Addressing Overfitting in Medical Image Classification through Variance Penalization: A Novel Approach

Athukorala S.A.A.J.M.
Department of Electrical and Electronic Engineering,
Faculty of Engineering, University of Sri Jayewardenepura
Sri Lanka
janinduathukorala@ieee.org

***Abstract*—Medical image classification is a crucial application of deep learning in healthcare, aiding in the diagnosis and treatment planning of various diseases. Convolutional Neural Networks (CNNs) have shown considerable potential in this field, especially regarding the automated detection of abnormalities in medical images such as X-rays, MRIs, and CT scans. However, overfitting remains a critical challenge, where models learn patterns specific to training datasets but fail to generalize to unseen data. Existing methods, including data augmentation and regularization techniques, have limitations, particularly in preserving essential anatomical features in medical images, necessitating a more reliable approach. This study proposes a novel approach to address overfitting through the implementation of a variance-penalized loss function within CNNs by computing variance across convolutional layers and penalizes high variance, promoting the model's focus on stable and relevant features. Experiments carried out on two different datasets - the NIH Chest X-ray dataset and the Skin Cancer MNIST: HAM10000 - showed that the proposed method led to a notable reduction in overfitting, as evidenced by improved generalization performance, more stable training-validation accuracy curves, and lower loss compared to baseline and data-augmented models. This study highlights the potential of the proposed solution to enhance the reliability of AI-assisted diagnostics in clinical environments, advancing feature selection and model optimization. Findings show the effectiveness of the variance-penalized loss function in improving generalization and reducing overfitting, making it a valuable contribution to the field of medical image analysis.**

I. Introduction

A. Importance of Medical Image Classification

Medical image classification plays a crucial role in modern healthcare, enabling automated disease diagnosis, treatment planning, and clinical decision-making. Machine learning (ML) algorithms have significantly impacted medical image classification across various modalities like MRI, CT, and X-ray. Convolutional Neural Networks (CNNs) have shown exceptional performance in image analysis tasks, outperforming traditional methods in classification, segmentation, and detection (1). These models can detect patterns in imaging data that human radiologists might miss, improving early disease detection and reducing diagnostic errors enabling disease detection, and prognosis prediction. However, the effectiveness of these models depends on their ability to generalize well to new, unseen data.

B. Overfitting in Medical Image Classification

Still there are some challenges in algorithm selection, data utilization, and model generalization across different demographic groups (2). One of the key challenges in medical image classification is overfitting, where a model learns patterns specific to the training dataset rather than general medical features. Overfitting occurs due to continuous gradient updating and scale sensitiveness of cross-entropy loss (3). This ends up with high accuracy in training data but poor performance on validation and test data, making the model unreliable for real-world clinical applications. Generally, medical datasets have fewer samples, especially for rare diseases, making it limited data resources and imbalance between classes. Larger, more complex models struggle to converge on smaller datasets, while simpler models perform better increasing the risk of overfitting (4). Differences in imaging devices, patient demographics, and acquisition protocols create dataset biases, which models may inadvertently learn instead of general medical features also causes less generalize on unseen image data.

C. Existing Methods to Prevent Overfitting and their Limitations

Several regularization techniques have been developed to mitigate overfitting in deep learning models. Transfer learning, data augmentation, and generative adversarial networks are commonly used to tackle the small data problem (5). Transfer learning, while common, may not always address overfitting effectively and not always be effective for small medical datasets. Data Augmentation is an another method which is artificially increasing dataset size by applying transformations like rotations, flips, and contrast adjustments (6). However, in medical imaging, excessive augmentation can introduce unrealistic variations, altering critical diagnostic features. Normalizing feature distributions across mini batches is also a common technique, but it does not specifically control feature variance and may not sufficiently prevent overfitting in small datasets. Another popular technique is early stopping, which

is halting training when validation loss stops improving, but does not directly address feature-level overfitting.
Multiple Instance Learning (MIL) has shown promise in improving medical image analysis, especially for lesion detection, despite challenges in dataset annotation and computational resources (6). A novel approach combining large pretrained vision transformers with adversarial and contrastive learning techniques has shown promise in overcoming both underfitting and overfitting (4), (7). These details highlight, overfitting remains a persistent issue and the importance of reducing overfitting by addressing the limitations in existing methods. In simple, improve model generalization. Therefore, a more targeted regularization technique is needed.

D. Variance Penalization: A More Effective Approach

To overcome the limitations of several existing methods, this study introduces a variance-penalized loss function that dynamically regulates feature importance by penalizing high variance across convolutional layers. Unlike traditional techniques, variance penalization directly reduces reliance on dataset-specific patterns while preserving clinically relevant features. Proposed approach encourages the model to focus on stable features across samples rather than noise or artifacts. Unlike static methods like dropout, proposed method dynamically adjusts based on feature variance during training.

The remainder of this paper is structured as follows. Section II (Literature Review) provides an overview of existing research on medical image classification, with a focus on overfitting prevention strategies. Section III (Methodology) describes the proposed variance penalization technique. Section IV (Experiments and Results) describes the experimental setup to validate the proposed method and a comparative analysis of the proposed approach against baseline methods. Section V (Discussion) includes the findings. Finally, Section VI (Conclusion and Future Work) summarizes the key insights of this study and suggests directions for further research.

II. Literature Review

A. Background

Deep Learning has made significant progress in medical image based cancer diagnosis, including in image classification, reconstruction, detection, segmentation, registration, and synthesis. However, the lack of high-quality labelled datasets limits the role of deep learning and presents challenges in diagnosing rare cancers, multimodal image fusion, model interpretability, and generalization (8). Recent studies have explored effective methods to prevent overfitting and improve classification accuracy in deep learning models for medical image diagnosis. Common approaches include batch normalization, dropout, weight initialization, and data augmentation (8).

B. Data Augmentation and Dropout Techniques

Eric J. Snider et al.(9) describes data augmentation techniques, such as affine transformations and MixUp, improved the generalizability of the machine learning models for shrapnel detection in ultrasound images. Despite a decrease in training accuracy, the model's performance on blind tests increased from 68% to more than 85%. Eduardo Castro et al.(10) proposed a new method of performing rotation-based data augmentation within the CNN architecture itself, by randomly rotating the weights of the convolutional layers in each training batch. Validates the proposed method by showing its usefulness in different scenarios. Feng Li et al.(11) presented that dropout technique has been particularly beneficial in diagnosing Alzheimer's disease, improving classification accuracies by 5.9% compared to conventional deep learning methods. Additionally, the researchers included other tecniques like stability selection, adaptive learning, and multitask learning within the deep learning framework to improve its performance even further.

C. Transfer Learning Techniques

R. Sangeetha et al. (12) explores the use of transfer learning to improve the accuracy of breast cancer classification in medical imaging. Transfer learning models demonstrate increased computational efficiency, reduced overfitting, and the ability to learn useful representations from smaller datasets. Ahmad Al-Qerem et al. (13) showed that transfer learning approach performed significantly better than the classification-based data augmentation approach on the same dataset. Also saved considerable time and achieved competitive accuracy compared to the data augmentation approach.

D. Novel Regularization Techniques and Existing Variance Penalization Techniques

Walid Abdullah Al & I. Yun(14) proposed a reinforced classifier using generalization-feedback from a subset of training data has shown promising results by improving generalization on small datasets. The performance of the reinforced classifier was assessed across three distinct classification tasks and outperformed conventional deep classifiers with overfitting prevention techniques. Hao Li et al.(15) founded a history-based approach can both identify and prevent overfitting in deep learning models without modifying the model architecture. This method achieves an F1 score of 0.91 for overfitting detection, which is at least 5% higher than the current best-performing non-intrusive overfitting detection approach. Also, it can terminate from the training process to avoid overfitting at least 32% sooner than traditional early stopping, while maintaining the same or better rate of returning the best model. Harangi et al.(16) proposed a method to create diverse CNN ensembles by introducing a new Pearson correlation penalty term in the loss function, improving classification accuracy. Shubin et al.(17) introduced Variance Aware Training (VAT), which directly reduces variance error in the loss function, achieving comparable or better performance than self-supervised methods. This method requires selecting only one hyperparameter and matches or improves the performance of state-of-the-art self-supervised methods while achieving a significant decrease in GPU training time. Qiu et al.(18) developed CompNet, a CNN-based model that integrates image and designed features, significantly reducing overfitting in medical image datasets. The CompNet model outperformed other comparable methods that combine images and designed features, both on the LIDC dataset (19) and on the datasets used in other studies. Simpson et al.(20) presented GradMask, a new regularization method that penalizes saliency maps - if they do not align with the actual lesion segmentation, preventing the model from incorrectly linking non-tumor related features with the classification of unhealthy samples. This study shows that the application of the GradMask method has improved test accuracy by 1-3% when compared to the baseline model, suggesting that it is effective in mitigating overfitting. Y. Yang et al.(21) proposed a two-stage selective ensemble of CNN branches using a novel deep tree training (DTT) approach to address overfitting and the training difficulties of deep CNNs for medical image classification. Zhi-Fei Lai et al.(22) proposes a deep learning framework that integrates high-level features from a deep convolutional neural network with selected traditional features to achieve high classification accuracy on medical image datasets.

The insights from this literature review highlight the need for more effective regularization techniques, setting the stage for the proposed variance penalization approach discussed in the following section.

III. Methodology

The proposed loss function is aimed to reduce overfitting in medical image classification by adding a variance penalty in addition to the conventional loss function. Instead of using binary or categorical cross-entropy, this custom variance penalized can be replaced. This approach works by adding regularization on the model's learned feature representations, discouraging excessive reliance on specific feature patterns and fosters better generalization.

At its core, the function accepts as input the model's predictions as well as the corresponding feature representations. To reduce computational complexity, it samples a subset of features instead of accessing all at once. Then the variance of the sampled features is computed, ensuring that feature importance is more uniformly distributed across different samples. A variance term is added to the classification loss, avoiding the model from becoming unduly sensitive to a small subset of features.

The function can be used with both binary and multi-class classification tasks, allowing it to be versatile for various medical image classification tasks. The final total loss (eqn 1) is a weighted sum of the classification loss and the variance penalty, controlled by three hyper-parameters:

- *Variance Weight (variance_weight):* Determines the strength of the variance penalty. Higher values enforce greater regularization.
- *Feature Sampling Ratio (feature_sampling_ratio):* Controls the proportion of features sampled for variance calculation, balancing computational efficiency and regularization effectiveness.
- *Lambda (lambda_):* A scaling factor for the classification loss, ensuring that the contribution of cross-entropy loss remains balanced with the variance penalty.

Finally, total loss is computed as follows.

$$\begin{aligned}\text{Total Loss} = {} & lambda\_ \times \text{CrossEntropyLoss} \\ & + (\text{variance\_weight} \times \text{FeatureVariance})\end{aligned} \tag{1}$$

By fine-tuning these hyperparameters, the model can achieve better generalization, reducing overfitting while still performing well in classification.

IV. Experiments and Results

To validate the performance and effectiveness of the developed variance-penalized loss function, several experiments were carried out on two publicly accessible medical image datasets, the NIH Chest X-ray dataset(23) and the Skin Cancer MNIST (HAM10000) dataset (24). The purpose of these tests was to evaluate how well the suggested loss function reduced overfitting in comparison to the most common and conventional technique, which is data augmentation.

A. Experiment Set 1 - NIH Chest X-ray Dataset: Binary Classification

The first dataset used was the NIH Chest X-ray dataset which consists of two classes: pneumonia and normal. The preprocessing of the images included resizing to a standard input size, normalization, and contrast enhancement. A baseline convolutional neural network (CNN) model was designed using several convolutional and max-pooling layers and was initially trained. Binary cross-entropy was used as the loss function. In both experiments, early stopping callbacks were implemented. The training process was monitored, and training was terminated if there was no improvement in validation loss over 25 epochs. This ensured that models were not trained unnecessarily beyond the stage of optimal generalization. The Adam optimizer was used as the optimization algorithm in both.

Model Variant	Accuracy	Loss
Baseline Model	0.2984	0.8814
Variance Penalized Model	0.7164	0.3025
Data Augmented Model	0.7099	0.5667

TABLE I: Accuracies and Losses on test set in Experiment Set 1

Comparison	Accuracy Improvement	Loss Reduction
Baseline Model → Variance Penalized Model	+41.8%	-65.70 %
Data Augmented Model → Variance Penalized Model	+0.65%	-46.62%

TABLE II: Accuracy Improvement and Loss Reduction in Experiment Set 1

Initial Model Results (without overfitting prevention):

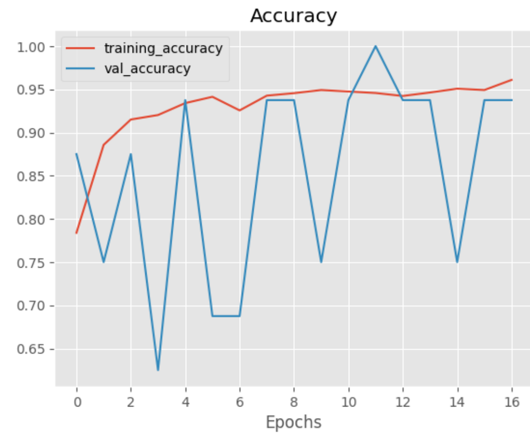

Fig. 1: Chest X-ray Baseline Model - Training/Validation Accuracy Curves

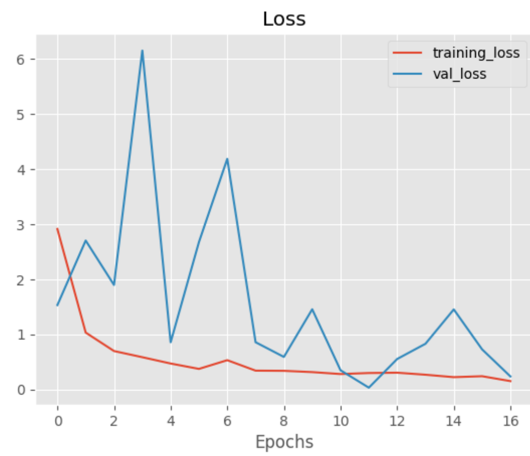

Fig. 2: Chest X-ray Baseline Model - Training/Validation Loss Curves

The figures (Fig. 1 & Fig. 2) clearly showed considerable overfitting, as the model performed well on the training data but poorly on validation and test data.

To address this, the binary cross-entropy loss was replaced with the custom variance-penalized loss function. This function adjusts the loss dynamically according to the variance of extracted features, minimizing dependence on high variance features that may lead to overfitting. However, the hyperparameters variance_weight, feature_sampling_ratio, and lambda_ have a considerable effect on its performance. Instead of manually adjusting them, a grid search was performed across various values to determine the optimal set. The model was then re-trained using these best parameters identified for a total of 20 epochs. The number of epochs was reduced due to limited computational resources. Training can be extended for more epochs based on resource availability.

Variance-Penalized Model Results:

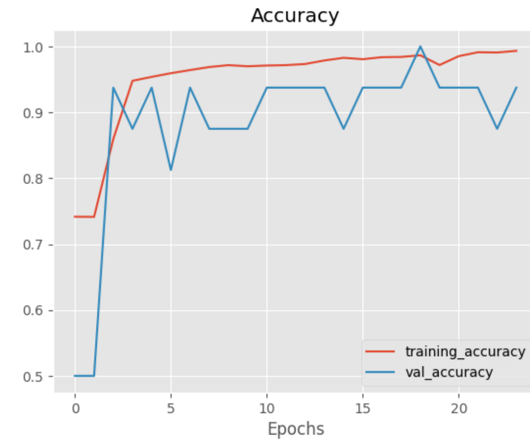

Fig. 3: Chest X-ray Variance Penalized Model - Training/Validation Accuracy Curves

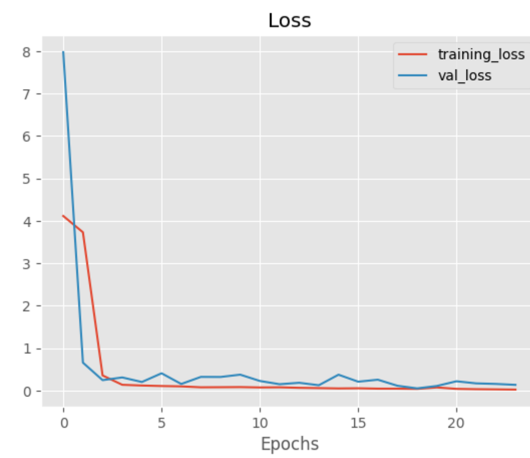

Fig. 4: Chest X-ray Variance Penalized Model - Training/Validation Loss Curves

The figures (Fig. 3 & Fig. 4) and the compared results on TABLE II indicated a notable reduction in overfitting, with considerable enhancements observed in test accuracy and loss. Looking at the graphs shows that overfitting has been considerably minimized, as they illustrate convergence and a narrower gap between training and validation curves.

For further validation, a widely used technique to avoid overfitting, which is data augmentation, was also applied. The training set was expanded by applying two transformations - random zoom (0.1) and random translation (10% shift in both height and width). The same model architecture was trained on these augmented images for 25 epochs, and the accuracy and loss curves were recorded.

Data Augmentation Model Results:

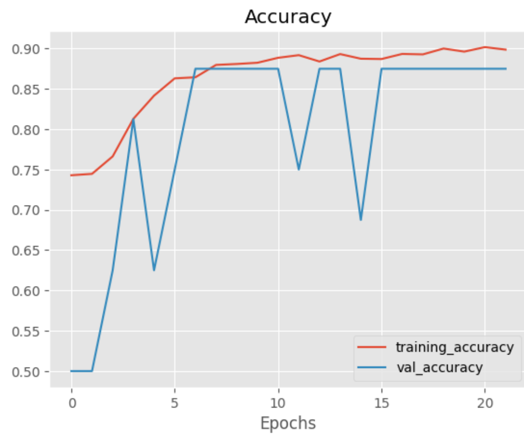

Fig. 5: Chest X-ray Data Augmented Model - Training/Validation Accuracy Curves

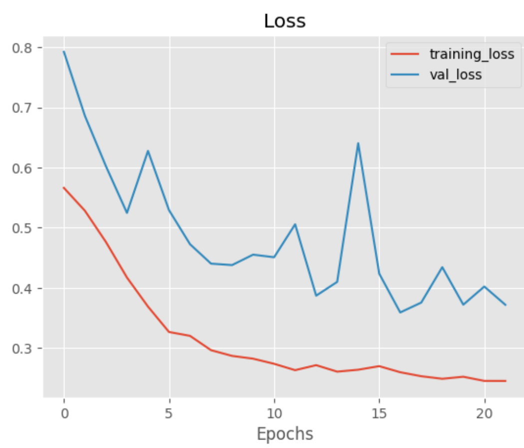

Fig. 6: Chest X-ray Data Augmented Model-Training/Validation Loss Curves

When comparing the obtained results as in TABLE II, the variance-penalized model outperformed the data-augmented model, showing that the proposed loss function is more effective in reducing overfitting.

B. Experiment Set 2 - Skin Cancer Dataset: HAM10000 Multiclass Classification

To further validate the effectiveness of the variance-penalized loss function, a second experiment was carried out using the Skin Cancer MNIST: HAM10000 dataset. This dataset consists of seven classes and has a larger number of images compared to the chest X-ray dataset. After loading, cleaning (Exploratory data analysis - EDA), and preprocessing the images (contrast enhancement, resizing, normalization), a train-test split executed to generate subsets of the data.

Model Variant	Accuracy	Loss
Baseline Model	0.6809	0.8325
Variance Penalized Model	0.8571	0.4167

TABLE III: Accuracies and Losses on test set in Experiment Set 2

Comparison	Accuracy Improvement	Loss Reduction
Baseline Model → Variance Penalized Model	+17.62%	-49.96%

TABLE IV: Accuracy Improvement and Loss Reduction in Experiment Set 2

A CNN model was designed with a deeper architecture than in the first experiment, to capture more complex patterns within this multi-class dataset. Initially, the model was trained for 25 epochs using categorical cross-entropy as the loss function, and the accuracy and loss curves were recorded.

Initial Model Results (without overfitting prevention):

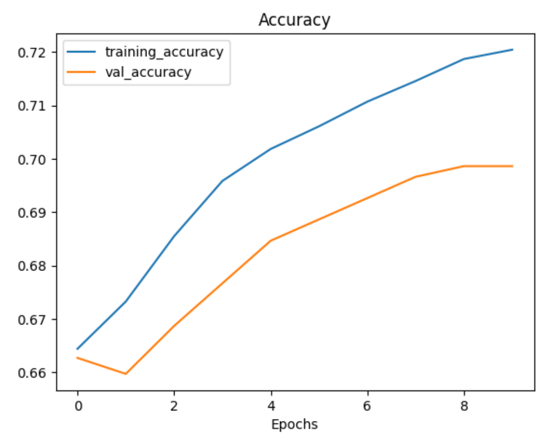

Fig. 7: Skin Cancer Baseline Model - Training/Validation Accuracy Curves

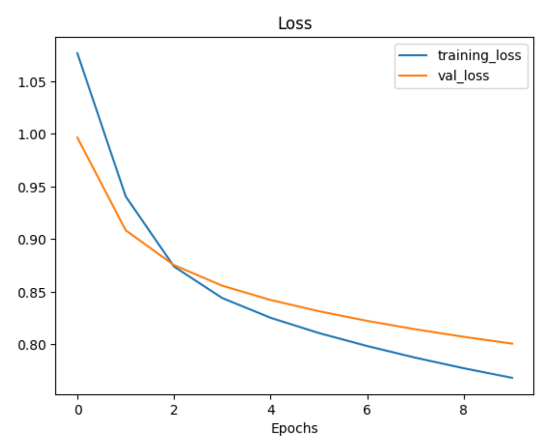

Fig. 8: Skin Cancer Baseline Model - Training/Validation Loss Curves

Similar to the previous dataset, the model showed overfitting, as indicated by the gap between loss and accuracy curves for training and validation phases in Fig. 7 and Fig. 8.

The standard categorical cross-entropy was substituted with the custom variance-penalized loss function, and the best hyper-parameters were determined through grid search as described in experiment set 1. The model was retrained using optimal parameter settings.

Variance-Penalized Model Results:

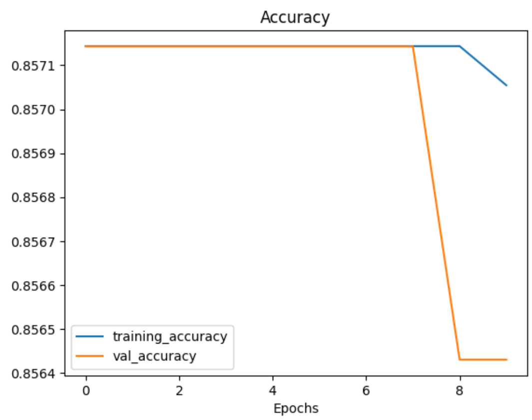

Fig. 9: Skin Cancer Variance Penalized Model - Training/Validation Accuracy Curves

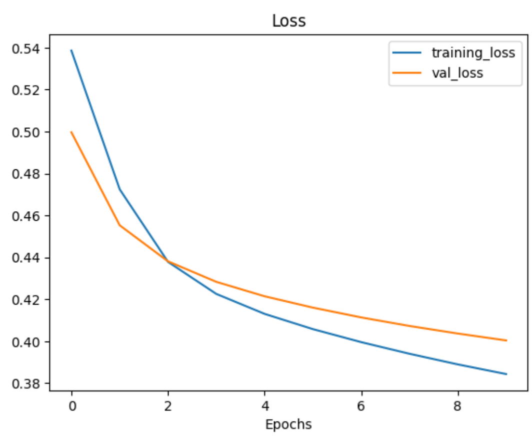

Fig. 10: Skin Cancer Variance Penalized Model - Training/Validation Loss Curves

The variance-penalized model achieved a notable increase in accuracy and a significant reduction in loss, showcasing its capability to reduce overfitting. Analysing the graphs reveals that overfitting has been greatly minimized, as they illustrate convergence and a narrower gap between training and validation performance, further validating its effectiveness. The improvement in accuracy and loss reduction are summarized in TABLE IV.

Due to restricted computational resources, advanced model architectures and complex transfer learning techniques were not used to evaluate the performance of the proposed method. Nevertheless, the proposed method showed promising improvements even with the conventional CNN architectures.

The results across both datasets confirm that the variance-penalized loss function successfully minimizes overfitting and enhances generalization when compared to traditional cross-entropy loss and conventional overfitting prevention techniques like data augmentation.

V. Discussion

The results obtained from the experiments indicate that the proposed variance-penalized loss function effectively reduces overfitting in the medical image segmentation domain. Both datasets - NIH Chest X-ray and HAM10000 Skin Cancer - initially exhibited considerable overfitting, resulting in poor generalization to unseen data. However, after applying the custom loss function, the models showed enhanced accuracy, reduced loss, and more consistent training-validation curves, indicating better generalization.

In contrast to conventional techniques such as data augmentation, the proposed loss function demonstrated better performance, proving its ability to dynamically regulate feature learning. The approach effectively penalized high-variance features, ensuring the model does not overly depend on noise or biases specific to the dataset. Additionally, implementing grid search for hyperparameter tuning and early stopping further enhanced the training stability.

Overall, these findings highlight that the variance-penalized loss function is a promising approach for reducing overfitting in medical image classification tasks, ensuring more reliable and robust model performance across various datasets.

VI. Conclusion and Future Work

This study introduced a variance-penalized loss function to address overfitting in medical image classification, showing its efficacy across different datasets. The proposed approach dynamically modifies the loss depending on feature variance, allowing the model to focus on more reliable patterns instead of random noise.

Although the method has been effective, it does have certain limitations. The method depends on precise tuning of hyperparameters, which may vary from one dataset to another. Additionally, the increased computational overhead from variance calculations could be a concern for large-scale datasets.

As future implementations, this approach could be further improved by including adaptive variance thresholds, thus increasing its independence from specific datasets. Exploring its integration with self-supervised learning or transformer-based architecture may also enhance its ability to generalize. Combining this technique with other regularization methods could further enhance performance. Finally, applying this approach across a wider range of medical imaging applications, such as 3D imaging and multi-modal data, could further validate its effectiveness.

References

[1] E. Ushaa and E. Vishal, "Unlocking clinical insights from medical images using deep learning," *i-manager's Journal on Artificial Intelligence Machine Learning*, vol. 1, p. 37, 2023.

[2] M. Toma and G. Husain, "Algorithm selection and data utilization in machine learning for medical imaging classification," pp. 1–6, 11 2024.

[3] S. Salman and X. Liu, "Overfitting mechanism and avoidance in deep neural networks," 01 2019.

[4] V. Khobragade, J. Nirmal, and S. Chedda, "Revaluating pretraining in small size training sample regime," *International Journal of Electrical and Electronics Research*, vol. 10, pp. 694–704, 09 2022.

[5] S. Piffer, L. Ubaldi, S. Tangaro, A. Retico, and C. Talamonti, "Tackling the small data problem in medical image classification with artificial intelligence: a systematic review," *Progress in Biomedical Engineering*, vol. 6, p. 032001, 06 2024.

[6] D. Barbosa, M. Ferreira, G. B. Junior, M. Salgado, and A. Cunha, "Multiple instance learning in medical images: A systematic review," *IEEE Access*, vol. 12, pp. 78409–78422, 01 2024.

[7] P. I. Khan, A. Dengel, and S. Ahmed, "Medi-cat: Contrastive adversarial training for medical image classification," *arXiv (Cornell University)*, 01 2023.

[8] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep learning for medical image-based cancer diagnosis," *Cancers*, vol. 15, pp. 3608–3608, 07 2023.

[9] E. J. Snider, S. I. Hernandez-Torres, and R. Hennessey, "Using ultrasound image augmentation and ensemble predictions to prevent machine-learning model overfitting," *Diagnostics*, vol. 13, p. 417, 01 2023.

[10] E. Castro, J. C. Pereira, and J. S. Cardoso, "Weight rotation as a regularization strategy in convolutional neural networks," *PubMed*, pp. 2106–2110, 07 2019.

[11] F. Li, L. Tran, K.-H. Thung, S. Ji, D. Shen, and J. Li, "A robust deep model for improved classification of ad/mci patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, pp. 1610–1616, 09 2015.

[12] R. Sangeetha, R. P. Shukla, S. Vats, P. Vishwakarma, and J. Logeshwaran, "Transfer learning for accurate classification of breast cancer in medical imaging," pp. 1–6, 11 2023.

[13] A. Al-Qerem, A. A. Salem, I. Jebreen, A. Nabot, and A. Samhan, "Comparison between transfer learning and data augmentation on medical images classification," *2021 22nd International Arab Conference on Information Technology (ACIT)*, pp. 1–7, 12 2021.

[14] W. A. Al and I. D. Yun, "Reinforcing medical image classifier to improve generalization on small datasets," *arXiv (Cornell University)*, 01 2019.

[15] H. Li, G. K. Rajbahadur, D. Lin, C.-P. Bezemer, and Z. M. Jiang, "Keeping deep learning models in check: A history-based approach to mitigate overfitting," *IEEE access*, pp. 1–1, 01 2024.

[16] B. Harangi, A. Baran, M. Beregi-Kovacs, and A. Hajdu, "Composing diverse ensembles of convolutional neural networks by penalization," *Mathematics*, vol. 11, p. 4730, 11 2023.

[17] D. Shubin, D. Eytan, and S. D. Goodfellow, "About explicit variance minimization: Training neural networks for medical imaging with limited data annotations," *arXiv (Cornell University)*, 01 2021.

[18] B. Qiu, D. Raicu, J. Furst, and R. Tchoua, "Compnet: A designated model to handle combinations of images and designed features," *arXiv (Cornell University)*, 01 2022.

[19] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. Van Beeke, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batrah, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Casteele, S. Gupte, M. Sallamm, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, and B. Y. Croft, "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, pp. 915–31, 2011.

[20] B. Simpson, F. Dutil, Y. Bengio, and J. P. Cohen, "Gradmask: Reduce overfitting by regularizing saliency.," 04 2019.

[21] Y. Yang, Y. Hu, X. Zhang, and S. Wang, "Two-stage selective ensemble of cnn via deep tree training for medical image classification," *IEEE Transactions on Cybernetics*, vol. 52, p. 9194–9207, 09 2022.

[22] Z. Lai and H. Deng, "Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–13, 09 2018.

[23] "Nih chest x-ray dataset — cloud healthcare api."

[24] P. Tschandl, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Harvard Dataverse*, 01 2018.