left(\left\{\mathbf{h}_j^{(k)} : v_j \in \mathcal{N}_5(v_i)\right\}\right)\right), \quad (3.20)$$

The GCN, GAT, GraphSAGE, GINE and GraphSAGE–GAT architectures are investigated. Phase 1 optimizes cross-entropy on labeled nodes; The quality depends simultaneously on the node representation, similarity measure and k .



Hình 3.8: K -NN graph with $k = 5$ built from a unified representation based on cosine similarity.

The k -NN graph is generated from the unified representation and cosine similarity. A k value that is too small can cause the graph to become fragmented; Too large a value can connect regions of different layers and cause excessive smoothing. The message passing process aggregates neighbor information across multiple layers, then the node classifier creates a probability distribution. Therefore, the choice of k , representation dimension and GNN depth need to be considered simultaneously.

The edge set is constructed without using the label of the destination node, thereby

limiting information leakage. When the neighbor relationship is not symmetrical, edges can be symmetrical to ensure bidirectional message transmission. GCN, GAT, GraphSAGE and GINE variants [20, 21, 22, 91] were compared to evaluate the aggregation mechanism; The difference between them reflects how to normalize, sample, or weight neighboring information.

3.4. Confidence-based pseudo-labeling

With prediction probability p_{jc} of unlabeled sample x_j , the prediction entropy is:

$$H(\hat{\mathbf{y}}_i) = - \sum_{c=1}^C \hat{y}_{i,c} \log(\hat{y}_{i,c}), \quad (3.24)$$

Two thresholds α_1, α_2 divide the sample into High, Medium and Low Trust:

$$T(\hat{\mathbf{y}}_i) = \begin{cases} \text{High,} & H(\hat{\mathbf{y}}_i) \leq \alpha_1, \\ \text{Medium,} & \alpha_1 < H(\hat{\mathbf{y}}_i) \leq \alpha_2, \\ \text{Low,} & H(\hat{\mathbf{y}}_i) > \alpha_2. \end{cases} \quad (3.26)$$

Only High Trust samples that satisfy the entropy condition are labeled pseudo:

$$\tilde{y}_j = \arg \max_c \hat{y}_{j,c}, \quad \text{if } H(\hat{\mathbf{y}}_j) \leq \alpha_1. \quad (3.28)$$

The extended training set is:

$$D_{\text{new}} = D_L \cup D_{\text{pseudo}}^{\text{High}}, \quad (3.29)$$

This selection method reduces the risk of including unreliable predictions in re-training but does not guarantee that all High Trust labels are correct.

Predicted entropy measures the dispersion of probabilities between classes. Samples with low entropy are classified as High Trust and are newly eligible to receive a pseudo-label; The Medium and Low Trust groups were kept out of the retraining set. The selection threshold helps trade off between quantity and quality of pseudo-labels. If the threshold is too loose, the error is amplified; If it's too tight, the amount of additional data won't be enough to make a significant change.

Pseudolabels are taken from the class with maximum probability but are accepted only after the entropy condition. These two criteria complement each other: argmax determines the class, while entropy evaluates the concentration of the entire distribution. A relatively high maximum probability can still be associated with a competitive distribution among many classes; Entropy helps identify that situation better than just using a single probability threshold.

3.5. Supervised phase setup and results

ISIC 2020 is organized into labeled and unlabeled volumes; The main problem is classifying benign and malignant under imbalanced conditions. Missing metadata handled, z-score normalized continuous variable, one-hot categorical variable. Feature groups are selected using Chi-square in the Chapter 3 setup itself.

Phase 1 results show that traditional machine learning models have high Specificity but low Sensitivity in some configurations. GNNs integrate metadata that changes the balance between the two metrics; The efficiency depends on the graph structure. So the evaluation uses Macro F1, Macro Recall, Macro AUC, Sensitivity and Specificity

instead of just Accuracy.

Bảng 3.10: Comparison of supervised training phase results between traditional machine learning models and GNN-based architectures on the ISIC 2020 dataset.

Model	Sensitivity	Specificity	Macro F1	Macro AUC
midrule Logistic Regression	53.8 $\pm$ 2.9	99.1 $\pm$ 0.1	75.6 $\pm$ 1.2	92.6 $\pm$ 1.1
XGBoost	51.3 $\pm$ 3.7	99.7 $\pm$ 0.1	80.6 $\pm$ 1.6	96.0 $\pm$ 0.8
GCN	56.7 $\pm$ 20.4	70.8 $\pm$ 8.4	44.2 $\pm$ 2.5	71.2 $\pm$ 3.6
GINE	76.0 $\pm$ 22.7	61.2 $\pm$ 13.7	40.8 $\pm$ 5.0	75.9 $\pm$ 5.2
GraphSAGE-GAT	70.2 $\pm$ 17.7	65.8 $\pm$ 6.9	42.8 $\pm$ 2.1	74.8 $\pm$ 5.7
bottomrule				

Supervised phase creates a baseline and provides a pseudo-label generation model. Comparison between traditional machine learning models and GNN architectures shows the value of inter-sample relations when node representation is good enough. The Sensitivity, Specificity, Macro F1, and Macro AUC metrics are reported together to reflect both the ability to detect positive classes and control mispredictions in imbalanced data.

Traditional models use the same input representation to create a fair counterpoint at the classifier level. With GNN, the graph structure adds neighborhood information but also increases inter-sample dependencies; Therefore, data division and graph construction must be controlled. The phase 1 results are the required reference point to determine whether the pseudo-label phase produces consistent improvement or changes solely due to training fluctuations.

3.6. Results of the pseudo-labeling phase

After phase 2, GINE improved several class balance metrics, notably Macro F1, Macro Recall and Macro AUC. Paired t-tests over 15 validations were used to evaluate the change before and after retraining. Results must be interpreted on the labeled data section; ISIC 2020 Test does not have a public classification label, so the pseudo-label is not the real label.

Bảng 3.11: Compare the results of the GINE model before and after the confidence-based pseudo-labeling phase.

Index	Phase 1 (%)	Phase 2 (%)
midrule Accuracy	78.5 $\pm$ 1.8	89.2 $\pm$ 1.1
Macro AUC	72.6 $\pm$ 2.1	88.5 $\pm$ 1.4
Macro F1	43.0 $\pm$ 1.9	60.9 $\pm$ 1.7
Macro Recall	65.0 $\pm$ 2.3	82.1 $\pm$ 1.6
Precision	97.4 $\pm$ 0.6	95.7 $\pm$ 0.7
Recall	67.2 $\pm$ 2.5	85.1 $\pm$ 1.8
bottomrule		

After adding the High Trust group, the model is retrained on the expanded set but still evaluated on the independent test set. The amount of change between the two phases indicates the actual benefit of the unlabeled data. Selecting only about a third of the samples shows that the approach prioritizes quality over quantity. Pseudo-labels should not be reported as authentic labels and do not replace the expert labeling process.

The increase in Macro F1 and Macro Recall is more important than the increase in Accuracy alone because these two metrics give more balanced weighting between

classes. The slight decrease in Precision in some configurations shows that retraining creates a trade-off: the model detects more positive samples but may also generate false positive predictions. Therefore, the entropy threshold needs to be chosen on the validation set and reported with the same range of variation across validation runs. Phase 2 results should only be considered evidence to support the sampling mechanism, not evidence that all unlabeled data can be automatically converted into training data.

3.7. Reliability analysis and discussion

Of the 10,982 unlabeled samples, High Trust has 3,623 samples, Medium Trust 3,621 and Low Trust 3,738. Only about 33% of unlabeled data was included in retraining; about two-thirds were eliminated in exchange for scale for reliability.

Bảng 3.13: *Distribution of number of samples according to confidence level on unlabeled set D_U*

Confidence level	Number of samples	Ratio	Usage
midrule High Trust	3,623	33.0%	Yes
Medium Trust	3,621	33.0%	No
Low Trust	3,738	34.0%	No

Entropy only reflects the relative certainty of the model, not replacing ECE, Brier score or calibration curve. Remaining limitations include missing metadata, graph edges that may connect patterns with different semantics, not fully surveying the modern backbone, and not multicenter testing.

Conclusion of Chapter 3

The proposed framework combines multimodal representation, graph relations, and entropy-filtered pseudo-labels. The results confirm the potential of data organization at the sample level but are not evidence of clinical diagnostic effectiveness.

Confidence distribution provides additional information beyond Accuracy. A large proportion of Medium or Low Trust samples may indicate class intersection, missing metadata, domain bias, or insufficiently distinct representation. Therefore, entropy is both a filter and an error analysis tool. In future implementations, it will be necessary to incorporate probability correction to confirm that the predicted confidence level corresponds to the observed probability of being correct.

The three confidence levels also provide a priority mechanism for expert labeling. The Low Trust group may be moved into the review queue because it contains many difficult or out-of-distribution cases; Medium Trust groups are suitable for active learning; The High Trust team serves controlled coaching extensions. This usage exploits the entire unlabeled set without forcing every sample to receive a false label.

3.8. Comparison with related methods and scope of application

The proposed model achieved Accuracy 89.2% in the thesis’s experimental protocol, competing with compared methods. However, differences in data division, preprocessing, backbone and reporting metrics do not allow to conclude absolute superiority. The main value of the method is to combine three components in one process: multimodal representation, inter-sample graph, and entropy-controlled pseudo-labels.

The new method has been tested on publicly available data, has not been tested in

Bảng 3.14: Compare the proposed model with some related methods on ISIC 2020 according to the Accuracy index

Method	Approach	Accuracy (%)
midrule Gupta et al	Knowledge distillation	87.3
Alhudhaif et al	Deep CNN Network	85.5
Choudhary et al	Mask-RCNN combined with GRU	88.0
Rana et al	Stacked CNN model	86.7
Proposed model	Multimodal + graph + confidence-filtered pseudo-labels	89.2

a multicenter setting, and has not been clinically evaluated. Therefore, the model only has research significance and potential screening support, not an independent diagnostic tool.

Comparisons with related studies are made with caution because each work may use different data division and preprocessing methods. The results of the thesis demonstrate the feasibility of multimodal fusion chain, graph learning and pseudo-label selection, but do not prove the absolute advantage beyond the used protocol. Multicenter external testing and evaluation under distributional shifts are necessary steps before considering clinical application.

Additional analysis of graph learning and pseudo-labels

Node representation directly determines graph quality. Backbones provide complementary visual descriptions, while hand-crafted features and metadata help distinguish samples that are morphologically close to each other. After normalization and attention-based feature fusion, cosine similarity is used to identify neighbors. If the graph is built before feature fusion, the relationship may only reflect the image and ignore age, gender or lesion location.

The k value controls the connection level. Graphs that are too sparse limit information transmission; Graphs that are too thick increase interlayer edges and risk smoothing the representation. The GCN, GAT, GraphSAGE and GINE architectures perform neighbor aggregation by different mechanisms, and should be compared within the same protocol. The supervised phase results establish the background model before any unlabeled samples are included in training.

In the second phase, the entropy is calculated over the entire probability distribution. Compared to using only maximum probability, entropy also reflects the level of competition between the remaining classes. High Trust groups have centralized distribution and are labeled according to argmax; The remaining two groups did not participate in retraining. This mechanism reduces the amount of data used but limits the risk of confirmation bias, in which an initially incorrect prediction is relearned as the truth.

The increases in Macro F1, Macro Recall, and Macro AUC show that the benefits do not just come from the majority class. However, Precision can decrease slightly when the model detects more positive samples. This trade-off needs to be interpreted along with the standard deviation across validation runs. Pseudo-labels are only a training tool and should not be used as reference labels when calculating metrics on the test set.

Reliability stratification also supports data analysis. Low Trust samples may be

difficult cases, lacking metadata, or out of distribution; The Medium Trust template is suitable for an active, expert-led learning process. In future research, entropy should be combined with ECE, Brier score, temperature scaling or ensemble to evaluate both sharpness and probability correction.

The important limitation is that the graph is built from a learned representation on a specific data domain. When switching medical facilities or acquisition devices, the neighbor structure may change even though the classifier still gives a high probability. Multicenter external testing, distribution shift testing, and error analysis by patient group are necessary steps before considering integration into the screening support process.

In addition to prediction performance, future research needs to measure edge stability when backbones, augmentations, or metadata change. The proportion of neighbors that are preserved and the degree of label uniformity within the neighborhood can be used to diagnose graph quality. Combining these measures with predicted confidence will help distinguish samples that are difficult due to the image itself from samples that are difficult due to inappropriate graph relationships.

Implications of the results for unlabeled data mining

Entropy stratification shows that the unlabeled data is not a homogeneous block. The High Trust team provides signals that are focused enough to extend coaching; The Medium Trust group contains cases where further evidence is needed; The Low Trust group may reflect a difficult pattern or shift in distribution. Using only the High Trust group reduces the size of the additional data but provides a clearer control mechanism than pseudo-labeling the entire set.

For a complete evaluation, it is necessary to monitor the pseudo-label error rate by class, the change in the confusion matrix, and the stability over multiple runs. If erroneous pseudo-labels are concentrated mainly in the majority layer, the overall performance may increase while the imbalance becomes more severe. Therefore, the entropy threshold can be calibrated by class or incorporate quotas to prevent one class from dominating the extended set. These measures need to be independently verified before being included in the next development version.

In the context of clinical support, samples that do not meet the threshold should not be removed from the pipeline but can be prioritized for expert review. This organization turns reliability into a streaming mechanism: strong samples support retraining, intermediate samples serve active learning, and difficult samples are moved to in-depth evaluation. This is a safer exploit than treating all automatic predictions as real labels.

The resulting workflow therefore separates automatic learning from expert decision-making. High Trust samples provide a controlled source for model updating, whereas Medium and Low Trust samples remain available for active learning, error analysis, and distribution-shift monitoring. This separation preserves the value of the unlabeled set without assuming that confidence is equivalent to correctness, and it provides a practical basis for future multicenter validation.

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Conclusion

The thesis has solved two linking problems in classifying skin lesion images. BlockNet reorganizes the representation at the feature map level by entropy selection, fusion with handcrafted features, and balance in the feature space. The second framework organizes data at the sample level by merging images with metadata, constructing a k -NN graph, and mining unlabeled data in a controlled manner.

In the first contribution, the activation entropy is calculated from the histogram of each map and used to filter according to the block’s own interval. The retained maps were vectorized and fused with HSV, GLCM, and fractal features. On VNCancer, the best Accuracy of the selected test increased from 80.97% to 88.75%; F1-score increased from 82.11% to 88.16%. On HAM10000, the increase is smaller but appears in the F1-score and AUC of some difficulty classes. The results confirm the validity of the full procedure within the experimental protocol, without attributing the entire improvement to entropy alone.

In the second contribution, image representations from VGG19, DenseNet121, and MobileNet, together with handcrafted features, are fused with metadata using a two-branch attention mechanism. Graph modeling inter-sample relationships; Predicted entropy filtering out erroneous pseudo-labels. Only 3,623/10,982 of High Trust samples are used for retraining. Some Macro metrics improved after phase 2, but the pseudo-label was not considered a reference label.

Limitations and future research directions

Key limitations include not fully separating the effects of each BlockNet component; The imbalance mechanisms have not been fully controlled; missing metadata; Graph quality depends on node representation and k ; Entropy has not yet replaced correction evaluation; No multicenter external examination yet. The next direction is to perform full ablation, minority class analysis, develop cross-attention or multimodal Transformer, investigate graph transformers and heterogeneous graphs, combine entropy with ECE, Brier score and deep ensembles, integrate XAI, independent external testing and clinical appraisal.

Overall, the two contributions are united in the goal of organizing and exploiting more reliable multimodal representations. The model is the result of methodological research, not an independent diagnostic tool.

Future work should retain the same evaluation discipline while extending both contributions. For BlockNet, controlled ablation should quantify the separate effects of entropy selection, handcrafted features, and feature-space balancing. For the graph framework, external validation should examine the stability of node representations, neighbor relations, and Trust-level thresholds under changes in acquisition devices and clinical populations. Calibration metrics and expert review must accompany predictive metrics so that an apparent performance gain is not confused with reliable clinical behavior. These extensions will clarify the generalizability of the proposed methods while preserving their intended role as methodological and screening-support research.

LIST OF PUBLICATIONS RELATED TO THE DISSERTATION

- [CT1]. Nguyen, Thi Trang; Nguyen, Ha Phuong; Vu, Van Hieu. “Hybrid Feature Fusion: Enhancing Scene Image Classification with Handcrafted and Deep Learning Approaches.” In: Proceedings of the 17th National Conference on Fundamental and Applied Information Technology Research (FAIR), Hanoi, Vietnam, 8–9 August 2024. 2024. DOI: 10.15625/vap.2024.0226.
- [CT2]. Vu, Van Hieu; Nguyen, Thi Trang; Nguyen-Thi, Thao. “Ensemble Deep Learning Models for Enhanced Skin Lesion Diagnosis Using the HAM10000 Dataset.” In: Proceedings of the 27th National Workshop on Selected Problems of Information and Communication Technology, Nha Trang, Vietnam, 11–12 October 2024. 2024.
- [CT3] Nguyen, T.T., Vu, V.H. “Entropy-based feature fusion via BlockNet for robust skin cancer classification.” *Iran Journal of Computer Science*. Volume 8, pages 2871–2891 (2025). DOI: 10.1007/s42044-025-00351-8 (Scopus, Q2).
- [CT4] Nguyen, T.T., Vu, V.H. & Nguyen, V.A. “Semi-supervised GNN for Multi-modal Skin Lesion Classification.” *International Journal of Computational Intelligence Systems*. Volume 19, article number 208 (2026). <https://doi.org/10.1007/s44196-026-01296-x> (SCI-E/Q1).