# Addressing Overfitting in Medical Image Classification through Variance Penalization

**Research and Design Project (EE 4301)**

Department of Electrical and Electronic Engineering

University of Sri Jayewardenepura

Batch 5

**S.A.A.J.M. Athukorala**

20/ENG/007

## 1.0 Background and Motivation

Medical image classification is a crucial application of deep learning in healthcare, assisting in the diagnosis and treatment planning of various diseases. Convolutional Neural Networks (CNNs) come into place and have demonstrated considerable performance in medical image analysis, enabling automated detection of abnormalities in X-rays, MRIs, CT scans, and histopathological images. However, one of the major challenges in medical image classification is overfitting, where a model learns patterns specific to the training dataset but fails to generalize well to unseen data.

Unlike any other images, medical images contain many anatomical structures and clinically significant features that can be easily distorted by excessive preprocessing, aggressive data augmentation. Existing techniques such as dropout layer initialization, batch normalization, enabling callbacks while training and data augmentation have been widely applied to mitigate overfitting, but they come with limitations. For instance, some augmentation techniques can introduce unrealistic variations that do not exist in real clinical scenarios. Addressing overfitting without compromising essential medical features remains a critical research problem.

Standard loss functions like binary cross-entropy or categorical cross-entropy optimize prediction accuracy but do not consider the impact of feature variance in convolutional layers. When variance is high across multiple layers, the model may be relying on unstable or irrelevant features, leading to poor generalization.

By penalizing high variance in convolutional layers, the model is encouraged to focus on stable, generalizable features rather than noise or dataset-specific patterns. This technique provides an alternative to existing regularization methods.

In clinical environments, unreliable AI systems can lead to incorrect diagnoses, which may have serious implications for patient care. The method proposed enhances the reliability of AI-assisted diagnostics by improving generalization. It also has the potential to drive further advancements in feature selection and model optimization.

By focusing on these essential factors, the study seeks to provide a practical, scalable, and clinically applicable solution to improve the dependability of deep learning models in medical imaging use cases.

## 2.0 Description of the proposed solution

The Optimized Variance-Penalized loss function aims to reduce overfitting by adaptively penalizing features with high variance within convolutional layers. Rather than using binary or categorical cross-entropy, this tailored variance penalized loss function ought to be utilized instead. Key features can be briefly explained as below.

- Variance Calculation for Feature Stability – The function processes input features by separating them from the final predictions. It computes the mean and variance of a sampled subset of features.
- Variance Penalization Mechanism –

A variance weight parameter controls the degree to which high-variance features are penalized. The total loss is computed as,

$$Total\ Loss = lambda \times CrossEntropy\ Loss + (variance\_weight\ \times Feature\ Variance)$$

When variance is low, the model focuses more on these features; when variance is high, the loss increases, thereby diminishing focus on those features.

- Sampling Procedure for computational efficiency – Instead of using all feature dimensions (which can be computationally expensive), a feature sampling ratio is introduced. This ratio determines how many features are randomly selected for variance computation.
- Applicability to binary and multi-class classification – The function supports both binary and categorical cross-entropy loss, allowing it to be versatile for various medical image classification tasks.

### *2.1 Hyperparameters in the developed custom loss function*

The proposed loss function from the study incorporates three critical hyperparameters that control its behavior.

1. *variance_weight* – This parameter determines the strength of the variance penalty. Higher values increase penalization of high-variance features, effectively discouraging the model from relying on these potentially noisy or overfitted representations. Conversely, lower values reduce this penalty.

2. *feature_sampling_ratio* – This parameter controls the proportion of features sampled for variance calculation. It has two main functions - it minimizes computational complexity by working with only a limited number of features and introduces randomness into the learning process, as various features are randomly chosen in every iteration.

3. *lambda* – This balancing factor controls the relative importance of the classification loss compared to the variance penalty.

### *2.2 Novelty of the Approach*

Unlike conventional loss functions that do not explicitly consider feature stability, this function integrates variance penalization dynamically into the training process. Also, this method prevents CNN models from fixating on highly variant, dataset-specific patterns, thereby enhancing generalization to unseen medical images.

## 3.0 Literature Survey

### *3.1 Background*

Deep learning has achieved great success 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 poses challenges in rare cancer diagnosis, multimodal image fusion, model explainability, 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].

### *3.2 Data augmentation and dropout techniques*

Eric J. Snider et al. [18] describes data augmentation techniques, such as affine transformations and MixUp, improved the generalizability of the machine learning models for shrapnel detection in ultrasound images, even though they reduced the training accuracy improved the model's blind test accuracy from 68% to over 85%. Eduardo Castro et al. [20] 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. [9] presented that dropout technique has been particularly effective in Alzheimer's disease diagnosis, improving classification accuracies by 5.9% compared to classical deep learning methods. The researchers also incorporated other techniques, such as stability selection, adaptive learning, and multitask learning, into the deep learning framework to further improve its performance.

### *3.3 Transfer learning techniques*

R. Sangeetha et al. [21] 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. [22] 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 performance accuracy compared to the data augmentation approach.

### *3.4 Variance penalization and novel regularization techniques*

Walid Abdullah Al & I. Yun [10] proposed a reinforced classifier using generalization-feedback from a subset of training data has shown promise in improving generalization on small datasets. The reinforced classifier was evaluated on three different classification problems and outperformed standard deep classifiers with overfitting prevention techniques. Hao Li et al. [11] founded a history-based approach can both detect and prevent overfitting in deep learning models without modifying the model structure. The proposed approach achieves an F1 score of 0.91 for detecting overfitting, which is at least 5% higher than the current best-performing non-intrusive overfitting detection approach. Also, it can stop training to avoid overfitting at least 32% of the times earlier than early stopping, while maintaining the same or better rate of returning the best model. Harangi et al. [12] 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. [13] introduced Variance Aware Training (VAT), which explicitly minimizes 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 an order of magnitude reduction in the GPU training. Qiu et al. [14] developed CompNet, a CNN-based model that combines image and designed features, significantly reducing overfitting in medical image datasets. The CompNet model outperformed other similar approaches that combine images and designed features, both on the LIDC dataset and on the datasets used in other studies. Simpson et al. [15] presented GradMask, a new regularization method that penalizes saliency maps when they are not consistent with the actual lesion segmentation, preventing the model from incorrectly associating non-tumor related features with the classification of unhealthy samples. The paper demonstrates that using the GradMask method can improve test accuracy by 1-3% compared to the baseline model, indicating that it is effective at reducing overfitting. Y. Yang et al. [17] proposed a two-stage selective ensemble of CNN branches using a novel deep tree training (DTT) strategy to address overfitting and the training difficulties of deep CNNs for medical image classification. Zhi-Fei Lai et al. [19] proposes a deep learning model that combines high-level features from a deep convolutional neural network with selected traditional features to achieve high classification accuracy on medical image datasets.

## 4.0 Methodology

The process begins with preprocessing, and initial model training to establish baseline performance metrics. The next phase involves implementing the custom loss function, tuning hyperparameters through grid search, and comparing results against existing overfitting prevention methods to further validation.

### *4.1 Datasets*

Two distinct medical image datasets were chosen to ensure the robustness and generalizability of the findings. The NIH Chest X-ray Dataset served as the primary testbed for initial experimentation due to its well-defined binary classification task, containing normal and pneumonia X-ray images sourced from Kaggle. For more complex validation, the Skin Cancer MNIST (HAM10000) dataset was chosen, comprising seven different skin lesion categories.

### *4.2 Data Preprocessing*

All images were resized to a common dimension to ensure uniform input to the neural networks. Pixel values were normalized to the range [0,1] to improve training stability and convergence. Each dataset was then partitioned into training, validation, and test sets using stratified sampling to maintain class distribution across all subsets.

### *4.3 Model Architecture*

For each dataset, a convolutional neural network (CNN) architecture was designed to match the complexity of the classification task. The Chest X-ray model consists of relatively simple CNN comprising convolutional layers, max-pooling layers, and dense layers, appropriate for the binary classification task. In contrast, the Skin cancer model featured a more complex architecture with additional convolutional and pooling layers. Both architectures used ReLU activation functions for intermediate layers and appropriate activation functions for the output layer (sigmoid for binary classification, softmax for multiclass classification).

### *4.4 Early Stopping Implementation*

To ensure fair comparison and prevent unnecessary computation, early stopping was implemented across all experimental conditions. Validation loss was continuously monitored during training, with training terminated if validation loss failed to improve for a specified number of consecutive epochs.

### *4.5 Baseline Model Training*

Initial models were trained without any overfitting prevention techniques. Standard cross-entropy loss functions were used - binary cross-entropy for the chest X-ray dataset and categorical cross-entropy for the skin cancer dataset. All models utilized the Adam optimizer with a learning rate of 0.001. In each setting both chest X-ray models, and  skin cancer models ran for 25 epochs.

### *4.6 Hyperparameter Optimization for Variance-Penalized Loss*

A grid search approach was employed to identify optimal hyperparameters for the proposed variance-penalized loss function. A predefined grid of potential values was initialized for three key parameters - variance_weight, feature_sampling_ratio, and lambda_. For each parameter combination, models were run for 20 epochs to assess performance without excessive computational cost. Parameter sets were evaluated based on validation performance.

### *4.7 Custom accuracy metric implementation*

To properly evaluate models trained with the custom loss function, a specialized accuracy metric was developed, because the variance-penalized loss function operates on a concatenated tensor containing both feature activations and predictions.

*4.8 Comparative experiments*

The Baseline model served as a control, trained without any specific overfitting prevention techniques except early stopping callbacks. For comparison with conventional techniques, a Data augmentation model was implemented using random zoom (±10%) and random translation (height and width shifts of up to 10%) to artificially expand the training dataset.

*4.9 Evaluation procedure*

A thorough evaluation step was implemented to assess model performance across all conditions. Training and validation accuracy/loss curves were plotted and analyzed to identify overfitting patterns and convergence behavior. Final models were evaluated on the test set to assess generalization performance.

## 5.0 Results and Discussion

For each scenario, the model was evaluated on the test set which is unseen data. Table 1 shows the accuracy, loss and the improvement of accuracy with the loss reduction as percentages with compared to the baseline results.

*Table 1: results comparison*

| Dataset | Method | Test accuracy | Test loss | Accuracy improvement | Loss reduction |
|---|---|---|---|---|---|
| Chest X-ray | Baseline | 0.2984 | 0.8814 | - | - |
| Chest X-ray | Variance penalized | 0.7164 | 0.3025 | +41.80% | -65.70% |
| Chest X-ray | Data augmented | 0.7099 | 0.5667 | +41.15% | -35.70% |
| Skin cancer | Baseline | 0.6809 | 0.8325 | - | - |
| Skin cancer | Variance penalized | 0.8571 | 0.4167 | +17.62% | -49.94% |

*5.1 Chest X-ray Classification Results*

*5.1.1 Baseline Model Performance*

The baseline model, trained without specific overfitting prevention techniques, exhibited clear signs of overfitting. The training accuracy increased steadily while validation accuracy plateaued and eventually decreased. This performance confirms the presence of severe overfitting by looking at the plots.

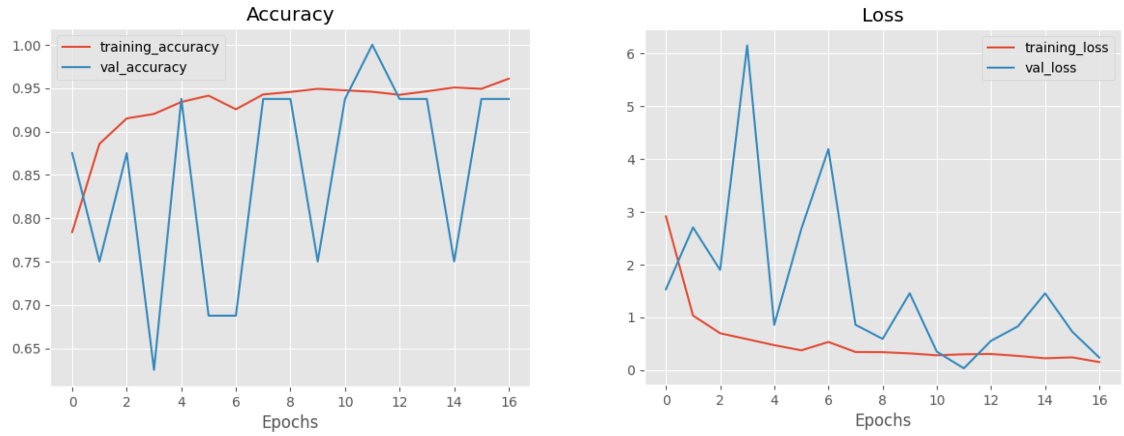

*Figure 1: Chest X-ray - baseline model performance curves*

### *5.1.2 Variance-Penalized Model Performance*

After implementing the proposed variance-penalized loss function with optimized hyperparameters, a significant improvement in model performance was observed. The training and validation curves showed markedly smaller gaps compared to the baseline model, indicating substantially reduced overfitting.

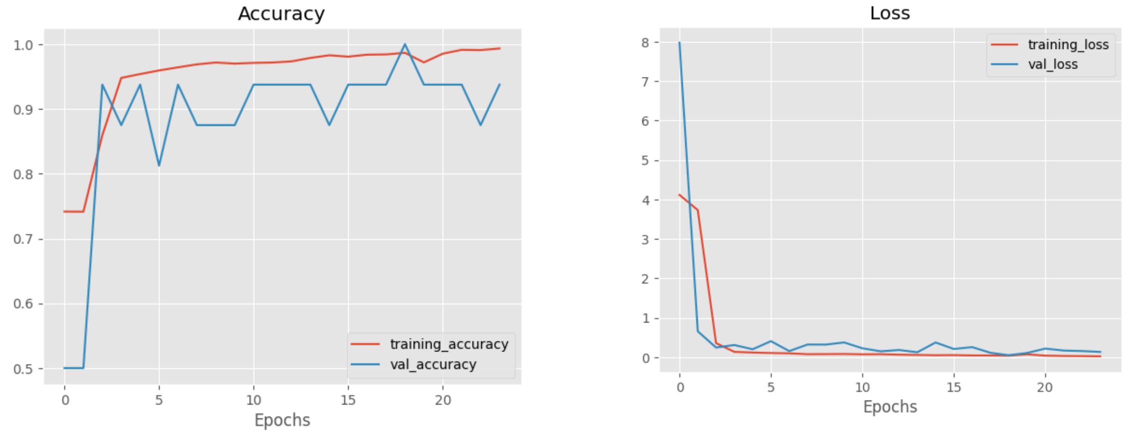

*Figure 2: Chest X-ray - variance penalized model performance curves*

### *5.1.3 Data Augmentation Comparison*

The data augmentation approach, utilizing random zoom and translation, also improved performance compared to the baseline.

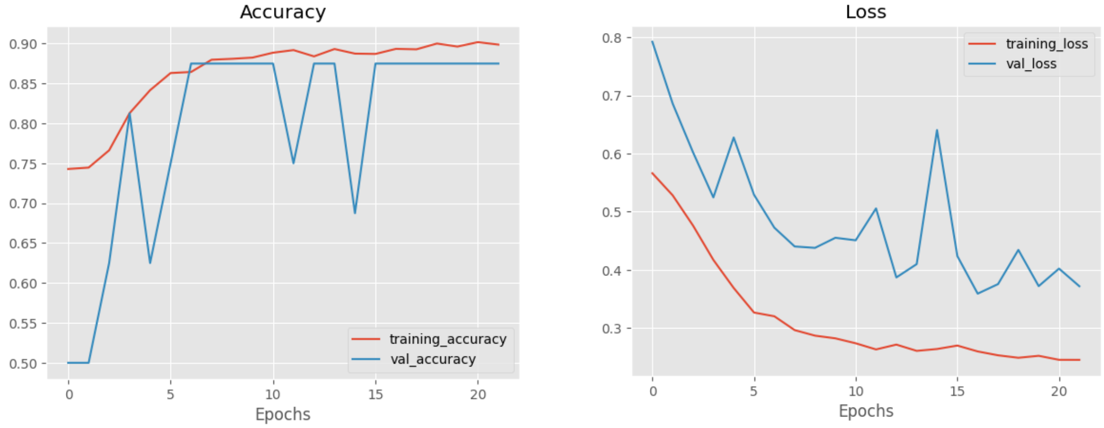

*Figure 3: Chest X-ray - data augmented model performance curves*

## *5.2 Skin Cancer Classification Results*

### *5.2.1 Baseline Model Performance*

The baseline model for the more complex skin cancer classification task also showed overfitting. The divergence between training and validation curves indicated significant overfitting issues that limited the model's ability to generalize.

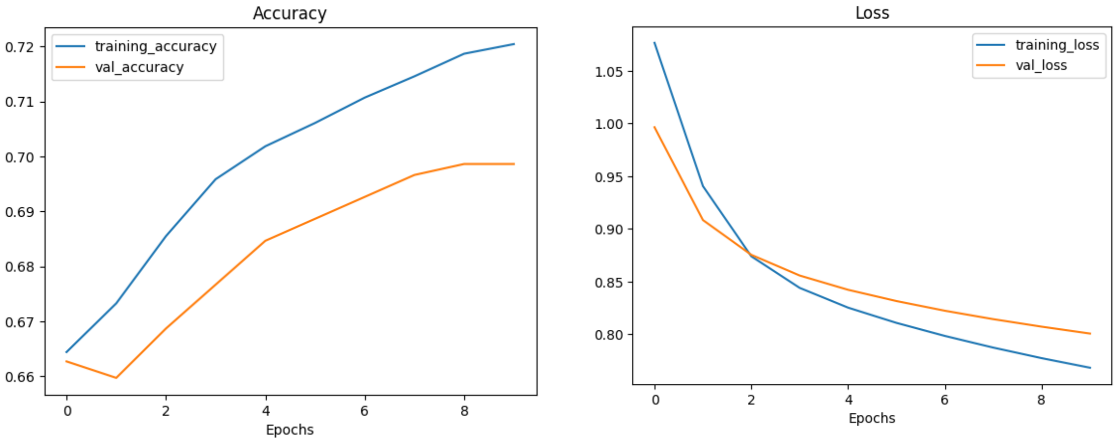

*Figure 4: Skin cancer - baseline model performance curves*

### *5.2.2 Variance-Penalized Model Performance*

The convergence patterns of training and validation curves showed significantly reduced overfitting, with the model maintaining consistent performance across both training and validation sets throughout the training process.

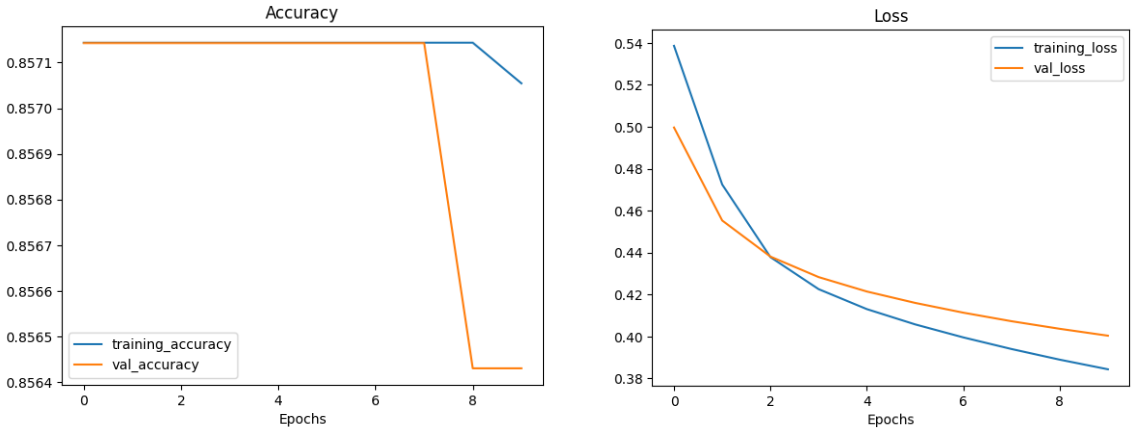

*Figure 5: Skin cancer - variance penalized model performance curves*

*5.3 Discussion of Findings*

The experimental results support several important conclusions about the proposed variance-penalized loss function. First, the consistent performance improvements across two distinct medical imaging datasets suggest that the approach is particularly well-suited to medical image classification tasks. Second, the feature sampling mechanism effectively reduces computational requirements while maintaining performance benefits. Third, the approach demonstrated effectiveness across different model architectures and classification tasks (binary and multiclass), highlighting its flexibility.

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.

## 6.0 Conclusion

This research successfully developed and validated a variance-penalized loss function that effectively addresses overfitting in medical image classification models. The key findings demonstrate that the developed function can outperform traditional approaches such as data augmentation in preventing overfitting while maintaining high model accuracy.

Future research directions could include testing the approach on larger and more diverse medical datasets, optimizing the hyperparameter selection process, inspecting the interpretability of features with different variance patterns, and combining this technique with other regularization methods for potentially greater performance improvements.

## 7.0 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, no. 2, p. 37, 2023, doi: https://doi.org/10.26634/jaim.1.2.20044.

[2] M. Toma and G. Husain, “Algorithm Selection and Data Utilization in Machine Learning for Medical Imaging Classification,” pp. 1–6, Nov. 2024, doi: https://doi.org/10.1109/lisat63094.2024.10807895.

[3] S. Salman and X. Liu, “Overfitting Mechanism and Avoidance in Deep Neural Networks,” *arXiv.org*, Jan. 19, 2019. https://arxiv.org/abs/1901.06566

[4] P. I. Khan, A. Dengel, and S. Ahmed, “Medi-CAT: Contrastive Adversarial Training for Medical Image Classification,” *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2311.00154.

[5] V. Khobragade, J. Nirmal, and S. Chedda, “Revaluating Pretraining in Small Size Training Sample Regime,” *International Journal of Electrical and Electronics Research*, vol. 10, no. 3, pp. 694–704, Sep. 2022, doi: https://doi.org/10.37391/ijeer.100346.

[6] 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, no. 3, p. 032001, Jun. 2024, doi: https://doi.org/10.1088/2516-1091/ad525b.

[7] 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, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3403538.

[8] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, “Deep Learning for Medical Image-Based Cancer Diagnosis,” *Cancers*, vol. 15, no. 14, pp. 3608–3608, Jul. 2023, doi: https://doi.org/10.3390/cancers15143608.

[9] 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, no. 5, pp. 1610–1616, Sep. 2015, doi: https://doi.org/10.1109/jbhi.2015.2429556.

[10] W. A. Al and I. D. Yun, “Reinforcing Medical Image Classifier to Improve Generalization on Small Datasets,” *arXiv (Cornell University)*, Jan. 2019, doi: https://doi.org/10.48550/arxiv.1909.05630.

[11] H. Li, Gopi Krishnan Rajbahadur, D. Lin, Cor-Paul Bezemer, and Zhen Ming Jiang, “Keeping Deep Learning Models in Check: A History-Based Approach to Mitigate Overfitting,” *IEEE access*, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3402543.

[12] B. Harangi, A. Baran, M. Beregi-Kovacs, and A. Hajdu, “Composing Diverse Ensembles of Convolutional Neural Networks by Penalization,” *Mathematics*, vol. 11, no. 23, p. 4730, Nov. 2023, doi: https://doi.org/10.3390/math11234730.

[13] 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)*, Jan. 2021, doi: https://doi.org/10.48550/arxiv.2105.14117.

[14] B. Qiu, D. Raicu, J. Furst, and R. Tchoua, “CompNet: A Designated Model to Handle Combinations of Images and Designed features,” *arXiv (Cornell University)*, Jan. 2022, doi: https://doi.org/10.48550/arxiv.2209.14454.

[15] B. Simpson, F. Dutil, Yoshua Bengio, and J. P. Cohen, “GradMask: Reduce Overfitting by Regularizing Saliency.,” Apr. 2019.

[16] A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, “An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification,” *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 31–40, Jan. 2017, doi: https://doi.org/10.1109/jbhi.2016.2635663.

[17] 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, no. 9, pp. 9194–9207, Sep. 2022, doi: https://doi.org/10.1109/TCYB.2021.3061147.

[18] 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, no. 3, p. 417, Jan. 2023, doi: https://doi.org/10.3390/diagnostics13030417.

[19] 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, Sep. 2018, doi: https://doi.org/10.1155/2018/2061516.

[20] E. Castro, J. C. Pereira, and J. S. Cardoso, “Weight Rotation as a Regularization Strategy in Convolutional Neural Networks,” *PubMed*, pp. 2106–2110, Jul. 2019, doi: https://doi.org/10.1109/embc.2019.8856448.

[21] R. Sangeetha, R. P. Shukla, Satvik Vats, Pramod Vishwakarma, and J. Logeshwaran, “Transfer Learning for Accurate Classification of Breast Cancer in Medical Imaging,” pp. 1–6, Nov. 2023, doi: https://doi.org/10.1109/rmkmate59243.2023.10368665.

[22] A. Al-Qerem, A. A. Salem, Issam 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, Dec. 2021, doi: https://doi.org/10.1109/acit53391.2021.9677144.