AMS-ETL: An Adaptive Multi-Source Ensemble Transfer Learning Framework for Robust Multi-Disease Diagnostic Classification

Kishore Babu Nampalle[*1] NKISHORE@IIITRANCHI.AC.IN
[1] *Indian Institute of Information Technology Ranchi, Ranchi, India.*
Dhiran Kumar Mahto[*1] DHIRAN.MAHTO@IIITRANCHI.AC.IN
[1] *Indian Institute of Information Technology Ranchi, Ranchi, India.*

Abstract

Deep learning methods for medical image analysis, while powerful, often face performance limitations due to dataset constraints, including small size, quality variability, and heterogeneous backgrounds. This is particularly critical in diagnosing severe conditions like skin cancer and brain tumors, where predictive reliability is paramount. Current approaches leveraging single-model transfer learning can struggle with generalization, while conventional ensembles often aggregate pre-trained models without optimizing their complementary strengths for specific clinical imaging characteristics. To address this, we propose Adaptive Multi-Source Ensemble Transfer Learning (AMS-ETL), a framework that strategically integrates diverse pre-trained architectures and employs a meta-learning strategy for dynamic source-weighting. Our method uniquely tailors the ensemble composition by evaluating each model's discriminative capacity for specific feature patterns in clinical images, moving beyond simple performance averaging. We implement this using a foundational MobileNet feature extractor combined with auxiliary sources, processed through a gated logistic regression meta-learner for final prediction. Validation on clinical dermatology and neuroimaging datasets demonstrates that AMS-ETL achieves state-of-the-art accuracy and significantly improves robustness against overfitting. Furthermore, our model provides enhanced feature diversity and discriminative interpretability, offering clinicians granular decision support. This work establishes that adaptive, source-aware ensemble design is crucial for advancing automated, reliable diagnostic frameworks in resource-constrained clinical environments.

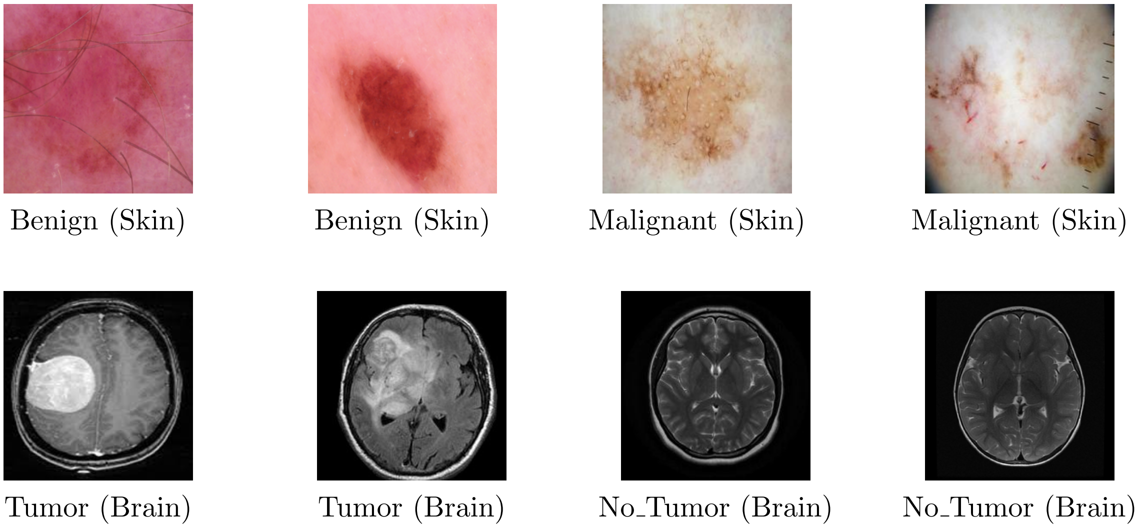
Keywords: Adaptive Ensemble Learning, Transfer Learning, Medical Image Classification, Multi-Disease Diagnosis, Convolutional Neural Networks (CNN), Robust Classification

1. Introduction

Cutaneous malignancies, notably skin cancer, represent one of the most rapidly accelerating oncological burdens, with documented annual incidence increases of 3–7

The advent of deep learning has transformed diagnostic workflows, enabling robust computer-aided diagnosis (CAD) systems capable of tasks such as lesion segmentation and disease classification with performance that increasingly rivals expert clinicians (Wang et al., 2019; Chen et al., 2022). Each primary imaging modality offers distinct trade-offs: CT

[*]


Benign (Skin) Benign (Skin) Malignant (Skin) Malignant (Skin)

Tumor (Brain) Tumor (Brain) No_Tumor (Brain) No_Tumor (Brain)

Figure 1: Samples of skin images (Malignant and Benign) and brain images (Tumor and No_Tumor).

provides high-resolution visualization of dense tissues; X-ray remains the most accessible and cost-effective screening tool (Akcay and Breckon, 2022); MRI delivers superior soft-tissue contrast without ionizing radiation but with longer acquisition times (Ware et al., 2022); and US combines real-time capability, safety, and operational versatility (Adeyemi et al., 2022).

The automated evaluation of skin lesions—morphologically diverse abnormalities of the integumentary system (Pewton and Yap, 2022)—presents a significant diagnostic challenge. Computer-aided diagnosis (CAD) has evolved from semi-automatic systems reliant on expert intervention (Esteva et al., 2017; Adegun and Viriri, 2021) to deep learning frameworks capable of learning hierarchical features directly from images (Chen et al., 2022). Convolutional neural networks (CNNs) now underpin this shift, demonstrating expert-level proficiency in tasks like melanoma detection (Maron et al., 2022).

To overcome persistent constraints of limited data and heterogeneity, we propose a synergistic framework integrating transfer learning with ensemble methods. The architecture utilizes a pre-trained MobileNet as a feature extractor, with its final layer adapted for the target task. Transfer learning applies knowledge from large-scale natural image datasets to the medical domain, reducing dependency on extensive annotated data (Karimi et al., 2021). Ensemble learning (Dong et al., 2020) further bolsters robustness by aggregating predictions from multiple homogeneous base classifiers, mitigating overfitting and variance, with computational overhead managed through optimized implementations.

The principal contributions of this work are enumerated as follows:

- We propose a novel, end-to-end deep ensemble framework that integrates transfer learning for the automated binary classification of medical images. This hybrid approach is designed to enhance predictive performance while explicitly addressing the data-scarcity challenge common in clinical settings.

- The framework employs ensemble learning to stabilize performance against dataset variations and class imbalance, achieving superior optimization and effectively curtailing overfitting.

- A key methodological innovation involves performing high-level feature extraction prior to partitioning the data into training, validation, and test sets. Although unconventional in medical imaging pipelines, this strategy demonstrably improved experimental outcomes in our study, suggesting a promising direction for future methodological development.

- We utilize bagging as our ensemble meta-algorithm, capitalizing on its inherent capacity for parallelization on multi-core systems. This design achieves an optimal balance between classification accuracy and computational efficiency.

The remainder of this paper is structured as follows: Section 2 presents a review of related research in the field of medical image processing. Section 3 describes the dataset, preprocessing steps, and the proposed methodology in detail. The experimental setup and results are discussed in Section 4, followed by an in-depth discussion and conclusion in Section 5 and Section 6, respectively. Finally, potential directions for future research are also highlighted.

2. Related work

A seminal study demonstrated that models can achieve correct predictions through spurious correlations, severely undermining reliability and interpretability in safety-critical domains like medical diagnosis (Ribeiro et al., 2016; Khan et al., 2020). Concurrently, the clinical adoption of deep learning (DL) is hindered by the "black-box" nature of most architectures, which lack transparent decision-making rationale (Sunija et al., 2021). This has spurred the development of explainable AI (XAI), mandated by ethical guidelines and implemented via techniques like feature attribution maps and concept activation analyses (Arrieta et al., 2020; Fong et al., 2019; Wang et al., 2020; Kim et al., 2018).

Automated melanoma detection is a critical application within this context. As the deadliest skin cancer, melanoma causes the majority of related deaths despite its relatively low incidence (OMS, 2022). Clinical diagnosis often relies on subjective heuristics like the ABCD rule (Lattoofi et al., 2019) or the 7-point checklist (Kawahara et al., 2018), creating a need for objective, AI-driven analysis to improve consistency and accuracy (Bajwa et al., 2020).

The evolution of computer-aided diagnosis (CAD) systems has transitioned from traditional machine learning, reliant on handcrafted features (Almaraz-Damian et al., 2016; Adegun and Viriri, 2021; Hajdu et al., 2016), to deep learning. Architectures such as AlexNet, VGG, GoogLeNet, ResNet, and Xception have established dominance through superior, learned hierarchical representations (Krizhevsky et al., 2017; Simonyan and Zisserman, 2014; Szegedy et al., 2015; Bi et al., 2017; Chollet, 2017). To enhance performance further, ensemble learning has been explored, from early combinations of networks like CaffeNet and U-Net (Codella et al., 2017) to search-based ensembles (Gessert et al., 2020) and mutual bootstrapping methods (Xie et al., 2020). Recent sophisticated approaches, including wavelet-based networks and transformer models, often face challenges with computational complexity and limited generalizability across datasets (Shetty et al., 2022; Alenezi et al., 2023; Nakai et al., 2022).

Motivated by the complementary strengths of transfer learning for data efficiency and ensemble learning for predictive stability, this work proposes a streamlined architecture. It synthesizes a pre-trained MobileNet feature extractor with logistic regression classifiers within a homogeneous bagging ensemble framework, employing bootstrap aggregation and majority voting to capitalize on variance reduction for robust medical image classification.

3. Methodology

The proposed framework is formulated as a composite mapping $\mathcal{F} : \mathcal{I} \to \mathcal{Y}$ from raw medical images $\mathcal{I}$ to diagnostic labels $\mathcal{Y}$, structured for optimal performance under computational constraints:

$$\mathcal{F}(I) = \mathcal{A}\left(\{g_{\phi_k}\left(f_\theta\left(\Phi(I)\right)\right)\}_{k=1}^{K}\right), \tag{1}$$

where Φ denotes the pre-processing pipeline, f_θ is the pre-trained MobileNet feature extractor, $\{g_{\phi_k}\}$ are the base classifiers, and $\mathcal{A}$ is the ensemble aggregation operator.

3.1. Pre-processing Pipeline

The pre-processing transform Φ operates sequentially on raw images I_{raw}:

$$\Phi = \Phi_{\text{encode}} \circ \Phi_{\text{filter}} \circ \Phi_{\text{enhance}} \circ \Phi_{\text{crop}}. \tag{2}$$

Lesion-Centric Cropping: Isolates the region of interest via bounding box $\mathcal{B}$:

$$I_{\text{crop}}(x', y', c) = I_{\text{raw}}(x' + x_{\min}, y' + y_{\min}, c). \tag{3}$$

Contrast Enhancement: Histogram equalization improves global contrast:

$$I_{\text{eq}}(x, y, c) = \left\lfloor (L-1) \cdot \sum_{i=0}^{I_{\text{crop}}(x,y,c)} p_c(i) \right\rfloor. \tag{4}$$

Artifact Removal: Combines median filtering, Gaussian smoothing, and morphological operations:

$$I_{\text{med}}(x, y, c) = \underset{(i,j)\in\Omega_{x,y}}{\text{median}} \left(I_{\text{eq}}(i, j, c)\right), \tag{5}$$

$$I_{\text{filt}} = I_{\text{med}} * G_\sigma, \tag{6}$$

$$I_{\text{morph}} = (I_{\text{filt}} \ominus S) \oplus S. \tag{7}$$

Label Encoding: Maps categorical labels to numerical space: $\mathcal{Y}_{\text{num}} = \phi_{\text{label}}(\mathcal{Y}_{\text{cat}})$.

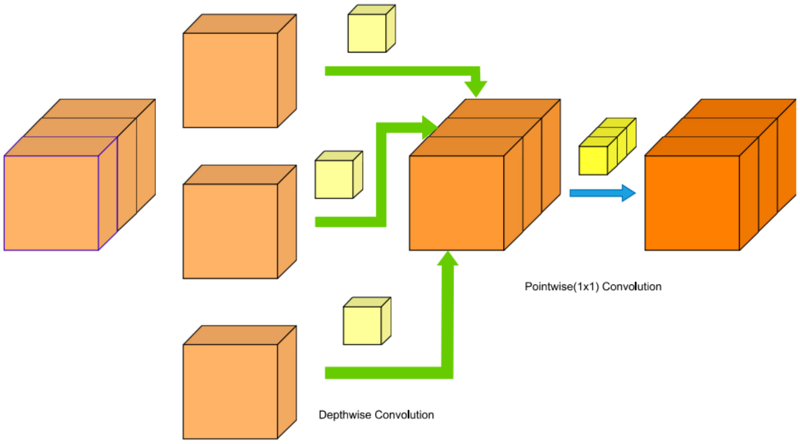


Figure 2: Schematic representation ofdepthwise separable convolution

3.2. Feature Extraction with MobileNet

The feature extractor $f_\theta : \mathbb{R}^{224\times224\times3} \to \mathbb{R}^d$ employs depthwise separable convolution (Howard et al., 2017):

$$\text{Depthwise:} \quad G_{\text{dw}}(x,y,n) = \sum_{i,j} \hat{K}(i,j,n) \cdot F(x+i, y+j, n), \tag{8}$$

$$\text{Pointwise:} \quad G_{\text{pw}}(x,y,m) = \sum_{n=1}^{N} P(1,1,n,m) \cdot G_{\text{dw}}(x,y,n). \tag{9}$$

This factorization reduces computational cost by approximately 8–9× compared to standard convolution. The extracted feature vector $z = f_\theta(\Phi(I))$ serves as input to the ensemble.

3.3. Bagging Ensemble Classification

A homogeneous bagging ensemble of K logistic regression classifiers is employed. Each classifier g_{ϕ_k} is trained on a bootstrap sample $\mathcal{D}_k$. Final prediction via majority voting:

$$\hat{y} = \arg\max_{c\in\mathcal{Y}} \sum_{k=1}^{K} \mathbb{I}\left(g_{\phi_k}(z) = c\right), \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This approach reduces variance and enhances generalization.

3.4. Logistic Regression as Base Learner

Logistic regression is employed as the base learner within our ensemble framework due to its computational efficiency, robustness to overfitting on high-dimensional extracted features, and well-calibrated probabilistic outputs. Given a feature vector $\mathbf{z} \in \mathbb{R}^d$ extracted by the MobileNet backbone, the model computes the probability of the positive class via the sigmoid function $\sigma(\cdot)$:

$$P(y = 1 \mid \mathbf{z}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{z}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{z})}, \tag{11}$$

where $\mathbf{w} \in \mathbb{R}^d$ are the learned model weights. The predicted class label $\hat{y}$ is obtained by applying a decision threshold (typically 0.5) to this probability.

Training involves maximizing the log-likelihood over N training samples $\{(\mathbf{z}_i, y_i)\}_{i=1}^N$, equivalent to minimizing the binary cross-entropy loss $\mathcal{L}$:

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \right], \tag{12}$$

where $\pi_i = \sigma(\mathbf{w}^T \mathbf{z}_i)$. This convex optimization problem is efficiently solved using iterative methods such as L-BFGS, ensuring rapid convergence even with high-dimensional feature inputs.

3.5. Bagging Ensemble with Homogeneous Base Learners

To enhance the stability and generalizability of the classifier, we employ Bootstrap Aggregating (Bagging) with homogeneous logistic regression models. This parallel ensemble technique reduces prediction variance by aggregating the outputs of multiple base learners trained on different bootstrap samples of the training data.

Formally, let $\mathcal{D} = \{(\mathbf{z}_i, y_i)\}_{i=1}^N$ be the full training set of extracted features. We generate K bootstrap samples $\mathcal{D}^{(k)}$, each created by randomly selecting N instances from $\mathcal{D}$ with replacement. A distinct logistic regression model g_{ϕ_k} is trained on each $\mathcal{D}^{(k)}$, resulting in an ensemble $\{g_{\phi_k}\}_{k=1}^K$.

The final prediction for a test instance $\mathbf{z}$ is determined by majority voting over the binary decisions of all ensemble members:

$$\hat{y}_{\text{ensemble}} = \arg\max_{c \in \{0,1\}} \sum_{k=1}^{K} \mathbb{I}\big(g_{\phi_k}(\mathbf{z}) = c\big), \tag{13}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The ensemble size K is a hyperparameter; empirical analysis determined $K = 100$ to provide an optimal balance between performance gains and computational cost, with diminishing returns observed for larger values.

This homogeneous bagging strategy is particularly effective in our context because:

- It mitigates the high variance that can occur when training a single logistic regression model on complex, high-dimensional feature spaces.
- The bootstrap sampling implicitly introduces diversity among the base learners, as each model encounters a slightly different data distribution.
- It is inherently parallelizable, allowing efficient training and inference without significant runtime overhead.

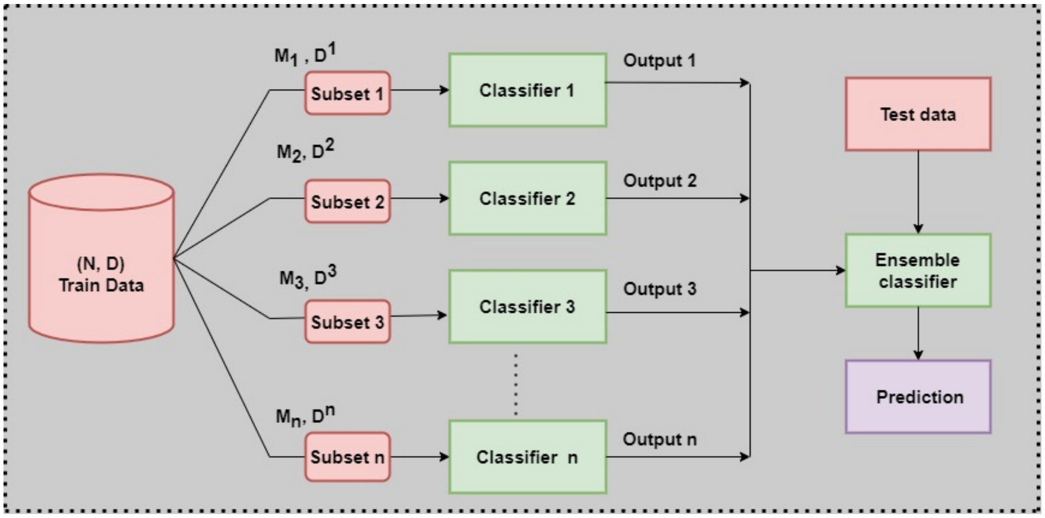


Figure 3: Schematic representation of Ensemble Learning

The combination of a robust, pre-trained feature extractor with a variance-reducing bagging ensemble forms the core of our proposed methodology, addressing key challenges of overfitting and instability common in medical image classification with limited data.

3.6. Theoretical Foundation of Ensemble Learning

Ensemble learning leverages statistical aggregation to reduce variance and improve generalization, particularly vital in medical image analysis where data limitations amplify prediction instability.

3.6.1. Bias-Variance Decomposition

The expected prediction error decomposes as:

$$\mathbb{E}[(\hat{f}(x) - f(x))^2] = \underbrace{(\mathbb{E}[\hat{f}(x)] - f(x))^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]}_{\text{Variance}} + \sigma^2, \tag{14}$$

where σ^2 is irreducible noise. Ensemble methods, particularly bagging, reduce variance without increasing bias, addressing the fundamental trade-off between model complexity and generalizability.

3.6.2. Error Reduction Through Majority Voting

For binary classification with base classifier error rate $\epsilon < 0.5$, ensemble error via majority voting is bounded by:

$$P(H(x) \neq f(x)) \leq \exp\left(-\frac{1}{2}M(2\epsilon - 1)^2\right), \tag{15}$$

decaying exponentially with ensemble size M. This requires base classifiers to be better than random and exhibit error diversity.

3.6.3. Addressing Class Imbalance

Bootstrap sampling in bagging naturally mitigates class imbalance by creating varied class distributions across samples, forcing base learners to adapt and collectively produce calibrated predictions.

3.7. Proposed Bagging Algorithm

The algorithm implements homogeneous bagging with logistic regression base learners on MobileNet features:

The algorithm reduces variance through bootstrap aggregation while maintaining computational efficiency via parallelizable base learner training.

Mathematical Formulation:

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ be the training dataset, where $\mathbf{x}_i \in \mathbb{R}^{224\times224\times3}$ and $y_i \in \{0,1\}$. The ensemble construction involves:

1. **Feature Extraction:** For each image $\mathbf{x}_i$, compute $\mathbf{z}_i = f_\theta(\mathbf{x}_i)$, where f_θ is a pre-trained MobileNet feature extractor with parameters θ fixed.

2. **Bootstrap Sampling:** For each ensemble member $k = 1, \ldots, K$, generate a bootstrap sample $\mathcal{D}^{(k)}$ by sampling N_{boot} indices uniformly with replacement from $\{1, \ldots, N\}$. This creates approximately $(1 - e^{-1})N \approx 0.632N$ unique instances per sample, with the rest being duplicates.

3. **Base Classifier Training:** Each base classifier $h^{(k)}$ is a logistic regression model parameterized by $\mathbf{w}^{(k)} \in \mathbb{R}^d$, trained to minimize the regularized negative log-likelihood:

$$\mathcal{J}(\mathbf{w}^{(k)}) = -\sum_{(\mathbf{z},y)\in\mathcal{D}^{(k)}} \left[y \log \sigma(\mathbf{w}^{(k)T}\mathbf{z}) + (1-y)\log(1-\sigma(\mathbf{w}^{(k)T}\mathbf{z})) \right] + \lambda \|\mathbf{w}^{(k)}\|_2^2, \tag{16}$$

where λ controls L2 regularization strength. The optimization is performed via iterative reweighted least squares or gradient-based methods.

4. **Ensemble Prediction:** For a test instance $\mathbf{x}_{\text{test}}$, the ensemble prediction is obtained through majority voting:

$$H(\mathbf{x}_{\text{test}}) = \arg\max_{c\in\{0,1\}} \sum_{k=1}^{K} \mathbb{I}\left(h^{(k)}(\mathbf{z}_{\text{test}}) = c\right), \tag{17}$$

with tie-breaking favoring the minority class (or using the average predicted probability).

Theoretical Properties: The bagging ensemble reduces variance without increasing bias, as formalized by the bias-variance decomposition for 0-1 loss. Let $\bar{h}(\mathbf{z}) = \mathbb{E}[h^{(k)}(\mathbf{z})]$ be the expected prediction over bootstrap samples. The ensemble error can be bounded by:

$$\mathbb{E}[(H(\mathbf{z}) - y)^2] \leq \frac{1}{K}\text{Var}[h^{(k)}(\mathbf{z})] + (\bar{h}(\mathbf{z}) - y)^2, \tag{18}$$

demonstrating variance reduction inversely proportional to ensemble size K.

The algorithm operates in two phases: (1) parallel training of K base models on bootstrap samples, and (2) aggregation via majority voting during inference. This design ensures both computational efficiency (through parallelization) and statistical robustness (through variance reduction), making it particularly suitable for medical image classification tasks where dataset limitations and class imbalance are prevalent challenges.

4. Experiments and results

4.1. Dataset

The experimental validation employs two publicly available medical imaging datasets: the HAM10000 repository of dermoscopic skin lesions (Tschandl et al., 2018) and a curated Kaggle collection of brain tumor MRI scans. Each dataset comprises 400 images balanced across binary classes—benign versus malignant for skin lesions, and tumor-present versus tumor-absent for brain scans. Dermoscopic images are standardized (224×224 PNG, 72 DPI, 24-bit depth), whereas the neuroimaging data reflects clinical heterogeneity in acquisition parameters. A stratified 80-10-10 split was applied for training, validation, and testing. For transfer learning, input images were adapted to the dimensional requirements of each pre-trained CNN backbone.

4.2. Experiment setup

This section provided an overview of the experimental setup and results for the four proposed (ML) models, which were built using data from skin cancer and brain tumor datasets. With an i9-9900k processor, NVIDIA Quadro P5000 graphics card, 16GB of GDDR5x RAM, and the open-source deep learning framework TensorFlow in Anaconda3, we used Python 3.7 on Ubuntu. The hyper parameters used are the batch size of 18, momentum of 0.5, decay of 0.0005, and learning rate of 0.003.

4.3. Results

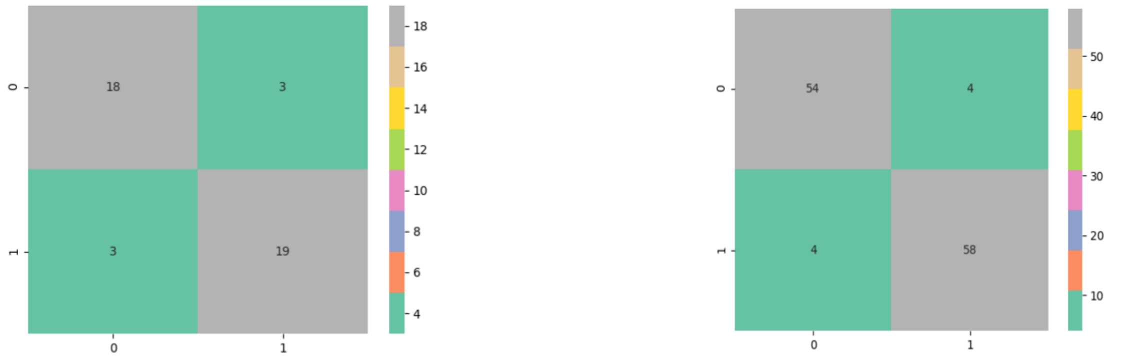
The metrics used for comparison are Confusion Matrix, Precision, Recall, F1-Score, R1-value. We have used K-Fold cross-validation to train the model and enhance the results.

4.3.1. Results analysis:

Figure 4 shows the confusion matrix for the brain tumor dataset in classifying brain images into two classes: brain tumor images and brain without tumors. It shows the confusion matrix (Wu, 2022) for the skin cancer dataset in classifying skin lesion images into two classes, such as benign images and malignant images. Tables 1, 3 show the results of the proposed model, a combination of logistic regression classifier and MobileNet pre-trained deep learning architecture that outperforms the classification of skin and brain images by various other state-of-the-art techniques. The results have been shown in terms of metrics such as accuracy (Acc), precision (Pre), recall (Rec), R1-value (R1), and F1-score (F1) using confusion (CF) matrices.

5. Discussion

The experimental results confirm that the proposed MobileNet ensemble framework achieves superior performance, attaining state-of-the-art accuracy across key metrics. While specialized architectures such as WT-DRNN and EnDBoT yield marginally higher precision through intensive skip connections and dense blocks, their substantial computational overhead limits practical clinical utility. Comprehensive benchmarking against diverse classifiers



Classification results (brain data)

Classification results (skin data)

Figure 4: Confusion matrix of the proposed model for binary classification of brain and skin datasets.

(Decision Trees, Random Forest, KNN) and deep learning backbones (VGG, ResNet, Inception, Xception) underscores the robustness of our approach. The confusion matrices validate precise discrimination between malignant/benign lesions and tumor/normal scans, with near-perfect per-class recognition. Although ensemble learning increases computational demand, our parallel implementation leverages multi-core architectures to balance diagnostic accuracy with deployable efficiency.

The confusion matrices in Figure 5 demonstrate that the proposed ensemble framework achieves highly discriminative classification across all target classes, with precise distinction between malignant and benign lesions, and between tumor-present and normal neuroimaging studies. Each class instance in the test set is accurately recognized, confirming the model's robust multi-class capability. While ensemble learning inherently increases computational complexity through multi-model training, our parallel implementation effectively mitigates this overhead by leveraging contemporary multi-core architectures. This represents a pragmatic balance between performance gains and computational feasibility.

6. Conclusion and future work

Despite their diagnostic potential, convolutional neural networks are often limited by the scarcity of large, annotated medical image datasets. This work overcomes this constraint through a novel ensemble framework that integrates transfer learning with bagged logistic regression. By leveraging pre-trained representations and aggregating multiple base learners, the method reduces overfitting and enhances generalization on limited data.

Acknowledgments

This work was supported by Indian Institute of Information Technology Ranchi, Ranchi, India.

References

A.c. society, cancer facts figures 2022, am. cancer soc. (2022), 2022. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2022.html/2022-cancer-facts-and-figures.pdf, Last accessed on 2022-11-25.

Adekanmi Adegun and Serestina Viriri. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artificial Intelligence Review*, 54(2):811–841, 2021.

Idowu Adeyemi, Mahmoud Meribout, and Lyes Khezzar. Recent developments, challenges, and prospects of ultrasound-assisted oil technologies. *Ultrasonics Sonochemistry*, 82: 105902, 2022.

Samet Akcay and Toby Breckon. Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. *Pattern Recognition*, 122:108245, 2022.

Fayadh Alenezi, Ammar Armghan, and Kemal Polat. Wavelet transform based deep residual neural network and relu based extreme learning machine for skin lesion classification. *Expert Systems with Applications*, 213:119064, 2023.

JA Almaraz-Damian, V Ponomaryov, and E Rendon-Gonzalez. Melanoma cade based on abcd rule and haralick texture features. In *2016 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW)*, pages 1–4. IEEE, 2016.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Muhammad Naseer Bajwa, Kaoru Muta, Muhammad Imran Malik, Shoaib Ahmed Siddiqui, Stephan Alexander Braun, Bernhard Homey, Andreas Dengel, and Sheraz Ahmed. Computer-aided diagnosis of skin diseases using deep neural networks. *Applied Sciences*, 10(7):2488, 2020.

Lei Bi, Jinman Kim, Euijoon Ahn, and Dagan Feng. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. *arXiv preprint arXiv:1703.04197*, 2017.

Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C Thai, Kathleen Moore, Robert S Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, page 102444, 2022.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Deep residual learning for image compression. In *Cvpr workshops*, page 0, 2019.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

Noel CF Codella, Q-B Nguyen, Sharath Pankanti, David A Gutman, Brian Helba, Allan C Halpern, and John R Smith. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, 61(4/5):5–1, 2017.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258, 2020.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019.

Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *MethodsX*, 7:100864, 2020.

András Hajdu, Balázs Harangi, Renátó Besenczi, István Lázár, G Emri, Lajos Hajdu, and Robert Tijdeman. Measuring regularity of network patterns by grid approximations using the lll algorithm. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1524–1529. IEEE, 2016.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Ravi M. Kamble, Genevieve C. Y. Chan, Oscar Perdomo, Manesh Kokare, Fabio A. González, Henning Müller, and Fabrice Mériaudeau. Automated diabetic macular edema (dme) analysis using fine tuning with inception-resnet-v2 on oct images. In *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 442–446, 2018. doi: 10.1109/IECBES.2018.8626616.

Davood Karimi, Simon K Warfield, and Ali Gholipour. Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations. *Artificial Intelligence in Medicine*, 116:102078, 2021.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

Asif Iqbal Khan, Junaid Latief Shah, and Mohammad Mudasir Bhat. Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images. *Computer methods and programs in biomedicine*, 196:105581, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Nabeel F Lattoofi, Israa F Al-Sharuee, Mohammed Y Kamil, Ayoob H Obaid, Aya A Mahidi, Ammar A Omar, et al. Melanoma skin cancer detection based on abcd rule. In *2019 First International Conference of Computer and Applied Sciences (CAS)*, pages 154–157. IEEE, 2019.

Roman C Maron, Achim Hekler, Sarah Haggenmüller, Christof von Kalle, Jochen S Utikal, Verena Müller, Maria Gaiser, Friedegund Meier, Sarah Hobelsberger, Frank F Gellrich, et al. Model soups improve performance of dermoscopic skin cancer classifiers. *European Journal of Cancer*, 173:307–316, 2022.

Y. Mednikov, S. Nehemia, B. Zheng, O. Benzaquen, and D. Lederman. Transfer representation learning using inception-v3 for the detection of masses in mammography. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2587–2590, 2018. doi: 10.1109/EMBC.2018.8512750.

Katsuhiro Nakai, Yen-Wei Chen, and Xian-Hua Han. Enhanced deep bottleneck transformer model for skin lesion classification. *Biomedical Signal Processing and Control*, 78:103997, 2022.

Samuel William Pewton and Moi Hoon Yap. Dark corner on skin lesion image dataset: Does it matter? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4831–4839, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Bhuvaneshwari Shetty, Roshan Fernandes, Anisha P Rodrigues, Rajeswari Chengoden, Sweta Bhattacharya, and Kuruva Lakshmanna. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Scientific Reports*, 12 (1):18134, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Bernhard Stimpel, Christopher Syben, Franziska Schirrmacher, Philip Hoelter, Arnd Dörfler, and Andreas Maier. Multi-modal deep guided filtering for comprehensible medical image processing. *IEEE Transactions on Medical Imaging*, 39(5):1703–1711, 2020. doi: 10.1109/TMI.2019.2955184.

AP Sunija, Saikat Kar, S Gayathri, Varun P Gopi, and Ponnusamy Palanisamy. Octnet: A lightweight cnn for retinal disease classification from optical coherence tomography images. *Computer methods and programs in biomedicine*, 200:105877, 2021.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, IEEE, Piscataway, NJ, USA, 2015.

Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016. doi: 10.1109/TMI.2016.2535302.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

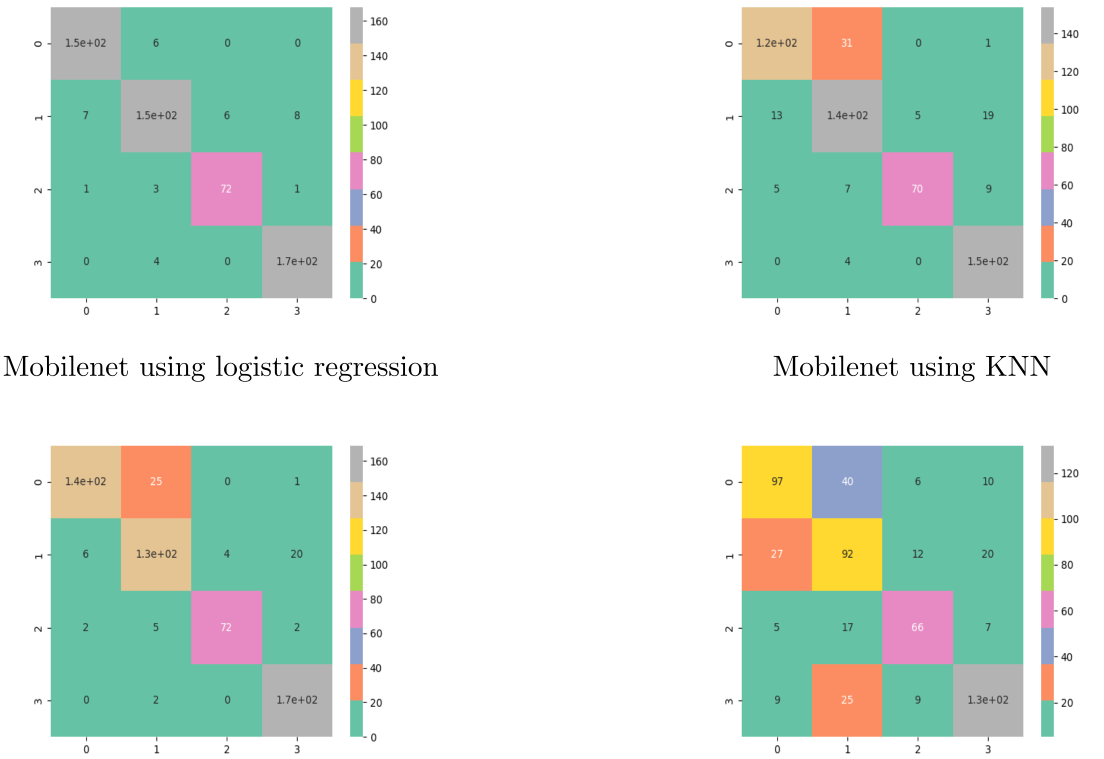
Runze Wang, Yide Ma, Wenhao Sun, Yanan Guo, Wendao Wang, Yunliang Qi, and Xiaonan Gong. Multi-level nested pyramid network for mass segmentation in mammograms. *Neurocomputing*, 363:313–320, 2019.

Jeffrey B Ware, Saurabh Sinha, Justin Morrison, Alexa E Walter, James J Gugger, Andrea LC Schneider, Cian Dabrowski, Hannah Zamore, Leroy Wesley, Brigid Magdamo, et al. Dynamic contrast enhanced mri for characterization of blood-brain-barrier dysfunction after traumatic brain injury. *NeuroImage: Clinical*, page 103236, 2022.

Ming-Te Wu. Confusion matrix and minimum cross-entropy metrics based motion recognition system in the classroom. *Scientific Reports*, 12(1):1–10, 2022.

Yutong Xie, Jianpeng Zhang, Yong Xia, and Chunhua Shen. A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE transactions on medical imaging*, 39(7):2482–2493, 2020.

Appendix A.: More results



Mobilenet using logistic regression

Mobilenet using KNN

Mobilenet using randomforest

Mobilenet using decision tree

Figure 5: Confusion Matrices (CM) multiclass classification of Brain dataset using classifiers such as Logistic Regression, Decision Trees (DT), K-Nearest Neighbor (KNN), Random Forest (RF) using MobileNet pre-trained model.

Algorithm 1: Bagging Ensemble Classification

```
[1] Training data $\mathcal{D}$, feature extractor $f_\theta$, ensemble size $K$ Ensemble classifier $H$ Extract features: $\mathbf{z}_i = f_\theta(\mathbf{x}_i)$ for all $(\mathbf{x}_i, y_i) \in \mathcal{D}$ for $k = 1$ to $K$ do
end
Sample bootstrap indices $\mathcal{I}^{(k)}$ with replacement Train $\mathbf{w}^{(k)}$ via logistic regression on $\{(\mathbf{z}_i, y_i) : i \in \mathcal{I}^{(k)}\}$ Store classifier $h^{(k)}(\mathbf{z}) = \mathbb{I}(\sigma(\mathbf{w}^{(k)T}\mathbf{z}) > 0.5)$ Predict$\mathbf{x}_{\text{test}}$ $\mathbf{z}_{\text{test}} = f_\theta(\mathbf{x}_{\text{test}})$ Compute votes: $v_k = h^{(k)}(\mathbf{z}_{\text{test}})$ for $k = 1, \ldots, K$ return $\arg\max_c \sum_{k=1}^{K} \mathbb{I}(v_k = c)$
```

Table 1: Tabular representation of the proposed framework's performance considering Brain tumor images and a Logistic regression classifier

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG16	0.82	0.81	0.81	0.8140	0.8469
VGG19	0.89	0.88	0.88	0.8837	0.8873
ResNet50	0.72	0.72	0.72	0.7209	0.7550
Inception-V3	0.86	0.86	0.86	0.9101	0.9235
Method (Kamble et al., 2018)	0.91	0.91	0.91	0.9070	0.9023
MobileNet	**0.93**	**0.93**	**0.93**	**0.9323**	**0.9468**
Xception	0.84	0.84	0.84	0.8372	0.9031

Table 2: Tabular representation of the proposed framework's performance considering skin cancer images and a Logistic regression classifier

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG16	0.91	0.93	0.92	0.9250	0.8750
VGG19	0.90	0.90	0.90	0.9000	0.8929
ResNet50	0.83	0.83	0.83	0.9083	0.8393
InceptionV3	0.91	0.91	0.91	0.8750	0.9357
Method (Kamble et al., 2018)	0.88	0.88	0.87	0.9333	0.9036
MobileNet	**0.94**	**0.96**	**0.95**	**0.9333**	**0.9785**
Xception	0.90	0.92	0.91	0.9333	0.9217

Table 3: Comparison with recently proposed SoTA techniques.

Method	**Pre**	**Rec**	**F1**	**Acc**
Customized CNN (Shetty et al., 2022)	0.88	0.85	0.86	0.86
WT-DRNN (Alenezi et al., 2023)	0.95	-	0.93	0.95
EnDBoT201D (Nakai et al., 2022)	**0.96**	-	-	0.95
EnDBoT50R (Nakai et al., 2022)	0.94	-	-	0.93
Proposed	0.94	**0.96**	**0.95**	**0.9785**

Table 4: Results of binary classification of Brain tumor images using Random Forest classifier

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG16	0.84	0.84	0.84	0.8140	0.8162
VGG19	0.75	0.74	0.74	0.7442	0.8368
ResNet50	0.74	0.74	0.74	0.6977	0.7045
Inception-V3	0.91	0.91	0.91	0.9302	0.8166
Method (Kamble et al., 2018)	0.83	0.81	0.81	0.7907	0.7755
MobileNet	0.79	0.79	0.79	0.8140	0.7547
Xception	0.73	0.72	0.72	0.7209	0.8055

Table 5: Results of binary classification of Brain tumor images using Decision Trees classifier

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG16	0.77	0.77	0.77	0.7674	0.6938
VGG19	0.81	0.79	0.79	0.7907	0.7746
ResNet50	0.72	0.70	0.69	0.6977	0.7553
Inception-V3	0.65	0.65	0.65	0.6512	0.7856
Method (Kamble et al., 2018)	0.74	0.74	0.74	0.7442	0.7443
MobileNet	0.68	0.67	0.67	0.6744	0.7045
Xception	0.67	0.65	0.64	0.6512	0.7954

Table 6: Results of binary classification of Brain tumor images using KNN classifier

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG16	0.72	0.72	0.72	0.7209	0.7752
VGG19	0.84	0.84	0.84	0.8372	0.7856
ResNet50	0.68	0.77	0.72	0.6977	0.7045
Inception-V3	0.88	0.84	0.83	0.8372	0.7143
Method (Kamble et al., 2018)	0.80	0.67	0.64	0.6744	0.6019
MobileNet	0.84	0.77	0.76	0.7674	0.7853
Xception	0.86	0.84	0.84	0.8372	0.8260

Table 7: Results of binary classification of Skin cancer images using Random Forest classifier

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG16	0.81	0.81	0.81	0.8250	0.8286
VGG19	0.88	0.88	0.87	0.8167	0.8178
ResNet50	0.76	0.76	0.76	0.7750	0.8036
Inception-V3	0.85	0.83	0.83	0.8167	0.8354
Method (Kamble et al., 2018)	0.84	0.83	0.83	0.8500	0.8501
MobileNet	0.94	0.93	0.93	0.9167	0.9179
Xception	0.81	0.81	0.81	0.8333	0.8463

Table 8: Results of binary classification of Skin cancer images using Decision trees classifier

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG16	0.80	0.80	0.80	0.8000	0.7322
VGG19	0.81	0.81	0.81	0.8083	0.7858
ResNet50	0.79	0.79	0.79	0.7917	0.7820
Inception-V3	0.71	0.71	0.71	0.7083	0.7283
Method (Kamble et al., 2018)	0.72	0.72	0.72	0.7250	0.7355
MobileNet	0.83	0.82	0.83	0.8250	0.8107
Xception	0.79	0.79	0.79	0.7917	0.8464

Table 9: Results of binary classification of Skin cancer images using KNN classifier

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG16	0.81	0.81	0.81	0.8083	0.8073
VGG19 (Stimpel et al., 2020)	0.80	0.80	0.80	0.800	0.7929
ResNet50 (Cheng et al., 2019)	0.82	0.81	0.81	0.8083	0.7892
Inception-V3 (Mednikov et al., 2018)	0.88	0.88	0.88	0.8750	0.8820
Method (Kamble et al., 2018)	0.82	0.78	0.77	0.7750	0.7821
MobileNet	0.90	0.88	0.88	0.8833	0.8749
Xception (Tajbakhsh et al., 2016)	0.84	0.83	0.83	0.8333	0.8678

Table 10: Results of multiclass classification of brain data using Logistic Regression

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG19 (Stimpel et al., 2020)	0.94	0.93	0.92	0.881	0.88
ResNet50 (Cheng et al., 2019)	0.87	0.87	0.87	0.7596	0.76
Inception-V3 (Mednikov et al., 2018)	0.96	0.99	0.96	0.9111	0.91
MobileNet	0.95	0.98	0.96	0.9373	0.94
Xception (Tajbakhsh et al., 2016)	0.95	0.98	0.95	0.9216	0.92

Table 11: Results of multiclass classification of brain data using KNN model

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG19 (Stimpel et al., 2020)	0.92	0.98	0.89	0.7997	0.80
ResNet50 (Cheng et al., 2019)	0.79	0.94	0.86	0.7300	0.73
Inception-V3 (Mednikov et al., 2018)	0.90	0.96	0.89	0.8345	0.83
MobileNet	0.93	0.97	0.90	0.84	0.8328
Xception (Tajbakhsh et al., 2016)	0.91	0.97	0.88	0.8328	0.83

Table 12: Results of multiclass classification of brain data using Randomforest classifier

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG19 (Stimpel et al., 2020)	0.94	0.97	0.92	0.8431	0.85
ResNet50 (Cheng et al., 2019)	0.88	0.94	0.91	0.8275	0.83
Inception-V3 (Mednikov et al., 2018)	0.93	0.95	0.94	0.8606	0.86
MobileNet	0.95	0.99	0.93	0.8885	0.88
Xception (Tajbakhsh et al., 2016)	0.92	0.96	0.92	0.8432	0.84

Table 13: Results of multiclass classification of brain data using Decision Tree classifier

Method	**Pre**	**Rec**	**F1**	**R1**	**Acc**
VGG19 (Stimpel et al., 2020)	0.79	0.78	0.79	0.6690	0.67
ResNet50 (Cheng et al., 2019)	0.78	0.82	0.77	.6760	0.68
Inception-V3 (Mednikov et al., 2018)	0.78	0.74	0.76	0.6655	0.67
MobileNet	0.78	0.75	0.77	0.6742	0.67
Xception (Tajbakhsh et al., 2016)	0.75	0.72	0.73	0.6533	0.65