Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Multi-label classification analysis with modified C-Tran on SCIN dataset

Hasih Pratiwi[a*], Fauzi Nafi'udin[a], Sri Sulistijowati Handajani[a], Respatiwulan[a], Yuliana Susanti[a] and Muhammad Bayu Nirwana[a]

[a]*Research Group of Statistics and Data Science in Environment and Health, Statistics Department, Universitas Sebelas Maret, Surakarta, Indonesia*

A B S T R A C T

Article history:
Received August 5, 2024
Received in revised format August 29, 2024
Accepted November 18 2024
Available online
November 18 2024

Keywords:
Multi-label classification
C-Tran
Skin condition
SCIN dataset
Metadata

Skin conditions affect millions of people globally, with symptoms appearing in different body areas. Technological advancements have brought diverse data types, including situations where an image depicting a skin condition can be assigned multiple labels. The Classification Transformer (C-Tran) method, which utilizes transfer learning and transformers, was developed for multi-label classification. Recently, Google introduced a new dataset called SCIN (Skin Condition Image Network), which aims to provide diverse data on skin conditions. This research aimed to use the C-Tran method for the multi-label classification of skin conditions with the SCIN dataset while incorporating additional metadata inputs to improve the metric results. The results show that the multi-label classification process using metadata is far superior to the model without metadata. For example, In the mAP metric, models that utilized metadata scored 82.37, whereas models without metadata only scored 47.02. Similarly, models with metadata achieved 70.83% in the accuracy metric, while models without metadata achieved only 34.72%. Out of the 10,379 data points available with metadata in the SCIN dataset, only 718 were actually utilized for the classification task. It is thought that the inaccurate prediction outcomes are due to unreliable data, even with a confidence level of 4. In this analysis, two metadata categories stood out the most in terms of different measurements: the body part and symptoms metadata categories from the SCIN dataset. With just the body part and symptoms metadata groups, the mAP results achieved a 74.23%, accuracy at 63.89%, CF1 at 68.79%, and OF1 at 73.13%.

1. Introduction

Skin conditions impact millions globally, displaying symptoms in various body areas (Hay et al., 2014; Lim et al., 2017; Richard et al., 2022). Sometimes, a solitary patient could simultaneously suffer from multiple skin conditions (Creadore et al., 2022; Cohen et al., 2023). Hence, creating classification models that can accurately recognize and categorize various skin conditions is crucial. Due to technological advancements, there is now a growing variety of data types, including the scenario where an image, such as one depicting a skin condition, may be assigned multiple labels (Omeroglu et al., 2023). This trend has led to an increased focus on multi-label classification, requiring models to identify and categorize multiple labels from one input (Bi et al., 2020; Tang et al., 2022; Han et al., 2023). Medical data is becoming more varied with helpful metadata like body location, visible symptoms, and patient demographics (Ward et al., 2024). This additional data can help the model comprehend the skin condition being examined within the scope of dermatology. Transfer learning methods have been extensively employed to enhance the effectiveness of classification models (Hosna et al., 2022; Asif et al., 2023; Zhu et al., 2023). Transfer learning involves taking models trained for one task and adjusting them for other related tasks, accelerating training time and improving precision. Transfer learning is very beneficial in the medical field, as extensive training data is often unavailable. We can use well-trained models like ResNet (He et al., 2016), trained on extensive datasets like ImageNet (Deng et al., 2009), ISIC (Rotemberg et al., 2021) or Fitzpatrick 17k (Groh et al., 2021) to improve feature representations for skin disease classification. However, transformer techniques have gained significant popularity recently, not least in image data processing (Chen et al., 2021; Li et al., 2023; Kameswari et al., 2023; He et al., 2023). Transformers utilize an attention

* Corresponding author
E-mail address hpratiwi@mipa.uns.ac.id (H. Pratiwi)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2024.11.003

mechanism that enables the model to dynamically concentrate on specific input sections, enhancing its capacity to comprehend intricate contexts (Vaswani et al., 2017). These techniques are also starting to be used in computer vision, with models like the Vision Transformer (ViT) demonstrating strong performance in different image recognition assignments (Khan et al., 2022). The Classification Transformer (C-Tran) method, which utilizes transfer learning and transformers, has been developed for multi-label classification (Lanchantin et al., 2021). C-Tran employs a transformer design for analyzing image data, incorporating transfer learning benefits for enhanced efficiency and precision. This model excels in managing the intricacy of multi-label classification more effectively than conventional methods due to its capacity to comprehend the context and connections among labels. Additionally, Google recently introduced a new dataset named the SCIN (Skin Condition Image Network), aiming to offer a broader range of data on skin conditions (Ward et al., 2024). It encompasses various skin types and medical conditions, enhancing its representativeness for creating impartial and precise classification models. Additionally, the SCIN dataset incorporates beneficial metadata to aid in the classification procedure. We intend to utilize the C-Tran method for the multi-label classification of skin conditions with the SCIN dataset while incorporating additional input metadata to enhance the metric outcomes. Furthermore, we analyze the impact of the metadata categories.

2. Literature Review

This section examines the different studies carried out by researchers on image data processing for multi-label classification. Different journal experts have identified several effective methods to enhance accuracy and efficiency in classification. In a recent journal article by Gour and Khanna (2021), scientists studied the detection of multi-class, multi-label ophthalmology diseases using a convolutional neural network with transfer learning. They suggested two methods for categorizing fundus images using this technique. The outcomes demonstrated that when utilizing the SGD optimizer, VGG16 achieved AUC and F1 scores of 84.93 and 85.57, respectively, in the two-input approach. Referring to Model-2, the input-concatenated method using the VGG16 architecture demonstrated superior AUC and F1 score figures of 68.88 and 85.57 when compared to alternative architectures. Mahbod et al. (2020) investigated the impact of different input image sizes on skin lesion classification performance using pre-trained CNNs and transfer learning. Datasets from the ISIC challenges in 2016 (Gutman et al., 2016), 2017 (Codella et al., 2017), and 2018 (Codella et al., 2019) were utilized to analyze six different image sizes (224×224, 240×240, 260×260, 300×300, 380×380, and 450×450 pixels) using cropping and resizing techniques. The study selected three models from the SeNet (Hu et al., 2018) and EfficientNet (Tan & Le, 2019) families, known for their strong classification capabilities in both natural and medical image domains. Their research indicated that cropping was more effective than resizing, with consistent performance regardless of image size for cropping and better performance with higher resolution for resizing. Additionally, they proposed a multi-scale multi-CNN (MSM-CNN) fusion approach, which showed superior classification performance on the ISIC 2018 dataset compared to state-of-the-art algorithms. Moreover, Tabbakh and Barpanda (2023) have also published a journal on using Transfer Learning and Vision Transformer (TLMViT) in the deep feature extraction model for plant disease classification. This article suggests a fresh method for extracting deep features and categorizing diseased plant leaves. This blended approach involves a Transfer Learning model paired with a Vision Transformer (ViT). TLMViT was evaluated using five pre-trained models, with each one being followed by ViT. TLMViT demonstrated excellent performance in plant disease categorization, obtaining validation accuracies of 98.81% and 99.86% using the VGG19 model and ViT on the PlantVillage and wheat datasets, respectively. The results of the comparison indicated that TLMViT enhanced the validation accuracy by 1.11% and 1.1% while also decreasing the validation loss by 2.576% and 2.92% in contrast to the Transfer Learning-based model for the PlantVillage and wheat datasets. The article also points out that using data augmentation helps to address image shortcomings that lead to overfitting and minimizes the impact of uneven datasets. The pre-trained model is utilized for initial leaf feature extraction and dimension reduction of the original image before being fed into the inner layers for deep feature extraction. Furthermore, the ViT model demonstrates the capability to derive in-depth features from the features extracted by the CNN model. The paper authored by Lanchantin et al. (2021) explores multi-label image classification, where the goal is to predict a group of labels that match objects in an image. Their method involves training a Transformer encoder to forecast a group of target labels using masked input labels and visual characteristics from a convolutional neural network. The label mask training objective, using ternary coding, is a crucial part of the method, representing label states as positive, negative, or unknown during training. Because this model explicitly captures the label state when training, it is more versatile and leads to improved outcomes when dealing with images that have incomplete or extra-label annotations during inference. This function was examined on the COCO, Visual Genome, News-500, and CUB datasets. The findings demonstrate that this method is effective in both standard multi-label classification and multi-label classification with partially or additionally observed labels. C-Tran surpassed cutting-edge methods in several situations. In the paper written by Cai et al. (2023), they describe the use of multimodal transformers to combine images and metadata in skin disease classification. They proposed a novel multimodal transformer with two encoders for images and metadata, along with a single decoder for merging the multimodal information. This model was created to classify skin diseases using both images and metadata. One decoder is utilized to integrate multimodal features, with two encoders extracting features from images and metadata separately. Studies demonstrate that the suggested framework attains noteworthy results in categorizing skin diseases. The model performs better than other popular networks, with an accuracy of 0.816 on a private dataset. The method obtained a 0.9381 accuracy and a 0.99 AUC on the ISIC 2018 dataset. This model demonstrates efficient performance and advancement in skin disease diagnosis when compared to cutting-edge methods.

2. Material and methods

2.1 Data Description

This research uses the (Skin Condition Image Network: SCIN) Dataset, published by Google, containing images of various skin diseases (Ward et al., 2024). Google obtained this dataset through crowdsourcing using its advertising system. Compared to other datasets, this dataset is unique because most existing dermatology datasets tend to overrepresent malignant skin diseases and underrepresent darker skin colors. This may cause bias in the deep learning model developed. Examples of some images from the SCIN dataset can be seen in Fig. 1.

Fig. 1. Example Images from SCIN Dataset (Ward et al., 2024)

This SCIN dataset includes 5033 patient data, including disease images and metadata. Each data has 1 to 3 images. Adults from the United States contributed to the images in this dataset. The size of the images in this dataset also varies. In the original dataset, the metadata consists of various variables, which were then grouped into 9 categories. The nine categories of metadata are General metadata, Skin type metadata, Ethnicity metadata, Texture metadata, Body parts metadata, Symptoms metadata, Other metadata, Shot type metadata, and Label metadata. Details of each metadata category can be found in Appendix 1. These metadata groups assist the neural network's training process, except for the label metadata group, which acts as a label.

2.2 Data Preparation

Data preparation from the SCIN dataset involves various complex steps, starting with aligning the images with their respective metadata. Out of 5033 data, there may be 1-3 images for each data, bringing the total to 10379. Next, since this is a classification, it is essential to determine which columns act as labels. In this case, the group label metadata becomes the target of the classification, where the group metadata contains the disease label and the dermatologist's confidence level. The processing of this variable is crucial because the target variable is the determining variable in the classification process. The variable dermatologist_skin_condition_on_label_name contains a list of labels of several diseases simultaneously because this dataset is multi-label. Therefore, the contents of the variable dermatologist_skin_condition_confidence, a list of each disease mentioned in the label, must be aligned. Data was removed if one of the two variables was misaligned or missing. In addition, redundancy is found in labeling diseases in the same image, which requires cleaning. Data that had no values in both variables or had missing values in both variables were removed immediately. The process of eliminating duplicate data was also performed. All these steps are done to sort and use the data with images. If the data does not have images, then the data is immediately deleted. Furthermore, categorical data is converted into numbers in processing the other 8 metadata groups. Data that has a missing value is replaced with -1. For example, if in one variable, there are 8 types of categories, including the unknown category, and there are missing values, then the data contains values 1-7 and -1. Data with boolean type is converted to 0 and 1. Numeric variables are left alone. After preprocessing, 6503 image data and their metadata are obtained. From all 6503 data, several versions of data were formed according to the dermatologist's confidence level in labeling the disease from the image. Since the confidence level of the dermatologist is in the range of 1-5, 5 versions of data were formed. Version 1 is only for data with a confidence level of 5, version 2 for confidence levels 4-5, version 3 for confidence levels 3-5, version 4 for confidence levels 2-5, and version 5 for confidence levels 1-5. In addition, each label used must have a minimum of 50 data. Otherwise, the data will be deleted. In the process of dividing data for neural network training, this data is divided into two parts, namely 80% for training data and 20% for testing data. Details of the number and distribution of data can be seen in Table 1.

Table 1
Data Distribution Each Version

	Dataset Version				
	Version 1 (Conf. 5-5)	Version 2 (Conf. 4-5)	Version 3 (Conf. 3-5)	Version 4 (Conf. 2-5)	Version 5 (Conf. 1-5)
Total Data	718	2253	3655	5501	6232
Num of Label	7	16	27	44	54
Data Training	574	1802	2924	4400	4985
Data Testing	144	451	731	1101	1247

2.3 Architecture Model

Baseline Model

The baseline model used in this research is the Classification Transformer (C-Tran) (Lanchantin et al., 2021). C-Tran is a general framework for multi-label image classification that utilizes Transformers to identify and exploit possible relationships between image features and the data labels. This model approach uses a Transformer encoder trained to predict target labels based on input-masked labels and visual features from a convolutional neural network. In addition, within the Transformer structure is a backbone model that uses transfer learning, explicitly ResNet-101, as part of the feature extraction process from images. The general architecture of C-Tran can be seen in Fig. 2.

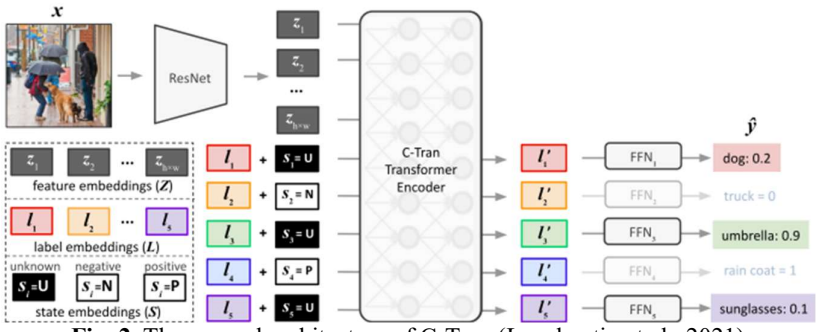

Fig. 2. The general architecture of C-Tran (Lanchantin et al., 2021)

In the original C-Tran, 3 embeddings were used. First, feature embeddings are vectors extracted by transfer learning, such as ResNet-101, that represent the visual features of the image. Second, label embeddings describe vector representations that show the relationship between labels used in multi-label classification. Finally, state embeddings refer to vector representations that use a ternary coding system. State embeddings help users avoid relying on prior information by incorporating unknown embeddings, leveraging both negative and positive embeddings to incorporate partially labeled or additional information, and experimenting with interventions in the model by observing the impact of label changes on predictions.

2.4 Proposed Model

The modification applied to C-Tran in this study involves adding input in the form of metadata embedding. This vectorized metadata is inserted into the linear layer so that it can be easily adapted to various possible concepts. We carried out multiple experiments to find the optimal method for incorporating metadata embedding within the current framework. There are multiple options available during this insertion procedure. First, the metadata embedding is summed with the sum of the label embedding and state embedding. Second, the metadata embedding is incorporated as a transformer input. Third, the metadata embedding is summed into the label embedding that has passed through the transformer architecture. A more in-depth clarification of these three options can be observed in Fig. 3.

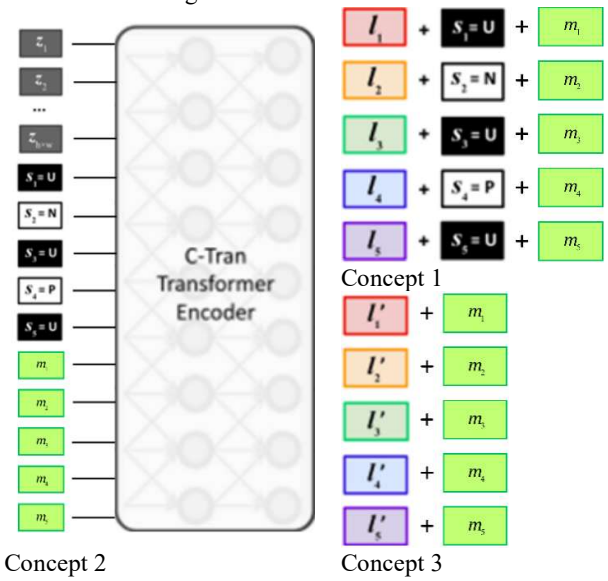

Fig. 3. Three Concepts of Potential Metadata Input Insertion in C-Tran Architecture

For this study, the parameters used in Modified C-Tran followed the model's defaults. The training system uses label-mask training, with optimization using Adam's algorithm and a learning rate of 0.00001. In addition, a freeze backbone was used, where the parameters of the transfer learning were not retrained. The length of the vectors in the embedding metadata varies, with the maximum vector length reaching 51. The model training process is carried out for 25 epochs. A batch size of 16 was used, with gradient accumulation steps set to 2. The model used 3 transformer layers, each with 4 attention heads, and sine positional embeddings were applied. This research uses computing resources in the form of a 12th-generation Intel(R) Core(TM) i9-12900K processor with 24 CPUs at 3.2 GHz. In addition, this research is supported by 64 GB of RAM and RTX 3060 Ti GPU. The software includes Python version 3.12 and TensorFlow version 2.14.0, with Windows 11 operating system.

2.5 Evaluation Metrics

The modified C-Tran model for multi-label classification is evaluated using various metrics. These metrics include Mean Average Precision (mAP), Accuracy, Example-Based F1 Score, average per-class precision (CP), average per-class recall (CR), average per-class F1 Score (CF1), average overall precision (OP), average overall recall (OR), average overall F1 Score (OF1), Top-1 F1 Score, and Top-3 F1 Score. These metrics match previous journal studies' references (Lanchantin et al., 2021). Furthermore, the researchers included another measure known as the Jaccard Index. This measurement evaluates the similarity between two sets of data. The Jaccard Index is calculated by dividing the intersection of two sets by their union (Hamers et al., 1989). A higher value indicates improved performance of the classification model. Each metric provides a unique view of a particular aspect of model performance. In addition, according to Chen et al. (2019), among the various metrics, OF1, CF1, and mAP are generally considered more important for performance evaluation.

3. Results and Discussion

In this section, we present the implementation results of the proposed model and the findings obtained in the multi-label classification process in detail. This section also demonstrates and evaluates the performance results and presents the analysis of all versions of the dataset and a graphical visualization of the results. To begin, we aim to determine which of the three potential concepts is the most optimal. The data used in this comparison process is version 1 of the dataset. The metrics used for this comparison are mAP, Accuracy, CF1, and OF1. The comparison results are illustrated in the graph in Fig. 4.

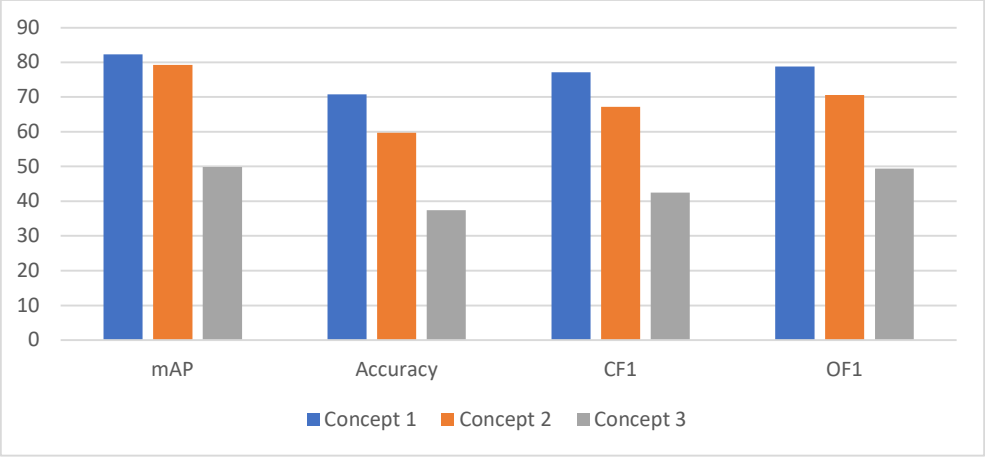

Fig. 4. Model comparison results with three potential concepts

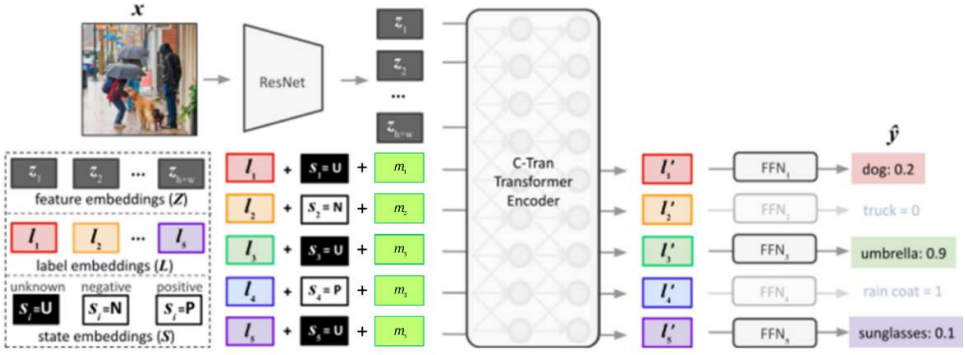

Fig. 5. Modification of C-Tran Model Architecture

The results in Fig. 4 demonstrate that Concept 1 outperforms the other two concepts in several metrics. Concept 2 yields result that are competitive, although they are not as impressive as those of concept 1. In the meantime, concept 3 demonstrates the poorest outcomes and differs significantly in comparison to the other concepts. Based on these results, concept 1 was selected to modify the C-Tran architecture. The results of this C-Tran architecture modification can be seen in Figure 5. The next step is to compare whether the metadata embedding input actually affects the optimal metric results. Therefore, the model with the additional metadata embedding is compared with the original C-Tran architecture that does not use metadata embedding input. This comparison is presented in Fig. 6.

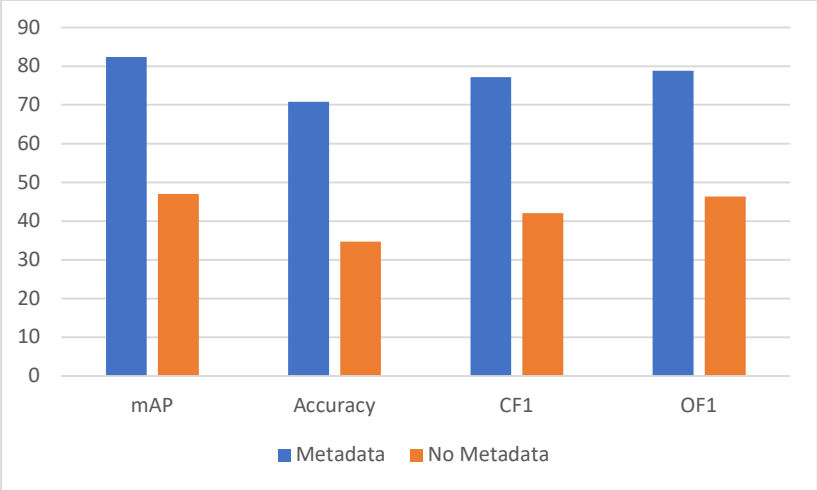

Fig. 6. Comparison results model with metadata and without metadata using version 1 data

Figure 6 shows that the multi-label classification process with the model using metadata is far superior to that without metadata. For example, on the mAP metric, there is a significant difference where the model with metadata achieves a score of 82.37. In contrast, the model without metadata only scores 47.02, showing a difference of 35 points between the two models. Similarly, models with metadata achieved 70.83% in the accuracy metric, while models without metadata achieved only 34.72%. After performing various steps to determine the optimal architecture model, the next step is to apply the model to multiple versions of the available datasets. Table 2 details the metric results of the model implementation on various datasets. Table 2 indicates that in version 1, the model performed well in various metrics, achieving an mAP of 82.37 and an accuracy of 70.83. In general, the model's performance noticeably declined with every iteration of the dataset. The decrease is observed in the mAP measurement, falling from 82.37 in Version 1 to 44.32 in Version 5. It is confirmed that an expert's level of certainty when labeling a disease is crucial in the classification process. Still, even a tiny difference in confidence between scores of 4 and 5 can lead to significantly different outcomes.

Table 2
Model results with various data versions

Data	Metrics									
							Top-3		Top-1	
	mAP	Acc	JI	ebF1	CF1	OF1	CF1	OF1	CF1	OF1
Version 1	82.37	70.83	71.88	72.22	77.19	78.81	77.19	78.81	77.89	79.25
Version 2	63.37	41.02	45.31	46.70	54.09	56.45	54.09	56.45	49.55	53.88
Version 3	52.81	33.52	39.27	41.20	41.19	51.41	41.19	51.41	36.30	47.48
Version 4	48.34	18.07	32.23	37.21	31.97	44.31	31.51	44.01	22.87	36.54
Version 5	44.32	10.59	29.33	36.12	32.30	43.62	31.48	42.95	20.04	32.28

Despite containing 10,379 data with metadata, only 718 were utilized for the classification task from the SCIN dataset. In this situation, the unsatisfactory prediction results are believed to be impacted by uncertain data, even with a confidence level of 4. Moreover, cutting a significant portion of the dataset has diminished the intended goal of creating a diverse range of skin colors to address a societal problem. By removing this data, the strength of the identity's relevance diminishes. Furthermore, a majority of the remaining information has been assigned only one label despite the presence of some with multiple labels. However, the SCIN dataset continues to be a valuable resource for researching multi-label classification. It is also highly beneficial for investigating variations in skin color in neural network applications and other techniques. Additionally, we show the model findings using only one metadata group in each trial. This experiment identifies the metadata group with the most significant impact on the multi-label classification. Fig. 7 displays the outcomes of the experiment.

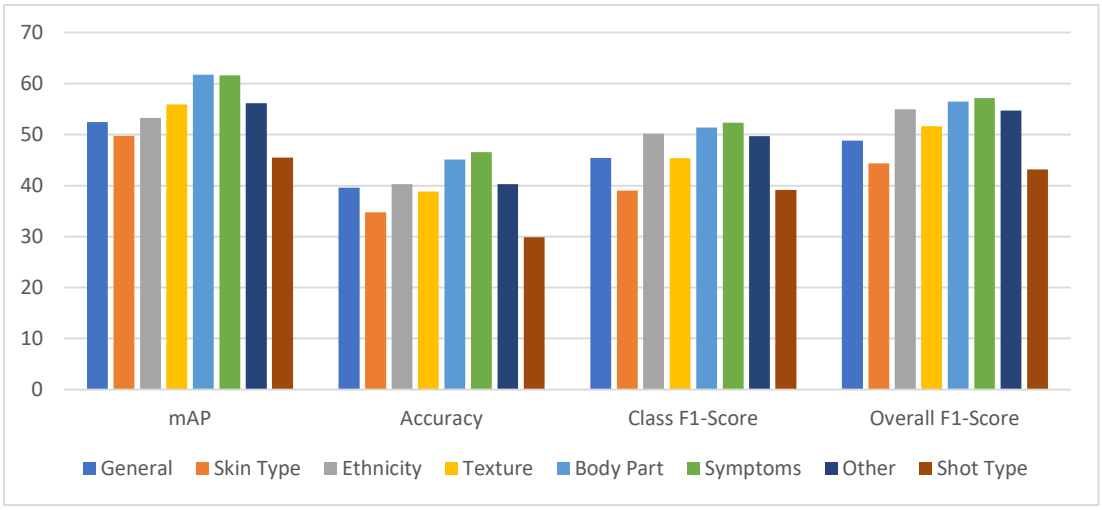

Fig. 7. Comparison results model with one group metadata using version 1 data

Fig. 7 shows that two metadata groups dominate this comparison process in various metrics: the body part and symptoms metadata groups. While it cannot be said that these two metadata groups are the most important, it shows that metadata can also be essential in the classification process. Interestingly, the skin type metadata group has the most minor metrics apart from the shot type metadata group. This is surprising, considering that the shot type metadata group has the most minor continuity with skin. In general, skin type is closely related to the skin, which can be seen in the image and recognized by deep learning models. Still, in this study, skin type metadata did not significantly impact the classification results compared to others. From the experiments using two groups of metadata only, namely the body part and symptom metadata groups, the mAP results were 74.23%, 63.89% accuracy, 68.79% CF1, and 73.13% OF1.

4. Conclusions

This research demonstrates that utilizing modified C-Tran architecture with added metadata embedding results in significantly better multi-label classification compared to using the model without metadata. On the mAP metric, the model with metadata scores 82.37, while the model without metadata scores 47.02, showing a 35-point difference between them. Also, the accuracy of the model using metadata reached 70.83%, whereas the model without metadata only reached 34.72%. This research also validated the significance of an expert's confidence level when assigning a disease label in the classification procedure. A slight variance in confidence level from a score of 4 to 5 can result in widely varying outcomes. While the SCIN dataset has 10,379 data with metadata, only 718 data were utilized for the classification task. Nevertheless, the SCIN dataset continues to be a valuable resource for researching multi-label classification. It is also highly beneficial for investigating variations in skin color in neural network applications and other techniques. The research also revealed that employing a single metadata group in every trial pinpointed the metadata group with the most significant impact. The body part and symptom groups were the main metadata groups in this comparison. Although these two metadata groups may not be the most crucial, they demonstrate that metadata can play a significant role in the classification process. Interestingly, the skin type metadata category has minimal influence on these measurements, just like the shot type metadata category. To sum up, incorporating metadata into the C-Tran framework demonstrates that this extra data can significantly enhance results in multi-label classification, considering the significance of confidence level and choosing the proper metadata categories.

Acknowledgments

The authors would like to thank the Indonesian Ministry of Education, Culture, Research and Technology, for providing financial support through Research Grant No. 086/E5/PG.02.00.PL/2024

References

Asif, S., Zhao, M., Tang, F., & Zhu, Y. (2023). An enhanced deep learning method for multi-class brain tumor classification using deep transfer learning. *Multimedia Tools and Applications, 82*(20), 31709-31736.

Bi, L., Feng, D. D., Fulham, M., & Kim, J. (2020). Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network. *Pattern Recognition, 107*, 107502.

Cai, G., Zhu, Y., Wu, Y., Jiang, X., Ye, J., & Yang, D. (2023). A multimodal transformer to fuse images and metadata for skin disease classification. *The Visual Computer, 39*(7), 2781-2793.

Chen, C. F. R., Fan, Q., & Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. *In Proceedings of the IEEE/CVF international conference on computer vision* (pp. 357-366).

Chen, Z. M., Wei, X. S., Wang, P., & Guo, Y. (2019). Multi-label image recognition with graph convolutional networks. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5177-5186).

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., ... & Halpern, A. (2018, April). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic*). In 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) (pp. 168-172). IEEE*.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., ... & Halpern, A. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368.

Cohen, B., Cadesky, A., & Jaggi, S. (2023). Dermatologic manifestations of thyroid disease: a literature review. *Frontiers in Endocrinology, 14*, 1167890.

Creadore, A., Desai, S., Alloo, A., Dewan, A. K., Bakhtiar, M., Cruz-Diaz, C., ... & Mostaghimi, A. (2022). Clinical characteristics, disease course, and outcomes of patients with acute generalized exanthematous pustulosis in the US. *JAMA dermatology, 158*(2), 176-183.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. *In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee*.

Gour, N., & Khanna, P. (2021). Multi-class multi-label ophthalmological disease detection using transfer learning-based convolutional neural network. *Biomedical signal processing and control, 66,* 102329.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., ... & Badri, O. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1820-1828).*

Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., & Halpern, A. (2016). Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397.

Han, M., Wu, H., Chen, Z., Li, M., & Zhang, X. (2023). A survey of multi-label classification based on supervised and semi-supervised learning. *International Journal of Machine Learning and Cybernetics, 14*(3), 697-724.

Hay, R. J., Johns, N. E., Williams, H. C., Bolliger, I. W., Dellavalle, R. P., Margolis, D. J., ... & Naghavi, M. (2014). The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *Journal of investigative dermatology, 134*(6), 1527-1534.

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., ... & Shen, D. (2023). Transformers in medical image analysis. *Intelligent Medicine, 3*(1), 59-78.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., & Azim, M. A. (2022). Transfer learning: a friendly introduction. *Journal of Big Data, 9*(1), 102.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

Kameswari, C. S., Kavitha, J., Reddy, T. S., Chinthaguntla, B., Jagatheesaperumal, S. K., Gaftandzhieva, S., & Doneva, R. (2023). An overview of vision transformers for image processing: A survey. *International Journal of Advanced Computer Science and Applications, 14*(8).

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR), 54*(10s), 1-41.

Lanchantin, J., Wang, T., Ordonez, V., & Qi, Y. (2021). General multi-label image classification with transformers. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16478-16488).

Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., & Zhou, S. K. (2023). Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical image analysis, 85*, 102762.

Lim, H. W., Collins, S. A., Resneck Jr, J. S., Bolognia, J. L., Hodge, J. A., Rohrer, T. A., ... & Moyano, J. V. (2017). The burden of skin disease in the United States. *Journal of the American Academy of Dermatology, 76*(5), 958-972.

Mahbod, A., Schaefer, G., Wang, C., Dorffner, G., Ecker, R., & Ellinger, I. (2020). Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. *Computer methods and programs in biomedicine, 193*, 105475.

Omeroglu, A. N., Mohammed, H. M., Oral, E. A., & Aydin, S. (2023). A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification. *Engineering Applications of Artificial Intelligence, 120*, 105897.

Richard, M. A., Paul, C., Nijsten, T., Gisondi, P., Salavastru, C., Taieb, C., ... & EADV Burden of Skin Diseases Project Team. (2022). Prevalence of most common skin diseases in Europe: a population-based study. *Journal of the European Academy of Dermatology and Venereology, 36*(7), 1088-1096.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., ... & Soyer, H. P. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data, 8*(1), 34.

Tabbakh, A., & Barpanda, S. S. (2023). A deep features extraction model based on the transfer learning model and vision transformer “tlmvit” for plant disease classification. *IEEE Access, 11*, 45377-45392.

Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

Tang, P., Yan, X., Nan, Y., Xiang, S., Krammer, S., & Lasser, T. (2022). FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. *Medical Image Analysis, 76,* 102307.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Ward, A., Li, J., Wang, J., Lakshminarasimhan, S., Carrick, A., Campana, B., ... & Rao, P. (2024). Crowdsourcing Dermatology Images with Google Search Ads: Creating a Real-World Skin Condition Dataset. arXiv preprint arXiv:2402.18545.

Zhu, Z., Lin, K., Jain, A. K., & Zhou, J. (2023). Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Appendices

Appendix 1

1.	General Metadata	:	It consists of 2 variables, namely age_group and sex_at_birth
2.	Skin Type Metadata	:	It consists of 3 variables, namely fitzpatrick_skin_type, monk_skin_tone_label_india, and monk_skin_tone_label_us
3.	Ethnicity Metadata	:	Consists of 10 variables, namely race_ethnicity_american_indian_or_alaska_native, race_ethnicity_asian, race_ethnicity_black_or_african_american, race_ethnicity_hispanic_latino_or_spanish_origin, race_ethnicity_middle_eastern_or_north_african, race_ethnicity_native_hawaiian_or_pacific_islander, race_ethnicity_white, race_ethnicity_other_race, race_ethnicity_prefer_not_to_answer, and race_ethnicity_two_or_more_after_mitigation
4.	Textures Metadata	:	Consists of 4 variables, namely textures_raised_or_bumpy, textures_flat, textures_rough_or_flaky, and textures_fluid_filled
5.	Body Parts Metadata	:	Consists of 12 variables, namely body_parts_head_or_neck, body_parts_arm, body_parts_palm, body_parts_back_of_hand, body_parts_torso_front, body_parts_torso_back, body_parts_genitalia_or_groin, body_parts_buttocks, body_parts_leg, body_parts_foot_top_or_side, body_parts_foot_sole, and body_parts_other
6.	Symptoms Metadata	:	Consists of 15 variables, namely condition_symptoms_bothersome_appearance, condition_symptoms_bleeding, condition_symptoms_increasing_size, condition_symptoms_darkening, condition_symptoms_itching, condition_symptoms_burning, condition_symptoms_pain, condition_symptoms_no_relevant_experience, other_symptoms_fever, other_symptoms_chills, other_symptoms_fatigue, other_symptoms_joint_pain, other_symptoms_mouth_sores, other_symptoms_shortness_of_breath, and other_symptoms_no_relevant_symptoms
7.	Others Metadata	:	It consists of 2 variables, namely related_category and condition_duration
8.	Shot Type Metadata	:	Consists of 3 variables, namely shot_type_AT_AN_ANGLE, shot_type_AT_DISTANCE, and shot_type_CLOSE_UP
9.	Labels Metadata	:	It consists of 2 variables, namely dermatologist_skin_condition_confidence and dermatologist_skin_condition_on_label_name