PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY 20 AOUT 1955 - SKIKDA

Faculty of Sciences

Department of Computer Science

Master's Degree Thesis in Computer Science

Option: Artificial Intelligence

Subject

Transformers-based Ensemble Methods for Medical Imaging: A Theoretical and Experimental Study

Presented by: **Selma ZIANE**

Supervised by: **Dr. Samira HAZMOUNE**

Academic year 2023/2024

Acknowledgements

First and foremost, I thank God Almighty for giving me the strength, courage, and perseverance to complete this thesis. Without His blessings, none of this would have been possible.

I would like to express my sincere thanks to my advisor, Dr. Samira Hazmoune, for her invaluable guidance, encouragement, and patience throughout this research. Her expertise and insights have been crucial in shaping this work. She has been a mentor and a source of inspiration, always pushing me to achieve my best and providing me with the necessary tools and knowledge to navigate this challenging journey.

I am also deeply grateful to the examination committee for their time, effort, and valuable feedback.

Secondly, I would like to express my deepest gratitude to my family. To my parents, Ziane Rachid and Mouat Yamina, your unwavering support, encouragement, and belief in me have been the bedrock of my journey. Your love and sacrifices have been invaluable, providing me with the foundation to pursue and achieve my dreams. Your constant words of wisdom and faith in my abilities have been a continuous source of motivation and strength.

To my siblings, your constant support and understanding have been a source of great strength. Your encouragement and patience, even during my most challenging times, have been deeply appreciated. Each one of you has played a significant role in my academic journey, and for that, I am eternally grateful.

Last but not least, I owe a special debt of gratitude to my colleagues and friends, whose collaboration, discussions, and support have been a source of great help and motivation.

Finally, thank you all for your tremendous support and encouragement. This achievement would not have been possible without each one of you. Your contributions have been invaluable, and I am deeply grateful for your presence and support.

Abstract

Medical image classification accuracy significantly aids doctors in diagnosing diseases and planning treatments.

Although Transformers, a cutting-edge technology in artificial intelligence, perform exceptionally well in various image classification tasks, their ability to capture the subtle details within medical images can be limited. To address this challenge, we propose, in this thesis, a novel ensemble method that leverages Transformers to improve medical image classification. We employ eight diverse medical imaging datasets and implement eight different Transformer-based ensemble methods on each dataset, covering several medical imaging areas such as magnetic resonance, computed tomography, magnetic resonance, dermatoscopic, chest X-ray, screening Mammography of medical imaging classification.

A deep experimental study of the proposed approach is conducted to evaluate its effectiveness in the medical image classification domain. Furthermore, an ablation study is performed to identify the optimal combination of base models for each ensemble method across different datasets. Our experiments encompass datasets of varying sizes, acknowledging the ongoing challenges of limited data availability in medical imaging. Despite this limitation, our ensemble method approach consistently outperforms state-of the-art methods across multiple datasets. This demonstrates its effectiveness in alleviating limitations associated with data size and diversity. These findings highlight the potential of Transformers-based ensemble methods to revolutionize medical image classification. This paves the way for improved diagnostic accuracy and treatment decision-making in clinical settings.

Key words: Medical image classification, Magnetic resonance, Computed tomography, Dermatoscopic, Chest X-ray, Screening Mammography, Artificial Intelligence, Transformers, ViT, DEiT, BEiT, Swin, Ensemble methods, Voting, Boosting, Bagging, Stacking, Ablation study.

Résumé

La précision de la classification des images médicales aide considérablement les médecins à diagnostiquer les maladies et à planifier les traitements. Bien que les Transformers, une technologie de pointe en intelligence artificielle, excellent dans diverses tàches de classification d'images, leur capacité à capturer les détails subtils des images médicales peut étre limitée.

Pour relever ce défi, nous proposons, dans cette thèse, une nouvelle méthode d'ensemble qui exploite les Transformers pour améliorer la classification des images médicales. Nous employons huit ensembles de données diversifiées d'imagerie médicale et implémentons huit différentes méthodes d'ensemble basées sur les Transformers sur chaque ensemble de données, couvrant plusieurs domaines de l'imagerie médicale tels que la résonance magnétique, la tomodensitométrie, la résonance magnétique, la dermatoscopie, la radiographie thoracique et la mammographie de dépistage.

Une étude expérimentale approfondie de l'approche proposée est menée pour évaluer son efficacité dans le domaine de la classification des images médicales. De plus, une étude d'ablation est réalisée pour identifier la combinaison optimale de modéles de base pour chaque méthode d'ensemble à travers différents ensembles de données. Nos expériences englobent des ensembles de données de tailles variées, reconnaissant les défis permanents liés à la disponibilité limitée des données en imagerie médicale. Malgré cette limitation, notre approche par méthode d'ensemble surpasse constamment les méthodes à la pointe de la technologie sur plusieurs ensembles de données.

Cela démontre son efficacité à atténuer les limitations associées à la taille et à la diversité des données. Ces résultats mettent en lumière le potentiel des méthodes d'ensemble basées sur les Transformers pour révolutionner la classification des images m´edicales. Cela ouvre la voie `a une meilleure précision diagnostique et `a une prise de d´ecision thérapeutique dans les environnements cliniques.

Mots clés : Classification d'image medécale, Résonance magnétique, Tomodensitométrie, Dermatoscopie, Radiographie thoracique, Mammographie de dépistage, Intelligence artificielle, Transformers, ViT, DEiT, BEiT, Swin, Méthodes d'ensemble, Vote, Boosting, Bagging, Stacking, Ablation study.

ملخص

تساعد دقة تصنيف الصور الطبية الأطباء بشكل كبير في تشخيص الأمراض والتخطيط للعلاجات على الرغم من أن المحولات، وهي تكنولوجيا متقدمة في الذكاء الاصطناعي، تتفوق في مهام تصنيف الصور المتنوعة، إلا أن قدرتها على التقاط التفاصيل الدقيقة داخل الصور الطبية يمكن أن تكون محدودة. لمواجهة هذا التحدي، نقترح في هذه الأطروحة طريقة تجميع جديدة تستفيد من المحولات لتحسين تصنيف الصور الطبية. نستخدم ثمانية مجموعات بيانات متنوعة للتصوير الطبي وننفذ ثمانية طرق تجميع مختلفة تعتمد على المحولات على كل مجموعة بيانات، تغطي العديد من مجالات التصوير الطبي مثل الرنين المغناطيسي، التصوير المقطعي المحوسب، الرنين المغناطيسي، التصوير الجلدي، التصوير بالأشعة السينية للصدر، وتصوير الثدي الشعاعي للكشف.

تم إجراء دراسة تجريبية عميقة للنهج المقترح لتقييم فعاليته في مجال تصنيف الصور الطبية. بالإضافة إلى ذلك، تم إجراء دراسة إزالة لتحديد التركيبة المثلى للنماذج الأساسية لكل طريقة تجميع عبر مجموعات البيانات المختلفة. تشمل تجاربنا مجموعات بيانات بأحجام مختلفة، معترفين بالتحديات المستمرة المتعلقة بتوفر البيانات المحدود في التصوير الطبي. على الرغم من هذا القيد، فإن نهجنا باستخدام طريقة التجميع يتفوق باستمرار على الطرق الحديثة عبر مجموعات بيانات متعددة. يوضح ذلك فعاليته في تخفيف القيود المرتبطة بحجم وتنوع البيانات. تبرز هذه النتائج الإمكانية الهائلة لطريقة التجميع المعتمدة على المحولات في إحداث ثورة في تصنيف الصور الطبية. يفتح ذلك الطريق نحو دقة تشخيصية أفضل واتخاذ قرارات علاجية محسنة في البيئات السريرية.

كلمات مفتاحية: تصنيف الصور الطبية، الرنين المغناطيسي، التصوير المقطعي المحوسب، تنظير الجلد، تصوير الصدر بالأشعة السينية، تصوير الثدي الماموغرافي، الذكاء الاصطناعي، المحولات، طرق التجميع، التصويت، التعزيز، التجميع، التراص، دراسة الإزالة.

Contents

2

List of Figures

General introduction

Medical imaging plays a pivotal role in modern medicine [1], providing invaluable insights into the intricate structures and abnormalities within the human body. The explosion of digital imaging technologies has resulted in a vast amount of complex medical image data. While this data holds immense potential for improving patient care, it also presents challenges in accurately interpreting and extracting valuable information.

Despite advancements in imaging techniques and computational tools, accurately interpreting medical images remains a demanding task. Doctors face challenges like inter-observer variability in interpretations, subtle disease manifestations, and imbalanced datasets where certain conditions are less prevalent. In this complex setting, machine learning (ML) offers promise as a transformative tool for enhancing diagnoses and patient outcomes. ML techniques [2], particularly deep learning (DL) [3], have demonstrated remarkable capabilities in analyzing medical images. Deep learning models, such as Convolutional Neural Networks (CNNs), have achieved impressive results in tasks like image classification, segmentation, and detection.

Recently, transformers [4] have emerged as a powerful addition to the DL landscape, offering effective solutions for processing sequential data. Known for their success in natural language processing (NLP), transformers have also shown remarkable efficacy in various computer vision tasks, including medical image analysis. These models excel at capturing relationships and context within data sequences, making them well-suited for medical image interpretation, where understanding spatial relationships and meaning is crucial. However, even with powerful individual DL models, achieving consistent and reliable diagnostic accuracy remains challenging, especially in the face of data variability and uncertainty.

Ensemble method offers a compelling solution by leveraging the combined wisdom of diverse models[5]. It provides a systematic approach to combine multiple models, aiming to improve overall performance and prediction robustness. By aggregating insights from various sources [6], ensemble method alleviates the biases and uncertainties inherent in individual models, leading to more reliable and interpretable medical image analysis.

In this research, we explore the potential of ensemble learning, particularly when combined with transformer models, for medical image classification. Medical image data often exhibits high dimensionality, class imbalance, and noise, posing significant challenges for single models. Ensemble method addresses these challenges by combining predictions from multiple models, thereby enhancing accuracy, robustness, and generalization performance. Ensemble method also encourages model diversity by incorporating different architectures, training data subsets, or optimization strategies. This allows for integrating diverse perspectives and complementary information, enabling ensemble models to capture a wider range of diagnostic features and reduce the risk of overfitting to specific data patterns.

Our study comprehensively explores ensemble method techniques for medical image

classification. We investigate eight diverse ensemble strategies and four state-of-the-art transformer models (DEiT [7], BEiT [8], Swin [9], and ViT [10]) . We utilize datasets covering various medical imaging modalities (DermaMNIST, BloodMNIST, OrgansMNIST, OrganaMNIST, OrgancMNIST, BreastMNIST, pneumoniaMNIST) from the MedMNIST repository and Chest dataset from the Hugging Face library . Through experimentation on these datasets, we evaluate the performance of ensemble approaches in boosting classification accuracy. Furthermore, an ablation study delves into the impact of adding or removing individual models within the ensemble, shedding light on the synergistic effects that contribute to improved classification performance. This investigation not only informs us about the optimal ensemble composition but also clarifies the underlying mechanisms leading to enhanced accuracy.

By using Transformers in an ensemble method framework, this research aims to push the boundaries of medical image classification accuracy. We seek to confirm whether ensemble learning, particularly with transformers, can outperform single models in classification tasks. The findings gained from this research are expected to guide future efforts in utilizing ensemble method for better healthcare diagnostics and decision-making.This thesis organized as follows:

- **Chapter 1 -An overview of machine learning techniques**: This chapter covers fundamental concepts in machine learning and deep learning. It begins with an overview of machine learning principles, discussing various types and algorithms. The focus then shifts to deep learning, emphasizing Artificial Neural Networks (ANNs) as foundational components. Various ANN models such as Transformers, Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs) are explored in detail. Additionally, ensemble learning techniques in deep learning for improving model performance are discussed.

- **Chapter 2 -State of the art of medical imaging**: This chapter provides a comprehensive overview of medical imaging analysis using deep learning techniques. It starts by introducing different medical imaging modalities and essential tasks like classification and segmentation. Key datasets used in the field are discussed to facilitate model development and evaluation. The chapter also covers commonly employed performance metrics for evaluating model effectiveness. The application of CNNs for medical image classification is explored, followed by recent advancements in transformer-based models. Both pure transformer-based approaches and hybrid models are presented, alongside effective ensemble learning techniques that enhance predictive accuracy and model robustness.

- **Chapter 3 -Design of a transformers-based ensemble framework for medical imaging**: This chapter outlines our ensemble methods for improving medical image classification using four transformer models: ViT, DEiT, BEiT, and Swin. The approach involves meticulous tuning and explores various ensemble methods from hard voting to boosting. The chapter emphasizes transparency in evaluation metrics and experimental setup.

- **Chapter 4 -Implementation and experimental study of the proposed ensemble methods**: In this chapter, we experimentally evaluate proposed ensemble methods for medical image classification. We begin with implementation setup details, including hardware and hyperparameter tuning for fine-tuning individual models. The results of individual models are analyzed to establish a baseline for

comparison. Eight ensemble methods using the four fine-tuned models are implemented and assessed across multiple datasets. The findings demonstrate consistent improvements in classification accuracy and robustness. An ablation study provides deeper insights into the contributions of different model combinations. Comparisons with individual models and state-of-the-art methods highlight significant accuracy improvements.

Chapter 1

An overview of machine learning techniques

1.1 Introduction

Machine learning, a dynamic field within computer science, has revolutionized technology by enabling computers to learn from data rather than rely solely on programmed instructions. Algorithms form the backbone of this technology, allowing computers to process data autonomously and improve their performance over time. This has led to advancements across industries such as agriculture, banking, robotics, and healthcare, where machine learning aids in tasks like disease diagnosis and treatment planning. Beyond specialized domains, machine learning has permeated everyday life through applications like object detection in self-driving cars, facial recognition, data sorting in search engines, and virtual assistant functionalities like audio-to-text translation. Within machine learning, deep learning stands out as a particularly potent subset, drawing inspiration from the human brain's structure and learning mechanisms. Deep learning models, trained on vast datasets, acquire sophisticated capabilities akin to human cognition, reshaping how computers predict outcomes and solve complex problems. As we explore machine learning and deep learning further, their potential becomes increasingly evident, promising a future where technology continuously learns, adapts, and solves complex problems [11].

In this chapter, we will explore various types of machine learning and deep learning algorithms, including supervised and unsupervised learning. We will discuss classical machine learning algorithms such as Linear regression ,KNNs clustering, artificial neural networks (ANNs), transfer learning, as well as modern deep learning architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers. Moreover, we will explore the concept of ensemble learning, which leverages multiple models to enhance prediction accuracy and robustness. Our goal is to provide a comprehensive understanding of both traditional machine learning and modern deep learning paradigms.

1.2 Artificial intelligence

In recent years, artificial intelligence (AI) has emerged as a transformative force within computer science. This field focuses on developing intelligent machines that can mimic and even surpass human cognitive abilities. AI has become a leading area of technological innovation, offering solutions to complex problems that were once considered beyond the reach of machines. This significant progress in AI is largely driven by advancements in machine learning (ML) and deep learning (DL). These subfields allow machines to learn and adapt from data, enabling them to tackle challenges previously deemed too difficult and automate tasks once thought impossible [11].

1.3 Machine learning

Machine learning (ML) is a core subfield of artificial intelligence (AI). It empowers computers to learn from data and make increasingly accurate predictions on new, unseen data. Introduced by Arthur Samuel in 1959, ML represents a paradigm shift. Unlike traditional programming with explicit rules and predefined inputs, ML algorithms analyze data sets to identify patterns and relationships. This allows them to learn and adapt, leading to improved predictions. The effectiveness of a ML model depends heavily on the quality and suitability of the data used, as well as the chosen algorithms and techniques. As ML continues to evolve, it is reshaping various fields and pushing the boundaries of what machines can learn and achieve [2].

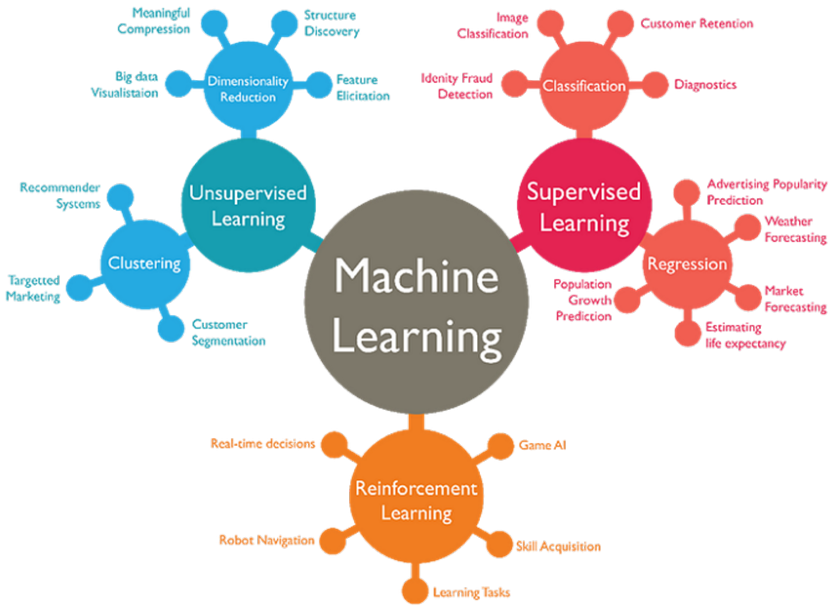


Figure 1.1: Machine learning and its types
[2]

1.3.1 Supervised learning

Machine learning as seen in Figure 1.1 employs supervised learning to train algorithms using labeled data. This data consists of features, which are the informative parts used for prediction, and labels, which act as the correct answers the algorithm aims to learn. By analyzing numerous paired examples of features and labels, the algorithm identifies patterns and improves its ability to predict labels for unseen data. This process is analogous to a student learning from solved problems with answer keys, enabling them to tackle similar problems independently. Supervised learning is particularly useful when labeled data is available, as it allows algorithms to learn effective prediction models for future tasks [12].

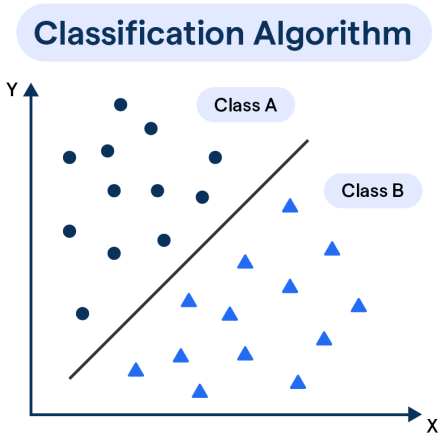


Figure 1.3: Classification algorithms

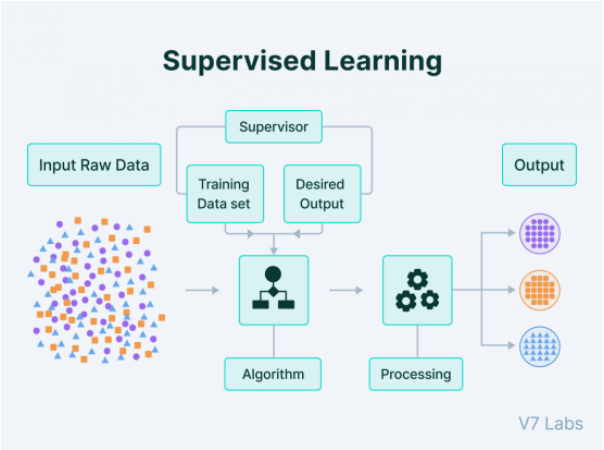


Figure 1.2: Supervised learning [12]

1.3.1.1 Types of supervised machine learning algorithms

Classification: Classification in machine learning allows computers to categorize data points into distinct groups. Similar to sorting objects based on their characteristics, classification algorithms learn from labeled examples. These examples contain both the data itself (object features) and the corresponding category label (bin label). By analyzing these paired examples, the model establishes a relationship between the data and its category usually as seen in Figure 1.3. This learned relationship empowers the model to predict the most likely category for new, unseen data points. Classification algorithms can handle various data types, including numerical (e.g., age, height) and non-numerical (e.g., color, text) inputs. However, their core objective remains consistent: classifying data points into predefined categories. This versatility makes them valuable tools across diverse applications, ranging from image recognition and fraud detection to medical diagnosis and customer segmentation [13].

Regression: It empowers machines to predict continuous values based on input data. Imagine you want to predict the price of a house based on its square footage and location. A regression algorithm would analyze existing data to uncover the underlying relationship between these input features and the continuous output variable usually as

seen in Figure 1.4, which in this case is the house price. This relationship becomes the foundation of the model, allowing it to estimate the price for any new house within the domain it has been trained on. However, unlike classification models that predict discrete categories, regression models are specifically suited for tasks involving continuous outputs like temperature, income, scores, or even the probability of an event occurring. This makes them valuable tools for tasks like predicting sales figures, analyzing stock market trends, and even estimating weather patterns [13].

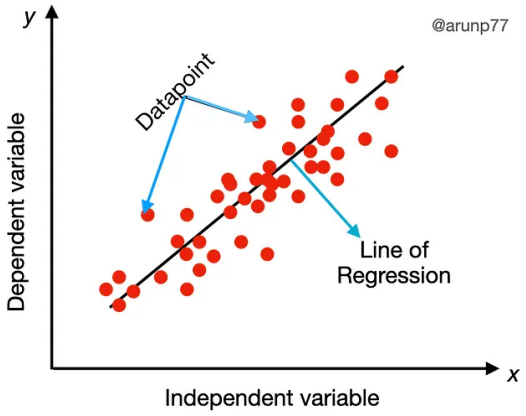


Figure 1.4: Regression algorithms

1.3.1.2 Common supervised learning algorithms

Linear Regression: Linear regression is a cornerstone technique in supervised machine learning. It excels at predicting continuous values. This method achieves this by fitting a straight line that best approximates a set of data points as seen in Figure 1.7. Linear regression leverages labeled data, where each data point consists of an independent variable (input) and a dependent variable (output). The core objective is to identify a linear equation that minimizes the overall discrepancy between the predicted and actual values of the dependent variable across the entire dataset [14].

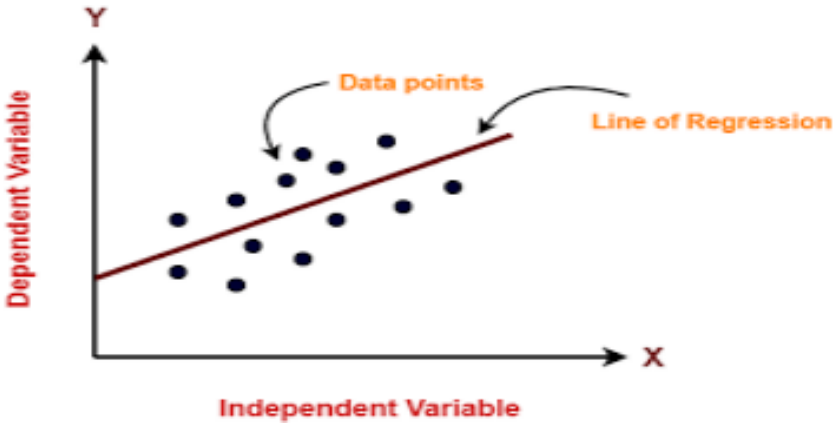


Figure 1.5: Linear Regression

Naïve bayes algorithm: Naive Bayes[14] is a supervised learning algorithm for classification tasks. It falls under the category of probabilistic classifiers, meaning it

leverages Bayes' theorem to make predictions based on the probability of an object belonging to a particular class as seen in Figure 1.6.

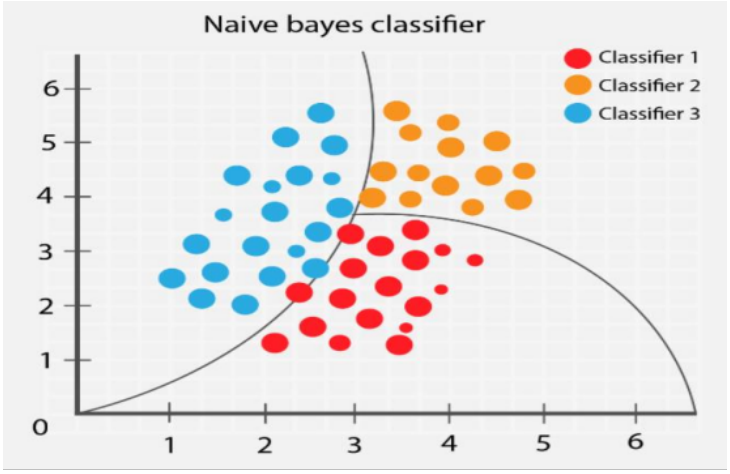


Figure 1.6: Naïve bayes
[15]

Decision trees: Decision trees are a versatile supervised learning method capable of tackling both classification and regression problems. They function by building a tree-like model that predicts the target variable as seen in Figure 1.7. This model is constructed through a series of decision rules learned from the data's features. Similar to a flowchart, these rules ask questions about specific features, leading to further questions or the final prediction. The algorithm iteratively partitions the data into increasingly homogeneous subsets based on these rules, aiming to achieve a point where the target variable is consistent within each subset or further splitting is unnecessary [14]. It's important to acknowledge that decision trees are prone to "greedy learning." This means they make optimal choices at each step without considering the overall impact on the model's performance. This can limit their accuracy and ability to capture intricate relationships within the data.

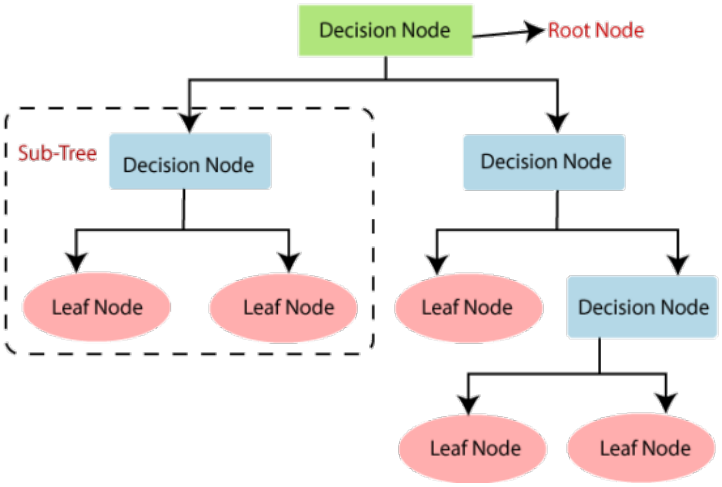


Figure 1.7: Decision Tree Algorithm
[16]

K-Nearest neighbors: K-Nearest Neighbors (KNN) is a widely used supervised learning algorithm for both classification and regression tasks [14]. It works by identifying the k closest data points (neighbors) in the training data to a new, unseen data point Figure 1.8. The labels (in classification) or values (in regression) of these neighbors are then used to predict the class or value of the new data point. In essence, KNN leverages the characteristics of the closest data points to make predictions for new data.

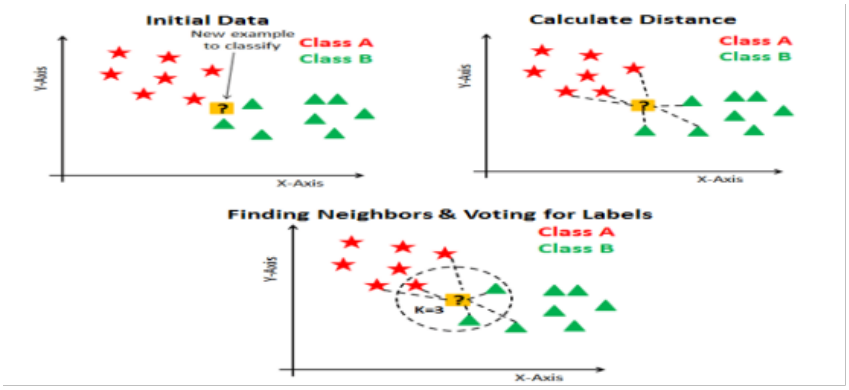


Figure 1.8: K-Nearest Neighbors Algorithm [17]

Support vector machines: Support Vector Machines (SVMs) are a versatile supervised learning method applicable to classification, regression, and even outlier detection tasks [14]. SVMs focus on identifying the most significant training data points that define the optimal separation boundary (hyperplane) between different classes as seen in Figure 1.9. This approach prioritizes maximizing the margin, the distance between the hyperplane and the closest data points from each class [14].

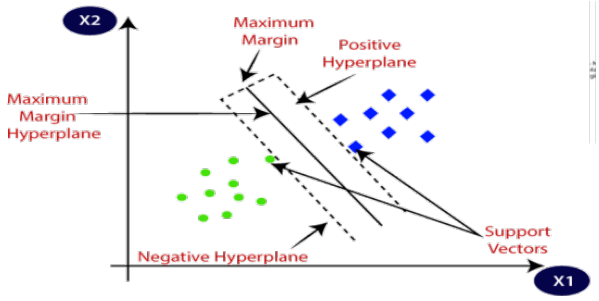


Figure 1.9: Support Vector Machine Algorithm [18]

1.3.2 Unsupervised learning

Unsupervised learning deals with data that lacks labeled outputs (Y). Unlike supervised learning, where algorithms are trained with pre-defined answers, unsupervised methods aim to uncover the inherent structure or patterns within the data itself [14]. This allows researchers to gain valuable findings and enhance their understanding of the data. Algorithms in unsupervised learning operate autonomously, exploring the data to identify and highlight these interesting patterns or structures.

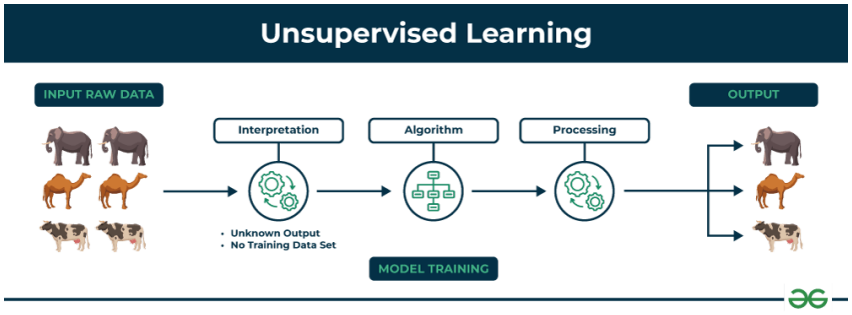


Figure 1.10: Unsupervised Learning
[19]

1.3.2.1 Types of unsupervised learning

Clustering: a popular unsupervised learning technique, groups data points into distinct clusters Figure 1.11. This process aims to organize data points with similar features together, while points in different clusters exhibit significant dissimilarity [20]. Similar to classification tasks, clustering assigns a cluster label to each data point, indicating its membership within a specific group.

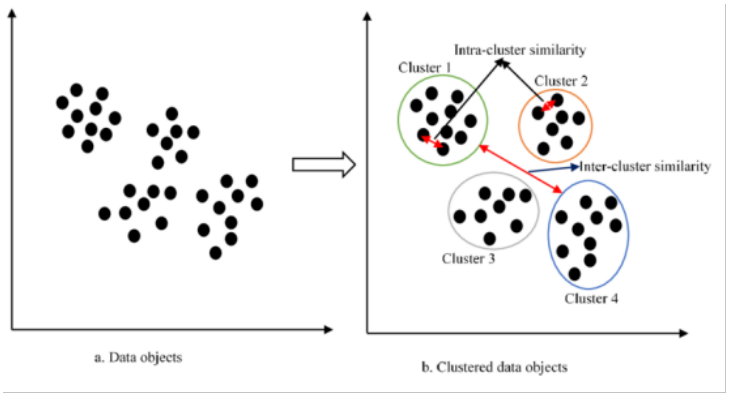


Figure 1.11: Clustering Algorithms

Dimensionality reduction: It is a method of analyzing high-dimensional data by reducing the number of variables by removing ambiguous characteristics or combining multiple features into single elements, making the data more understandable. Dimensional reduction has advantages as it eliminates unnecessary aspects in the data and enables easier visualization, such as in two dimensions, but combined features may become less interpretable, and some information will inevitably be lost. The most basic dimensional reduction approaches are unsupervised algorithms that work on the entire dataset by retaining the most significant features or discovering new ones [21].

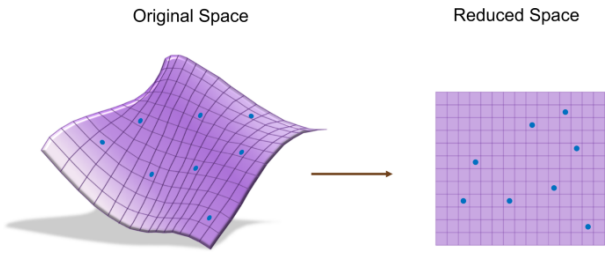


Figure 1.12: Dimensionality Reduction
[22]

1.3.2.2 Common unsupervised learning algorithms

K-Means: K-means clustering, a popular unsupervised learning method, efficiently partitions data points into distinct clusters as seen in Figure 1.13. Data points are assigned to the closest cluster center, which represents the average of its member points. This iterative process of assignment and recalculation continues until clusters stabilize, leading to well-defined groupings [20]. While K-means is known for its simplicity and efficiency, it requires pre-specifying the desired number of clusters, which can be challenging if the optimal number is unclear.

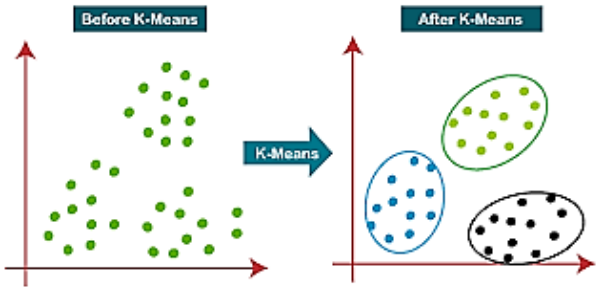


Figure 1.13: K-Means clustering

Principal component analysis (PCA): Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms data to a new coordinate system where features are uncorrelated as seen in Figure 1.14. This allows for selecting a smaller set of features with high explanatory power, resulting in a more compact representation of the data while retaining essential information. PCA facilitates data analysis and visualization in lower dimensions [20].

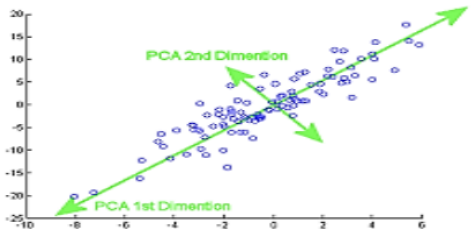


Figure 1.14: Principal component analysis
[23]

1.3.3 Semi-supervised learning

Semi-supervised learning bridges the gap between supervised and unsupervised learning by leveraging both labeled and unlabeled data [24]. This approach is particularly useful when labeled data, a common challenge in machine learning, is limited. Semi-supervised methods exploit the abundance of readily available unlabeled data alongside the scarce labeled data. It functions similarly to a student learning from a combination of solved problems and practice exercises. Unsupervised techniques identify hidden patterns within the unlabeled data, while labeled data guides initial predictions for the unlabeled points. These initial predictions, even if imperfect, are then used to refine existing models or generate new hypotheses, ultimately enhancing learning effectiveness. By strategically combining labeled and unlabeled data, semi-supervised learning offers a powerful approach to tackling real-world problems [24].

1.3.4 Reinforcement learning

Reinforcement learning (RL) offers a distinct approach to machine learning, centered on optimal decision-making within an environment [25]. Similar to a child exploring their surroundings, RL agents learn through trial and error, interacting with the environment, receiving feedback (rewards), and adapting their actions to achieve specific goals [25]. Unlike supervised learning with pre-defined data, RL thrives on continuous interaction. This focus on experience and adaptation makes RL well-suited for dynamic environments where goals require navigating uncertainty [25]. Its success spans various applications, from robot training [26] to self-learning AI in games. This unique learning paradigm paves the way for intelligent agents that can interact with the world and achieve desired outcomes in evolving environments.

1.4 Deep learning

For decades, accessing information from computers required users to adapt to complex protocols and specialized languages, limiting information retrieval to those with specific skill sets. Deep learning (DL), a powerful branch of machine learning (ML), has revolutionized this landscape. Inspired by the human brain's structure and function, particularly the layered architecture of the neocortex, DL algorithms as seen in Figure 1.15 excel at automatically extracting intricate patterns and representations directly from raw data, eliminating the need for manual feature engineering. This bio-inspired approach facilitates more intuitive and natural interactions with technology by automating the feature extraction process. Unlike traditional ML techniques that rely on carefully engineered features, DL models can learn hierarchical representations through successive non-linear transformations, enabling them to discover complex patterns within data without human intervention [3]. This capability has transformed the way users interact with computers, making information retrieval more accessible and intuitive.

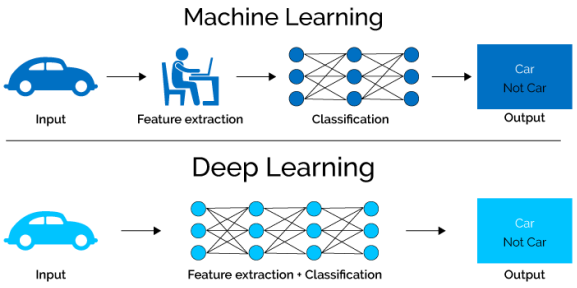


Figure 1.15: Deep learning

1.4.1 Artificial neural networks

Deep learning (DL) algorithms rely on artificial neurons modeled after biological neurons. These artificial neurons process information through weighted connections and activation functions, mimicking the interconnected network in the brain. However, unlike traditional rule-based systems, DL models can learn and adapt. They adjust these connections based on experience, similar to biological learning processes [27]. This dynamic learning capability empowers DL to move beyond pre-programmed rules. It allows them to uncover hidden patterns and relationships within data. This capability not only facilitates more natural user interactions but also opens doors to tackling complex problems previously deemed unsolvable.

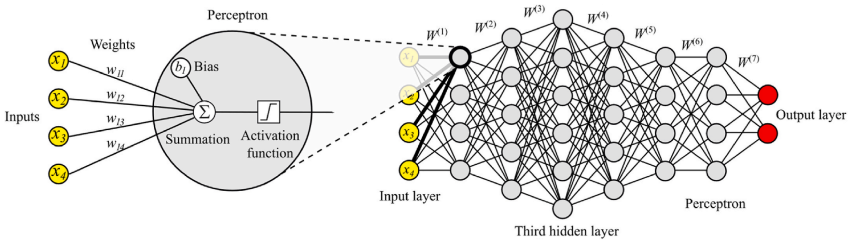


Figure 1.16: Artificial neural perceptron and multilayer perceptron [28]

Neural network components

Input layer: Haykin, (2009). The input layer serves as the entry point for data in a neural network, receiving raw data and transmitting it to the next layer without modification.

Hidden layers: Goodfellow et al, (2016) Hidden layers are the core processing units within a neural network, transforming inputs through activation functions to extract features and build intermediate representations of the data .

Output layer: Goodfellow et al.(2016) The output layer receives processed information from the hidden layers and generates the network's final prediction or interpretation. It employs an activation function tailored to the specific task, such as classification or regression .

Weights: Weights represent the influence of one neuron's output on another and are adjusted during training to optimize the model's performance. Larger weights indicate a stronger influence, while smaller weights reflect a weaker contribution [29].

Biases: Biases are constant values added within each neuron, allowing a neuron to activate even when inputs are zero, adding flexibility and preventing inactivity [30].

Activation functions: Activation functions process information within each neuron, determining whether the weighted sum of inputs is sufficient to activate the neuron and influence the network's output [29]. Different activation functions are suited for various tasks:

Rectified linear unit (ReLU): Offers speed and avoids vanishing gradients, making it a popular choice for diverse tasks, as shown in Figure 1.17.

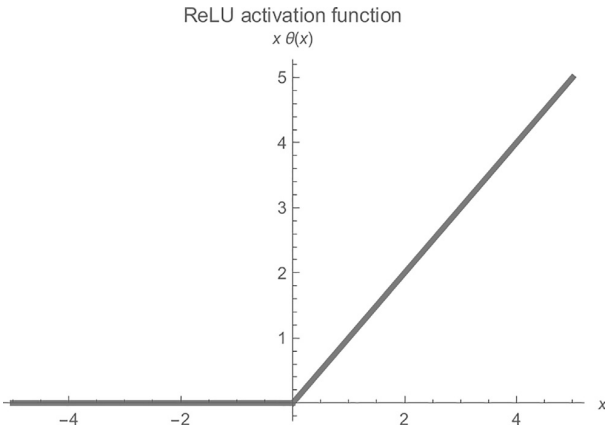


Figure 1.17: ReLU

Sigmoid function: Squeezes values between 0 and 1, ideal for tasks like predicting probabilities, as seen in Figure 1.18.

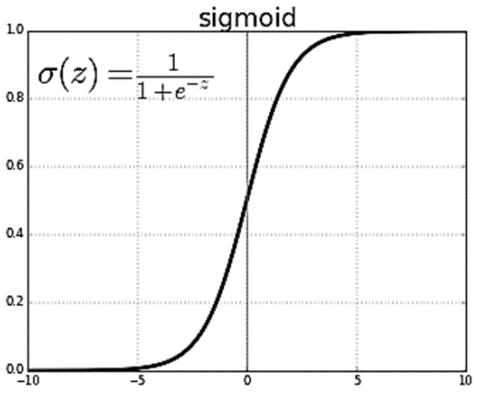


Figure 1.18: Sigmoid function

Tanh function: Similar to Sigmoid but maps values to -1 and 1, offering advantages in certain scenarios, as seen in Figure 1.19.

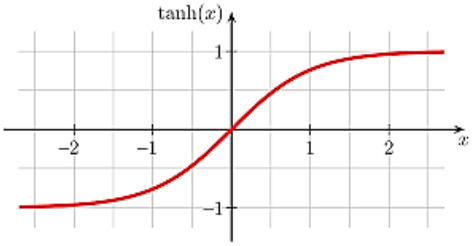


Figure 1.19: Tanh Function

Leaky reLU (LReLU): Addresses the issue of "dead neurons" in ReLU by allowing some negative influence, as shown in Figure 1.20.

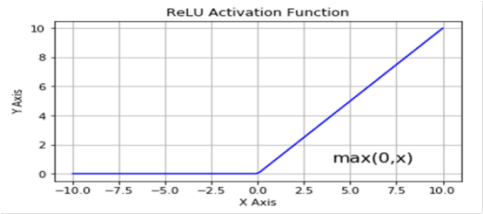


Figure 1.20: Leaky ReLU (LReLU)

Softmax function: Crucial for multi-class classification tasks, transforming outputs into probability distributions, as shown in Figure 1.21.

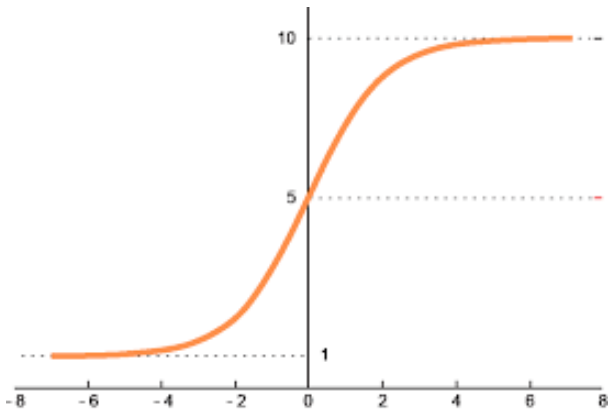


Figure 1.21: Softmax Function

Loss function: The loss function quantifies the discrepancy between the model's predictions and the actual targets [29]. Various forms exist depending on the task and architecture:

Mean squared error (MSE): Used for regression problems, measures the average squared difference between predicted and actual values.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1.1}$$

Binary cross-entropy (BCE): Used for binary classification, calculates the cross-entropy between the predicted probabilities and the true labels.

$$\text{BCE} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{1.2}$$

Categorical cross-entropy (CCE): Used for multi-class classification, computes the cross-entropy between the predicted probability distribution and the one-hot encoded true label.

$$\text{CCE} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij}) \tag{1.3}$$

Hyperparameters Hyperparameters are external settings that influence a neural network's training process [31]. These parameters are set by the researcher and play a crucial role in the learning algorithm's progression, impacting aspects like weight updates. Choosing appropriate hyperparameters is essential for effective learning, avoiding underfitting (limited learning) or overfitting (memorizing the data) [31].

Learning rate The learning rate controls the magnitude of adjustments made to the network's internal parameters during training. It directly affects the speed and direction of the optimization process, aiming to minimize prediction errors [31]. Selecting an appropriate learning rate is vital for efficient learning [31].

Regularization Regularization techniques help mitigate overfitting by imposing constraints on parameter growth during training [31]. Methods like L1 and L2 regularization penalize large weights, encouraging simpler solutions and better generalization to unseen data.

Momentum Momentum is used during training to help the learning algorithm escape local minima [31]. It assists the optimizer in navigating the error landscape by identifying downward slopes leading towards the global minimum, improving training efficiency [31].

Optimizers Optimizers guide the adjustment of model parameters to minimize errors during neural network training. Common examples include Gradient Descent, Stochastic Gradient Descent, and Adam, each offering unique advantages suited for specific tasks [31].

1.4.2 Types of ANN

1.4.2.1 Feed-forward neural network (FNN)

Feedforward neural networks (FNNs) are a type of artificial neural network characterized by a single direction of information flow. Unlike recurrent neural networks (RNNs) that allow for loops, information in FNNs travels in one direction only, passing through multiple processing layers sequentially. This forward-only architecture distinguishes FNNs from RNNs, making them well-suited for tasks where data has a natural order or sequence.

1.4.2.2 Single layer perceptron model (SLP)

The single-layer perceptron (SLP) is a fundamental building block of artificial neural network [32]. Often referred to as the simplest form of neural network, it serves as the foundation for more complex architectures used in deep learning. SLPs excel at classification tasks, assigning labels (like categories) to data points based on their features. As presented in Figure 1.22 an SLP consists of an input layer, a hidden layer with a single neuron, and an output layer. Data is fed into the input layer, then multiplied by weights and summed with a bias term in the hidden layer. This combined value is transformed by an activation function, influencing the output layer's classification decision [32].

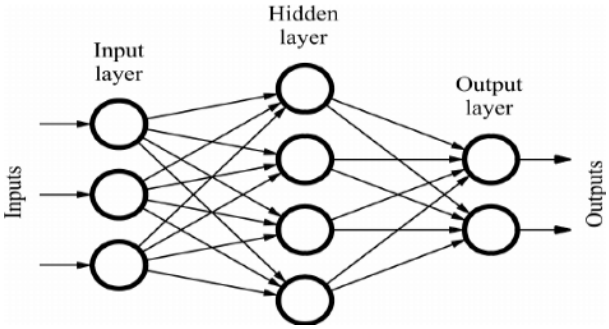


Figure 1.22: Sample of a feed forward neural network [28]

1.4.2.3 Multi-layer perceptron model (MLP)

The multi-layer perceptron (MLP) as presented in Figure 1.23 builds upon the single-layer perceptron (SLP) by introducing multiple hidden layers, forming a more complex feed-forward architecture . Neurons within each layer connect to neurons in subsequent layers, allowing for more intricate information processing . Unlike SLPs, MLPs leverage the back-propagation algorithm, a powerful training method that enables them to learn and adapt to complex data patterns . This increased network complexity empowers MLPs to tackle a wider range of problems compared to SLPs [32].

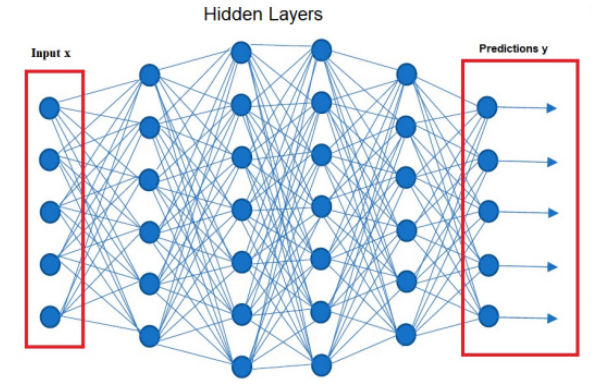


Figure 1.23: Representation of a neural network [33]

1.4.3 Convolutional neural networks (CNNs)

Convolutional Neural Networks (CNNs) have become a dominant force in image classification [34]. Unlike traditional methods requiring manual feature engineering, CNNs excel

at automatically learning features directly from raw image data. This capability is driven by their unique architecture comprised of convolutional layers and pooling operations.

Convolutional layers form the core of a CNN. They employ learned filters to extract relevant features from input images, reducing dimensionality in the process [34]. This generates feature maps, capturing progressively more abstract representations of the image content. Pooling layers further reduce the size of these feature maps, summarizing key information while maintaining efficiency.

This hierarchical structure, as presented in Figure 1.24, empowers CNNs to learn increasingly complex features, from basic edges and textures to higher-level object representations. This multi-stage approach allows them to effectively classify objects within images based on these learned features, achieving impressive results on various tasks.

Additionally, CNNs require minimal pre-processing compared to other algorithms. Their ability to automatically learn relevant features eliminates the need for extensive manual engineering, saving time and effort while potentially leading to superior performance [34].

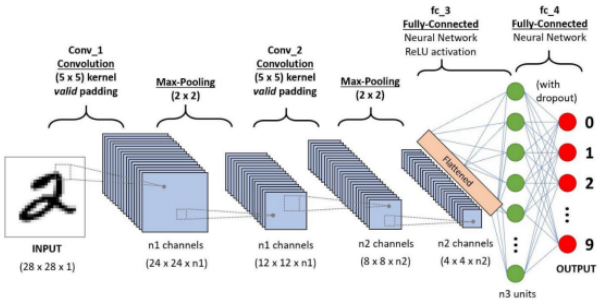


Figure 1.24: Convolution Neural Network [34]

Layers in a convolutional network

A. Convolution layer: Convolutional layers are responsible for feature extraction from image data [35]. These layers utilize filters, or kernels, that slide across the image to identify and capture specific features, as presented in Figure 1.25. The convolution operation involves multiplying corresponding elements between the filter and the image, followed by summation, generating a feature map that highlights the presence and strength of particular features.

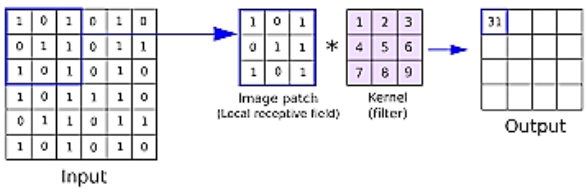


Figure 1.25: Convolution Operation

B. Pooling mayer: Pooling layers, such as those illustrated in Figure 1.26, address the sensitivity of convolutional layers to the precise location of features within the image. They summarize information within feature maps using pooling functions like average pooling and max pooling, reducing the size of feature maps while retaining essential

information [35]. This process enhances the network's robustness to small positional variations in the input image.

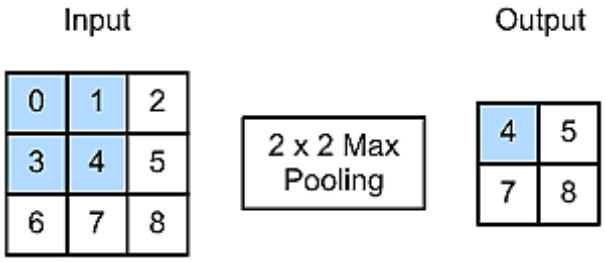


Figure 1.26: Max Pooling

C. Fully connected layer: Following feature extraction and reduction, fully connected layers, as described in [35], bridge the gap between the learned feature representation and the network's final predictions. These layers transform the multi-dimensional feature maps into a one-dimensional vector and perform complex calculations to generate final predictions. Non-linear activation functions like ReLU or Softmax are applied to ensure interpretable and meaningful outputs.

Convolutional neural network models

LeNet: LeNet-5, introduced by LeCun et al., is an early CNN architecture designed for improved pattern recognition. It uses a seven-layer structure with convolutional and subsampling layers, followed by fully connected layers and a softmax classifier [36].

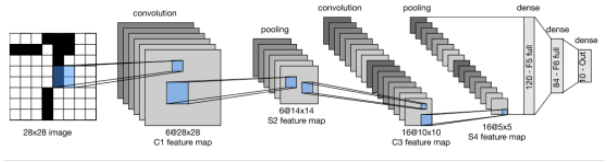


Figure 1.27: LeNet Architecture
[36]

AlexNet: The AlexNet architecture, developed by Krizhevsky et al., significantly advanced CNNs by incorporating greater depth and complexity. It consists of five convolutional layers and three fully connected layers, achieving high accuracy on image classification tasks [36].

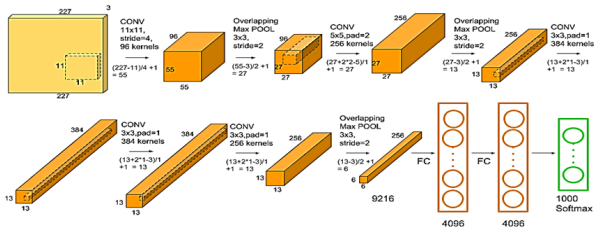


Figure 1.28: AlexNet
[36]

VGG: VGGNet, created by Simonyan and Zisserman, employs small 3x3 filters throughout its convolutional layers and emphasizes depth to achieve high accuracy. It consists of convolutional layers with ReLU activation followed by a final Softmax layer [36].

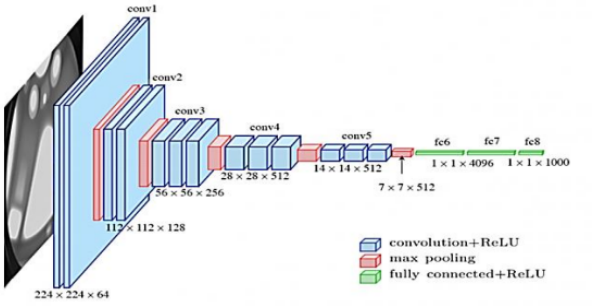


Figure 1.29: VGGNet
[36]

ResNet: ResNet, introduced by Kaiming He, addresses the vanishing gradient problem in deep networks, allowing for the training of very deep networks, such as ResNet-50, which has 50 layers [36].

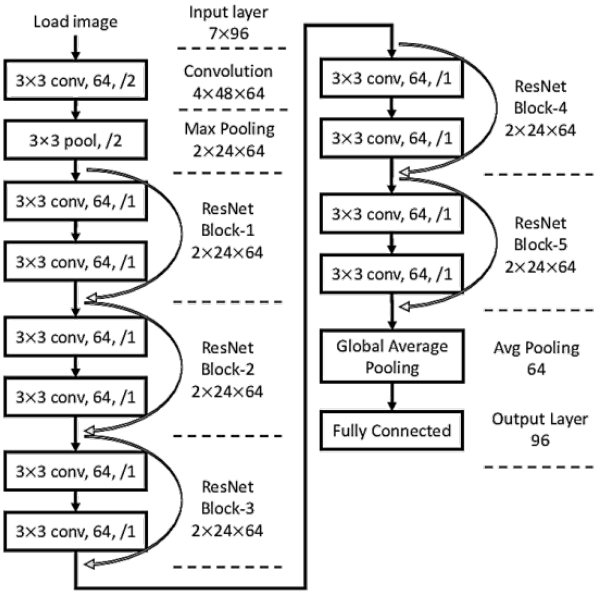


Figure 1.30: ResNet
[36]

Inception: Inception, building on GoogLeNet, merges multiple convolutional filters of varying sizes into a unified filter, reducing the number of trainable parameters and computational complexity [36].

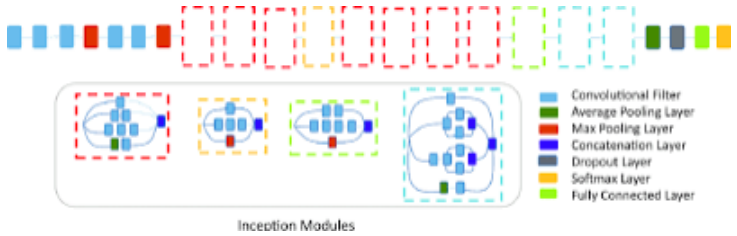


Figure 1.31: Inception
[36]

1.4.4 Recurrent neural networks

Recurrent Neural Networks (RNNs) are designed to handle sequential data by leveraging their temporal memory.

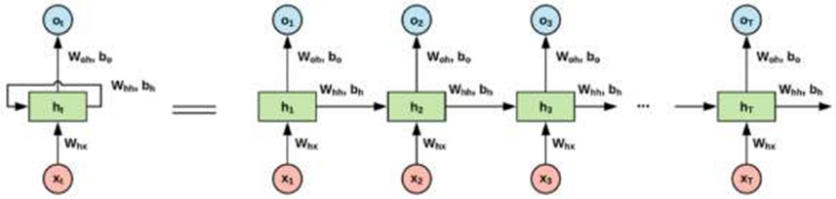


Figure 1.32: RNN architecture unfolded for every time step

Unlike Feed-forward Neural Networks (FFNNs), which process each input independently, RNNs use their hidden state (h_t) to capture information from previous inputs. This is mathematically represented as:

$$h_t = g_1(W_{hh}h_{t-1} + W_{hx}x_t + b_h)$$

$$o_t = g_2(W_{oh}h_t + b_o)$$

where h_t is the hidden state, x_t is the input, b_h and b_o are biases, o_t is the output, and g_1 and g_2 are activation functions. This ability to retain past information makes RNNs suitable for tasks such as language processing and time-series analysis. However, when stacked into deeper architectures, RNNs suffer from vanishing and exploding gradient problems, which hinder their ability to learn long-term dependencies.

To address the limitations of traditional RNNs, more advanced architectures like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) have been developed.

LSTM Long Short-Term Memory (LSTM) networks, introduced by Hochreiter and Schmidhuber in 1997, mitigate the vanishing gradient problem through the use of gating mechanisms. These gates control the information flow within the network:

- **Forget gate:** Decides which information should be discarded from the cell state.
- **Input gate:** Determines which new information is added to the cell state.
- **Output gate:** Selects the information to be output at each time step.

The cell state in LSTMs allows them to retain information over longer periods, making them effective for tasks requiring long-range dependencies.

GRU Gated Recurrent Units (GRUs), proposed by Cho et al. in 2014, simplify the LSTM architecture by combining the forget and input gates into a single update gate and using a reset gate to control the flow of information:

- **Update gate:** Manages the amount of past information to keep.
- **Reset gate:** Controls the incorporation of new information.

GRUs have fewer parameters than LSTMs and can be more efficient while still effectively handling the vanishing gradient problem.

Both LSTM and GRU networks extend the capabilities of traditional RNNs, enabling more robust performance on complex sequential tasks.

1.4.5 Transformer

The Transformer, introduced by Vaswani et al. in 2017, has become a cornerstone of deep learning, particularly in natural language processing (NLP). Initially designed for machine translation, its effectiveness has been demonstrated across various tasks, especially with pre-trained Transformer models achieving state-of-the-art performance. This success has established the Transformer as a preferred architecture, particularly in NLP research. Beyond language, the Transformer's reach extends to computer vision, audio processing, and even scientific fields like chemistry,life sciences and medical imaging. The core principle behind the Transformer's success lies in its use of the attention mechanism, a key concept that warrants further exploration to fully understand how the Transformer operates [4].

1.4.5.1 Attention mechanism

Drawing inspiration from human information processing, researchers have developed attention mechanisms for deep learning models. These mechanisms mimic human focus by selectively attending to relevant parts of vast datasets. One such approach, Bahdanau attention [37], was introduced for machine translation. It calculates attention scores based on a weighted sum of all encoder outputs ("annotations") in relation to the decoder's current state.

1.4.5.2 Attention mechanism in computer vision

In the domain of computer vision (CV), analogous methodologies have been devised. For instance, Hu et al [38]. introduced a novel attention mechanism known as Squeeze-and-Excitation. This mechanism serves to recalibrate features by emphasizing informative aspects pertinent to a specific visual task, while deeming less relevant features as comparatively less significant.

A. Self-attention: As described in [4], reframed the attention mechanism by introducing the concept of queries, keys, and values extracted from the input vectors of the module, distinct from Bahdanau attention. The resultant output is characterized as a weighted combination of values, with the weight assigned to each value determined by the attention computed between queries and keys. The self-attention operation is usually performed in matrix form to accelerate calculation in parallel. To briefly illustrate the concept of self-attention, we first describe it in an element-wise form. For each input $x_i \in \mathbb{R}^c$, $i = 1, \ldots, n$, the corresponding query $q_i \in \mathbb{R}^d_q$, key $k_i \in \mathbb{R}^d_k$, and value $v_i \in \mathbb{R}^d_v$

vectors are generated through the parameters W_q, W_k, and W_v, respectively. d_q, d_k, d_v are the sizes of q_i, k_i, v_i and also the number of features that are learned from x_i.

$$q_i = x_i \times W_q, \quad W_q \in \mathbb{R}^{c \times d_q}, \tag{1.4}$$
$$k_i = x_i \times W_k, \quad W_k \in \mathbb{R}^{c \times d_k}, \tag{1.5}$$
$$v_i = x_i \times W_v, \quad W_v \in \mathbb{R}^{c \times d_v}, \tag{1.6}$$
$$d_q = d_k. \tag{1.7}$$

The output is also a probability calculated as the weighted +sum of the calculated weighting values:

$$\alpha_{ij} = \text{Softmax}\left(\frac{\alpha'_{ij}}{\sqrt{d_k}}\right) = \frac{\exp(\alpha'_{ij}/\sqrt{d_k})}{\sum_j \exp(\alpha'_{ij}/\sqrt{d_k})}, \tag{1.8}$$

where $\alpha'_{ij} = q_i \times k_j^T$, (3) measures the contribution of the jth element of the input to the ith element of the output. Through this operation, α'_{ij} can be regarded as the attention assigned to the element v_i. Thereby, the final output attentions can be computed as a weighted sum of all values as follows:

$$z_i = \sum_j \alpha_{ij} \times v_j. \tag{1.9}$$

The element-wise self-attention can be feasibly extended to matrices. In most cases, the query q_i, key k_i, and value v_i for each input x_i are generated using parallel matrix computation. x_i, q_i, k_i, v_i can be stacked together to matrices, respectively. Let $X \in \mathbb{R}^{s \times c}$ denote the input matrix, Q denote the query matrix, K denote the key matrix, and V denote the value matrix, where s is the number of samples and each matrix consists of the elements, i.e., $X = [x_1; x_2; \ldots; x_s]^T$. Similarly, we compute the attention matrix A and output matrix Z as follows:

$$A = \text{Softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \in \mathbb{R}^{s \times s}, \tag{1.10}$$
$$Z = A \times V \in \mathbb{R}^{s \times d_v}. \tag{1.11}$$

.

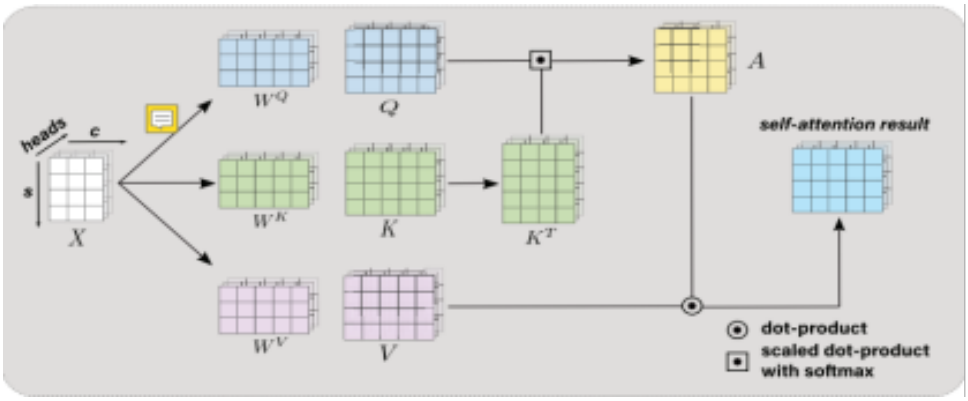


Figure 1.33: A brief illustration of a self-attention mechanism.
[4]

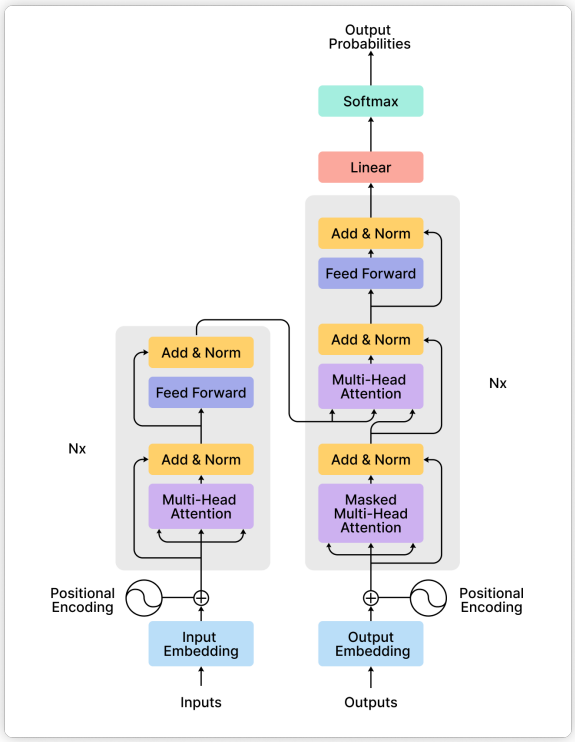


Figure 1.34: Transformer architecture [4]

B. Multi-head self-attention: As demonstrated in [4], employing multiple self-attentions on the same input could effectively capture hierarchical features. These layers of self-attention function akin to employing multiple kernels in convolutional layers Given h self-attentions (heads), the module outputs the final result by concatenating the calculated attentions:

$$Z_i = \text{Attention}(Q \times W_{Qi}, K \times W_{Ki}, V \times W_{Vi}), \quad (7) \tag{1.12}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(Z_1, \ldots, Z_h) \times W_O, \quad (8) \tag{1.13}$$

where W_{Qi}, W_{Ki}, W_{Vi} denote linear projection matrices that map matrices Q, K, V into different subspaces, respectively. W_O is an output projection matrix that concatenates self-attention outputs of all attention heads.

1.4.5.3 Transformer architecture

A transformer architecture as presented in Figure 1.34 is a highly influential framework in deep learning, widely acclaimed for its adaptability and performance across diverse tasks like natural language processing, computer vision, and audio processing. Fundamentally, the transformer comprises an encoder-decoder architecture usually it includes

The encoder: Maps an input sequence $\{x_1, \ldots, x_n\}$ to an output sequence $\{y_1, \ldots, y_m\}$ of the same length, while the decoder generates the output $\{y_1, \ldots, y_m\}$ from the encoded representation z in an element-wise manner, incorporating the previous output as an additional input. In a standard transformer model, the encoder as shown inFigure 1.35 consists of 6 stacked blocks, each containing two main layers: multi-head attention and

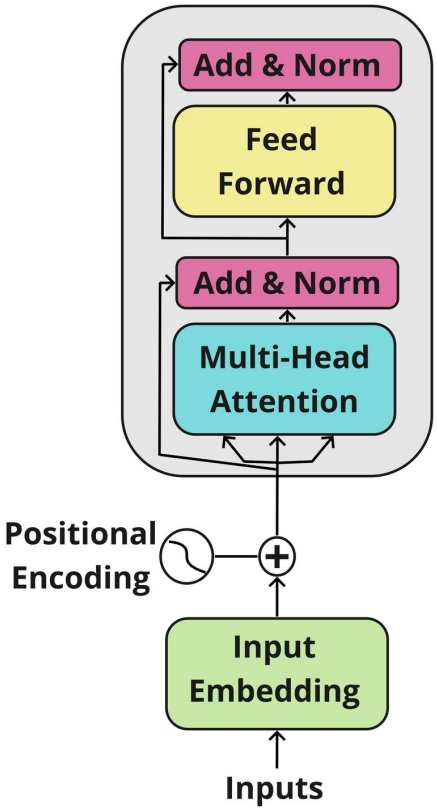


Figure 1.35: Detailed view of a transformer encoder block [39]

feed-forward layers. These are supplemented with residual connections and layer normalization. Within each block, multi-head attention is computed first, followed by layer-wise normalization, where the input and output of the attention layer are summed. Subsequently, a feed-forward layer is applied, followed by another layer-wise normalization step [4].

The decoder: Similarly as seen in Figure 1.36 , the decoder also comprises 6 blocks similar to the encoder, with some adjustments. An additional self-attention layer is added on top of the encoded output. In the first self-attention layer, masking is applied to prevent subsequent contributions to the previous state, considering predictions rely on a known state. Finally, a linear layer and a Softmax layer are included after the decoder's output to produce the final sequence [4].

1.4.5.4 Vision transformers

The impressive performance of transformers in Natural Language Processing (NLP) has led computer vision (CV) researchers to explore their potential in visual tasks. This has resulted in the rapid development of transformer-based models specifically designed for vision, including prominent examples like DETR [41] for object detection, ViT [10] and DeiT [7] for image classification, and Swin-Transformer [9] and BEiT for various vision tasks.

DETR: Carion et al [41] introduced DETR (DEtection TRansformer), a novel approach for object detection in computer vision. DETR leverages transformers, a powerful

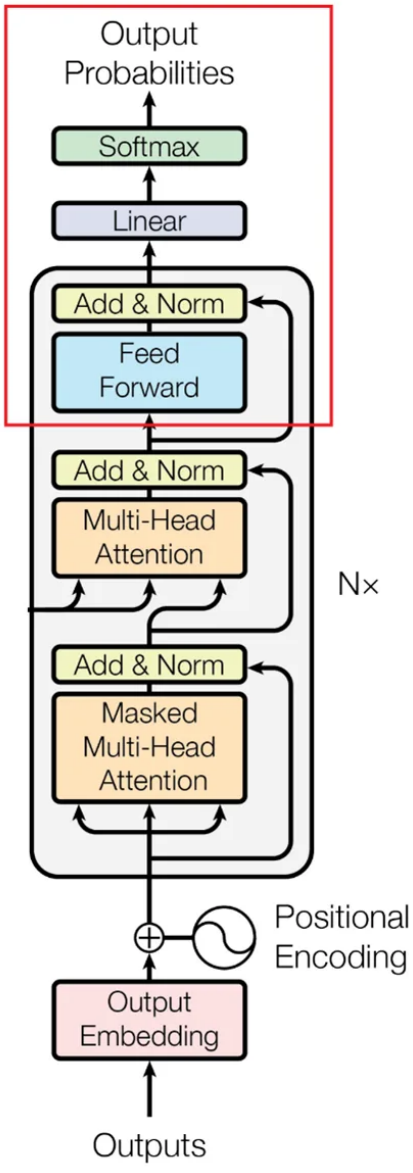


Figure 1.36: Detailed view of a transformer decoder block [40]

architecture from natural language processing, for the task. Unlike traditional methods with manual feature engineering, DETR adopts an end-to-end strategy. First, a convolutional neural network (CNN) extracts informative features from the input image. These features are then processed by a transformer encoder to capture intricate relationships between them. Subsequently, a transformer decoder generates object queries, and a dedicated network assigns class labels and bounding boxes to the detected objects.

VIT: Building on DETR's success with transformers for object detection, Dosovitskiy et al. presented Vision Transformer (ViT) for image classification. ViT [10] operates by dividing input images into fixed-size patches. Each patch is embedded into a vector representation, and additional positional encoding is applied to retain spatial information within the image. These patch embeddings, along with a special class token, are fed into a transformer encoder with multiple layers of Multi-Head Self-Attention (MHSA). The MHSA allows the model to learn relationships between different image patches. Finally, the final state of the class token is used for image classification via a Multi-Layer Perceptron (MLP). Notably, ViT can be further enhanced by incorporating pre-trained convolutional neural network (CNN) feature maps, which can improve the model's ability to capture complex relationships within the image.

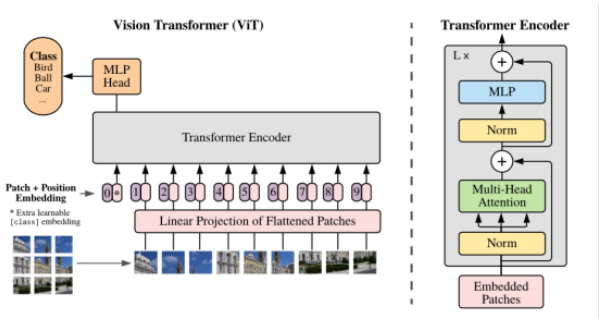


Figure 1.37: ViT transformer [10]

DEIT: To address the significant computational resources required for training Vision Transformers (ViT), Touvron et al. [7] proposed Data-efficient Image Transformers (DeiT). DeiT achieves comparable performance on smaller datasets while requiring less training data. This is achieved through a knowledge distillation framework. Here, a pre-trained, larger ViT model acts as a "teacher," and a smaller ViT-based model, the "student," learns from the teacher's outputs. To facilitate this knowledge transfer, DeiT adds a special "distillation token" after the input sequence. This allows the student model to learn from the teacher's predictions during training. This distillation process effectively transfers knowledge from the teacher to the student, enabling DeiT to achieve good performance on smaller datasets with less training data.

SWIN: Building on Vision Transformers (ViT), Liu et al. [9] introduced Swin-Transformer specifically for processing high-resolution images. Swin-Transformer tackles the challenge of handling large feature variations within an image by employing a novel window-based self-attention mechanism. This approach breaks down the image into smaller local regions (windows) for more efficient processing. Additionally, shifted window attention allows the model to capture relationships across windows, maintaining global context. Swin-Transformer further incorporates patch merging to progressively combine information and build hierarchical feature representations across various scales. This combined strategy allows Swin-Transformer to efficiently handle high-resolution images

by focusing on relevant local information while maintaining consistency across different scales.

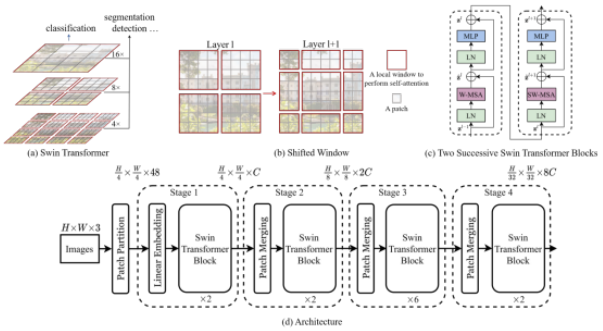


Figure 1.38: Swin-transformer architecture
[9]

BEIT: BEiT, or Behavioral Transformers, is a recent advancement in transformer-based models designed to address the challenges faced during training Vision Transformers (ViT). Developed by Jin et al. in 2023, BEiT offers a solution to the computational and data inefficiencies of ViT by introducing a technique called behavioral cloning. In this approach, BEiT leverages knowledge from a pre-trained Convolutional Neural Network (CNN) teacher model to guide the learning process of a more efficient BEiT student model. Additionally, BEiT employs Masked Image Modeling (MIM), a technique where it predicts masked portions of the input image, further enhancing its learning capabilities. This innovative strategy improves training efficiency, reduces memory usage, and maintains competitive performance across various computer vision tasks.

1.4.6 Transfer learning

Similar to humans who leverage past experiences for new tasks, transfer learning allows machine learning models to benefit from knowledge gained in one domain to improve performance in a related one. This is particularly advantageous when dealing with limited data, as it avoids retraining a model entirely. Transfer learning holds significant potential for various modern neural network architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformers, and other artificial neural networks (ANNs).

Transfer Learning Workflow: The transfer learning workflow typically involves several key steps

Data preprocessing: This crucial step ensures both the original training data and our new dataset are prepared for the model. It can involve normalization, scaling, data cleaning, and potentially data augmentation (artificially increasing the size and diversity of our dataset).

Load the pre-trained model: Choose a pre-trained model from a public repository (e.g., TensorFlow Hub, PyTorch Hub) that aligns with our new task.

Feature Extraction (Freezing Layers): Here, we leverage the pre-trained model's ability to extract valuable features from the data. We typically freeze the weights of the earlier layers in the pre-trained model. These layers are assumed to have captured general-purpose features applicable to various tasks.

Fine-tuning (training new layers): Adding new layers on top of the pre-trained model's frozen layers. These new layers are specific to our new task. We then train these

new layers along with a small portion of the pre-trained model's upper layers (unfrozen) using our new dataset. This allows the model to adapt the pre-extracted features to our specific task.

Fine-tuning with a low learning rate: Since the pre-trained model already holds valuable knowledge, a lower learning rate is used during fine-tuning. This prevents the model from drastically altering the pre-trained weights and focuses on adapting them for your new task Figure 1.39.

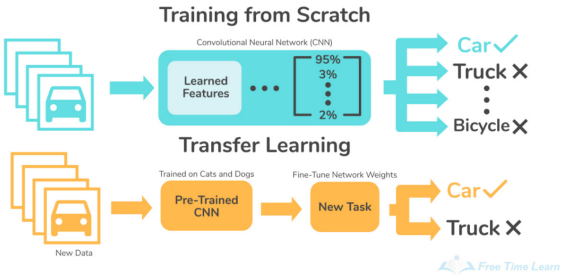


Figure 1.39: Transfer learning

1.5 Ensemble learning

Ensemble learning is a widely adopted technique that significantly enhances the predictive accuracy of machine learning and deep learning algorithms. It achieves this by pooling the predictions from multiple models as seen in Figure 1.40. The core principle lies in the idea that aggregating the forecasts from multiple models often yields superior performance compared to any single model. This approach capitalizes on the diversity of models and their collective wisdom, effectively mitigating individual model weaknesses and boosting overall predictive accuracy (Brownlee, 2022). Ensemble learning offers a multitude of advantages. Firstly, it promotes robustness and stability in predictions by reducing the risk of overfitting to the training data. By leveraging the complementary strengths of diverse models, ensemble methods can generalize well to unseen data, thereby enhancing model reliability. Additionally, ensemble learning enables improved interpretability by providing findings into the consensus among individual models, thus aiding in model understanding and decision-making. Furthermore, ensemble learning facilitates scalability and versatility. It allows practitioners to employ a wide array of base learners and aggregation strategies tailored to the specific characteristics of the dataset and the problem domain. This adaptability empowers practitioners to tackle complex real-world challenges effectively, making ensemble learning a cornerstone technique in modern machine learning practice.

Method	Dependent	Fusion method	Heterogeneity
Bagging	Parallel	Weight Voting	Homogenous
Random	Forest	Parallel Weight	Voting Homogenous
Boosting	Sequential	Weight Voting	Homogenous
AdaBoost	Sequential	Weight Voting	Homogenous
Gradient Boosting	Sequential	Weight Voting	Homogenous
Extreme Gradient Boosting	Sequential	Weight Voting	Homogenous
Stacking	Parallel	Meta Learning	Heterogeneous
Hybrid Ensemble	Both	Both	Heter/Homogeneous

Table 1.1: Categorization of ensemble methods

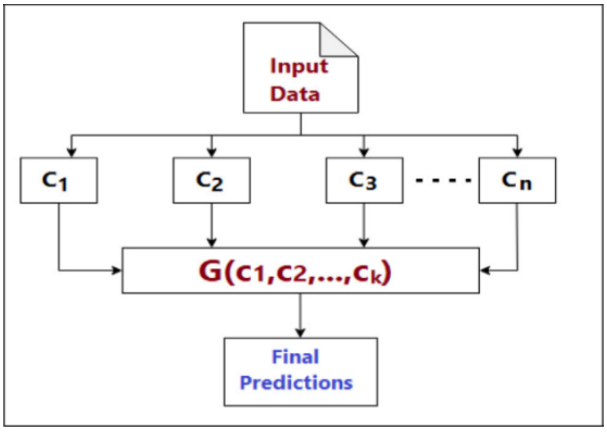


Figure 1.40: General Framework of Ensemble [42]

The fundamental structure of ensemble learning involves utilizing an aggregation function G to merge a set h of baseline classifiers, $c_1, c_2, \ldots, c_h$, in order to predict a single output. With a dataset of size n and features of dimension m, $D = (x_i, y_i)$ (where $1 \leq i \leq n$ and $x_i \in \mathbb{R}^m$), the prediction of the output using this ensemble method is represented by the equation $y_i = f(x_i) = G(c_1, c_2, \ldots, c_k)$. This illustrates the abstract framework of ensemble learning, where ensembles comprise a collection of baseline classifiers trained on input data, whose predictions are combined to yield an aggregate prediction.

All ensembles are made up of a collection of baseline classifiers (classifiers ensemble) that have been trained on input data that produce predictions who are combined to produce an aggregate prediction [5]. Ensemble strategies differ in terms of how they select and train the baseline classifiers. Two strategies, homogeneous and heterogeneous ensembles, generate diversity among the base classifiers based on their nature, either homogeneous or heterogeneous ensembles as seen in [43].

Homogeneous ensemble consists of baseline classifiers of the same type, with each classifier based on different data as seen in Figure 1.41. The feature selection method in this strategy is the same for different training data.However, they face challenges in generating diversity since they rely on the same learning algorithm [6].

Heterogeneous ensemble In contrast, incorporate different types and numbers of baseline classifiers, often trained on the same data but employing different feature selection methods as seen in Figure 1.41.,Homogeneous ensemble methods are favored by researchers for their simplicity and lower cost of construction compared to heterogeneous ones [6]. Ensemble frameworks can generally be characterized by three factors affecting their performance: the dependency on trained baseline models sequential as seen in Figure 1.42 or parallel as seen in Figure 1.43, fusion methods for combining classifier outputs (e.g. weighted voting or meta-learning), and the heterogeneity of the involved baseline classifiers (homogeneous or heterogeneous). These characteristics influence the effectiveness of ensemble methods, as summarized in Table 1, and will be discussed further in the subsequent sections.

1.5.1 Data sampling

The selection of a data sampling method is one of the most important factors affecting the performance of the ensemble system. Diversity in data sampling decisions among baseline classifiers is crucial for effective ensemble learning. Two main strategies for sampling methods within ensemble systems are independent datasets and dependent datasets [44].In the independent datasets strategy, subsets are not reliant on each other [45], while

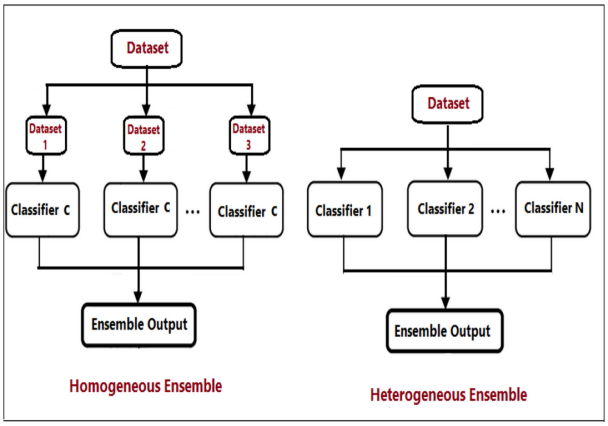


Figure 1.41: General framework of homogeneous and heterogeneous ensemble [42]

in the dependent datasets strategy, subsets are interdependent. An advantage of the independent datasets strategy is that the performance of one subset does not affect others, unlike the dependent datasets strategy where each subset's performance is influenced by preceding subsets [46].

1.5.2 Training baseline classifiers

The diversity among baseline classifiers stands as the second influential factor in ensemble systems. Within ensemble-based systems, two primary techniques for training individual ensemble members are utilized: the sequential ensemble technique and the parallel ensemble technique. In the sequential ensemble technique, different learners are trained sequentially due to data dependency. Consequently, errors made by the initial model are sequentially rectified by subsequent models, exploiting the interdependence between base learners. On the other hand, in the parallel ensemble technique, As demonstrated in figure 8 base learners are generated simultaneously without data dependency. Each data point in the base learner is generated independently, leveraging the independence between base learners. This approach's key advantage lies in its ability to account for differing errors made by independent models, enabling the ensemble model to average out errors effectively.

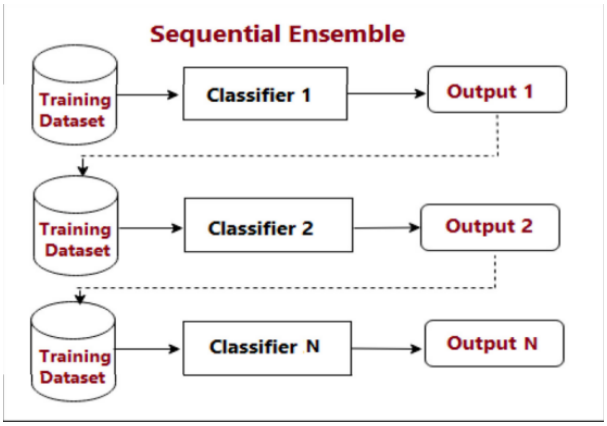


Figure 1.42: General framework of sequential ensemble [42]

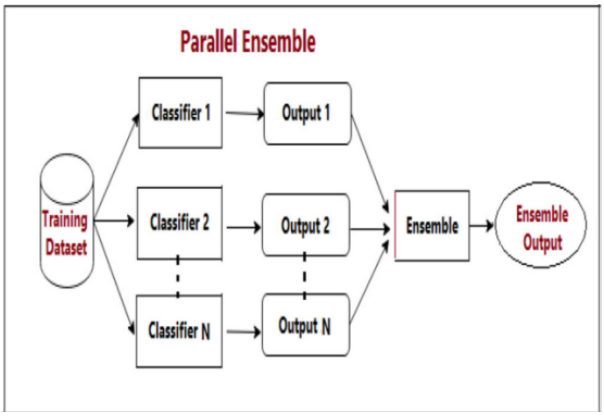


Figure 1.43: General framework of parallel ensemble [42]

1.5.3 Fusion methods

Output fusion involves combining the outputs of baseline classifiers into a single result. There are two common methods: the voting method and the meta-learning method. Each method has its way of integrating classifier outputs, with its own advantages and challenges. It's important to choose the right fusion method for each ensemble technique. These methods can work with different types of data samples and can be used with both parallel and sequential baseline classifiers.

Voting method: Ensemble learning often leverages voting to combine predictions from multiple constituent models, aiming to enhance overall classification or regression accuracy. Voting methods are particularly well-suited for integrating techniques like bagging and boosting, which generate diverse sets of base learners.

There are two main categories of voting methods: hard voting and soft voting. Each offers advantages and drawbacks:.

Hard voting: This is the simplest and most widely used voting method. It works by collecting the predicted class labels from each base learner and selecting the class with the most votes. For example, if three models (Model A, B, and C) predict [0, 0, 1] for a sample, the majority vote would be class 0, resulting in that classification. Hard voting is computationally efficient but can be less robust when dealing with weak learners (models with low individual accuracy) or when multiple base learners produce identical predictions [47].

Soft voting: This method combines predictions by calculating the average probability assigned to each class label across all base learners. It requires storing and utilizing all probability distributions from each model, making it computationally more expensive than hard voting. However, soft voting offers greater flexibility by allowing for various prediction calculation methods, such as using the maximum or average probability values. Soft voting can be more powerful than hard voting in reducing overfitting but assumes all base learners are equally effective, which may not always hold true [48].

Weighted voting: Both hard and soft voting methods can be extended to incorporate weights. These weights represent the relative importance or confidence assigned to each base learner [49].

- **Weighted Hard voting:** Similar to hard voting, the class label with the highest weighted vote is chosen. However, the weights influence the impact of each learner's prediction.
- **Weighted Soft voting:** Soft voting is extended by multiplying each predicted probability by its corresponding weight before calculating the average for each class.

Weighted voting methods offer the potential for improved accuracy by considering the strengths and weaknesses of individual learners. However, determining appropriate weights can be challenging and adds computational complexity. Choosing the most appropriate voting method depends on the specific ensemble learning application and characteristics of the base learners [48].

Meta learning method: The second fusion method is meta-learning [50], also known as "learning to learn," which involves learning from learners. Meta-learning encompasses learning based on prior experience with other tasks to enhance the performance of a learning algorithm by adjusting aspects of the algorithm based on experimental results. Unlike traditional machine learning models, meta-learning involves multiple learning stages, where individual inducer outputs serve as inputs to a meta-learner that generates the final output. Interest in meta-learning has surged in recent years, particularly since 2017 [51], due to the challenges of training advanced machine learning algorithms. Meta-learning addresses these challenges by improving learning algorithms, reducing the number of experiments required, enhancing adaptability to changing conditions, and optimizing hyperparameters. Common meta-learning methods include stacking [52]. Implementing meta-learning presents challenges in defining appropriate approaches and managing computation time complexity, especially with large datasets or multiple baseline models or levels of meta-learning [53].

1.5.4 Common ensemble methods

Four popular ensemble learning methods, namely bagging, boosting, random forest, and stacking, serve as powerful tools for enhancing the machine learning process. Each method operates uniquely, with distinct characteristics pertaining to data generation, training of baseline classifiers, and fusion methods. Our discussion will explorethe workings of each approach, highlighting its nuances and practical applications. Furthermore, we will examine the advantages, limitations, and implementation challenges associated with these methods, providing comprehensive findings into their effectiveness in various contexts.

A. Bagging: The bagging method [54], also referred to as bootstrap aggregating, is a data-specific algorithm designed to create multiple small subsets from the original dataset as seen in Figure 1.44 Its aim is to enhance predictive model diversity by adjusting the stochastic distribution of training datasets, where even minor alterations in the training data can result in significant changes in model predictions. Bagging combines bootstrapping, where ensemble models are trained on bootstrap replicates of the dataset, and aggregation, where the final result is determined by majority voting of the model predictions. This technique is beneficial for reducing variance and preventing overfitting, particularly on high-dimensional data. However, it comes with drawbacks, including high computational costs, increased bias, and reduced model interpretability [55]. Implementing the bagging method presents challenges such as determining the optimal number of base learners and subsets, as well as the fusion method for integrating the outputs of the base classifiers using various voting methods. In essence, bagging employs parallel ensemble techniques, where baseline learners are generated simultaneously without data

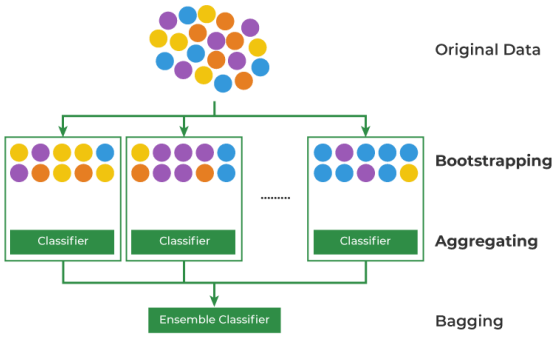


Figure 1.44: Bagging model
[54]

dependency, and the fusion methods rely on different voting techniques. The bagging function is represented as follows [54]:

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{1.14}$$

where $f_b(x)$ represents weak learners and 1_B generates bootstrapping sets.

B. Boosting: The boosting method, introduced by Freund and Schapire in 1997 [56], operates as a sequential process where each subsequent model aims to rectify the errors of the preceding one as seen in Figure 1.45. It involves sequentially incorporating multiple weak learners in an adaptive manner, giving more weight to observations that previous models handled poorly. Boosting, akin to bagging, is applicable to regression and classification tasks and encompasses various algorithms such as Adaptive Boosting (AdaBoost) [56] , Stochastic Gradient Boosting (SGB), and Extreme Gradient Boosting (XGB) [57]. Different types of boosting algorithms have been applied in various studies. For example, AdaBoost has been used for noise detection and speech feature extraction, while XGBoost has been applied in fake news classification and SGB for early prediction of safety accidents at construction sites. Boosting aids in model interpretation and helps mitigate variance and bias in machine learning ensembles. However, a drawback of boosting is that each classifier needs to rectify errors from its predecessors. Implementing boosting poses challenges such as scaling sequential training, computational costs, and susceptibility to overfitting with an increase in the number of iterations. The boosting method employs sequential ensemble techniques, where learners train sequentially with data dependency, and fusion methods rely on different voting methods. The boosting function is represented as follows:

$$f(x) = \sum_{t} a_t h_t(x) \tag{1.15}$$

where $f(x)$ creates a strong classifier from several weak classifiers $h_t(x)$, by building a model from the training data and then creating a second model to correct the errors of the first.

C. Stacking: The stacking method [58], also referred to as Stacked Generalization, is a model ensembling technique utilized to amalgamate information from multiple predictive models to generate a new model, known as a meta-model as seen in Figure 1.46. The architecture of a stacking model comprises two or more base models, termed as level 0

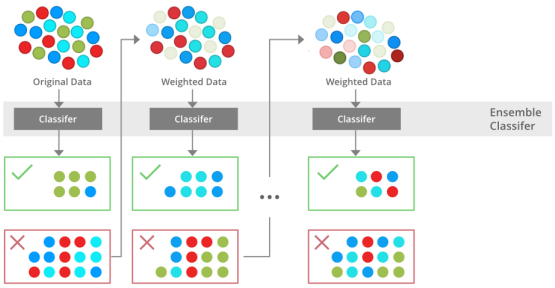


Figure 1.45: Boosting model [57].

models, and a meta-model that combines the predictions of the base models, denoted as a level-1 model. In level 0 models (base models), predictions are compiled from models fitted on the training data. Conversely, in the level 1 model (meta-model), the model learns the optimal combination of predictions from the base models. The stacking method typically outperforms all individual models. For example, Divina et al. (2018) proposed a stacking ensemble learning system to forecast electric energy usage in Spain, and Qiu et al. (2014) utilized stacking to forecast electric energy usage in Australia. Stacking offers the advantage of a deeper comprehension of the data, leading to increased precision and effectiveness. However, over-fitting is a major concern with model stacking, particularly when multiple predictors predict the same target that is merged. Additionally, multi-level stacking incurs high costs in terms of data and time, as each layer adds multiple models. Implementing stacking presents challenges in identifying the appropriate number of baseline models and interpreting the final model. Moreover, as the available data grows exponentially, computation time complexity increases significantly, particularly in highly complex models, which may take months to run. Finally, multi-label classification poses issues such as overfitting and the curse of dimensionality due to the high dimensionality of the data (Chatzimparmpas et al., 2020). the stacking method employs parallel ensemble techniques, where baseline learners are generated simultaneously without data dependency, and fusion methods depend on the meta-learning method. The stacking function is represented as follows (8):

$$f_s(x) = \sum_{i=1}^{n} a_i f_i(x) \tag{1.16}$$

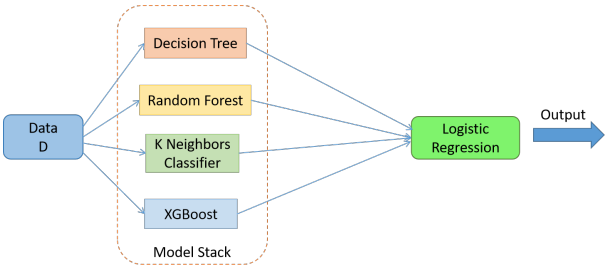


Figure 1.46: Stacking model [58]

D. Random forest: Random Forest is a versatile ensemble method that combines the concepts of bagging and decision trees. It constructs multiple decision trees using ran-

dom subsets of features and data samples. Each decision tree is trained independently, typically using a subset of the training data and a random subset of features as seen in Figure 1.47. This randomness helps to decorrelate the trees, reducing the risk of overfitting and improving the generalization performance of the ensemble. During training, each decision tree is built recursively by selecting the best split at each node based on a randomly selected subset of features. This process continues until a stopping criterion is met, such as reaching a maximum depth or minimum number of samples per leaf node.

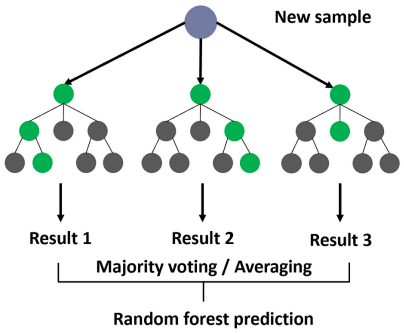


Figure 1.47: Random forest

Once all decision trees are constructed, the final prediction is determined by aggregating the predictions of all trees. For classification tasks, this is typically done through a majority vote among the trees' predictions, while for regression tasks, the predictions are averaged. Random Forests are renowned for their robustness, scalability, and resistance to overfitting. By averaging the predictions of multiple trees and introducing randomness in the tree-building process, Random Forests can effectively handle noisy data and high-dimensional feature spaces. Additionally, they require minimal hyperparameter tuning compared to individual decision trees, making them suitable for a wide range of applications.

1.6 Conclusion

This chapter covers fundamental concepts in machine learning and deep learning. It starts with an overview of machine learning principles, including different types and algorithms. Then, it explore deep learning, focusing on Artificial Neural Networks (ANNs) as foundational components. Various ANN models such as Transformers, Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs) are explored. Additionally, ensemble learning techniques in deep learning for improving model performance are discussed.

In the next chapter, we will explore medical imaging, including definitions, types, and tasks. Specifically, we'll investigate how deep learning techniques are transforming medical image analysis, with a focus on medical image classification. We'll also review related research in this area, examining how researchers are utilizing deep learning specially Transformers and ensemble learning for diverse medical imaging applications.

Chapter 2

State of the art of medical imaging

2.1 Medical imaging

Healthcare medical imaging is essential, for capturing images of organs to aid in diagnosis and treatment. It is vital, for detecting diseases tracking progress and planning treatments ultimately enhancing care and pushing forward research. The history of imaging dates back centuries. Has seen remarkable advancements and innovations that have transformed healthcare practices. With medical imaging doctors are able to diagnose and treat patients without any dangerous Side effects. By being able to observe what's happening inside the body without the requirement for surgery or other invasive measures [1]. The field traces its origins to the early days of X-ray imaging pioneered by Wilhelm Roentgen in 1895, which provided the first glimpse into the internal structures of the human body. Subsequent decades witnessed the development of various imaging modalities, including computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, positron emission tomography (PET), and single-photon emission computed tomography (SPECT), each offering unique insights into different aspects of anatomy and physiology.

Deep learning has emerged as a transformative force in medical image analysis, leveraging advanced computational techniques to extract meaningful information from complex imaging data. With the advent of deep convolutional neural networks (CNNs) and transformers, medical imaging has experienced unprecedented advancements in tasks such as image segmentation, feature extraction, and disease classification. These developments have been propelled by the growing availability of computational resources and large-scale medical image datasets, enabling researchers and clinicians to harness the full potential of artificial intelligence in healthcare. In this chapter will provide an overview of the various aspects involved in medical imaging analysis using deep learning techniques. We will start by introducing different medical imaging modalities and essential tasks such as classification and segmentation. To support the development and evaluation of deep learning models, we will discuss some widely used datasets in the field. We will also examine common performance metrics to assess the effectiveness of these models, ensuring a standardized and reliable evaluation process. The chapter will then explore the application of Convolutional Neural Networks (CNNs) for medical image classification tasks. Building on this, we will look at recent advancements in transformer-based models, which are gaining traction in the medical imaging domain due to their ability to capture long-range dependencies and model complex relationships within the data. We will present both pure transformer-based approaches and hybrid models. Additionally, we will discuss ensemble learning techniques, which enhance predictive accuracy and robustness in medical image analysis by aggregating predictions from multiple individual models.

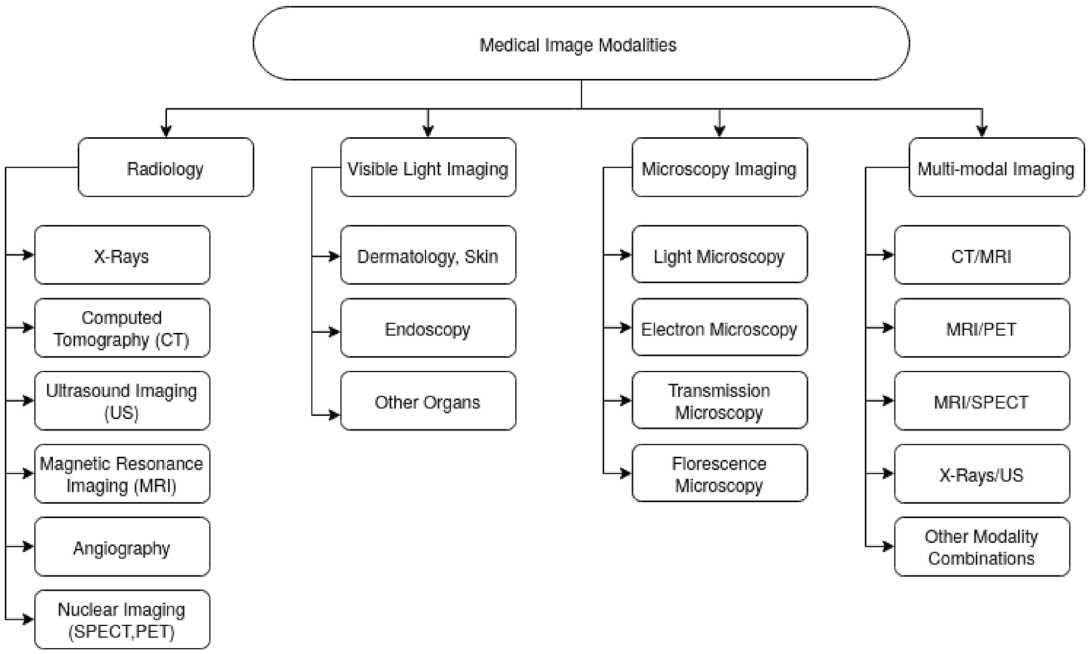


Figure 2.1: Classification of medical imaging modalities

2.1.1 Medical imaging modalities

Medical imaging modalities are numerous and ever-expanding Figure 2.1, driven by continuous technological advancements. All kinds operate diversely to develop images of what's occurring inside the body Figure 2.2. Some of the most popular types include:

2.1.1.1 Magnetic resonance imaging (MRI)

MRI uses magnetic fields and radio waves to examine internal organs and structures. It involves an MRI scanner with a powerful magnet that aligns hydrogen atoms' protons in the body. Radio waves then induce rotation of these protons. When the radio waves stop, the protons emit signals during their relaxation phase, which are detected to produce an image [59].

2.1.1.2 Ultrasound

Ultrasound imaging employs high-frequency sound waves reflected off tissues to create images of joints, muscles, organs, and soft tissues. It's a safe and economical method with no harmful effects, widely used in medical diagnostics [59].

2.1.1.3 Computed tomography (CT)

CT, also known as Computed Axial Tomography (CAT), produces 3D images using x-rays. A CT scanner rotates around the patient, emitting a narrow x-ray beam to create detailed images of bones, blood vessels, internal organs, and soft tissues. CT scans often replace the need for exploratory surgery [59].

2.1.1.4 X-ray

X-ray, or radiography, generates detailed images of internal structures, especially bones, using x-rays. Dense structures like bones appear white, while softer tissues appear darker.

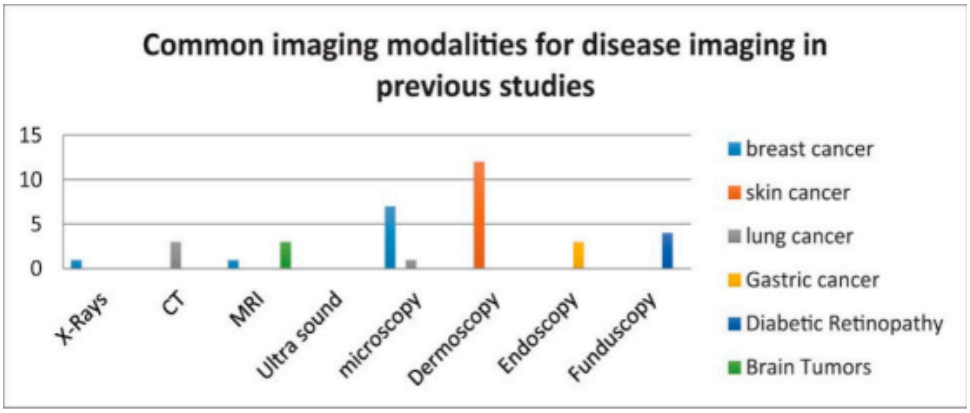


Figure 2.2: Common imaging modalities for disease imaging [60]

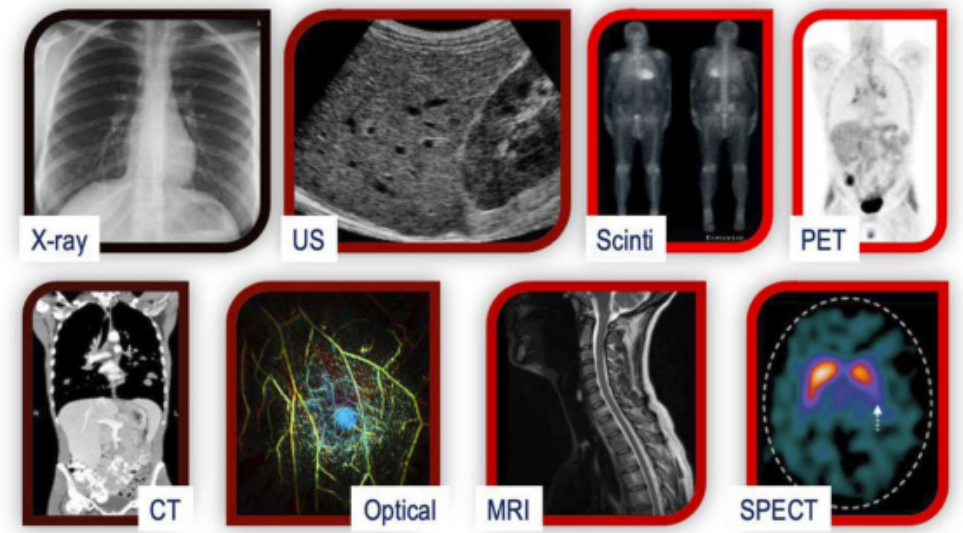


Figure 2.3: A comparison of various medical imaging modalities [60]

X-rays are created by high-speed electrons hitting a tungsten anode, and proper positioning and dosage are crucial for clear images [59].

2.1.1.5 Nuclear imaging

Nuclear imaging uses small amounts of radioactive material (radiotracers) to generate detailed images of internal processes. These radiotracers emit gamma rays, detected by a gamma camera to produce images. This technique provides functional information about blood flow, metabolism, and chemical activity, aiding in the evaluation of various conditions [60].

2.1.2 Tasks of medical imaging

2.1.2.1 Classification in medical imaging

In medical imaging analysis, classification tasks involve assigning labels to entire images or specific regions within them. This categorization is based on the extraction and analysis of features and characteristics present in the image data. The goal is to discriminate between various anatomical structures, pathological conditions, or other relevant factors

of interest. Machine learning and pattern recognition algorithms play a crucial role in this process, enabling applications such as disease diagnosis, tissue characterization, and treatment planning [61].

2.1.2.2 Segmentation in medical imaging

Segmentation in medical imaging involves partitioning or delineating images into distinct regions or objects of interest, such as organs, tissues, or lesions. This process aims to accurately identify boundaries and spatial extent of anatomical structures or pathological abnormalities within the images. Segmentation techniques may utilize various approaches, including thresholding, region growing, edge detection, and machine learning-based algorithms, to extract and separate specific features from the background or surrounding structures. Segmentation plays a crucial role in quantitative analysis, volumetric measurements, treatment planning, and image-guided interventions in medical imaging applications [61].

2.1.2.3 Registration in medical imaging

Registration in medical imaging refers to the process of aligning or matching multiple imaging datasets acquired from different modalities, time points, or imaging orientations to facilitate integration and comparison. This task aims to spatially align images to ensure accurate correspondence between anatomical structures or functional information across different imaging modalities or timeframes. Registration techniques may involve rigid, affine, or deformable transformations to correct for differences in position, orientation, and scale between images. Registration is essential for tasks such as multimodal image fusion, longitudinal studies, image-guided interventions, and treatment planning in medical imaging [61].

2.1.2.4 Detection in medical imaging

Detection in medical imaging involves identifying and localizing specific features, abnormalities, or regions of interest within images that may indicate the presence of disease, injury, or other pathological conditions. Detection tasks typically employ algorithms or methods capable of automatically identifying relevant structures or lesions based on characteristic image properties, such as intensity, texture, shape, or spatial location. Detection algorithms may utilize techniques such as template matching, machine learning-based classification, or deep learning-based object detection to identify abnormalities or anomalies within medical images. Detection plays a critical role in screening, diagnosis, and monitoring of various medical conditions across different imaging modalities [61].

2.1.2.5 Quantification

Quantification in medical imaging involves measuring and analyzing specific parameters or features within images to obtain numerical values or metrics. This task aims to provide objective and quantitative assessments of anatomical structures, physiological functions, or pathological changes present in the images. Quantification techniques may involve computing various metrics such as volume, area, intensity, density, or texture characteristics to characterize tissues, lesions, or other regions of interest within medical images. Quantitative analysis plays a crucial role in monitoring disease progression, treatment response, and outcome prediction in clinical practice [61].

2.1.2.6 Image enhancement

Image enhancement in medical imaging refers to the process of improving the visual quality, clarity, and interpretability of images to facilitate accurate diagnosis and analysis. This task aims to enhance relevant image features, suppress noise, and improve contrast to optimize image visualization and perception. Image enhancement techniques may include filtering, de-noising, histogram equalization, contrast stretching, or adaptive enhancement methods to enhance specific image characteristics or structures while preserving diagnostically relevant information. Image enhancement is essential for improving image quality, reducing artifacts, and enhancing diagnostic confidence in medical imaging applications [61].

2.1.2.7 Image reconstruction

Image reconstruction in medical imaging involves generating high-quality images from raw data acquired by imaging systems, such as CT, MRI, or PET scanners. This task aims to reconstruct images with high spatial resolution, contrast, and fidelity while minimizing artifacts and noise inherent in the acquired data. Image reconstruction techniques may utilize mathematical algorithms, iterative reconstruction methods, or advanced image processing approaches to reconstruct images from raw projection data or signal measurements. Image reconstruction is crucial for generating clinically useful images for diagnosis, treatment planning, and research purposes in medical imaging [61].

2.2 Related work

2.2.1 Datasets

Datasets in the context of medical imaging refer to collections of medical images that are compiled for various research and diagnostic purposes. These datasets typically consist of images obtained from medical imaging modalities such as X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound. Each image within a dataset is annotated with relevant clinical information, such as the presence or absence of certain conditions or pathologies, to facilitate training and evaluation of machine learning algorithms. Medical imaging datasets play a crucial role in advancing research and development in the field of healthcare. These datasets are utilized by researchers, clinicians, and machine learning practitioners to develop and evaluate algorithms for tasks such as disease diagnosis, treatment planning, and patient monitoring. Medical imaging datasets encompass a wide range of medical conditions and imaging modalities, providing a diverse and comprehensive resource for studying various diseases and disorders.

Among the plethora of medical imaging datasets available, certain datasets have gained prominence due to their size, diversity, and relevance to specific research areas.

COVID-19 database: A public database was created by the authors combining several public databases and also by collecting images from recently published articles. The database contains a mixture of 423 COVID-19, 1485 viral pneumonia, and 1579 normal chest X-ray images [62].

Chest-xray-classification: This dataset was exported via roboflow.ai on March 31, 2022 includes 5824 images. Pneumonia are annotated in folder format.The following pre-processing was applied to each image:Auto-orientation of pixel data (with EXIF-orientation stripping) Resize to 640x640 (Stretch),No image augmentation techniques were applied.

RSNA intracranial hemorrhage dataset: The RSNA Brain Hemorrhage CT Dataset [63] is the largest public dataset of its kind containing a very large and heterogeneous collection of brain CT studies from multiple scanner manufacturers, institutions, and countries. It is also a "real-world" dataset containing complex examples of cerebral hemorrhage in both the inpatient and emergency setting. The dataset, released under a noncommercial license, has representation of a large variety of cerebral pathologic states for use in future machine learning applications.

COV19-CT-DB: The dataset used in this paper for training and validation is the COV19-CT-DB dataset, which consists of 5,000 chest CT scans that are annotated as COVID-19. Data was collected in the period from September 1, 2020 to March 31, 2021. The data were aggregated from different hospitals, containing anonymized human lung CT scans with signs of COVID-19 and without signs of COVID19.

OASIS-3: Is a longitudinal multimodal neuroimaging, clinical, cognitive, and biomarker dataset for normal aging and Alzheimer's Disease. OASIS-4 contains MR, clinical, cognitive, and biomarker data for individuals that presented with memory complaints[64]. The OASIS datasets hosted by central.xnat.org provide the community with open access to a significant database of neuroimaging and processed imaging data across a broad demographic, cognitive, and genetic spectrum an easily accessible platform for use in neuroimaging, clinical, and cognitive research on normal aging and cognitive decline.

POCUS: dataset is more than 250 recordings (92 COVID-19, 73 bacterial pneumonia and 90 healthy controls). The dataset comprises data from various sources, including unpublished clinical data collected in a hospital by our collaborator in Northumbria (UK), as well as recordings from healthy volunteers scanned in Neuruppin (Germany), but also data from other publications and educational websites.

Busi: The data collected at baseline include breast ultrasound images among women in ages between 25 and 75 years old [65].This data was collected in 2018. The number of patients is 600 female patients. The dataset consists of 780 images with an average image size of 500 $\times$ 500 pixels. The images are in PNG format. The images are categorized into three classes, which are normal, benign, and malignant.

The HAM10000: ("Human Against Machine with 10000 training images") dataset is a significant resource in the field of automated diagnosis of pigmented skin lesions. It addresses the challenge of limited dataset size and lack of diversity by providing a collection of 10,015 dermatoscopic images sourced from various populations and acquired through different modalities.

Dermnet: The dataset comprises around 19,500 images of 23 skin diseases sourced from , with approximately 15,500 images allocated to the training set and the rest to the test set. Additionally, there are 3,152 images of colorectal polyps, each sized 224 by 224 pixels, sourced from the Department of Pathology and Laboratory Medicine at Dartmouth-Hitchcock Medical Center (DHMC). These images are labeled by seven pathologists from DHMC and aim to facilitate the development of new methodologies for histology image analysis in digital pathology.

The NCT-CRC-HE-100K: dataset is a set of 100,000 non-overlapping image patches extracted from 86 HE stained human cancer tissue slides and normal tissue from the NCT biobank (National Center for Tumor Diseases) and the UMM pathology archive

(University Medical Center Mannheim). While the dataset Colorectal Cacner-Validation-Histology-7K (CRC-VAL-HE-7K) consist of 7180 images extracted from 50 patients with colorectal adenocarcinoma and were used to create a dataset that does not overlap with patients in the NCT-CRC-HE-100K dataset. It was created by pathologists by manually delineating tissue regions in whole slide images into the following nine tissue classes: Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM).

MedMNIST: It is a comprehensive dataset tailored for biomedical images, akin to MNIST but specifically designed for biomedical applications. It includes 12 datasets for 2D and 6 datasets for 3D images, with images standardized to sizes of 28×28 (2D) or 28×28×28 (3D) and accompanied by classification labels. This dataset, comprising a total of 708,069 2D images and 9,998 3D images, supports various classification tasks across different dataset sizes and objectives Figure 2.2. One of its key features is its diversity across data modalities, dataset scales, and task complexities, making it suitable for evaluating machine learning and deep learning algorithms in diverse scenarios. With standardized formats and train-validation-test partitions provided for all datasets, MedMNIST v2 facilitates straightforward algorithmic comparisons.it also emphasizes efficiency, with its compact image sizes conducive to evaluating machine learning algorithms efficiently. Moreover, its availability under the Creative Commons (CC) License ensures easy accessibility for educational purposes, benefiting interdisciplinary researchers without prerequisite expertise in related fields [66].

MedMNIST-2D datasets

1. PathMNIST: Analyzes patches of colorectal cancer histology slides, consisting o100,000 training images and 7,180 test images, to predict patient survival.

2. ChestMNIST: Studies over 100,000 X-ray images to detect and classify various chest diseases, including pneumonia, tuberculosis, and lung cancer.

3. DermaMNIST: Examines 10,015 dermatoscopic images depicting common pigmented skin lesions like melanoma, nevus, and seborrheic keratosis.

4. OCTMNIST: Focuses on diagnosing retinal diseases such as macular degeneration and diabetic retinopathy by analyzing over 100,000 optical coherence tomography images.

5. PneumoniaMNIST: Identifies cases of pneumonia from a dataset of nearly 6,000 pediatric chest X-ray images, aiding in early diagnosis and treatment.

6. RetinaMNIST: Assesses the severity of diabetic retinopathy using 1,600 retina fundus images, categorized into different levels of disease progression.

7. BreastMNIST: Helps detect breast abnormalities, including benign and malignant tumors, from 780 ultrasound images of breast tissue.

8. BloodMNIST: Classifies different types of blood cells, such as red blood cells, white blood cells, and platelets, from over 17,000 images, aiding in blood-related disorder diagnosis.

9. TissueMNIST: Analyzes images of kidney cortex cells, categorized into different tissue types, to study cellular abnormalities and diseases like renal carcinoma.

MedMNIST-3D datasets:

1. OrganMNIST-3D: Utilizes 3D CT images to classify various organs, such as the liver, lungs, and kidneys, aiding in organ-specific disease diagnosis.

2.NoduleMNIST-3D: Focuses on lung nodule detection and malignancy classification using 3D images from CT scans, crucial for early lung cancer diagnosis.

3. AdrenalMNIST-3D: Helps diagnose adrenal gland conditions, such as tumors or hyperplasia, by analyzing 3D shape masks derived from abdominal CT scans.

4.FractureMNIST-3D: Assists in classifying different types of rib fractures, including buckle, nondisplaced, and displaced fractures, using 3D CT images.

5. VesselMNIST-3D: Identifies intracranial aneurysms and other vascular abnormalities from 3D brain vessel models reconstructed from medical imaging data.

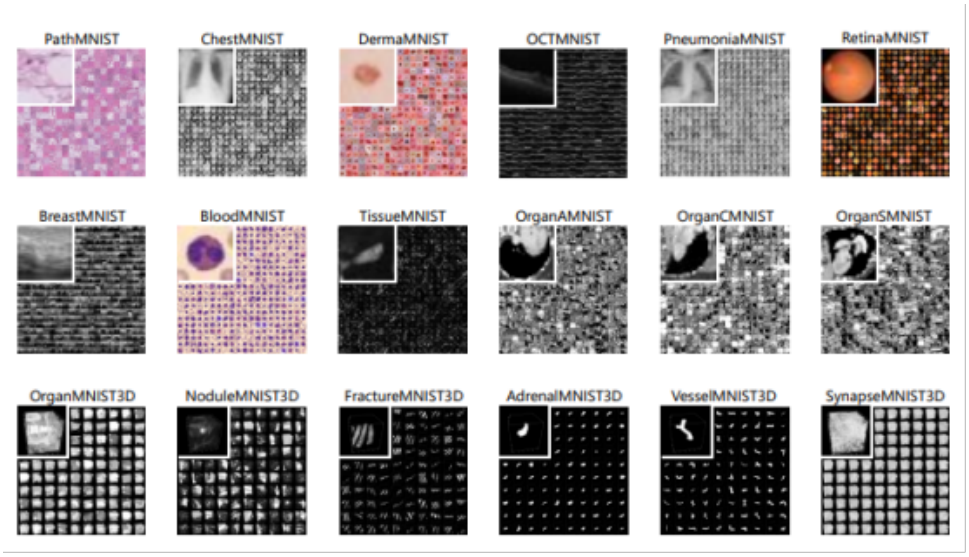


Figure 2.4: An overview of MedMNIST v2 [66]

2.2.2 Performance evaluation metrics

Performance metrics as presented in Table 2.2, as quantitative measures, serve as crucial tools for evaluating the effectiveness and accuracy of medical image understanding algorithms. They provide insights into the algorithm's performance by quantifying various aspects of its behavior, such as its ability to correctly classify images and identify different types of errors. These metrics are derived from the confusion matrix, which acts as a foundation for performance evaluation. The confusion matrix as shown in Table 2.1 , a visual representation of the algorithm's performance, outlines actual and predicted results, including True Positives, True Negatives, False Positives, and False Negatives. From this matrix, performance metrics like accuracy, sensitivity, specificity, precision, F1

score, and area under the receiver operating characteristic curve (AUROC) are computed. Together, these metrics offer a comprehensive assessment of the algorithm's strengths and weaknesses, guiding further refinement and optimization efforts in medical image analysis.

Table 2.1: Confusion matrix

Total test samples	**Predicted positives**	**Predicted negative**
Actual positive	True positives (TP)	False negatives (FN)
Actual negative	False positives (FP)	True negatives (TN)

Table 2.2: Evaluation Metrics and Formulas

Evaluation metric	**To determine**	**Formula**	**Preferred value**
Accuracy	Overall how often the result is true	$\frac{TP+TN}{T}$	High (≈ 1)
Misclassification result (error rate)	Overall how often the result is false	$\frac{FP-FN}{T}$	Low (≈ 0)
True positive rate (TPR) (recall or sensitivity)	When it is actually true, how often does it predict true	$\frac{TP}{FN+TP}$	High (≈ 1)
False positive rate (FPR)	When it's actually false, how often does it predict true	$\frac{FP}{FP+TN}$	Low (≈ 0)
False negative rate (FNR)	When it's actually true, how often does it predict true	$\frac{FN}{FN+TP}$	Low (≈ 0)
Specificity	When it's actually false, how often does it predict false	$\frac{TN}{FP+TN}$	High (≈ 1)
Precision (positive predictive value (PPV))	When it predicts true, how often is it correct?	$\frac{TP}{FP+TP}$	High (≈ 1)
F-Score (Dice similarity coefficient)	Harmonic mean of recall and precision.	$2 \times \frac{PPV \times TPR}{PPV+TPR}$	High (≈ 1)
Receiver operating characteristic (ROC)	Commonly used graph to summarize the performance of a classifier over all possible thresholds.	Plot of recall versus FPR	High (≈ 1)
Area under ROC (AUC)	The area under ROC	Plots TPR versus FPR	High (≈ 1)

2.2.3 CNN in medical imaging classification

For the last few years, researchers have been using CNN-based models to extract unique and useful features for the diagnosis of various diseases, including but not limited to brain

cancer, heart disease, Alzheimer's disease (AD), COVID-19, Parkinson's disease, breast cancer [67], by using medical images. According to previous studies, using convolutional neural network-based models achieved a good level of accuracy when compared with traditional machine learning and volumetric techniques that are manually performed by physicians.
Hosseini-Asl et al. and Payan and Montana [32]both applied 3D convolutional neural networks based on an MRI(Magnetic resonance images) scans of the brain. The models were used to predict the disease status (cognitively normal older individuals, cognitive impairment, and Alzheimer's disease) of apatient, achieving an accuracy of 89.1% and 89.% respectively. However, the full characteristics of lesions related to Alzheimer's disease cannot be easily captured since the feature extraction withdown-sampling and the following classifier models are independent. To tackle this issue, a classification framework which combines convolutional (CNN) and recurrent neural networks (RNN) was presented by Cui et al. [33]. In their paper, using 3D-CNN, the spatial features are learned from constructing convolutional neural networks (CNNs) and then the longitudinal features for AD classi-fication are extracted by applying recurrent neural networks (RNN) with three cascaded bidirectional gated recurrent units (BGRU) layers. Apart from the application of 3D CNNs, 2D CNNs still work by analyzing 3D volumes slicebyslice [34–36]. Although the accuracy of 2D CNNs is lower than 3D CNNs, the applications of 2D networks can reduce the computation complexity or process the data sets with thick slices relative to in-plane resolution.

Alexander et al(2021). introduced a CNN model using MRI and diffusion-tensor imaging (DIO) was used for the classification of AD patients [68]. According to their study, the classification performance demonstrates that the size of the hippocampal region of interest (ROI) does not matter when bigger ROIs are combined with using CNN architecture for the classification. Using a six-layered convolutional neural network with $48 \times 48 \times 48$ ROI with a data fusion model achieved a good accuracy of 96.7% for their case (AD-Normal Control).

Villa-Pulgarin et al, the focus was to classify skin lesion cancersby using CNN-based models DensNet-201, Inception-ResNet-V2, and Inception-V3 [69]. In their work, they tested the models with different workflows, fine-tuning the optimization and using data augmentation. The best results of their model were obtained by using the HAM10000 dataset, with an accuracy of 98% using the data augmentation stage, and 93% by using the ISIC 2019 data-set using the optimized DenseNet-201 model.

Li et al. used the CNN model for the classification of lung image patches with interstitial lung disease (ILD) patterns [70]. Their proposed architecture can correctly identify the features of the image from the lung patches of ILD. The authors have compared their classification results with three different methods of feature extraction: the rotationinvariant local binary pattens (LBP) feature with three resolutions; the Scale-Invariant Feature Transform (SIFT) feature with a key point located at the patch center; as well as feature learning without supervision through the use of the unsupervised restricted Boltzmann machine (RBM). These three techniques are classified by using SVM. However,the proposed automatic CNN model did not use any extra classifier such as SVM, as their classifier model is trained by the three fully connected layers. Therefore, using the ANN model for the classification training has the advantage that the potential to use the backpropagation method to fine-tune the parameters in each of the layers may achieve a more accurate final classification approach. Out of all three techniques, their customized CNN model performed the best. A multiclass CNN architecture using MRI images was used for the detection of brain tumors. The presented model achieved an accuracy of 99% for classifying the four different classes (glioma, tumor, meningioma, and pituitary

tumors). The primary objective of this research was to get a faster learning rate with higher classification results while comparing with traditional deep learning models [71].

Yildirim et al. used the CNN-based MAColonNET model for identifying colon cancer using colon histopathological images [72]. The proposed model used 45 layers for classifying two classes of colon cancer with an accuracy rate of 99.75%. Additionally, this CNN-based model is applicable for pre-diagnosis purposes in non-specialist locations and reduces the workload pressure of experts, which can help them to avoid mistakes A VGG-19 model was trained on 3,797 chest X-ray images [73] for classification of Covid19, pneumonia and healthy cases. An accuracy of 97.11% on the test dataset was obtained. In addition, for further study, the original images and their matching categories were then stored in a Mango DB database. EfficientNet, GoogLeNet, and XceptionNet were integrated [74] in a study to classify patients as positive for COVID-19, pneumonia, or tuberculosis, or healthy. For a binary classifier the accuracy was 98%, while for multi-class, the accuracy was 99%.The dataset used for training and testing was taken from two sources. The authors also tried to keep the possible false predictions to a minimum, and hence obtained a better accuracy and generalized model. Another parameter was for no false positives, to have the model maintain a high specificity rate, which keeps the model much more reliable. Another study [75] assessed how well the transfer learning-based CNN models VGG- 16, ResNet-50, and Inception-v3 predict the presence of brain tumor cells. The models were trained and tested using a dataset of 233 MRIs. Accuracy was used to measure performance, and the findings revealed that the VGG-16 model gave results that were extremely accurate as compared to the other models. The trainable data for the VGG- 16 model, which employs 3×3 convolution kernels and 2×2 max-pool kernels and includes 138 million hyperparameters, was decreased by 44.9 percent. As a result, learning rates increased and overfitting was decreased. The ResNet50 model is a pretrained CNN model that permits training with more convolution layers without increasing training error rates. The Inception-v3 model uses parallel Inception modules to reduce depth in convolution layers.

Yang et al (2023) provides a comprehensive evaluation framework for "MedMNIST v2 - A Large-Scale Lightweight Benchmark [76]. provides . MedMNIST v2 includes 12 2D and 6 3D datasets from diverse medical imaging modalities, aiming to facilitate reproducible and accessible research. The study evaluates the performance of various CNN architectures, including ResNet, DenseNet, and EfficientNet, demonstrating their effectiveness across different tasks. Additionally, the paper explores the use of lightweight models designed for efficiency and the application of ensemble methods such as bagging and boosting to enhance model performance. The performance of these models is assessed using metrics like accuracy, precision, recall, F1-score, and ROC-AUC, providing a robust benchmark for future research in biomedical image analysis (Yang et al., 2023).

2.2.4 Transformers in medical imaging classification

Transformers, known for their attention-based mechanisms and ability to model long-range dependencies, have seen a surge in recent applications within the medical imaging domain. In this section, we will survey and explore recent works utilizing transformer-based architectures for medical image classification tasks related to disease diagnosis and prognosis. These classification methods using transformers can be divided into the following categories:

2.2.4.1 Pure Transformers-Based-Approaches

We call ViTs that are similar to the originally proposed one [10] pure transformers. These methods usually do not contain significant structural changes compared with the original

method. We introduce the literature of pure transformers by image modality, including X-ray [77], CT [78], magnetic resonance imaging (MRI) [79], ultrasound [80], and optical coherence tomography (OCT) [81]. During the COVID-19 pandemic in particular, X-ray has played a very important part in disease screening and is thus a popular modality for AI researchers to use when designing transformer-based methods.

Liu et al. [82] developed the vision outlooker (VOLO), a ViT model that replaced the original attention mechanism with the outlooker attention, as proposed in [83]. Their model achieved SOTA performance for the diagnosis of COVID-19 without pretraining on ImageNet. (In this study they trained and validated a new proposed model VOLO tailored with transfer learning technique using a larger X-ray dataset, compared with the related work. Early studies on detecting COVID-19 generally suffered from insufficient data and imbalanced data distribution. VOLO outperforms SOTA on the targeting datasets and an accuracy of 99.67% was achieved. The generality of VOLO was verified across datasets. However this study had limitation like small dataset, The study's dataset size is relatively small, potentially hindering generalizability and robustness. And also Limited validation: The model's performance was only evaluated on the same data it was trained on, raising concerns about overfitting.

Shome et al. [84] proposed a ViT-based model for COVID-19 diagnosis that was trained on a self-collected large dataset of COVID-19 chest X-ray images. They also used Grad-CAM [85] to show the progression of COVID-19. However it counts some limitation like data size while its larger than some studies, the aggregated dataset (30,000 images) might still not fully capture the variability of COVID-19 presentations. In top of that Combining data from different sources could introduce biases if not carefully addressed, impacting generalizability and accuracy.

Krishnan et al. [86] spearheaded the application of transformers for COVID-19 diagnosis on chest X-rays. They utilized a fine-tuned ViT-B/32 network pre-trained on ImageNet, achieving an impressive accuracy of 97.6% in classifying COVID-19 cases from the COVID-19 Lung COVID-19 X-ray database. Despite this success, they recognized that image noise could still hinder the model's performance, indicating potential avenues for further development.

Tanzi et al. [77] investigated the use of a ViT model for femur fracture classification, achieving an accuracy of 77.0%. They validated the model's feature extraction capabilities through clustering methods and compared its performance to CNNs. This study highlights the significance of large-scale datasets in enhancing transformer performance, as evidenced by the potentially higher performance observed in COVID-19 classification tasks that often utilize larger datasets. However, despite the promising results for ViTs in classifying femur fractures, limitations need to be addressed. These include data constraints such as limited dataset size and diversity, along with inconsistencies in image quality, which can impact model performance. Additionally, the study focused on a particular type of fracture, potentially limiting the model's generalizability to broader fracture classification tasks.

In CT imaging, **Than et al.** [87] investigated the effect of patch size on ViT performance for COVID-19 and diseased lung classification. They achieved an accuracy of 95.4% but observed a performance decline with larger patch sizes, highlighting a trade-off between capturing local details and global context. The optimal patch size was found to be 32 x 32 pixels. However, their study acknowledges limitations like a small dataset limiting generalizability, a narrow disease scope, and the need for further exploration and validation on larger and more diverse datasets.

Costa et al. [88] employed Vision Transformer (ViT) and its variants to differentiate COVID-19 pneumonia and other pneumonia cases from normal ones. They demonstrated that pretrained models, including DeiT [30], yielded competitive results. Furthermore, conventional ViT and its variants utilizing performer encoders also performed well, even without pretraining. Despite achieving an accuracy of 91.0%, several limitations need addressing. The study utilized a relatively small dataset of CT images, comprising approximately 300 images, potentially lacking representation of the full spectrum of COVID-19 presentations and variations. Additionally, the dataset suffered from data imbalance, with a majority of healthy images and only a minority of COVID-19 images, possibly biasing the model towards predicting healthy cases.

Li et al. [89] developed a COVID-19 diagnosis platform utilizing Vision Transformer (ViT). Their approach involved converting CT images into a sequence of flattened patches to align with ViT's input requirements for diagnosis. Additionally, they employed a teacher-student model to distill knowledge from a CNN pretrained on natural images. Despite achieving an impressive accuracy of 98.0%, several limitations need consideration. The study's dataset and disease scope were limited, warranting further exploration with larger and more diverse datasets to enhance generalizability and robustness. Furthermore, the absence of external validation and real-world evaluation raises concerns about overfitting and highlights the necessity for practical assessments to evaluate clinical applicability.

Gao et al. [90] applied Vision Transformer (ViT) to diagnose COVID-19 using both two-dimensional (2D) and 3D CT scans. They devised a method to construct image sub-volumes by extracting a fixed number of slices, aiming to 'normalize' imaging sequences with varying slice numbers. Additionally, they demonstrated ViT's superiority over DenseNet, a competitive CNN model, achieving an accuracy of 76.6%. However, the study confronts some limitations. The dataset's limited size may not fully represent all COVID-19 presentations, potentially compromising result generalizability.

Zhang et al. [91] employed the Swin-Transformer to train on CT images, adopting a strategy where the lung region is initially segmented via a Unet before being inputted into the feature extractor. This approach effectively reduced the computational burden of the transformer framework. These findings underscore the significance of pretraining for CT image classification tasks, given the challenges associated with acquiring CT images compared to X-ray images. Moreover, methods that mitigate computational complexity using attention mechanisms are essential for the classification of CT images, given their large volume. However, limitations persist. The size and potential imbalance of the training data may impede generalizability.

In MRI, **He et al** [79] proposed a two-pathway Global-Local Transformer architecture. This approach leverages separate pathways to capture both global contextual information and fine-grained local details from brain MRI scans. The extracted features are then fused and fed into a revised transformer, demonstrating improved brain age estimation accuracy compared to traditional CNN-based models, with a mean absolute error (MAE) of 2.70 years and a high correlation coefficient of 98.53%. However, the study acknowledges limitations. It focuses solely on healthy brains, limiting generalizability to patients with neurological disorders.

In ultrasound imaging, **Perera et al.** [80] propose POCFormer, a novel lightweight transformer architecture specifically designed for COVID-19 classification using point-of-care ultrasound (POCUS) images. POCFormer achieves a promising accuracy of 93.9% on POCUS image datasets, demonstrating potential for resource-limited environments.

However, the study acknowledges limitations. Firstly, the use of specific POCUS datasets might limit generalizability to other imaging modalities. Secondly, while the reported accuracy is encouraging, further clinical validation on larger and more diverse datasets is crucial for broader clinical adoption.

Gheflati et Rivaz [92] explored the application of Vision Transformers (ViTs) for breast ultrasound image classification. Their study investigated the effectiveness of various ViT configurations (R+Ti/16, S/32, B/32, Ti/16, R26+S/16) on two separate breast ultrasound (BUSI [63] Dataset B [64]) datasets, achieving promising accuracy results ranging from 85.0% to 86.7%. Notably, the B/32 configuration demonstrated generally superior performance compared to other tested configurations within this study. However, it's crucial to acknowledge limitations. Firstly, the study's findings are likely dependent on the specific datasets employed, potentially limiting generalizability to other datasets and real-world scenarios. Secondly, the reported accuracy values, while encouraging, but still necessite further validation on larger, more diverse datasets.

In addition to the above-mentioned imaging modalities, other imaging technologies have been adopted for the examination and diagnosis of specific diseases, e.g., using dermoscopy images [93], fundus images [94], or histopathology images [95].For instance:
Xie et al. [93] proposed a novel approach for melanoma detection using dermoscopy images, combining a Swin-Transformer with a parameter-free attention module (SimAM). This approach aimed to capture both high-level semantic information and detailed local features crucial for accurate classification. They achieved this by feeding the output of the first three Swin-Transformer blocks into separate SimAM blocks, then concatenating all SimAM outputs and the final feature map before feeding them into the final classification layer. While acknowledging limitations in data size, scope, and potential class imbalance bias, their work demonstrates the potential of transformer-based architectures with attention mechanisms for improving medical image classification tasks.

Ikromjanov et al. [95] investigated the utility of Vision Transformers (ViTs) in aiding pathologists to grade prostate cancer using whole-slide histopathology images. Their research yielded promising outcomes, suggesting the viability of transformer-based methodologies in this field. Nonetheless, they recognize certain limitations: the restricted size of the dataset, which may hinder its applicability to varied patient cohorts and disease manifestations, and the possibility of data imbalance, with an abundance of healthy slides potentially biasing the model towards predicting healthy cases.

Nejati, O. (2023) A Robust Vision Transformer for Generalized Medical Image Classification This study introduces MedViT (Medical Vision Transformer), a novel architecture tailored for medical image classification, and compares its performance with various CNN-based models across multiple datasets, including ChestX-ray8, CheXpert, MIMIC-CXR, and NIH ChestX-ray14. MedViT utilizes a Vision Transformer (ViT) backbone, incorporating self-attention mechanisms to effectively capture global image features. Results demonstrate competitive or superior performance compared to baseline CNN models across different tasks and datasets, with notable achievements such as an F1 score of 87.0% for pneumonia classification in ChestX-ray8 and an AUROC of 0.84 for cardiomegaly detection in CheXpert. However, the study acknowledges limitations, such as the predominantly focused evaluation on chest X-ray images, indicating the need for broader validation on diverse medical image modalities and addressing potential computational costs associated with training transformers [96].

2.2.4.2 Hybrid transformers-based-approaches

Wang et al(2021) introduced TransPath, a novel hybrid model that combines a convolutional neural network with a transformer architecture. This model undergoes pre-training on the TCGA and PAIP datasets using self-supervised learning techniques, followed by fine-tuning on the MHIST, NCT-CRCHE, and PatchCamelyon datasets. The reported accuracy rates are 89.68%, 95.85%, and 89.91%, respectively. While TransPath shows promise for histopathology image classification, it faces several limitations that warrant consideration. The dataset, although substantial in size, may not fully capture the diversity present in real-world scenarios, and training on unlabeled domain-specific data could potentially limit its generalizability [97].

Chen et al. (2022) [94] developed a model to differentiate between benign and malignant pulmonary nodules using computed tomography (CT) images. To address the unique challenges of small-scale medical image datasets, their approach builds upon the Swin Transformer architecture, combining both window-based multi-headed self-attention (W-MSA) and shifted-window-based multi-headed self-attention (SW-MSA) mechanisms. This combination enhances information exchange between windows, which is crucial for analyzing medical images. Utilizing the LIDC-IDRI dataset, their model achieved a high accuracy of 98.33%. However this use just one architecture and one dataset to train and test on so it may counts problem of overfitting .

Dai et al. (2021) [98] devised the TransMed model by integrating CNN and transformer architectures to capture low-level features from images, establish long-distance dependencies across modalities, and facilitate multimodal medical image classification. They asserted that their study represents one of the first attempts to apply transformers to this domain and evaluated TransMed on both the PGT dataset and the MRNet dataset, partitioning the data into training, validation, and test sets at ratios of 7:1:2 for PGT and 1130:120:120 for MRNet. Their model achieved an accuracy of 88.9% on the PGT dataset and 85% on the MRNet dataset.in tumor recognition.) However, the study uncovered several constraints through ablation experiments. Firstly, they observed a notable performance gap between pure transformers and conventional CNNs on small medical image datasets, indicating that the self-attention mechanism lacks the inherent inductive bias present in CNN structures. Secondly, they found that serializing two-dimensional images yielded only marginal improvements, and performance tended to deteriorate with increased serialization parameters. This implies that existing image serialization techniques, though effective in natural image classification, may not be suitable for medical images, posing a risk of compromising model performance, especially in tumor recognition.

Leamons et al, (2022) [99]developed three different deep-learning models to detect the presence of invasive ductal carcinoma, the most common form of breast cancer. These models include convolutional neural networks (CNNs), residual neural networks (RNNs), and visual transformers (VTs) used as baselines. The experimental results of the three models using breast cancer tissue image datasets for training and experiments show that the VT model is superior to CNN and RNN in different tasks, with a classification accuracy of 93%. In contrast, the highest classification rate of other models is 87%. however it is just used on one specific dataset and also its small not large enough to train this kind of deeplearning models so maybe overfitting problem .

Jang et Hwang, (2022) [100]proposed a 3D medical image classifier M3T. It consists of a transformer, a 3D convolutional neural network, and a 2D convolutional neural network. On the ADNI (Petersen et al., 2010), AIBL (Ellis et al., 2009), and OASIS datasets (Marcus et al., 2007), the accuracy achieved results of 3.21%, 93.27%,

and 85.26%. however While the study utilizes a publicly available dataset, it might not be representative of the entire population or encompass diverse disease presentations. Also the study primarily evaluates the model on the same data it was trained on, raising concerns about overfitting and the need for validation using independent datasets.

Jinwei Liu, Yan Li , (2022) [101] introduce a new architecture termed the Feature Pyramid Vision Transformer (FPViT) for classifying medical images. This approach leverages the MedMNIST Classification Decathlon, a dataset comprising diverse medical images spanning various organs and diseases. FPViT combines the strengths of Vision Transformers (ViTs), which excel at capturing global features, and Feature Pyramid Networks (FPNs), which are adept at integrating information across multiple scales. By integrating these two techniques, FPViT can extract detailed information from images at different scales, potentially resulting in improved performance. The authors report achieving competitive and potentially superior classification results compared to other methods on the MedMNIST dataset. Specifically, they achieve accuracies of 91.8% for PathMNIST, 94.8% for ChestMNIST, 76.6% for DermaMNIST, 81.3% for OCTMNIST, 89.6 for PneumoniaMNIST, 56.8% for RetinaMNIST, 89.1 for BreastMNIST, 93.5% for OrganMNIST ,90.3% for OrganMNIST C, and 78.5% for OrganMNIST S.However this study utlises just one model so it needs more models to train on it.

Manzari et al. (2023)introduce MedViT is A Robust Vision Transformer for Generalized Medical Image Classification [102].MedViT is a hybrid model that combines Convolutional Neural Networks (CNNs) and Vision Transformers to leverage the strengths of both architectures. This model utilizes the local feature extraction capabilities of CNNs and the global feature representation capabilities of transformers to improve robustness and performance. It incorporates an Efficient Convolution Block (ECB) to reduce the high computational complexity of transformers and employs a novel data augmentation technique called Patch Momentum Changer (PMC) to enhance adversarial robustness. Evaluated on the MedMNIST-2D dataset, MedViT demonstrates superior performance, highlighting its potential for clinical applications in medical imaging and reliable medical diagnosis.

2.2.4.3 Ensemble-learning-based-approaches

M. A. Cheema, M. A. Nawaz, M. Shoaib,(2023) researchers delved into the realm of ensemble learning to enhance the performance of established Convolutional Neural Networks (CNNs) for the crucial task of mammography classification, aiming to bolster accuracy and robustness in detecting breast cancer. Leveraging the widely accessible DDSM dataset, consisting of anonymized mammogram images alongside their respective labels denoting normal, benign, or malignant cases, the study employed a subset of 1,574 images for training and 400 images for testing purposes. The investigation focused on three prevalent pre-trained CNN architectures: VGG16, DenseNet169, and InceptionV3, each renowned for its distinct advantages in image classification tasks. Employing ensemble strategies such as Bagging, Boosting, and Stacking, the researchers amalgamated the predictive power of multiple CNNs to capitalize on diverse representations and mitigate overfitting. Results consistently showcased the superiority of ensemble methods over individual CNNs, with Stacking emerging as the most effective strategy, particularly evident in the VGG16 architecture, achieving an impressive accuracy of 92.5%. Nonetheless, the study acknowledged limitations stemming from the relatively modest dataset size, underscoring the potential impact on the broader generalizability of the findings. Moreover, the researchers meticulously explored various hyperparameter configurations for each ensemble strategy and CNN architecture, alongside an investigation into the efficacy of different data augmentation techniques, all aimed at optimizing performance,

robustness, and generalizability in the critical domain of mammography classification. How ever this study have some limitation like dataset used here is not large so maybe it will face overfitting problem [103].

Mousavi, S. M. A., Shirazian, M. (2022) researchers focused on improving medical image classification by combining ensemble learning and a new optimization algorithm called the Levy Flight-based Honey Badger Algorithm (LFHBA). They integrated these techniques into the emerging field of 6G-enabled Internet of Things (IoT). Using pretrained deep learning models like MobileNet and DenseNet, they extracted features from medical images and used LFHBA to select the most important ones for classification. This framework aimed to securely collect and analyze medical images in real-time over advanced networks. They tested their approach on two datasets: one containing chest X-ray images and another with Optical Coherence Tomography (OCT) images for diagnosing retinal diseases. Results showed improved accuracy, with 87.10% on chest X-ray images and 94.32% on OCT images. However, they noted limitations in the study's validation on small datasets and highlighted the need for further research on larger datasets. They also emphasized the potential of integrating 6G-enabled IoT for real-time medical image analysis but suggested addressing implementation challenges and ensuring data security in future studies [104].

Hengde Zhu, Wei Wang, (2023) In their groundbreaking work titled "MEEDNets: Medical Image Classification via Ensemble Bio-inspired Evolutionary DenseNets," researchers introduce an innovative approach that amalgamates ensemble learning with a bio-inspired evolutionary optimization algorithm for medical image classification. They leverage the DenseNet architecture as their foundational model and employ an evolutionary synthesis mechanism to automatically generate sparse and efficient offspring networks. This evolutionary process strategically prunes unnecessary parameters from the network, enhancing both accuracy and efficiency. Additionally, they propose an evolution-based ensemble learning mechanism to create a diverse ensemble of highly sparse DenseNets, further enhancing performance. The evaluation of MEEDNets is conducted on two publicly available medical image datasets: the Chest X-ray dataset, which classifies images into normal, bacterial pneumonia, or viral pneumonia categories, and the Optical Coherence Tomography (OCT) dataset, distinguishing images with retinal disease from those with a healthy status. The DenseNet architecture is chosen due to its dense connectivity, which aids in feature propagation and mitigating overfitting. Results demonstrate MEEDNets' superiority, achieving an accuracy of 87.10% on the Chest X-ray dataset and 94.32% on the OCT dataset, outperforming traditional single DenseNet models and ensemble learning methods. Additionally, the evolved networks exhibit significant parameter reduction compared to the original DenseNet architecture, showcasing improved efficiency. However, the study acknowledges limitations, such as its focus on only two datasets and the necessity for further validation on larger and more diverse datasets. Furthermore, the evolutionary optimization process requires meticulous parameter tuning for optimal results, and ensemble learning methods may incur computational expenses and reduced interpretability compared to single models. Overall, MEEDNets presents a promising avenue for advancing medical image classification tasks, but continued validation and enhancements in interpretability are crucial for wider adoption in the field. While the study evaluates the MEEDNets model on specific medical image datasets, such as SARS-CoV-2 CT scans and brain tumor datasets, the diversity of the datasets might be limited. This constraint raises concerns about the generalizability of the model's performance to broader datasets that encompass a wider range of medical conditions and imaging modalities [105].

2.3 Conclusion

This chapter has provided a comprehensive overview of the various aspects involved in medical imaging analysis using deep learning techniques. We began by introducing the different medical imaging modalities and the essential tasks associated with medical image analysis, such as classification, and segmentation. To facilitate the development and evaluation of deep learning models, we discussed the most widely used datasets in the field. Furthermore, we examined the commonly employed performance metrics used to assess the effectiveness of these models, ensuring a standardized and reliable evaluation process. The chapter then explore the application of (CNNs) for medical image classification tasks. Building upon this, we explored the recent advancements in transformer-based models, which have gained significant traction in the medical imaging domain due to their ability to capture long-range dependencies and model complex relationships within the data. We presented both pure transformer-based approaches and hybrid models. Additionally, we discussed ensemble learning techniques, which have proven effective in enhancing the overall predictive accuracy and robustness of medical image analysis models by aggregating the predictions of multiple individual models. Through this comprehensive exploration, we have provided a solid foundation for researchers and practitioners to understand the current landscape of deep learning techniques in medical imaging analysis.

In the following chapter, we will focus on our ensemble learning approaches. We will illustrate and explain each proposed ensemble learning approach, detailing the models we are using as base models and their fine-tuning processes.

Chapter 3

Design of a transformers-based ensemble framework for medical imaging

3.1 Introduction

This research aims to develop an ensemble learning method that combines multiple transformer models to improve accuracy in image classification tasks. By using diverse medical imaging datasets, we ensure that our approach is effective across various types of data. Ensemble learning techniques combine individual model outputs, potentially enhancing overall performance and proving our method works well with multiple datasets.

This chapter outlines our methodological approaches. It describes the four transformer models in the ensemble, detailing their architectures and component. We justify the selection of these models based on their ability to contribute diverse perspectives and address the limitations of individual models. We explore various ensemble learning methods, including hard voting, soft voting, weighted hard voting, weighted soft voting stacking, bagging, and boosting. The implementation details for these methods are provided, along with justifications for their use in capturing complementary strengths and addressing individual model weaknesses. Finally we discuss the evaluation metrics used, such as accuracy, precision, recall, and F1-score, explaining their relevance and appropriateness while avoiding potential biases from relying on a single metric.

3.2 General presentation of the proposed framework

Ensemble learning is a powerful machine learning paradigm designed to enhance the performance, robustness, and generalization of predictive models by combining multiple individual models. In our propose ensemble method as shown in Figure 3.1, we begin by fine-tuning transformer-based models (DeiT, BEiT, ViT, and Swin) on our specific medical imaging dataset. Each model is meticulously adapted to the dataset's characteristics, allowing them to specialize in each dataset. Once the models done fine-tunning ,they become individualized predictors, each offering unique insights and capabilities for image analysis. When new image data is introduced for prediction, we pass it through each of these fine-tuned models to generate individual predictions. This process harnesses the diverse perspectives and expertise embedded within each model. After obtaining predictions from all models, we employ a fusion method to combine these outputs into a comprehensive prediction.This approach not only improves prediction accuracy but also enhances the robustness and generalization capabilities of the models. In the next sections, we will explore the base models, the fine-tuning process, and the implementation of each ensemble method.

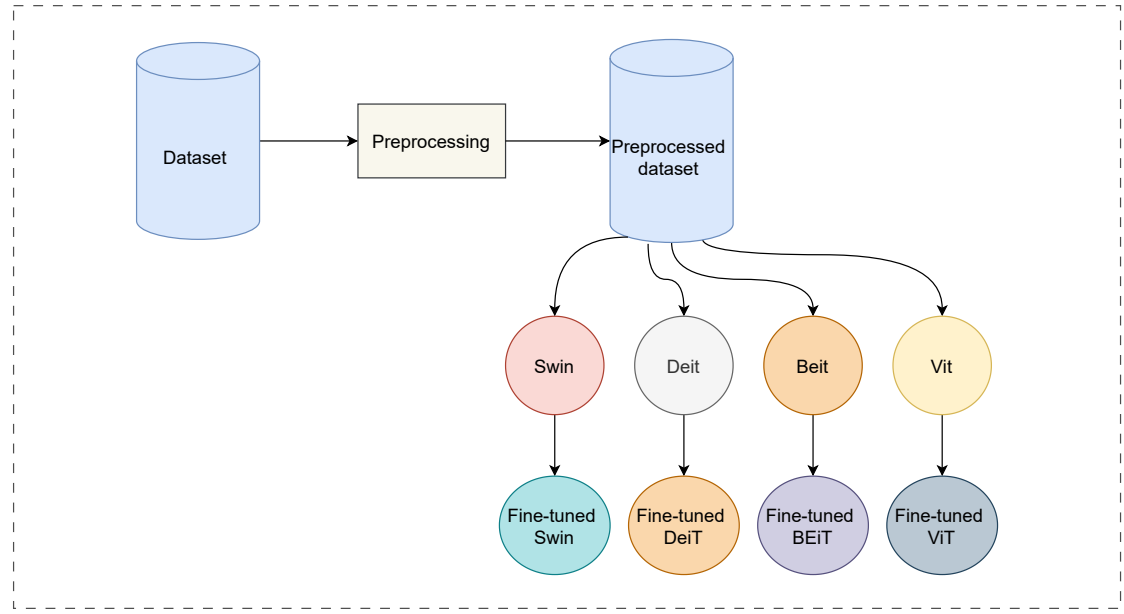


Figure 3.1: General schema of the proposed framework

3.3 Base models

We employed an ensemble learning approach that leveraged four state-of-the-art Vision Transformer models as base models, namely Vision Transformer (ViT), Bidirectional Encoder Representations from Transformers (BEiT), Swin Transformer, and Distilled Data-Efficient Image Transformer (DeiT). These models have made significant contributions to the field of computer vision and have demonstrated outstanding performance on various tasks.

3.3.1 ViT

The first base model in our ensemble learning approach is the Vision Transformer (ViT). Introduced by Google Brain in 2020, ViT adapts the transformer architecture for computer vision by splitting images into patches and treating them as tokens. This model captures global context and long-range dependencies, essential for understanding complex visual scenes. Despite its high performance on tasks like image classification and object detection, ViT requires substantial training data and computational resources. For our implementation, we used the ViT model (google/vit-base-patch16-224-in21k) pre-trained on ImageNet-21k, as provided by Hugging Face. This model is pre-trained on 14 million images across 21,843 classes at a resolution of 224x224, based on the work by Dosovitskiy et al [106].

3.3.2 BEiT

The second base model in our ensemble learning approach is the BEiT (Bidirectional Encoder representation from Image Transformers). Introduced by Microsoft Research and the Chinese University of Hong Kong in 2021, BEiT employs self-supervised pretraining on large-scale unlabeled image datasets. It combines a Vision Transformer encoder with a masked image modeling task, similar to masked language modeling in NLP. BEiT masks portions of input image patches and learns to reconstruct them, capturing rich visual representations transferable to downstream tasks. This self-supervised pretraining allows BEiT to achieve competitive performance with fewer labeled examples. However, it can be computationally expensive and may struggle with capturing fine-grained details. For our implementation, we used the BEiT model (microsoft/beiT-base-patch16-224-pt22k-ft22k) from Hugging Face. This model is pre-trained on ImageNet-22k (14 million images, 21,841 classes) at a resolution of 224x224 and fine-tuned on the same dataset [8].

3.3.3 Swin

The third base model in our ensemble learning approach is the Swin Transformer. This hierarchical Vision Transformer introduces the "Shifted Window" approach, combining the strengths of convolutional neural networks (CNNs) and transformers. It partitions the input image into non-overlapping windows and performs self-attention within each window, capturing local features. The model then uses a shifted window partitioning to enable cross-window connections, capturing global dependencies. This hierarchical approach, applied across multiple stages, gradually increases the window size and downsamples the feature maps. The Swin Transformer, introduced by Microsoft Research in 2021, excels in capturing both local and global context, achieving excellent performance on various vision tasks while maintaining computational efficiency. For our implementation, we used the Swin Transformer model (microsoft/swin-large-patch4-window7-224-in22k) from Hugging Face, pre-trained on ImageNet-21k (14 million images, 21,841 classes) at a resolution of 224x224 [9]. Swin is distinguished from the other base models as it is a large model, implying a larger parameter size and computational requirements.

3.3.4 DeiT

Finally, the fourth base model in our ensemble learning approach is the Data-efficient Image Transformer (DeiT). DeiT, designed for image classification tasks, builds upon the original Vision Transformer (ViT) architecture with modifications to enhance data efficiency and performance. Introduced by Facebook AI Research (FAIR) in 2021, DeiT employs a distillation mechanism where a high-capacity convolutional neural network serves as a teacher model to guide the training of the transformer model. This teacher-student approach enables DeiT to achieve competitive performance with fewer training examples compared to traditional convolutional models. DeiT also benefits from transformers' strengths, such as global context modeling and parallelization. However, like other Vision Transformers, it can be computationally expensive for large input resolutions and high-resolution feature maps. For our implementation, we used the DeiT base model (facebook/deit-base-patch16-224) from Hugging Face. This model is pre-trained and fine-tuned on ImageNet-1k (1 million images, 1,000 classes) at a resolution of 224x224 [7] .

3.4 Fine-tuning of base models on medical imaging datasets

While the pre-trained transformer models provide a strong foundation for visual understanding, we further fine-tune them on our specific medical imaging datasets to adapt them to our target domains and tasks. Initially, we obtain the medical imaging datasets from the MedMNIST v2 repository hosted on the Hugging Face platform and for the chest dataset, also from the Hugging Face platform. These datasets are already preprocessed to some extent, but we apply additional preprocessing steps to ensure compatibility and enhance model performance subsection 4.3.2. We create label-to-id and id-to-label dictionaries to map the class labels to numerical indices and vice versa. Each dataset is then split into training, validation, and test subsets, with the training set used for fine-tuning, the validation set for monitoring performance during training, and the test set for final evaluation.

To fine-tune the pre-trained models, we employ the Parameter-Efficient Fine-Tuning (PEFT) approach, specifically the Low-Rank Adaptation (LoRA) method. This technique allows for efficient fine-tuning by introducing small, trainable weight matrices instead of updating the entire model's weights **??**. For each combination of pre-trained model and medical imaging dataset, we execute the fine-tuning process. The pre-trained models are fine-tuned on the training subset of the respective dataset using the configured hyperparameters and optimization settings, with the goal of decreasing the cross-entropy loss on the training data. During and after the fine-tuning process, we evaluate the performance of the fine-tuned models on the validation and test subsets, respectively, by computing the relevant metrics. Throughout the fine-tuning process, we employ techniques such as experiment tracking and logging, integrating with Weights and Biases (wandb) for monitoring and visualization. After fine-tuning, we save the fine-tuned models, along with their configurations and relevant metadata, to be used for our main objective: Ensemble method of transformers in medical imaging datasets. For ensemble methods such as hard voting, soft voting, stacking, weighted hard voting, and weighted soft voting, we fine-tune each model on the entire training dataset. However, for bagging and boosting ensemble learning, the fine-tuning process differs:

- In **bagging**, each model is fine-tuned on a different subset of the dataset. This approach involves creating multiple subsets by random sampling with replacement and fine-tuning a separate model on each subset.

- In **boosting**, the fine-tuning process is sequential. The first model is fine-tuned on the original dataset. For each subsequent model, the dataset is adjusted by assigning higher weights to the misclassified samples from the previous model. This adjusted dataset is then used to fine-tune the next model, focusing on correcting the errors made by the previous model.

By employing these tailored fine-tuning strategies, we aim to maximize the performance and robustness of the ensemble models across different medical imaging datasets.

3.5 Ensemble methods

After fine-tuning the base models on the respective medical imaging datasets, we explored various ensemble learning techniques to combine their predictions and leverage their collective strengths. The ensemble approaches investigated in this study are as follows

3.5.1 Hard voting ensemble method

We employ a hard voting ensemble method to use the collective predictions of multiple pre-trained models for image classification. The ensemble consists of the four fine-tuned models.
The ensemble method, involves independently processing each test sample through the individual models. For each sample, the image is first preprocessed, then the preprocessed image is passed through the respective model to obtain the logits, which represent the raw output scores for each class.
The logits from each model are then converted to predicted class indices by taking the argmax operation, which selects the class with the highest logit value as the predicted class for that model. To combine the predictions from all models as shown in Figure 3.2, we implement a hard voting strategy. Specifically, the logits from each model are converted to predicted class. The class with the highest voted value is chosen as the final ensemble prediction.

3.5.2 Weighted hard voting ensemble method

We employ a weighted hard voting ensemble method to use the collective predictions of multiple pre-trained models for image classification. The ensemble method process, involves assigning weights to the four individual models based on their evaluated performance. These weights modulate the influence of each model's predictions on the final ensemble decision.
During the inference phase, each test sample is processed independently by the individual models. The image is first preprocessed, then the preprocessed image is passed through the respective model to obtain the logits, which represent the raw output scores for each class. The logits from each model are weighted according to the assigned model weights. These weighted logits are then converted to predicted class indices by taking the argmax operation. Finally, as shown in Figure 3.3, the class with the highest number of predictions is chosen as the ensemble prediction for the given test sample. The weighted hard voting the weights for this ensemble method were manually adjusted for each dataset. Due to time constraints, we opted for manual adjustment instead of an automated process. Therefore, all the results of the weighted approach reflect the best-performing weights we tried, as they are not randomly assigned. We experimented with many values to determine the high-performing weights.

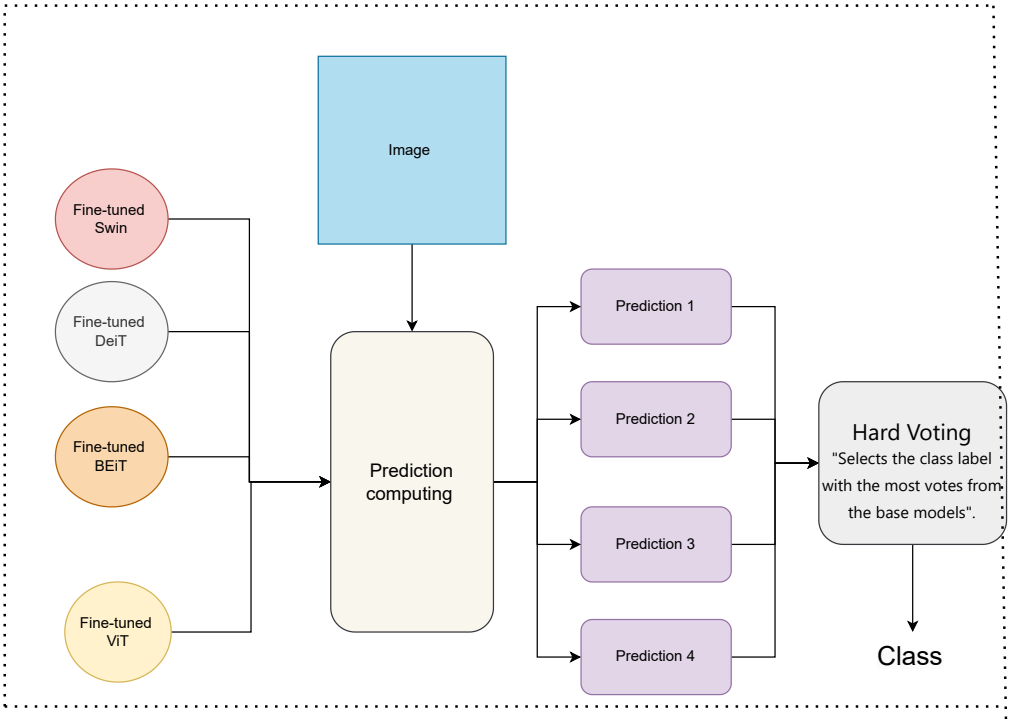


Figure 3.2: Illustration of the proposed hard voting ensemble method

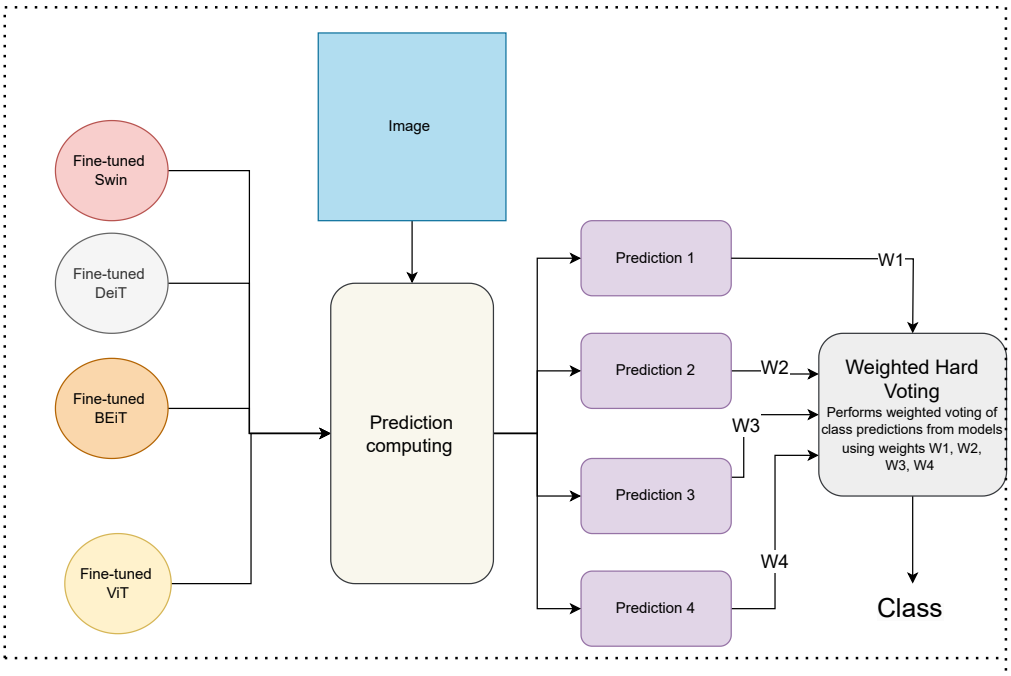


Figure 3.3: Illustration of the proposed weighted hard voting ensemble method

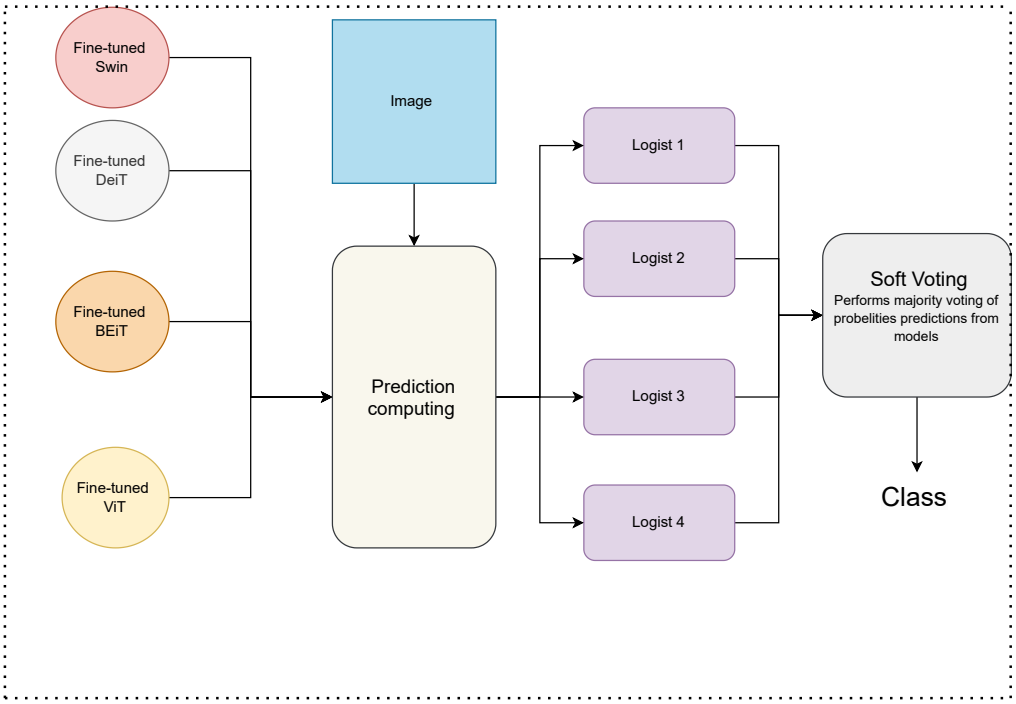


Figure 3.4: Illustration of the proposed soft voting ensemble method

3.5.3 Soft voting ensemble method

The soft voting ensemble method involves combining the predicted class probabilities from multiple models by selecting the class label with the highest summed predicted probability across all models. Unlike hard voting, where the majority vote of the predicted class labels is used, in soft voting, we consider the predicted probabilities from each model. The ensemble learning process, involves independently processing each test sample through the individual models. For each sample, the image is first preprocessed, then the preprocessed image is passed through the respective model to obtain the logits, which represent the raw output scores for each class.

The logits from each model are then converted to predicted probabilities using the softmax function. These predicted probabilities represent the models' confidence scores for each class, with higher values indicating a higher likelihood of the sample belonging to that class. To combine the predictions as shown in Figure 3.4, from all models, we employ a soft voting strategy. Specifically, the predicted probabilities from each model are summed across all models, resulting in a set of ensemble probabilities. The class with the highest ensemble probability is chosen as the final prediction for the given test sample.

By summing the predicted probabilities from multiple models, the soft voting approach aims to leverage the collective knowledge and strengths of the individual models, potentially improving the overall performance and robustness of the classification task.

3.5.4 Weighted soft voting ensemble learning

The weighted soft voting ensemble technique is similar to the soft voting approach, but it introduces weights for each model, allowing us to assign different levels of importance or

confidence to the predictions of individual models. we start by defining the list of model checkpoints and the corresponding weights for each model. The weights are specified as a Python list, and their sum should be equal to 1.

we then loads and initializes the models and corresponding image processors from the specified repository names. The process of loading the models and image processors is similar to previous ones. After that, we iterates over the test dataset and performs a set of steps for each sample it starts by obtainining the true label for the current sample, performs inference with each model by passing the input image through the corresponding image processor and the model, and then converts the logits to predicted probabilities using the softmax function, and multiplies the predicted probabilities from each model by their respective weights. The weighted predicted probabilities from each model are collected into a list.

The code then performs weighted soft voting by summing the weighted predicted probabilities across all models, creating a tensor ensemble_probs containing the weighted sum of predicted probabilities from all models for each class. The final ensemble prediction as shown in Figure 3.5 is determined by taking the argmax of the ensemble_probs tensor, corresponding to the class label with the highest weighted sum of predicted probabilities across all models. same with weighted hard voting the weights for this ensemble method were manually adjusted for each dataset. Due to time constraints, we opted for manual adjustment instead of an automated process. Therefore, all the results of the weighted approach reflect the best-performing weights we tried, as they are not randomly assigned. We experimented with many values to determine the high-performing weights.

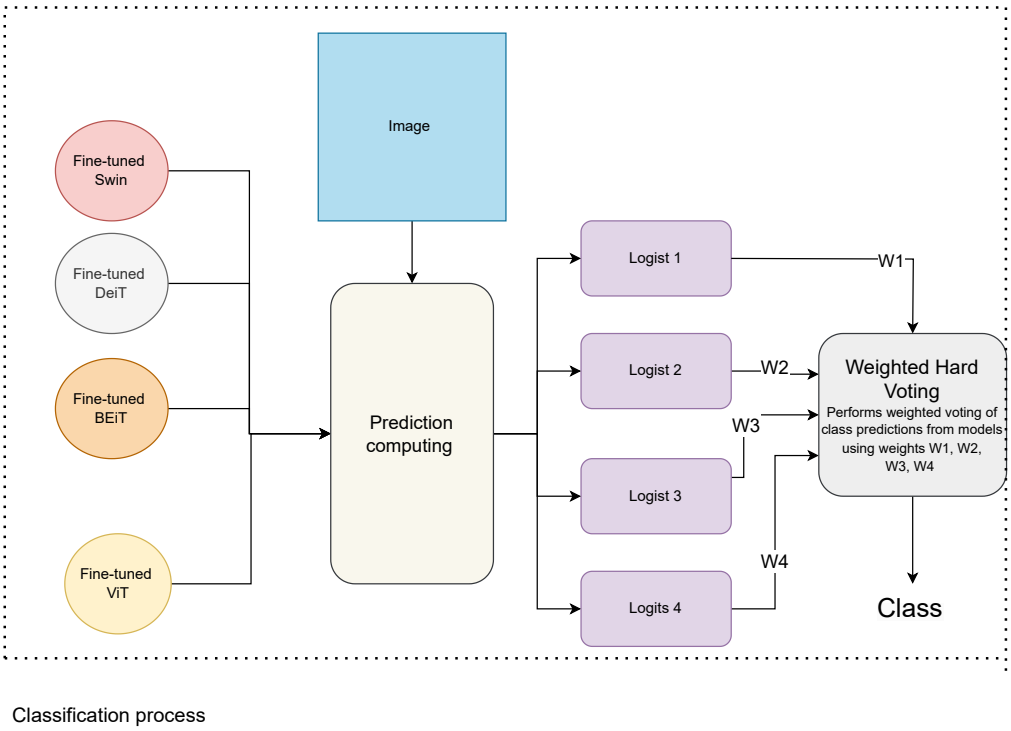


Figure 3.5: Illustration of the proposed weighted soft voting ensemble method

3.5.5 Stacking ensemble method

The stacking ensemble is a powerful technique that involves as shown in Figure 3.6 training a meta-model on the predictions (logits) of the base models. This meta-model learns to combine the outputs of the base models in a more sophisticated way, leveraging

their collective knowledge and complementary strengths. In our experiments, we explored stacking ensembles with two different meta-models: logistic regression and support vector machine.

3.5.5.1 Stacking ensemble with logistic regression as meta-model

We first investigate the use of logistic regression as the meta-model for the stacking ensemble. The process involves loading and initializing separate instances of the four fine-tuned base models and their corresponding image processors, as well as defining the dataset and its splits. For each validation sample, we obtain the logits from each base model and concatenate them into a single feature vector. We then perform hyperparameter tuning subsection 4.2.4 for the logistic regression meta-model using grid search method. The best hyperparameter combination is selected based on the highest accuracy score on the validation set. With the optimal hyperparameters, we train the logistic regression meta-model on the validation logits and labels. For each test sample, we obtain the logits from each base model and concatenate them into a single feature vector. We then make predictions on the test set using the trained logistic regression meta-model and the test logits as input features. Finally, we compute the accuracy, precision, recall, and F1-score of the stacking ensemble with logistic regression on the test set using the predicted labels and true labels.

3.5.5.2 Stacking ensemble with SVM as meta-model

In addition to the logistic regression meta-model, we also explore the use of a support vector machine (SVM) as the meta-model for the stacking ensemble [107]. The process follows a similar approach to the logistic regression meta-model, but with some differences in hyperparameter tuning and model training. For validation samples, we obtain the logits from each base model and concatenate them into a single feature vector. Subsequently, we perform hyperparameter subsection 4.2.4 tuning for the SVM meta-model using the grid search method. We conduct the grid search over the defined parameter grid, using the validation logits as input features and the validation labels as targets. The best hyperparameter combination is selected based on the highest accuracy score on the validation set. With the optimal hyperparameters, we train the SVM meta-model on the validation logits and labels. For test samples, we obtain the logits from each base model and concatenate them into a single feature vector. We then make predictions on the test set using the trained SVM meta-model and the test logits as input features. Finally, we compute the accuracy of the stacking ensemble with the SVM meta-model on the test set using the predicted labels and true labels.

3.5.6 Bagging ensemble method

We employ a bagging ensemble method to enhance the robustness and performance of our image classification models by utilizing multiple fine-tuned models. As illustrated in Figure 3.7, each model is trained on a different bootstrap sample of the training dataset. The four pretrained models are selected for the ensemble, and each is fine-tuned using Low-Rank Adaptation (LoRA) to reduce the number of trainable parameters while maintaining performance.

For each model in the ensemble, a bootstrap sample of the training dataset is created by randomly sampling with replacement, ensuring that each model is trained on a slightly different subset of the data. The models are then independently trained on their respective bootstrap samples using customized training arguments. After training, predictions on the test set are made by obtaining logits from the models. The final ensemble prediction for each test sample is determined by majority voting, where the most frequently predicted class across all models is chosen.

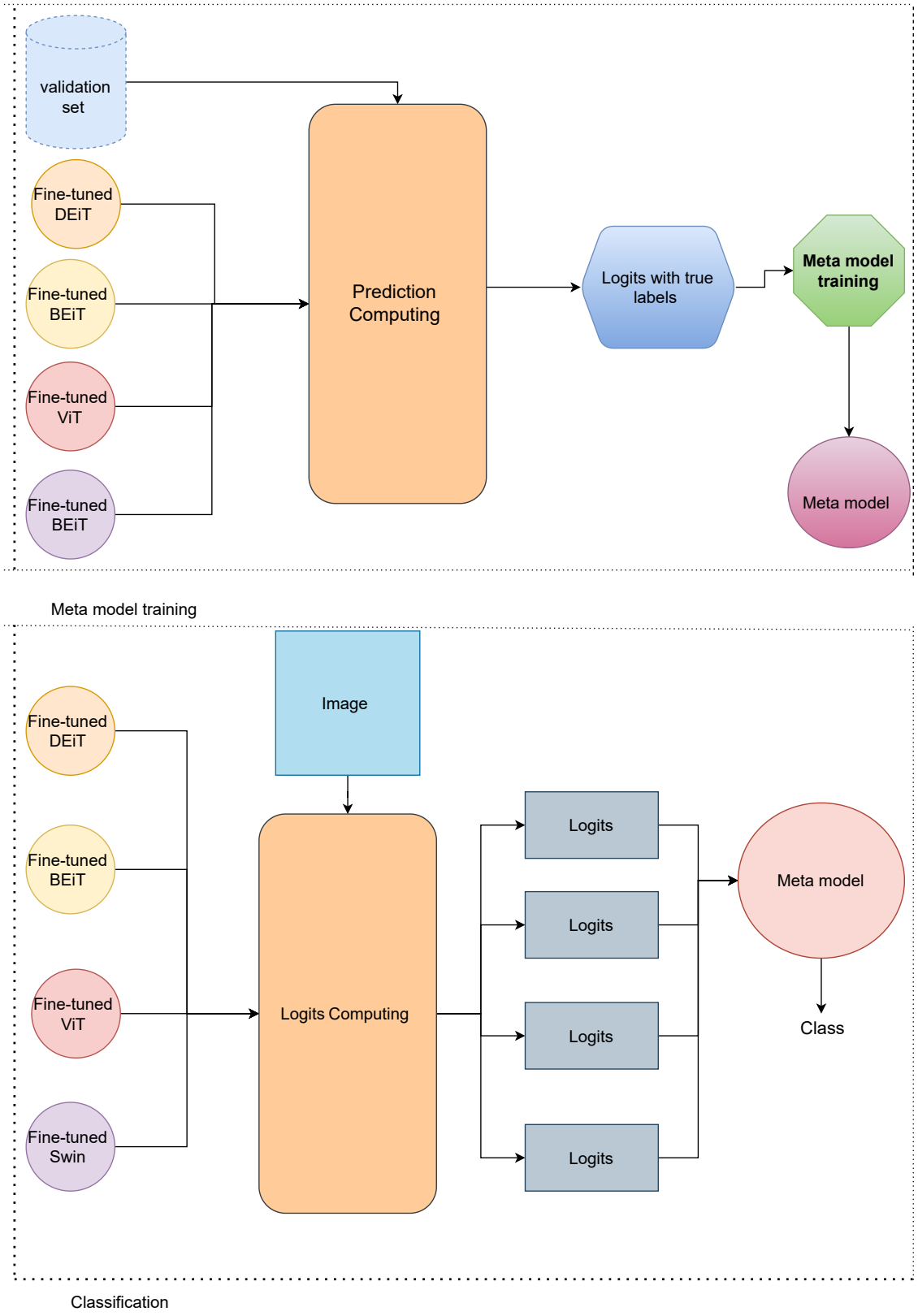
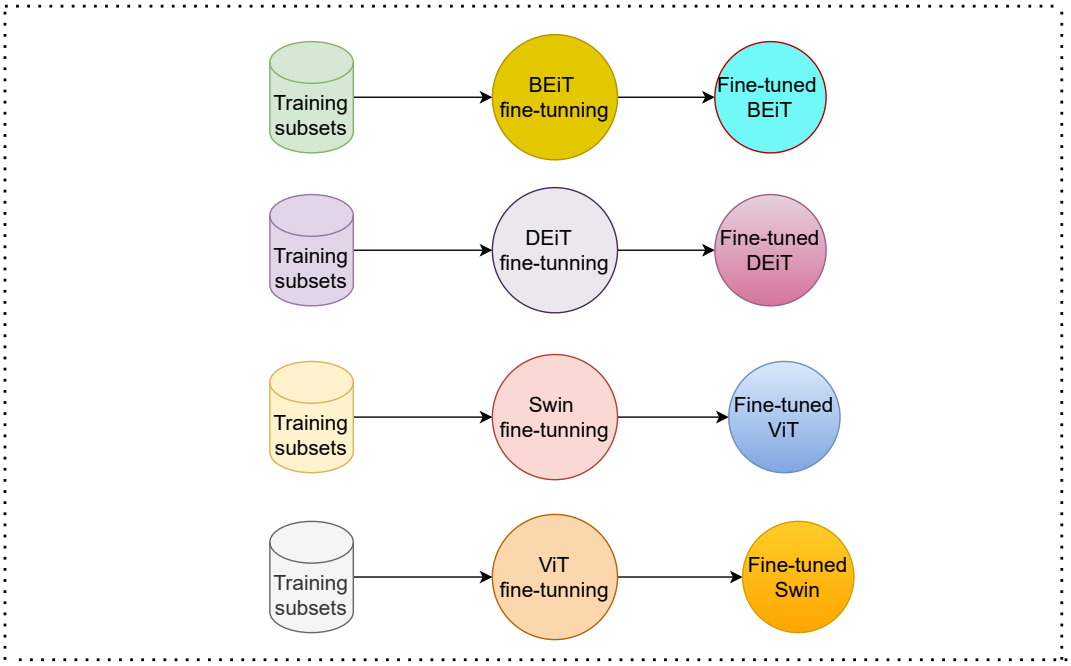
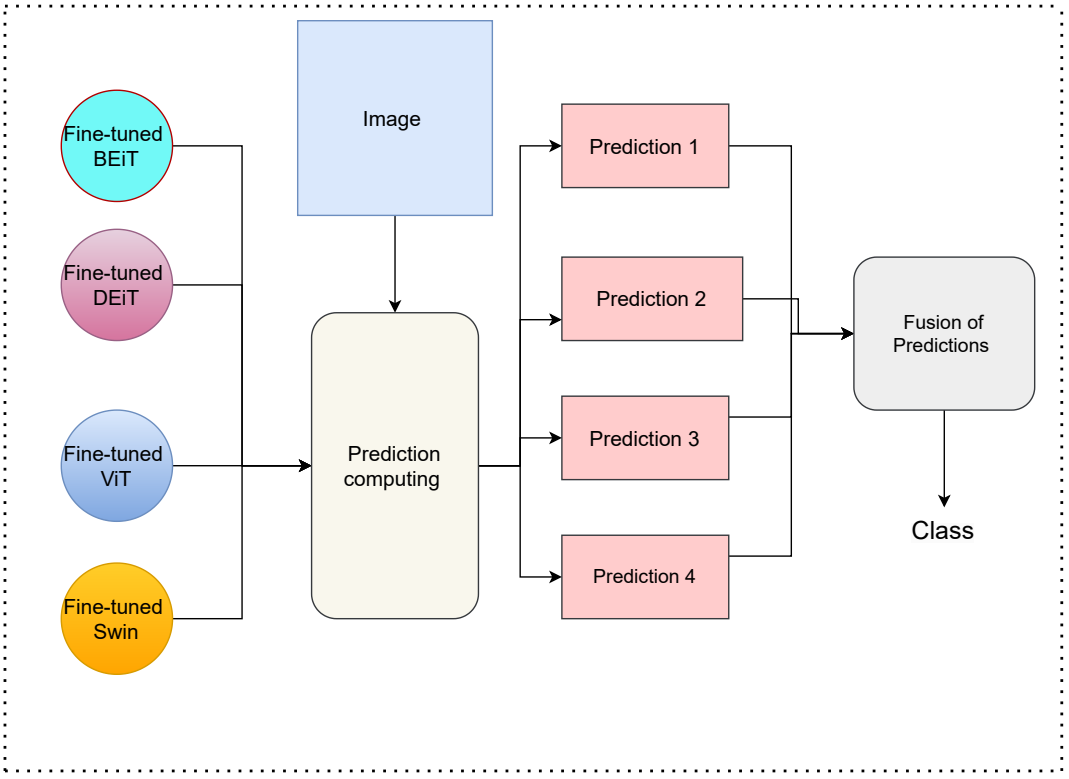


Figure 3.6: Illustration of the proposed stacking ensemble method



Training of bagging ensemble learning



Classification process

Figure 3.7: Illustration of the prepoced bagging ensemble method

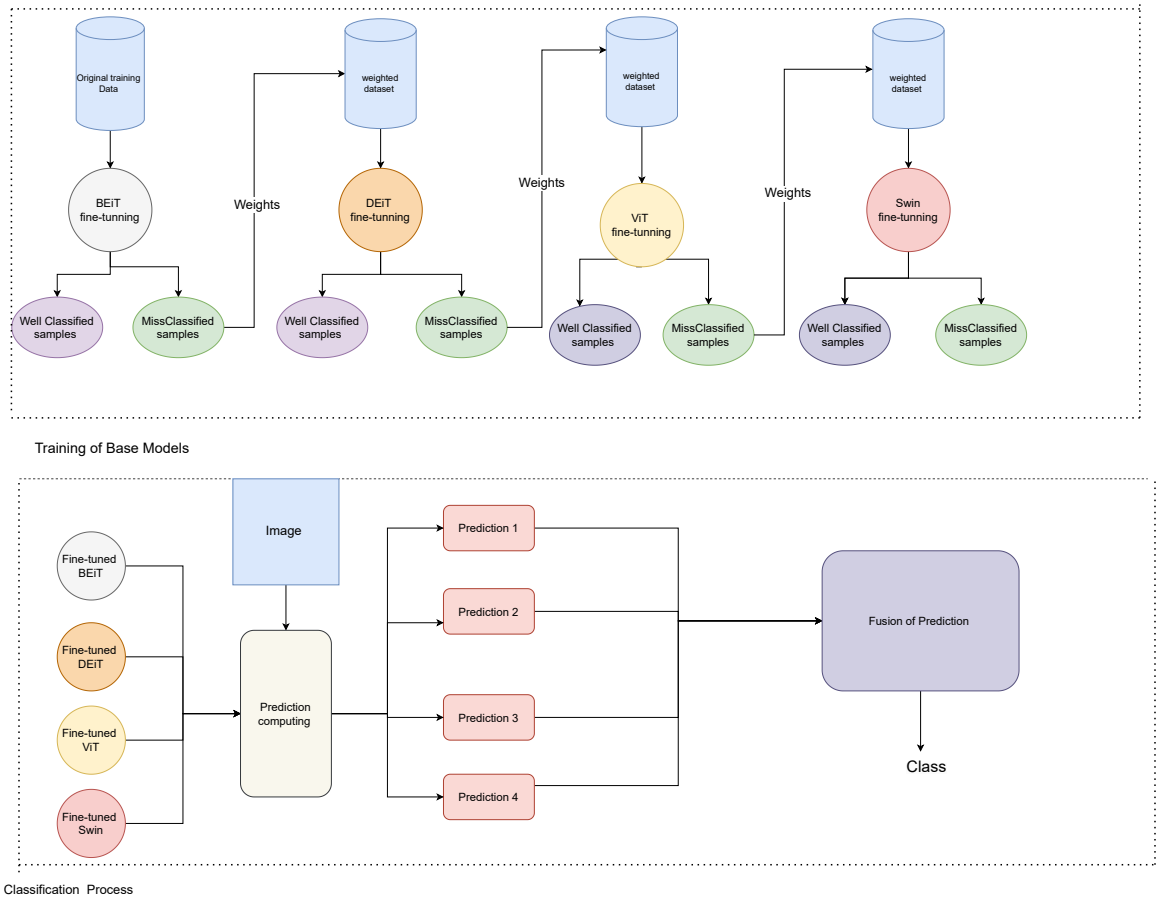


Figure 3.8: Illustration of the prepoced boosting ensemble method

3.5.7 Boosting ensemble method

We employ a boosting ensemble method to enhance the performance of our image classification models by sequentially training four models on the training dataset, each focusing on correcting the errors made by the previous models. The four pretrained models are selected for the ensemble, and each is fine-tuned using Low-Rank Adaptation (LoRA) to reduce the number of trainable parameters while maintaining performance. For each model, the training dataset is preprocessed using customized transformations, and the models are trained on the entire dataset with sample weights adjusted after each model's training phase to emphasize the samples that were misclassified. As illustrated in Figure 3.8, this iterative process increases the influence of misclassified samples in subsequent models.

Each model is trained, and after training, predictions on the test set are made by obtaining logits from each fine-tuned model. The final ensemble prediction for each test sample is determined by weighted majority voting, where the predictions of each model are weighted according to their performance, and the class with the highest weighted vote is chosen.

3.6 Evaluation metrics

To assess the performance of the base models, fine-tuned models, and ensemble models, we employed several widely adopted evaluation metrics tailored for classification tasks.

- **Accuracy**: Ratio of correctly classified samples to the total number of samples.

Provides an overall measure of the model's performance, suitable for balanced datasets. Chosen for its ability to offer a high-level overview.

- **Precision**: Ratio of true positive predictions to the total number of positive predictions made by the model. Crucial in medical applications to avoid unnecessary follow-up procedures, increased costs, and patient anxiety.
- **Recall**: Ratio of true positive predictions to the total number of actual positive samples. Essential for screening and diagnostic purposes to prevent missing true positive cases, which can have severe consequences.
- **F1-score**: Harmonic mean of precision and recall, providing a balanced measure suitable for class imbalance or when both precision and recall are important in medical image classification.

The selection of these evaluation metrics was based on their widespread usage and interpretability in the field of medical image classification. Specifically, accuracy provides a high-level overview, precision is crucial for avoiding false positives, recall is essential for preventing false negatives, and the F1-score balances precision and recall, making it suitable for comprehensive evaluation.

3.7 Conclusion

This chapter outlined our methods for improving medical image classification using four transformer models. We've selected ViT, DeiT, BEiT, and Swin for their strengths and ironed out our approach through meticulous tuning. From hard voting to boosting, we've explored diverse ensemble techniques, ensuring transparency in our evaluation metrics and experimental setup.Moving forward, our focus shifts to the implementation details and unveiling the results of our approach in the next chapter.

Chapter 4

Implementation and experimental study of the proposed ensemble methods

4.1 Introduction

In this chapter, we present the implementation details and experimental results of our study, aiming to advance the current state-of-the-art in medical imaging datasets. We begin by providing an overview of the implementation details including hard and soft ware implementation also hyperparameters tunning, then we discuss the datasets used in our research and the preprocessing steps undertaken to prepare the data for model training and evaluation. The datasets employed in our study include Chest, DermaMINIST, BloodMNIST, BreastMNIST, PneumoniaMNIST, OrganCMNIST , OrganSMNIST , and OrganAMNIST , each possessing unique characteristics and challenges.

After explaining the dataset description and preprocessing steps, we proceed to discuss the results of the base models after fine-tuning them on each individual dataset. We evaluate the performance of four different models: Vision Transformer (ViT), Swin Transformer (Swin), BEiT Transformer, and DEiT Transformer, on the respective datasets. The metrics used for evaluation include accuracy, precision, recall, and F1 score, providing a comprehensive understanding of the models' performance on each dataset.

Subsequently, we explore the results of our ensemble method implementation for each dataset type. Ensemble method combines the predictions of the four models to improve overall performance. We analyze the effectiveness of ensemble method by leveraging the diverse predictions of our base models, exploring the potential for enhanced accuracy and robustness.

Furthermore, we conduct an ablation study, systematically examining all possible combinations of the fine-tuned models. This study allows us to gain findings into the individual contributions of each model and determine the impact of their interactions on the overall performance. By selectively removing or combining specific models, we assess their effects on accuracy, precision, recall, and F1 score, unraveling the underlying dynamics within the ensemble.

Following the ablation study, we discuss the results of our best ensemble method, compared with individual models and compared to the state-of-the-art. Through this comprehensive analysis, we gain a deeper understanding of the strengths and weaknesses of our methodology, paving the way for meaningful conclusions.

4.2 Implementation

4.2.1 Hardware and Software Specifications

The experiments and model training in this study were primarily conducted using cloud-based computing resources and open-source software tools. Specifically:

- **Python**: Python is a high-level, interpreted programming language known for its simplicity, readability, and extensive ecosystem of libraries for various domains, including machine learning, data analysis, and scientific computing[108]. We implemented all the code for loading and fine-tuning the transformer models, implementing the ensemble techniques, and evaluating the models using Python. Python's extensive ecosystem of libraries, such as PyTorch, Scikit-learn, and NumPy, facilitated efficient implementation and experimentation.

- **Google Colab**: Google Colab is a cloud-based Jupyter notebook environment that provides free access to GPU-accelerated computing resources. It allows users to write and execute Python code through a web interface, making it a popular choice for machine learning and deep learning tasks that benefit from GPU acceleration[109]. We utilized Google Colab as our primary computing environment for coding, training, and evaluating the transformer models and ensemble techniques. The cloud-based Jupyter notebook interface of Google Colab allowed us to leverage GPU-accelerated computing resources, which were essential for efficient training of the large transformer models on the medical imaging datasets.

- **Kaggle Notebooks**: Kaggle is a platform for data science competitions, datasets, and code sharing. It offers a cloud-based notebook environment similar to Google Colab, where users can access GPU-accelerated resources for coding, training, and evaluating machine learning models [110]. In addition to Google Colab, we occasionally used Kaggle Notebooks for specific tasks, such as ensemble methodcoding and base models trainings and evaluation. The Kaggle platform provided an alternative cloud-based notebook environment with GPU-accelerated resources.

- **Hugging Face Transformers**: The Hugging Face Transformers library is an open-source Python library that provides pre-trained transformer models for various natural language processing (NLP) and computer vision tasks. These models are state-of-the-art and have been pre-trained on large datasets, enabling efficient transfer learning for downstream tasks [111]. We obtained the pre-trained BEiT, DEiT, ViT, and Swin Transformer models from the Hugging Face Transformers library. These models served as the base models for our experiments, and we fine-tuned separate instances of each model on the respective medical imaging datasets using the LoRA technique.

- **PyTorch**: PyTorch is an open-source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing. It provides a flexible and efficient computational framework for building and training deep neural networks and other machine learning models [112].

  We utilized PyTorch as the primary deep learning framework for implementing and training the transformer models and ensemble techniques. PyTorch's dynamic

computational graph and support for GPU acceleration made it well-suited for training the large transformer models on the medical imaging datasets. Its intuitive and Pythonic interface allowed for rapid prototyping and experimentation during the model development and fine-tuning process.

- **LoRA (Low-Rank Adaptation)**: LoRA (Low-Rank Adaptation) is a technique that introduces a small number of trainable parameters during fine-tuning, while keeping the pre-trained model weights frozen [113]. This approach significantly reduces the memory and computational requirements, making it suitable for adapting large models to specific tasks or datasets. We employed LoRA to fine-tune the pre-trained transformer models on the medical imaging datasets, benefiting from its parameter-efficient fine-tuning capabilities.

4.2.2 PEFT and LoRA Configuration

For fine-tuning the Vision Transformer models on medical imaging datasets, we utilized the Parameter-Efficient Fine-Tuning (PEFT) approach with Low-Rank Adaptation (LoRA). This method efficiently adapts pre-trained models by introducing small, trainable weight matrices. The configuration details for LoRA are as follows:

- **r**: The rank for the low-rank adaptation. We set $r = 16$ to balance between adaptation capability and computational efficiency.
- **lora alpha**: A scaling factor. We used `lora_alpha`=16 to ensure the adaptation significantly influences the model without overwhelming it.
- **target modules**: The modules to which LoRA is applied, specifically "■query" and "value" in this setup, which are critical components in the transformer's attention mechanism.

  lora dropout: Dropout rate set to 0.1 to prevent overfitting by randomly dropping some connections during training.

4.2.3 Training hyperparameters

The following hyperparameters were used for training the models with the Hugging Face Trainer API. These were selected based on extensive experimentation to balance training efficiency and model performance.

- **Learning_rate**: We set the learning rate to 5e3, providing a balance between fast convergence and stable updates.
- **Per_device_train_batch_size**: We set the batch size to 16, chosen after testing various sizes. Higher values caused out-of-memory errors, while lower values led to reduced accuracy.
- **Gradient_accumulation_steps**: We set this to 4, effectively increasing the batch size without requiring more GPU memory.
- **Per_device_eval_batch_size**: We also set this to 16, to match the training batch size for consistency.
- **Fp16**: We enabled this for faster computations and reduced memory usage.
- **Num_train_epochs**: We set this to 10, determined to be sufficient for convergence while preventing overfitting.

- **Logging_steps**: We set this to 10, ensuring frequent logging for better monitoring during training.
- **Optimizer**: We used the default Adam optimizer, which is well-suited for training transformer models due to its adaptive learning rate capabilities.

These hyperparameters were meticulously chosen after extensive experimentation. For instance, the batch size was adjusted based on memory constraints and model performance. Higher batch sizes resulted in out-of-memory errors, while lower batch sizes led to a drop in accuracy.Similarly to the number or ephocs heigher number of ephocs caused overfitting to models and also other hyperparameters were tuned to optimize the training process and achieve the best possible performance on the medical imaging datasets.

4.2.4 Stacking ensemble hyperparameter tuning

In addition to fine-tuning individual transformer models, when we employed stacking ensemble method it has some hyperparameters that we need to mention, which combines the predictions of multiple base models using a meta-model, such as Support Vector Machines (SVM) and Logistic Regression. Hyperparameter tuning was performed to optimize the performance of the meta-models. The following hyperparameters were tuned using grid search:

- For SVM:
  - **C**: The regularization parameter. We explored values in the range of [0.001, 0.01, 0.1, 1.0, 10.0, 100.0].
  - **kernel**: The kernel type. We experimented with linear, polynomial, radial basis function (RBF), and sigmoid kernels.
  - **gamma**: Kernel coefficient for 'rbf', 'poly', and 'sigmoid'. We tested with both 'scale' and 'auto'.
  - **degree**: Degree of the polynomial kernel function. We examined degrees 2, 3, 4, and 5.
- For Logistic Regression:
  - **C**: Inverse of regularization strength. We searched for optimal values in the range of [0.001, 0.01, 0.1, 1.0, 2.0, 10.0, 100.0].
  - **penalty**: The norm used in the penalization. We compared 'l1' and 'l2' penalties.
  - **solver**: Algorithm to use in the optimization problem. We explored 'liblinear' and 'saga' solvers.

These hyperparameters were selected to balance model complexity and performance, ensuring robustness across different datasets and tasks.

4.3 Experimental results and discussion

4.3.1 Datasets

Medical imaging datasets vary widely in complexity and content, encompassing X-rays, MRI scans, and more section 2.1. These datasets are crucial for diagnosis and treatment planning but present challenges due to their diversity.In this section, we introduce various datasets sourced from collections like MedMNIST databases. These datasets cover a range of medical conditions, from pulmonary diseases to DermaMNIST .We also outline our preprocessing steps, including resizing, normalization, and data augmentation. These steps ensure data consistency and enhance model performance across different tasks.

4.3.1.1 BloodMNIST

This dataset is based on the BloodMNIST dataset from the MedMNIST collection. It contains 17,092 images of individual normal cells, organized into 8 classes. The images are grayscale and have a resolution of 28x28 pixels section 2.2.1.

4.3.1.2 OrganCMNIST

This dataset, also known as OrganCMNIST, is based on 3D computed tomography (CT) images from the Liver Tumor Segmentation Benchmark. It contains 2D axial view images of 11 body organs, resized to 28x28 pixels, for a multi-class classification task section 2.2.1.

4.3.1.3 OrganSMNIST

This dataset, known as OrganSMNIST , is derived from 3D magnetic resonance imaging (MRI) scans. It contains 2D sagittal view images of various organs, resized to 28x28 pixels. The focus is on providing a diverse set of organ images for multi-class classification tasks.The dataset includes images labeled into 11 classes representing various OrganSMNIST (e.g., brain, heart, kidney) section 2.2.1.

4.3.1.4 OrganAMNIST

This dataset contains 2D axial view images of abdominal OrganSMNIST from 3D CT scans. The images are resized to 28x28 pixels.The dataset consists of images labeled into various classes corresponding to 11 abdominal OrganSMNIST (e.g., liver, spleen, pancreas) section 2.2.1.

4.3.1.5 DermaMNIST

The DermaMNIST dataset is based on the HAM10000 collection of dermatoscopic images from the MedMNIST article. It consists of 10,015 images categorized into 7 different skin disease classes, with images resized to 28x28 pixels for multi-class classification section 2.2.1.

4.3.1.6 BreastMNIST

The BreastMNIST dataset is part of the MedMNIST v2 collection and is based on the Curated BreastMNIST Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM). It consists of 780 mammogram images resized to 28x28 pixels for binary classification, categorized into malignant and benign classes section 2.2.1.

4.3.1.7 PneumoniaMNIST

The PneumoniaMNIST dataset is part of the MedMNIST v2 collection, derived from the Chest X-ray dataset of pediatric patients. It consists of 5,856 X-ray images resized to 28x28 pixels for binary classification, categorized into PneumoniaMNIST and normal classes section 2.2.1.

4.3.1.8 Chest

This dataset, consisting of 5,824 images with PneumoniaMNIST annotations, was exported via Roboflow and resized to 640x640 pixels. Unlike other datasets from MedMNIST v2, this dataset is sourced from Hugging Face. we used it to validate that our proposed ensemble methods can be effective across different datasets, not just MedMNIST.

4.3.2 Dataset preprocessing

We performed necessary preprocessing steps to prepare the medical imaging data for training, validation, and testing. First, we utilized the `AutoImageProcessor` from the Hugging Face Transformers library to load the image processor corresponding to the pre-trained model checkpoint. This image processor provided the mean and standard deviation values for normalization.

For all datasets, including training, validation, and testing, we applied a series of transformations to ensure consistency and prepare the data for model input.

- `Resized` the input images to the desired height without altering the aspect ratio.
- `Cropped` the input images to the desired height from the center.
- `Converted` the input images to PyTorch tensors.
- `Normalized` the input images using the mean and standard deviation values from the image processor.

These transformations were combined and applied to each example batch to ensure uniform processing across all datasets.

For the training data specifically, we additionally applied the following augmentations to enhance model generalization:

- `Resized and cropped` the input images to the desired height while applying random aspect ratios and scales.
- `Randomly flipped` the input images horizontally with a 50% probability.

By avoiding data augmentation techniques like random crops and flips for validation and testing data, we aimed to evaluate the model's performance on consistent, unaltered data samples for both validation and testing.

The preprocessed data, in the form of PyTorch tensors, was then used for training, validating, and evaluating the models.

4.3.3 Evaluation of base models

We present the results of evaluating individual base models on various medical imaging datasets. These base models, including Vision Transformer (ViT), Swin Transformer, BEiT, and DEiT transformers, have been fine-tuned on each dataset to perform specific classification tasks.The evaluation metrics include accuracy, precision, recall, and F1 score, providing a comprehensive understanding of each model's performance across different datasets. These results serve as a baseline for comparing the effectiveness of ensemble methods, we start by examining the Chest dataset, followed by analyses of the DermaMNIST, BloodMNIST, OrganCMNIST , OrganSMNIST , OrganAMNIST , BreastMNIST, and PneumoniaMNIST datasets. Through these analyses, we aim to gain findings into the strengths and limitations of each model on different medical imaging tasks and datasets, offering valuable perspectives for further research and application.

4.3.3.1 Chest dataset

Table 4.1: Results of individual models on Chest dataset

Dataset	Model	Accuracy	Precision	Recall	F1 Score
Chest	Vit	**96.22%**	95.06%	95.96%	95.49%
	Swin	95.88%	95.99%	94.01%	94.92%
	BEiT	91.07%	89.23%	89.23%	89.23%
	DEiT	**96.22%**	**95.31%**	**95.62%**	**95.46%**

A. ViT model: The ViT model achieved an accuracy of 96.22%, which is the highest among the models tested Table 4.1. The precision was recorded at 95.06%, indicating a high rate of correctly identified positive instances among the predicted positives. The recall was slightly higher at 95.96%, suggesting that the model effectively identified the majority of actual positive instances. The F1 score, which balances precision and recall, was 95.49%, reflecting robust overall performance.

B. Swin model: The Swin Transformer demonstrated strong performance as well, with an accuracy of 95.88% Table 4.1. It excelled in precision, achieving 95.99%, slightly outperforming ViT in this metric. However, its recall was slightly lower at 94.01%, indicating a marginally lesser capability in identifying all actual positives compared to ViT. The F1 score for Swin was 94.92%, showing a balanced performance, though slightly trailing ViT in overall effectiveness.

C. BEiT model: The BEiT model exhibited notably lower performance metrics compared to the other models Table 4.1. It achieved an accuracy of 91.07%, with both precision and recall at 89.23%. The F1 score matched these values at 89.23%, indicating that BEiT was consistently less effective across all metrics for this dataset. This suggests potential limitations in BEiT's ability to generalize well on the Chest dataset or possibly suboptimal fine-tuning.

D. DEiT model: The DEiT model matched ViT in terms of accuracy, also achieving 96.22% Table 4.1. Its precision was slightly higher at 95.31%, while recall was 95.62%, reflecting a strong capability in identifying positive instances both correctly and thoroughly. The F1 score for DEiT was 95.46%, closely aligning with ViT's performance. This indicates that DEiT is equally effective as ViT on this dataset, showcasing its robustness.

Among the four transformer models, ViT and DEiT demonstrated the highest accuracy at 96.22%, indicating their superior performance on the Chest dataset Table 4.1. While Swin also performed well, it slightly lagged behind in terms of recall and F1 score. BEiT, on the other hand, showed significantly lower performance across all metrics, suggesting it may not be as well-suited for this particular dataset or requires further optimization and hyperparameter tunning. The size and characteristics of the Chest dataset influenced these results. Both ViT and DEiT, with their high accuracy and balanced precision-recall metrics, appear to handle the dataset's characteristic effectively. Swin's slightly lower recall points to a minor trade-off, while BEiT's overall lower metrics indicate potential areas for improvement.

4.3.3.2 DermaMNIST dataset

Table 4.2: Results of individual models on DermaMNIST dataset

Dataset	Model	Accuracy	Precision	Recall	F1 Score
DermaMNIST	Vit	76.81%	68.21%	50.61%	54.97%
	Swin	78.75%	63.08%	61.01%	61.50%
	BEiT	75.61%	57.42%	53.53%	52.71%
	DEiT	**79.60%**	**66.43%**	**58.91%**	**61.64%**

A. ViT Model: The ViT model achieved an accuracy of 76.81%, indicating a moderate level of performance Table 4.2. Its precision was recorded at 68.21%, showing that a good portion of the predicted positive instances were correct. However, the recall was significantly lower at 50.61%, suggesting that the model struggled to identify all actual positive instances. The F1 score, balancing precision and recall, was 54.97%, reflecting the disparity between precision and recall.

B. Swin Model: The Swin Transformer model outperformed ViT in terms of accuracy, achieving 78.75% Table 4.2. However, its precision was lower at 63.08%, indicating a higher rate of false positives compared to ViT. The recall was notably higher at 61.01%, meaning it was more effective in identifying actual positives. The F1 score of 61.50% demonstrates a more balanced performance compared to ViT.

C. BEiT Model: The BEiT model showed the lowest performance metrics among the models tested Table 4.2. It achieved an accuracy of 75.61%, with a precision of 57.42%, and recall of 53.53%. The F1 score was 52.71%, indicating that BEiT struggled to balance precision and recall effectively on the DermaMNIST dataset, similar to its performance on the Chest dataset.

D. DEiT Model: The DEiT model achieved the highest accuracy among the models at 79.60%, indicating strong performance on the DermaMNIST dataset Table 4.2. Its precision was 66.43%, which is moderate, while the recall was 58.91%, showing that it was fairly good at identifying actual positives. The F1 score for DEiT was 61.64%, the highest among the models tested, indicating a good balance between precision .

On the DermaMNIST dataset, the DEiT model demonstrated the highest accuracy at 79.60% and the highest F1 score at 61.64%, indicating its superior performance. The Swin Transformer also performed well, with a balanced F1 score of 61.50%, despite a slightly lower accuracy Table 4.2. ViT showed moderate performance with a noticeable disparity between precision and recall, while BEiT had the lowest metrics, suggesting it was less effective on this dataset. The variations in performance metrics among the models attributed to the unique characteristics and complexity of the DermaMNIST dataset. DEiT and Swin models showed robustness in handling the dataset's challenges, whereas ViT and BEiT exhibited limitations.

4.3.3.3 BloodMNIST dataset

Table 4.3: Results of individual models on BloodMNIST dataset

Dataset	Model	Accuracy	Precision	Recall	F1 Score
BloodMNIST	Vit	**97.90%**	**97.72%**	**97.85%**	**97.78%**
	Swin	96.49%	96.27%	96.16%	96.19%
	BEiT	96.20%	96.18%	95.56%	95.83%
	DEiT	97.37%	97.30%	97.06%	97.18%

A. ViT model: The ViT model achieved an impressive accuracy of 97.90%, the highest among the models tested on the BloodMNIST dataset Table 4.3. The precision was 97.72%, indicating a very high rate of correctly identified positive instances among the predicted positives. The recall was slightly higher at 97.85%, showing the model's effectiveness in identifying almost all actual positive instances. The F1 score, balancing precision and recall, was 97.78%, reflecting exceptional overall performance.

B. Swin model: The Swin Transformer demonstrated strong performance with an accuracy of 96.49% Table 4.3. Its precision was 96.27%, indicating a high correctness rate among predicted positives. The recall was 96.16%, suggesting a slight drop in identifying all actual positives compared to precision. The F1 score of 96.19% indicates a well-balanced performance, though slightly lower than ViT.

C. BEiT model: The BEiT model showed an accuracy of 96.20%, similar to Swin's performance but marginally lower Table 4.3. The precision was 96.18%, and recall was 95.56%, indicating a minor imbalance between the two. The F1 score was 95.83%, reflecting a solid overall performance but slightly behind Swin and ViT.

D. DEiT model: The DEiT model achieved a high accuracy of 97.37%, closely following ViT. Its precision was 97.30%, showing a high rate of correct positive predictions. The recall was slightly lower at 97.06%, indicating excellent but slightly lesser ability to identify all actual positives compared to ViT Table 4.3. The F1 score was 97.18%, indicating a strong balance between precision and recall, and placing DEiT second in performance on the BloodMNIST dataset.

Among the four transformer models on the BloodMNIST dataset, ViT demonstrated the highest performance with an accuracy of 97.90% and F1 score of 97.78%, indicating its superior capability in handling this dataset Table 4.3. DEiT also performed excellently, with an accuracy of 97.37% and an F1 score of 97.18%, making it a strong competitor. Swin and BEiT, while performing well, lagged slightly behind ViT and DEiT. The high performance of all models on the BloodMNIST dataset suggests that this dataset is less challenging and more suitable for these transformer models compared to others. The slight variations in performance metrics highlight the nuanced differences in how each model processes and learns from the data.

4.3.3.4 OrganCMNIST dataset

Table 4.4: Results of individual models on OrganCMNIST dataset

Dataset	Model	Accuracy	Precision	Recall	F1 Score
OrganCMNIST	Vit	**92.83%**	**92.31%**	**91.60%**	**91.89%**
	Swin	91.99%	91.95%	90.67%	91.21%
	BEiT	92.56%	92.28%	91.37%	91.75%
	DEiT	92.40%	91.99%	91.23%	91.54%

A. ViT model: The ViT model achieved an accuracy of 92.83% on the OrganCMNIST dataset. Its precision was 92.31%, indicating a high rate of correctly identified positive instances among the predicted positives Table 4.4. The recall was 91.60%, suggesting that the model effectively identified the majority of actual positive instances. The F1 score, balancing precision and recall, was 91.89%, reflecting robust overall performance.

B. Swin model: The Swin Transformer achieved an accuracy of 91.99% on the OrganCMNIST dataset. Its precision was 91.95%, indicating a high rate of correctly identified positive instances among the predicted positives Table 4.4. The recall was 90.67%, suggesting that the model effectively identified a significant portion of actual positive instances. The F1 score of 91.21% demonstrates a balanced performance, though slightly lower than ViT.

C. BEiT model: The BEiT model showed an accuracy of 92.56% on the OrganCMNIST dataset. Its precision was 92.28%, indicating a high rate of correctly identified positive instances among the predicted positives Table 4.4. The recall was 91.37%, suggesting that the model effectively identified a majority of actual positive instances. The F1 score of 91.75% reflects a balanced performance, comparable to ViT.

D. DEiT model: The DEiT model achieved an accuracy of 92.40% on the OrganCMNIST dataset Table 4.4. Its precision was 91.99%, indicating a high rate of correctly identified positive instances among the predicted positives. The recall was 91.23%, suggesting that the model effectively identified a significant portion of actual positive instances. The F1 score of 91.54% demonstrates a balanced performance, though slightly lower than ViT and BEiT.

On the OrganCMNIST dataset, all four transformer models—ViT, Swin, BEiT, and DEiT performed exceptionally well, with accuracies ranging from 91.99% to 92.83%. ViT, BEiT, and DEiT demonstrated particularly high precision and recall, with balanced F1 scores, indicating robust overall performance. Swin also performed well, though slightly lower than the other models in terms of F1 score. The consistent high performance across all models suggests that the OrganCMNIST dataset is well-suited for transformer-based models.

4.3.3.5 OrganSMNIST dataset

Table 4.5: Results of individual models on OrganSMNIST dataset

Dataset	Model	Accuracy	Precision	Recall	F1 Score
OrganSMNIST	Vit	81.65%	78.62%	77.01%	76.59%
	Swin	82.3%	78.98%	77.86%	78.31%
	BEiT	**82.4%**	**78.95%**	**78.21%**	**78.52%**
	DEiT	80.8%	77.03%	76.86%	76.5%

A. ViT model: The ViT model achieved an accuracy of 81.65% on the OrganSMNIST dataset. Its precision was 78.62%, indicating a moderate rate of correctly identified positive instances among the predicted positives Table 4.5. The recall was 77.01%, suggesting that the model missed some actual positive instances. The F1 score, balancing precision and recall, was 76.59%, reflecting room for improvement in overall performance.

B. Swin model: The Swin Transformer achieved an accuracy of 82.30% on the OrganSMNIST dataset. Its precision was 78.98%, indicating a moderate rate of correctly identified positive instances among the predicted positives Table 4.5. The recall was 77.86%, suggesting that the model effectively identified a significant portion of actual positive instances. The F1 score of 78.31% demonstrates a balanced performance, slightly higher than ViT.

C. BEiT model: The BEiT model showed the highest accuracy among the models with 82.40% on the OrganSMNIST dataset Table 4.5. Its precision was 78.95%, indicating a moderate rate of correctly identified positive instances among the predicted positives. The recall was 78.21%, suggesting that the model effectively identified the majority of actual positive instances. The F1 score of 78.52% reflects a balanced performance, indicating robust overall effectiveness.

D. DEiT model: The DEiT model achieved an accuracy of 80.80% on the OrganSMNIST dataset. Its precision was 77.03%, indicating a moderate rate of correctly identified positive instances among the predicted positives Table 4.5. The recall was 76.86%, suggesting that the model missed some actual positive instances. The F1 score of 76.50% demonstrates room for improvement, with lower overall performance compared to the other models.

On the OrganSMNIST dataset, BEiT demonstrated the highest performance with an accuracy of 82.40% and an F1 score of 78.52%. Swin followed closely Table 4.5, with an accuracy of 82.30% and an F1 score of 78.31%. ViT also showed decent performance, but with slightly lower accuracy and F1 score compared to Swin and BEiT. DEiT performed the least well among the models, with an accuracy of 80.80% and an F1 score of 76.50%. The variations in performance among the models is attributed to the unique characteristics and complexities of the OrganSMNIST dataset. BEiT and Swin showed relatively better performance, suggesting their effectiveness in handling this dataset's challenges. ViT performed moderately well but has room for improvement, and DEiT showed the most limitations in its ability to handle the OrganSMNIST dataset effectively.

4.3.3.6 OrganAMNIST dataset

Table 4.6: Results of individual models on OrganAMNIST dataset

Dataset	Model	Accuracy	Precision	Recall	F1 Score
OrganAMNIST	Vit	93.27 %	93.99 %	93.01%	93.34%
	Swin	93.87%	94.3 %	93.43%	93.73%
	BEiT	93.29%	94.16%	92096%	93.40 %
	DEiT	**94.4** %	**94.64 %**	**93.95%**	**94.21 %**

A. ViT model: The ViT (Vision Transformer) model achieved an accuracy of 93.27% on the OrganAMNIST dataset. Its precision was 93.99%, indicating a high rate of correctly identified positive instances among the predicted positives Table 4.6. The recall was 93.01%, suggesting that the model effectively identified the majority of actual positive instances. The F1 score, balancing precision and recall, was 93.34%, reflecting robust overall performance.

B. Swin model: The Swin Transformer achieved an accuracy of 93.87% on the OrganAMNIST dataset. Its precision was 94.30%, indicating a high rate of correctly identified positive instances among the predicted positives Table 4.6. The recall was 93.43%, suggesting that the model effectively identified the majority of actual positive instances. The F1 score of 93.73% demonstrates excellent balanced performance, slightly higher than ViT.

D. BEiT model: The BEiT model showed an accuracy of 93.29% on the OrganAMNIST dataset. Its precision was 94.16% Table 4.6, indicating a high rate of correctly identified positive instances among the predicted positives. The recall was 92.96%, suggesting that the model effectively identified a significant portion of actual positive instances. The F1 score of 93.40% reflects a balanced performance, similar to ViT.

D. DEiT model: The DEiT model achieved the highest accuracy among the models with 94.40% on the OrganAMNIST dataset. Its precision was 94.64%, indicating a very high rate of correctly identified positive instances among the predicted positives Table 4.6. The recall was 93.95%, suggesting that the model effectively identified the majority of actual positive instances. The F1 score of 94.21% demonstrates robust overall performance, making it the best performer on this dataset.

On the OrganAMNIST dataset, DEiT demonstrated the highest performance with an accuracy of 94.40% and an F1 score of 94.21% Table 4.6. Swin followed closely, with an accuracy of 93.87% and an F1 score of 93.73%. ViT and BEiT also showed strong performance, with accuracies of 93.27% and 93.29% and F1 scores of 93.34% and 93.40%, respectively. The consistent high performance across all models suggests that the OrganAMNIST dataset is well-suited for transformer-based models because it has large number of samples so the models had learned well durring the finetunning process. DEiT showed the highest effectiveness, suggesting it have captured the complexities of this dataset slightly better than the other models. Swin, ViT, and BEiT also performed excellently, indicating robust capabilities in handling the dataset's challenges. By considering the dataset size and characteristics, we gain findings into the strengths of the individual model performances on the OrganAMNIST dataset, providing a more comprehensive understanding of the context in which these models operate.

4.3.3.7 BreastMNIST dataset

Table 4.7: Results of individual models on BreastMNIST dataset

Dataset	Model	Accuracy	Precision	Recall	F1 Score
BreastMNIST	ViT	**87.82%**	**89.71%**	**78.88%**	**82.32%**
	Swin	85.26%	81.62%	80.14%	80.82%
	BEiT	73.08%	36.54%	52.35%	42.22%
	DEiT	83.33%	80.79%	74.31%	76.53%

A. ViT model: The ViT (Vision Transformer) model achieved an accuracy of 87.82% on the BreastMNIST dataset. Its precision was 89.71%, indicating a high rate of correctly identified positive instances among the predicted positives Table 4.7. However, the recall was notably lower at 78.88%, suggesting that the model missed some actual positive instances. The F1 score, balancing precision and recall, was 82.32%, reflecting relatively good overall performance with room for improvement in recall.

B. Swin model: The Swin Transformer achieved an accuracy of 85.26% on the BreastMNIST dataset. Its precision was 81.62%, indicating a moderate rate of correctly identified positive instances among the predicted positives Table 4.7. The recall was 80.14%, suggesting a slightly lower ability to identify all actual positives. The F1 score of 80.82% reflects a relatively balanced performance but with a slight emphasis on precision.

C. BEiT model: The BEiT model showed the lowest performance metrics among the models tested on the BreastMNIST dataset Table 4.7. It achieved an accuracy of 73.08%, with a precision of 36.54% and recall of 52.35%. The F1 score was 42.22%, indicating a significant imbalance between precision and recall and overall poor performance on this dataset.

D. DEiT model: The DEiT (Data-efficient Image Transformer) model achieved an accuracy of 83.33% on the BreastMNIST dataset. Its precision was 80.79%, indicating a moderate rate of correctly identified positive instances among the predicted positives Table 4.7. However, the recall was relatively lower at 74.31%, suggesting that the model missed some actual positive instances. The F1 score of 76.53% reflects a balanced performance but with room for improvement in recall.

On the BreastMNIST dataset, ViT demonstrated the highest performance with an accuracy of 87.82% and an F1 score of 82.32%. Swin followed closely, with slightly lower accuracy and F1 score but relatively balanced performance metrics Table 4.7. DEiT also showed decent performance, with room for improvement in recall. However, BEiT performed poorly on this dataset, with a significant imbalance between precision and recall. It's important to note that the BreastMNIST dataset used for this analysis is relatively small, consisting of only 780 samples. The small size of the dataset had impacted the performance of the transformer models. With limited data, the models may not have had sufficient examples to learn complex patterns effectively, potentially leading to sub optimal performance, especially in terms of precision and recall.

4.3.3.8 PneumoniaMNIST dataset

Table 4.8: Results of individual models on PneumoniaMNIST dataset

Dataset	Model	Accuracy	Precision	Recall	F1 Score
PneumoniaMNIST	ViT	**93.59%**	**94.74%**	**91.79%**	**92.95%**
	Swin	88.78%	92.17%	85.13%	87.12%
	BEiT	84.46%	83.54%	83.12%	83.32%
	DEiT	86.22%	86.23%	84.02%	84.87%

A. ViT model: The ViT model achieved an accuracy of 93.59% on the PneumoniaMNIST dataset. Its precision was 94.74%, indicating a high rate of correctly identified positive instances among the predicted positives Table 4.8. However, the recall was slightly lower at 91.79%, suggesting that the model missed some actual positive instances. The F1 score, balancing precision and recall, was 92.95%, reflecting strong overall performance but with some room for improvement in recall.

B. Swin model: The Swin Transformer achieved an accuracy of 88.78% on the PneumoniaMNIST dataset. Its precision was relatively high at 92.17%, indicating a good rate of correct positive predictions Table 4.8. However, the recall was noticeably lower at 85.13%, indicating a higher number of missed positive instances. The F1 score of 87.12% reflects a relatively balanced performance but with a slight emphasis on precision.

C. BEiT model: The BEiT model showed an accuracy of 84.46% on the PneumoniaMNIST dataset. Its precision was 83.54%, indicating a moderate rate of correctly identified positive instances among the predicted positives Table 4.8. The recall was 83.12%, suggesting a somewhat balanced ability to identify actual positives. The F1 score of 83.32% reflects a decent overall performance but with room for improvement.

D. DEiT model: The DEiT model achieved an accuracy of 86.22% on the PneumoniaMNIST dataset. Its precision was 86.23%, indicating a relatively high rate of correctly identified positive instances among the predicted positives Table 4.8. However, the recall was 84.02%, suggesting that the model missed some actual positive instances. The F1 score of 84.87% reflects a balanced performance but with some room for improvement, particularly in recall.

On the PneumoniaMNIST dataset, ViT demonstrated the highest performance with an accuracy of 93.59% and an F1 score of 92.95% Table 4.8. Swin, DEiT, and BEiT followed with decreasing performance, indicating variations in their ability to handle the complexities of the dataset. Swin had a slightly higher emphasis on precision, while DEiT and BEiT showed more balanced performance metrics.
The variations in performance among the models may be attributed to the unique characteristics and small size of the PneumoniaMNIST dataset. ViT showed the best performance, suggesting its effectiveness in handling this dataset's challenges. Swin, DEiT, and BEiT showed decent performance but may require further optimization to achieve higher accuracy and F1 scores.

General analysis of individual model performances across all datasets

A. ViT model: ViT demonstrated strong performance across various datasets, particularly excelling in BloodMNIST (97.90% accuracy) and Chest (96.22% accuracy), while

displaying comparatively lower performance on DermaMNIST (76.81% accuracy) and OrganSMNIST (81.65% accuracy). Despite facing challenges with the small BreastMNIST dataset, ViT maintained high precision and recall on most datasets, indicating reliable detection of positive instances and low false positive rates. ViT showcased consistent performance and proved to be adept at handling diverse medical imaging tasks.

B. Swin model: Swin demonstrated high performance across various datasets, particularly excelling in BloodMNIST (96.49% accuracy) and OrganAMNIST (93.87% accuracy), with notable consistency observed in OrganCMNIST (91.99% accuracy). It maintained balanced precision and recall, indicating robustness in detecting and correctly identifying positive instances. However, Swin faced challenges similar to ViT, showing slightly lower performance on the DermaMNIST dataset (78.75% accuracy). Despite this, Swin's overall performance suggests its efficacy in handling diverse medical imaging tasks.

C. BEiT Model: BEiT demonstrated varying performance across datasets, high performer on OrganSMNIST (82.4% accuracy) and excelling on OrganAMNIST (93.29% accuracy) and OrganCMNIST (92.56% accuracy), while displaying poorer performance on DermaMNIST (75.61% accuracy) and BreastMNIST (73.08% accuracy). Despite these discrepancies, BEiT generally maintained a balanced F1 score, indicating good overall effectiveness in multiple scenarios. The challenges encountered with the DermaMNIST and BreastMNIST datasets underscored BEiT's limitations, potentially influenced by dataset size and complexity. Nonetheless, BEiT showcased its capability to handle diverse medical imaging tasks with reasonable consistency.

D. DEiT Model: DEiT demonstrated the highest performance among the models on several datasets, particularly excelling on OrganAMNIST (94.40% accuracy) and BloodMNIST (97.37% accuracy), consistently ranking as the top performer. It maintained high precision, recall, and F1 scores across datasets, showcasing its robustness and reliability. However, challenges were noted as DEiT's performance on the DermaMNIST (79.60% accuracy) and OrganSMNIST (80.80% accuracy) datasets was comparatively lower compared to others. Despite these challenges, DEiT's overall performance suggests its effectiveness in handling diverse medical imaging tasks with consistency and reliability.

The BreastMNIST dataset, comprising only 780 samples, posed challenges for all models, affecting their capacity to effectively learn complex patterns. This underscores the significance of larger and more diverse datasets for training robust models. Conversely, larger datasets like Organa, OrganCMNIST , and OrganSMNIST enabled the models to excel, showcasing the advantages of extensive training data in capturing intricate patterns and enhancing overall model performance.

The transformer models demonstrated strong performance across diverse medical imaging datasets, with DEiT emerging as the top performer in many datasets. The analysis underscores the importance of dataset size and complexity in influencing model performance and highlights the need for large, diverse training datasets to develop robust medical imaging models. By leveraging the strengths of these transformer models, significant advancements in medical imaging analysis and diagnostics can be achieved.

4.3.4 Evaluation of ensemble methods

We evaluate the performance of the proposed ensemble methods that use the strengths of the four base models (ViT, Swin, BEiT, and DEiT). The ensembles are constructed using various strategies, including soft voting, hard voting, weighted hard voting, weighted soft voting, bagging, boosting, and stacking with SVM and stacking with logistic regression.

4.3.4.1 Chest Dataset

Table 4.9: Enemble method results on Chest dataset

Dataset	Ensemble method	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	Soft Voting	96.74%	96.14%	95.98%	96.06%
	Weighted Soft Voting	96.80%	95.80%	96.30%	96.05%
	Hard Voting	96.74%	96.14%	95.98%	96.06%
	Weighted Hard Voting	96.91%	96.58%	95.93%	96.25%
	Stacking (SVM)	**97.25%**	**98.05%**	**98.05%**	**98.05%**
	Stacking (Logistic)	96.91%	96.42%	96.10%	96.26%
	Bagging	96.56%	96.60 %	96.56 %	96.58%
	Boosting	96.74 %	96.84 %	96.74 %	96.76 %

Voting methods: Both Soft Voting and Hard Voting exhibited robust performance Table 4.9, yielding high accuracy of approximately 96.74%. These methods effectively leveraged the collective decisions of base models, resulting in balanced precision and recall rates.

Weighted voting methods: Weighted Soft Voting and Weighted Hard Voting emerged as top performers among the voting methods Table 4.9, achieving the highest accuracy of approximately 96.80% and 96.91%, respectively. By assigning weights based on individual classifier performance, these methods optimized the ensemble's predictive power, leading to improved accuracy.

Stacking methods: Stacking with SVM emerged as the standout performer, boasting the highest accuracy of 97.25% among all ensemble methods Table 4.9. This approach leveraged the diverse predictions of base models, enabling more effective decision-making and yielding superior classification outcomes. Stacking with Logistic Regression also delivered respectable performance, achieving an accuracy of 96.91%.

Bagging and Boosting: While Bagging and Boosting methods showcased competitive performance Table 4.9, achieving accuracies around 96.56% and 96.74%, respectively, they required more computational resources and time for optimization. Despite this they demonstrated their efficacy in enhancing classification accuracy on the Chest dataset.

Summary: The ensemble methods showcased remarkable effectiveness in enhancing classification performance on the Chest dataset. Weighted Voting methods Table 4.9, particularly Weighted Hard Voting, and Stacking with SVM emerged as the top performers, surpassing individual base models in accuracy. Voting methods also proved reliable, offering a simpler yet effective approach to ensemble method.

4.3.4.2 DermaMNIST dataset

Table 4.10: Enemble method results on DermaMNIST dataset

Dataset	Ensemble methods	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	Soft Voting	79.00%	67.50%	67.50%	60.86%
	Weighted Soft Voting	**80.00%**	**67.88%**	**60.10%**	**62.75%**
	Hard Voting	78.85%	66.63%	57.55%	59.86%
	Weighted Hard Voting	79.70%	66.84%	59.60%	62.11%
	Stacking (SVM)	78.55%	58.75%	55.46%	56.61%
	Stacking (Logistic)	79.85%	69.80%	56.13%	60.28%
	Bagging	78.70%	78.01%	78.70%	78.02%
	Boosting	79.15%	78.17%	79.15%	78.04%

Voting methods: Both Soft Voting and Hard Voting yielded similar performance, with an accuracy of approximately 79.00% Table 4.10. However, their precision, recall, and F1 scores were relatively low, indicating suboptimal performance compared to other methods.

Weighted voting methods: Weighted Soft Voting outperformed other voting methods on the DermaMNIST dataset, achieving the highest accuracy of 80.00% Table 4.10. However, its precision, recall, and F1 scores were moderate, indicating room for improvement in correctly classifying positive instances. Also weighted hard voting showed comparable results indicating the efficacy or weighting the models.

Stacking methods: Stacking with SVM and Stacking with Logistic Regression showed mixed performance on the DermaMNIST dataset. Both methods achieved accuracies below 80.00% Table 4.10, with relatively low precision, recall, and F1 scores. This suggests that the stacking approach may not be as effective in improving classification performance on this particular dataset(DermaMNIST) specially when using traditional model like SVM and logistic regression.

Bagging and Boosting: Bagging and Boosting methods exhibited moderate performance on the DermaMNIST dataset, with accuracies around 78.70% and 79.15%, respectively Table 4.10. While these methods achieved balanced precision, recall, and F1 scores, their accuracy fell short compared to Weighted Soft Voting.

Summary: The ensemble methods demonstrated varied effectiveness in enhancing classification performance on the DermaMNIST dataset Table 4.10. Weighted Soft Voting emerged as the top performer, achieving the highest accuracy among all methods. However, its precision, recall, and F1 scores suggest the need for further optimization to improve its ability to accurately classify positive instances. Additionally Voting methods showed consistency in accuracy but struggled to achieve high precision and recall rates. Stacking methods, particularly with SVM and Logistic Regression, exhibited limited effectiveness, indicating that the stacking approach may not be well-suited for improving classification performance on the DermaMNIST dataset. Bagging and Boosting methods delivered moderate performance, with balanced precision, recall, and F1 scores.

4.3.4.3 BloodMNIST dataset

Table 4.11: Enemble method results on BloodMNIST dataset

Dataset	Ensemble method	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	Majority Soft Voting	97.81%	97.87%	97.64%	97.75%
	Weighted Soft Voting	97.90%	97.85%	97.81%	97.83%
	Hard Voting	97.78%	97.82%	97.62%	97.72%
	Weighted Hard Voting	**98.00%**	**97.97%**	**97.82%**	**97.89%**
	Stacking (SVM)	97.52%	97.64%	97.29%	97.44%
	Stacking (Logistic)	97.81%	97.88%	97.59%	97.73%
	Bagging	97.49%	97.51%	97.49 %	97.48 %
	Boosting	97.46%	97.47%	97.46%	97.45%

Voting methods: Both Soft Voting and Hard Voting achieved high accuracy of 97.81%, with consistent precision, recall, and F1 scores Table 4.11. These methods demonstrated robust performance in aggregating predictions from base models to make accurate classifications.

Weighted voting methods: Weighted Soft Voting slightly outperformed other voting methods, achieving an accuracy of 97.90%. Its precision, recall, and F1 scores were also commendable, indicating effective integration of base model predictions and weighting strategies to improve classification accuracy Table 4.11. Notably, Weighted Hard Voting emerged as the top performer, boasting the highest accuracy of 98.00%. This method effectively combines the predictions of base models while assigning weights based on their individual performance, leading to superior classification accuracy.

Stacking methods: Stacking with SVM and Stacking with Logistic Regression yielded accuracies around 97.52% and 97.81%, respectively Table 4.11. While these methods demonstrated competitive performance, their precision, recall, and F1 scores were slightly lower compared to other ensemble methods.

Bagging and Boosting: Bagging and Boosting methods exhibited moderate performance on the BloodMNIST dataset, with accuracies ranging from 97.46% to 97.49% Table 4.11. While these methods achieved balanced precision, recall, and F1 scores, their accuracy was slightly lower compared to Weighted Hard Voting.

Summary: The ensemble methods displayed strong performance in enhancing classification accuracy on the BloodMNIST dataset Table 4.11. Hard and soft Voting methods consistently achieved high accuracy, precision, recall, and F1 scores, indicating their effectiveness in aggregating predictions from diverse base models for BloodMNIST. Weighted Voting methods, particularly Weighted Hard Voting, emerged as the top performer with the highest accuracy of 98.00%. This highlights the importance of considering the individual performance of base models and implementing weighting strategies to optimize ensemble predictions. Stacking methods showed competitive performance but trailed slightly behind voting methods in accuracy. Bagging and Boosting methods offered moderate performance, demonstrating balanced precision, recall, and F1 scores the choice of ensemble method should be based on the dataset characteristics and the desired balance between accuracy and computational complexity. Weighted Hard Voting, in particular,

provides an effective approach to improve classification performance on the BloodMNIST dataset by leveraging the strengths of individual base models and incorporating weighting strategies to optimize ensemble predictions.

4.3.4.4 BreastMNIST dataset

Table 4.12: Enemble method results on BreastMNIST dataset

Dataset	Ensemble methods	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	Soft Voting	86.54%	88.75%	76.50%	80.04%
	Weighted Soft Voting	**88.46 %**	**90.18%**	**80.08%**	**83.42%**
	Hard Voting	86.54%	88.75%	76.50%	80.04%
	Weighted Hard Voting	88.46%	90.18%	80.08%	83.42%
	Stacking (SVM)	87.82%	88.53%	79.64%	82.68%
	Stacking (Logistic)	87.18%	89.23%	77.69%	81.20%
	Bagging	81.41%	80.56%	81.41%	79.96%
	Boosting	85.26%	84.78%	85.26%	84.65%

Voting methods: Both Soft Voting and Hard Voting yielded similar performance, with an accuracy of 86.54% Table 4.12. However, their precision, recall, and F1 scores were relatively low, indicating suboptimal performance compared to other methods.

Weighted voting methods: Weighted Soft Voting and Weighted Hard Voting outperformed other voting methods on the BreastMNIST dataset, achieving an accuracy of 88.46% Table 4.12. These methods demonstrated improved precision, recall, and F1 scores, indicating enhanced performance in classifying BreastMNIST -related medical imaging data.

Stacking methods: Stacking with SVM and Stacking with Logistic Regression yielded accuracies around 87.82% and 87.18%, respectively Table 4.12. While these methods exhibited competitive performance, their precision, recall, and F1 scores were slightly lower compared to Weighted Voting methods.

Bagging and Boosting: Bagging and Boosting methods exhibited moderate performance on the BreastMNIST dataset, with accuracies of 81.41% and 85.26%, respectively Table 4.12. While these methods achieved balanced precision, recall, and F1 scores, their accuracy was lower compared to other ensemble methods.

Summary: The ensemble methods displayed varied effectiveness in enhancing classification performance on the BreastMNIST dataset. Weighted Voting methods, particularly Weighted Soft Voting and Weighted Hard Voting, emerged as top performers, surpassing Voting methods in accuracy Table 4.12. Stacking methods showed competitive performance, with slightly lower accuracy compared to Weighted Voting methods. Bagging and Boosting methods offered moderate performance, demonstrating balanced precision, recall, and F1 scores but lower accuracy compared to other ensemble methods. The choice of ensemble method should consider the dataset (characteristics, small size) and the desired balance between accuracy and computational complexity. Weighted

Voting methods, in particular, provide an effective approach to improve classification performance on the BreastMNIST dataset by using the strengths of individual base models and incorporating weighting strategies to optimize ensemble predictions.

4.3.4.5 PneumoniaMNIST dataset

Table 4.13: Enemble method results on PneumoniaMNIST dataset

Dataset	Ensemble methods	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoiaMNIST	Soft Voting	92.95%	94.15%	91.03%	92.23 %
	Weighted Soft Voting	**94.07%**	**95.09%**	**92.44%**	**93.50%**
	Hard Voting	93.59%	94.61%	91.88%	92.96%
	Weighted Hard Voting	**94.07%**	**95.09%**	**92.44%**	**93.50%**
	Stacking (SVM)	90.87%	93.24%	87.99%	89.71%
	Stacking (Logistic)	91.67%	93.75%	89.06%	90.67%
	Bagging	93.43%	93.73%	93.43%	93.32%
	Boosting	92.79%	93.12%	92.79%	92.66%

Voting methods: Soft Voting and Hard Voting achieved moderate performance, with an accuracy of 92.95% and 93.59% respectively Table 4.13. These methods maintained balanced precision, recall, and F1 scores, indicating their effectiveness in ensemble predictions.

Weighted voting methods: Weighted Soft Voting and Weighted Hard Voting outperformed other voting methods on the PneumoniaMNIST dataset Table 4.13, achieving an accuracy of approximately 94.07%. These methods demonstrated improved precision, recall, and F1 scores, indicating enhanced performance in classifying pneumonia-related medical imaging data.

Stacking methods: Stacking with SVM and Stacking with Logistic Regression yielded accuracies around 90.87% and 91.67% , respectively Table 4.13. While these methods exhibited competitive performance, their precision, recall, and F1 scores were slightly lower compared to Weighted Voting methods.

Bagging and Boosting: Bagging and Boosting methods exhibited moderate performance on the PneumoniaMNIST dataset Table 4.13, with accuracies ranging from 92.79% to 93.43%. While these methods achieved balanced precision, recall, and F1 scores, their accuracy was slightly lower compared to Weighted Voting methods.

Summary: The ensemble methods displayed varied effectiveness in enhancing classification performance on the PneumoniaMNIST dataset Table 4.13. Weighted Voting methods, emerged as top performers, surpassing Voting methods in accuracy. Stacking methods showed competitive performance, with slightly lower accuracy compared to Weighted Voting methods. Bagging and Boosting methods offered moderate performance, demonstrating balanced precision, recall, and F1 scores but lower accuracy compared to other ensemble methods. The choice of ensemble method should consider the dataset characteristics and the desired balance between accuracy and computational complexity. Weighted Voting methods, in particular, provide an effective approach to improve classification performance on the PneumoniaMNIST dataset by leveraging the strengths

of individual base models and incorporating weighting strategies to optimize ensemble predictions.

4.3.4.6 OrganCMNIST dataset

Table 4.14: Enemble method results on OrganCMNIST dataset

Dataset	Ensemble methods	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	Soft Voting	93.49%	93.20%	92.27%	92.66%
	Weighted Soft Voting	**93.72%**	**93.33%**	**92.60%**	**92.90%**
	Hard Voting	93.42%	93.13%	92.17%	92.58%
	Weighted Hard Voting	93.67%	93.30%	92.54%	92.86%
	Stacking (SVM)	91.46%	90.63%	90.91%	90.53%
	Stacking (Logistic)	92.10%	91.66%	91.32%	91.27%
	Bagging	93.41%	93.44%	93.41%	93.38%
	Boosting	93.49%	93.54%	93.49%	93.47 %

Voting methods: Both soft voting and hard voting achieved similar performance, with an accuracy of approximately 93.49% Table 4.14. These methods demonstrated balanced precision, recall, and F1 scores, indicating their effectiveness in ensemble predictions.

Weighted voting methods: Weighted soft voting outperformed Weighted hard voting on the OrganCMNIST dataset, achieving an accuracy of approximately 93.72% compared to 93.67% Table 4.14. These methods exhibited improved precision, recall, and F1 scores, indicating enhanced performance in classifying organ-related medical imaging data.

Stacking methods: Stacking with SVM and stacking with logistic regression yielded accuracies around 91.46% and 92.10%, respectively Table 4.14. While these methods exhibited competitive performance, their precision, recall, and F1 scores were slightly lower compared to Weighted Voting methods.

Bagging and Boosting: Bagging and Boosting methods exhibited strong performance on the OrganCMNIST dataset, with accuracies ranging from 93.41% to 93.49% Table 4.14. These methods achieved balanced precision, recall, and F1 scores, demonstrating their effectiveness in ensemble predictions.

Summary: The ensemble method displayed varied effectiveness in enhancing classification performance on the OrganCMNIST dataset Table 4.14. Weighted Soft Voting emerged as the top performer among Weighted Voting methods, surpassing Weighted Hard Voting in accuracy. Bagging and Boosting methods also exhibited strong performance, demonstrating balanced precision, recall, and F1 scores. the choice of ensemble method should consider the dataset characteristics and the desired balance between accuracy and computational complexity. Weighted Soft Voting, Bagging, and Boosting provide effective approaches to improve classification performance on the OrganCMNIST dataset.

4.3.4.7 OrganSMNIST dataset

Table 4.15: Enemble method results on OrganSMNIST dataset

Dataset	Ensemble methods	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	Soft Voting	83.46 %	80.06 %	79.12%	79.29 %
	Weighted Soft Voting	83.72%	80.24%	79.43%	79.59%
	Hard Voting	83.58 %	80.13%	79.23 %	79.42%
	Weighted Hard Voting	**83.77%**	**80.06 %**	**79.36%**	**79.43%**
	Stacking (SVM)	79.82 %	76.98 %	76.60 %	75.46 %
	Stacking (Logistic)	79.48%	75.34 %	76.12 %	75.27 %
	Bagging	82.30 %	81.93%	82.30%	81.62%
	Boosting	82.26%	81.83%	82.26 %	81.78 %

Voting methods: Both Majority Soft Voting and Majority Hard Voting achieved a similar performance, with an accuracy of approximately 83.46% Table 4.15. These methods demonstrated balanced precision, recall, and F1 scores, indicating their effectiveness in ensemble predictions.

Weighted voting methods: Among Weighted Voting methods, Weighted Hard Voting emerged as the top performer on the OrganSMNIST dataset, achieving an accuracy of approximately 83.77% Table 4.15. This method exhibited improved precision, recall, and F1 scores compared to other voting methods, indicating enhanced performance in classifying OrganSMNIST dataset.

Stacking methods: Stacking with SVM and Stacking with Logistic Regression yielded accuracies around 79.82% and 79.48%, respectively Table 4.15. While these methods exhibited competitive performance, their precision, recall, and F1 scores were slightly lower compared to Weighted Voting methods.

Bagging and Boosting: Both Bagging and Boosting methods exhibited robust performance on the OrganSMNIST dataset, with accuracies ranging from 82.26% to 82.30% Table 4.15. These methods achieved balanced precision, recall, and F1 scores, demonstrating their effectiveness in ensemble predictions.

Summary: The ensemble method displayed varied effectiveness in enhancing classification performance on the OrganSMNIST dataset Table 4.15. Weighted Hard Voting emerged as the top performer among Weighted Voting methods, surpassing other voting methods in accuracy. Bagging and Boosting methods also exhibited robust performance, demonstrating balanced precision, recall, and F1 scores.The choice of ensemble method should consider the dataset characteristics and the desired balance between accuracy and computational complexity. Weighted Hard Voting, Bagging, and Boosting provide effective approaches to improve classification performance on the OrganSMNIST dataset.

4.3.4.8 OrganAMNIST dataset

Table 4.16: Enemble method results on OrganAMNIST dataset

Dataset	Ensemble methods	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Organa	Soft Voting	94.85%	95.49 %	94.57 %	94.91 %
	Weighted Soft Voting	**95.03%**	**95.56%**	**94.83%**	**95.10%**
	Hard Voting	94.86%	95.50 %	94.55%	94.90%
	Weighted Hard Voting	**95.03%**	**95.56%**	**94.83%**	**95.10%**
	Stacking (SVM)	92.50 %	92.66%	92.30 %	92.07%
	Stacking (Logistic)	93.80%	93.96 %	93.52%	93.46%
	Bagging	93.89%	94.27 %	93.89 %	93.89%
	Boosting	94.29 %	94.55 %	94.29 %	94.29%

Voting methods: Both soft voting and hard voting achieved a similar performance Table 4.16, with accuracies of approximately 94.85% and 94.86%, respectively. These methods demonstrated balanced precision, recall, and F1 scores, indicating their effectiveness in ensemble predictions.

Weighted voting methods: Both Weighted Voting methods, emerged as the top performer on the OrganAMNIST dataset, achieving an accuracy of 95.03% Table 4.16. This method exhibited improved precision, recall, and F1 scores compared to other voting methods. The results indicate that Weighted Voting has a slight edge in classifying OrganAMNIST dataset.

Stacking methods: Stacking with SVM and Stacking with Logistic Regression yielded accuracies of around 92.50% and 93.80%, respectively Table 4.16. While these methods exhibited competitive performance, their precision, recall, and F1 scores were slightly lower compared to the Weighted Voting methods, suggesting that stacking might not be as effective for this dataset.

Bagging and Boosting: Bagging and boosting showed moderate performance with an accuracy of 93.89% and 94.29 resepcively and high precision and recall values Table 4.16, reflecting their robustness and ability to reduce variance by combining multiple models .

Summary: The ensemble methods displayed varied effectiveness in enhancing classification performance on the OrganAMNIST dataset Table 4.16. Weighted Voting emerged as the top performers among ensemble methods, surpassing other voting methods in accuracy. Bagging also exhibited robust performance, demonstrating balanced precision, recall, and F1 scores.the choice of ensemble method should consider the dataset characteristics and the desired balance between accuracy and computational complexity. Weighted Soft Voting, along with Bagging, provides effective approaches to improve classification performance on the OrganAMNIST dataset. Further exploration and fine-tuning of ensemble methods, particularly Weighted Voting strategies, hold promise for achieving even higher accuracy and robust generalization across diverse medical imaging datasets.

4.3.5 Ablation study

We conducted an ablation study to systematically analyze how different combinations of models impact the performance of our ensemble methods in medical image classification tasks. Our goal was to gain findings into how the choice of base models influences the effectiveness of ensemble method. Ensemble techniques like Hard voting, Soft voting, Bagging, Boosting, and Stacking aim to boost predictive performance by merging the strengths of multiple models. However, the composition of the ensemble—specifically, the number of base models and their combinations—can significantly affect overall performance. Through our ablation study, we sought to pinpoint the optimal model combinations and ensemble setups that maximize performance metrics for each ensemble method and dataset. This approach allowed us to explore potential interactions and complementarity between different base models. By varying base model combinations systematically, we assessed whether specific combinations of vision transformer architectures (ViT, DEiT, BEiT, Swin) provided synergies or diverse perspectives, leading to improved ensemble performance. The ablation study was pivotal, providing valuable findings into optimal ensemble compositions and configurations for maximizing performance in medical image classification tasks using vision transformer models.

Our experimental setup involved exploring all possible combinations of base models (ViT, DEiT, BEiT, Swin) for each ensemble method including voting methods (hard and soft), weighted voting (hard and soft), Bagging, Boosting, stacking with SVM and logistic regression as meta-models. We considered ensemble sizes ranging from 2 to 4 models, covering all possible combinations of base model architectures. This exhaustive process was repeated across various medical imaging datasets (Chest, BloodMNIST, DermaMNIST, OrganCMNIST , OrganAMNIST , OrganSMNIST , BreastMNIST, PneumoniaMNIST), ensuring a thorough evaluation of ensemble methods and model combinations across diverse medical image classification tasks.

4.3.5.1 Experiments with hard voting ensemble

In the following paragraphs, we will conduct a detailed analysis of the results from our proposed hard voting ensemble method implementation for each dataset.

Chest dataset

Table 4.17: Hard voting results en Chest dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	ViT, DEiT, BEiT, Swin	96.74%	96.14%	95.98%	96.06%
	ViT, DEiT, BEiT	96.22%	95.45%	95.45%	95.45%
	ViT, DEiT, Swin	96.91%	96.27%	96.27%	96.27%
	ViT, BEiT, Swin	96.74%	96.29%	95.81%	96.00%
	DEiT, BEiT, Swin	96.39%	96.03%	95.23%	95.61%
	ViT, DEiT	**97.42%**	**96.82%**	**96.98%**	**96.90%**
	ViT, BEiT	95.36%	94.48%	94.32%	94.40%
	ViT, Swin	**97.42%**	**96.82%**	**96.98%**	**96.90%**
	DEiT, BEiT	96.05%	95.17%	95.32%	95.25%
	DEiT, Swin	96.74%	96.45%	95.64%	93.30%
	BEiT, Swin	95.19%	95.49%	92.84%	94.03%

The combinations of (ViT, Swin) and (ViT, DEiT) achieved the highest performance with an accuracy of 97.42%, precision of 96.82%, recall of 96.98%, and an F1 score of 96.90% Table 4.17. This indicates exceptional synergy and robustness between these two combinations of models. Other combinations like (ViT, DEiT, Swin) and (ViT, BEiT, Swin) showed strong performance with accuracies of 96.91% and 96.74%, respectively, demonstrating their effectiveness but increasing more complexity without increasing performance compared to using two models. Combinations involving all four models, as well as different subsets, generally maintained high accuracy around 96-97%, but precision, recall, and F1 scores varied, reflecting differences in their ability to balance true and false positives. Notably, combinations like (ViT, BEiT) and (BEiT,Swin) had relatively lower accuracy and F1 scores, indicating they were less effective and does not fit each other in this dataset. The study highlights that specific model combinations, particularly those (ViT, Swin) and (DEiT, ViT) lead to the highest performance on the Chest dataset, emphasizing the importance of selecting and chooseing models to achieve optimal results.

DermaMNIST dataset

Table 4.18: Hard voting results on DermaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	ViT, DEiT, BEiT, Swin	78.85%	66.63%	57.55%	59.86%
	ViT, DEiT, BEiT	78.35%	67.60%	55.74%	58.70%
	ViT, DEiT, Swin	79.55%	67.96%	59.64%	62.36%
	ViT, BEiT, Swin	78.05%	66.45%	55.31%	57.56%
	DEiT, BEiT, Swin	79.25%	66.74%	59.12%	61.34%
	ViT, DEiT	78.55%	65.12%	54.99%	58.22%
	ViT, BEiT	76.86%	64.98%	54.40%	56.55%
	ViT, Swin	78.55%	67.50%	57.58%	60.30%
	DEiT, BEiT	78.80%	64.60%	57.78%	59.37%
	DEiT, Swin	**80.10%**	**67.62%**	**61.96%**	**64.12%**
	BEiT, Swin	78.45%	64.91%	60.20%	61.01%

The combination of DEiT and Swin achieved the highest overall performance on the DermaMNIST dataset, boasting an accuracy of 80.10%, precision of 67.62%, recall of 61.96%, and an F1 score of 64.12% Table 4.18. This combination demonstrates robust synergy and effectiveness in capturing the dataset's complexities, resulting in superior classification outcomes. Both (ViT, DEiT, Swin) and (DEiT, BEiT, Swin) also performed commendably with accuracies of 79.55% and 79.25%, respectively. These configurations exhibit strong performance but slightly lower than the DEiT and Swin ensemble, indicating the critical role of Swin in enhancing model synergy and overall accuracy. The four-model ensemble (ViT, DEiT, BEiT, Swin) and other three-model combinations consistently delivered good performance, achieving accuracy values around 78-79%. However, variations in precision, recall, and F1 scores suggest differences in their ability to balance false positives and true positives, influenced by the complexity introduced by incorporating multiple models. Conversely, dual combinations such as (ViT, DEiT) and (ViT, BEiT) showed lower accuracy and F1 scores, suggesting reduced effectiveness in capturing the dataset's intricate patterns compared to their three- and four-model counterparts. This study underscores the significance of model selection and combination strategies in achieving optimal results on the DermaMNIST dataset. While most ensemble configurations perform reasonably well, the specific synergy observed between DEiT and Swin highlights their pivotal role in maximizing classification accuracy and efficiency.

BloodMNIST dataset

Table 4.19: Hard voting results on BloodMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	ViT, DEiT, BEiT, Swin	97.78%	97.82%	97.62%	97.72%
	ViT, DEiT, BEiT	97.72%	97.73%	97.55%	97.63%
	ViT, DEiT, Swin	**97.87%**	**97.94%**	**97.75%**	**97.84%**
	ViT, BEiT, Swin	97.78%	97.77%	97.65%	97.70%
	DEiT, BEiT, Swin	97.40%	97.43%	97.16%	97.29%
	ViT, DEiT	97.84%	97.77%	97.83%	97.80%
	ViT, BEiT	97.81%	97.72%	97.59%	97.65%
	ViT, Swin	97.84%	97.75%	97.75%	97.75%
	DEiT, BEiT	97.57%	97.60%	97.24%	97.41%
	DEiT, Swin	97.54%	97.47%	97.30%	97.37%
	BEiT, Swin	96.93%	96.94%	96.54%	96.71%

The combination of (ViT, DEiT, Swin) achieved the highest overall performance with an accuracy of 97.87%, precision of 97.94%, recall of 97.75%, and an F1 score of 97.84%, indicating excellent synergy and robustness Table 4.19. Close behind, the four-model ensemble (ViT, DEiT, BEiT, Swin) and the combination of (ViT, BEiT, Swin) both scored 97.78% in accuracy but they may add more complexity than the top performer, with minor variations in precision, recall, and F1 scores. Dual combinations also performed exceptionally well, with (ViT, DEiT) reaching an accuracy of 97.84% and (ViT, Swin) matching this accuracy, demonstrating their robustness and decreasing the complexity of the ensemble. However, the combination of (BEiT, Swin) had the lowest performance, with an accuracy of 96.93%, indicating a less effective synergy between these two models alone. The precision and recall values were consistently high across most combinations, reflecting a balanced performance in identifying true positives and minimizing false positives.

BreastMNIST dataset

Table 4.20: Hard voting results on BreastMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT, DEiT, BEiT, Swin	86.54%	88.75%	76.50%	80.04%
	ViT, DEiT, BEiT	85.90%	88.26%	75.31%	78.86%
	ViT, DEiT, Swin	**87.82%**	**88.53%**	**79.64%**	**82.68%**
	ViT, BEiT, Swin	87.18%	89.23%	77.69%	81.20%
	DEiT, BEiT, Swin	83.97%	83.98%	73.25%	76.24%
	ViT, DEiT	87.18%	88.00%	78.45%	81.58%
	ViT, BEiT	84.62%	89.04%	72.18%	75.85%
	ViT, Swin	86.54%	85.54%	78.76%	81.22%
	DEiT, BEiT	81.41%	81.22%	68.48%	71.16%
	DEiT, Swin	85.90%	83.56%	79.07%	80.87%
	BEiT, Swin	84.62%	84.62%	74.44%	77.43%

The combination of (ViT, DEiT, Swin) achieved the highest overall performance, with an accuracy of 87.82%, precision of 88.53%, recall of 79.64%, and an F1 score

of 82.68%, demonstrating strong synergy and robustness Table 4.20. Following closely, the combination of (ViT, BEiT, Swin) also performed well, achieving an accuracy of 87.18%, precision of 89.23%, recall of 77.69%, and an F1 score of 81.20%. While the four-model combination of (ViT, DEiT, BEiT, Swin) showed slightly lower performance with an accuracy of 86.54%, it still maintained high precision at 88.75%, though with a lower recall of 76.50%, indicating a trade-off between precision and recall. Combinations without Swin, such as (ViT, DEiT, BEiT) had slightly lower accuracy and F1 scores, suggesting that Swin's inclusion enhances performance. Notably, the combination of (DEiT, BEiT) showed the lowest performance with an accuracy of 81.41% and an F1 score of 71.16%, indicating less effective synergy between these two models alone. Precision and recall values varied significantly, reflecting differences in the ensemble' abilities to identify true positives and minimize false positives. The study underscores the importance of model selection and combination, highlighting that the inclusion of Swin generally boosts performance on the BreastMNIST dataset.

OrganCMNIST dataset

Table 4.21: Hard voting results on OrganCMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	ViT, DEiT, BEiT, Swin	93.42%	93.13%	92.17%	92.58%
	ViT, DEiT, BEiT	93.54%	93.16%	92.43%	92.73%
	ViT, DEiT, Swin	93.25%	93.06%	92.00%	92.44%
	ViT, BEiT, Swin	93.43%	93.18%	92.15%	92.59%
	DEiT, BEiT, Swin	93.18%	93.05%	91.91%	92.39%
	ViT, DEiT	93.20%	92.73%	92.09%	92.35%
	ViT, BEiT	**93.60%**	**93.18%**	**92.48%**	**92.76%**
	ViT, Swin	92.99%	92.83%	91.49%	92.05%
	DEiT, BEiT	93.18%	92.87%	92.01%	92.37%
	DEiT, Swin	92.74%	92.69%	91.40%	91.93%
	BEiT, Swin	92.82%	92.76%	91.43%	91.99%

The combination of (ViT, BEiT) emerged as the top performer with the highest accuracy of 93.60%, precision of 93.18%, recall of 92.48%, and F1 score of 92.76%, indicating strong synergy between these two models for OrganCMNIST dataset Table 4.21. The three-model combinations of (ViT, DEiT, BEiT) and (ViT, BEiT, Swin) also performed well, achieving accuracy scores of 93.54% and 93.43%, respectively, showcasing their effectiveness in leveraging multiple models. However, adding Swin generally resulted in slightly lower performance metrics compared to the combinations without it, suggesting that Swin may not significantly enhance the ensemble's performance on this dataset. Dual-model combinations involving DEiT, such as (DEiT, BEiT), and (ViT, DEiT) displayed robust results, with accuracies of 93.18% and 93.20%, respectively, emphasizing DEiT's substantial contribution. Precision and recall values were closely aligned across all combinations, reflecting balanced and effective performance in identifying true positives and minimizing false positives. The study highlights the notable impact of BEiT in enhancing ensemble performance and the importance of selecting complementary models to maximize accuracy and robustness on the OrganCMNIST dataset.

OrganSMNIST dataset

Table 4.22: Hard voting results on OrganSMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	ViT, DEiT, BEiT, Swin	**83.58%**	**80.13%**	**79.23%**	**79.42%**
	ViT, DEiT, BEiT	**83.58%**	**80.21%**	**79.30%**	**79.19%**
	ViT, DEiT, Swin	83.35%	79.97%	79.05%	79.11%
	ViT, BEiT, Swin	83.52%	80.23%	79.19%	79.56%
	DEiT, BEiT, Swin	83.34%	79.86%	79.10%	79.40%
	ViT, DEiT	82.61%	79.29%	78.27%	77.80%
	ViT, BEiT	83.40%	80.14%	79.06%	79.11%
	ViT, Swin	82.85%	79.59%	78.32%	78.67%
	DEiT, BEiT	83.04%	79.44%	78.82%	78.90%
	DEiT, Swin	83.10%	79.77%	79.06%	79.29%
	BEiT, Swin	82.81%	79.66%	78.64%	79.06%

The combinations of (ViT, DEiT, BEiT, Swin), and (ViT, DEiT, BEiT) achieved the highest metrics, each with 83.58% accuracy, over 80.13% precision, around 79.23% recall, and an F1 score of approximately 79.42%, demonstrating their strong synergy Table 4.22. Slightly lower performance was observed with the combination of ViT, DEiT, Swin, which scored 83.35% accuracy, indicating a minor decline when Swin is included instead of BEiT. Notably, the three-model combinations consistently outperformed the two-model pairs, such as (ViT, DEiT), which had an accuracy of 82.61%, suggesting that more diverse ensembles enhance performance. The dual combinations involving BEiT showed relatively strong results, such as (ViT, BEiT) with 83.40% accuracy, emphasizing BEiT's significant contribution. Generally, precision and recall values across combinations were balanced, reflecting effective identification of true positives and minimal false positives. The study underscores the importance of model diversity and the notable impact of BEiT in enhancing ensemble performance on the OrganSMNIST dataset.

OrganAMNIST dataset

Table 4.23: Hard voting results on OrganAMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganAMNIST	ViT, DEiT, BEiT, Swin	94.86%	95.49%	94.57%	94.91%
	ViT, DEiT, BEiT	**95.08%**	**95.65%**	**94.87%**	**95.16%**
	ViT, DEiT, Swin	94.80%	95.39%	94.46%	94.81%
	ViT, BEiT, Swin	94.32%	95.04%	93.93%	94.33%
	DEiT, BEiT, Swin	94.65%	95.27%	94.29%	94.65%
	ViT, DEiT	94.63%	95.15%	94.38%	94.66%
	ViT, BEiT	94.32%	95.09%	94.12%	94.47%
	ViT, Swin	94.07%	94.07%	93.65%	94.05%
	DEiT, BEiT	95.01%	95.60%	94.96%	95.20%
	DEiT, Swin	94.76%	95.32%	94.51%	94.83%
	BEiT, Swin	94.09%	94.78%	93.84%	94.14%

The top-performing combination, ViT, DEiT, and BEiT, achieved the highest metrics with 95.08% accuracy, 95.65% precision, 94.87% recall, and a 95.16% F1 score, indicating

strong complementary effects among these models Table 4.23. Including Swin generally resulted in slightly lower performance, suggesting it may not significantly enhance the ensemble when paired with (ViT, DEiT). Among the dual-model combinations, (DEiT, BEiT) stood out with impressive results, nearly matching the best three-model combination, highlighting that a well-chosen pair can be highly effective. ViT's consistent inclusion in high-performing combinations underscores its robustness, while DEiT's presence appears crucial for achieving top-tier accuracy and stability. Precision and recall values were closely aligned across combinations, reflecting balanced performance in identifying true positives and minimizing false positives.

PneumoniaMNIST dataset

Table 4.24: Hard voting results on PneumoniaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoniaMNIST	ViT, DEiT, BEiT, Swin	93.59%	94.61%	91.88%	92.96%
	ViT, DEiT, BEiT	92.63%	93.93%	90.60%	91.86%
	ViT, DEiT, Swin	92.63%	94.07%	90.51%	91.84%
	ViT, BEiT, Swin	**94.07 %**	**95.09%**	**92.44%**	**93.50%**
	DEiT, BEiT, Swin	91.51%	92.84%	89.27%	90.59%
	ViT, DEiT	91.99%	93.19%	89.91%	91.15%
	ViT, BEiT	93.59%	94.48%	91.97%	92.98%
	ViT, Swin	**94.07 %**	**95.22%**	**92.35%**	**93.49%**
	DEiT, BEiT	87.02%	87.00%	85.00%	85.80%
	DEiT, Swin	91.35%	92.73%	89.06%	90.40%
	BEiT, Swin	93.43%	94.49%	91.67%	92.78%

The combination of (ViT, Swin) achieved the highest performance with an accuracy of 94.07%, precision of 95.09%, recall of 92.44%, and an F1 score of 93.50%, indicating a robust balance between identifying true positives and minimizing false positives (tableTable 4.24. The combination of (ViT,BEiT, Swin) also performed well, matching the top accuracy of (94.07%) and showing strong precision (95.22%), recall (92.35%), and F1 score (93.49%) however it increase complexity by adding another model to the combination of (ViT, Swin) without increasing accuracy. The four-model ensemble (ViT, DEiT, BEiT, Swin) also demonstrated high performance with an accuracy of 93.59% and an F1 score of 92.96%, suggesting that including more models generally maintains high effectiveness for PneumoniaMNIST dataset. However, combinations like (DEiT, BEiT, Swin), and (DEiT, BEiT) showed lower accuracy and F1 scores, highlighting that certain pairings were less effective. The results underscore that strategic combinations of (ViT, BEiT, Swin) yield the highest performance on the PneumoniaMNIST dataset, emphasizing the value of careful model selection and combination to optimize diagnostic accuracy.

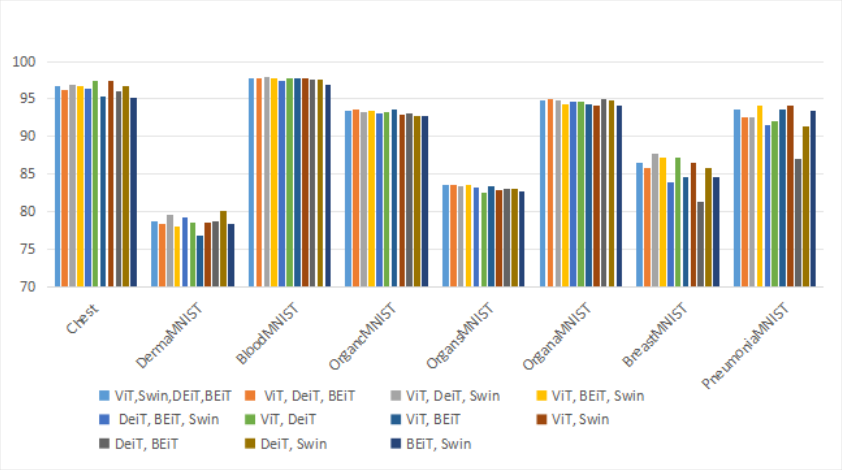


Figure 4.1: Illustration of Hard Voting ensemble method results results in terms of accuracy

Summary: Across various datasets as also illustrated in Figure 4.1, the inclusion of different models in hard voting ensemble method significantly impacts classification accuracy. While certain combinations consistently outperform others, the complementary nature of model architectures plays a crucial role in enhancing ensemble performance. ViT and DEiT often exhibit strong synergy, while Swin proves effective in improving ensemble predictions across different datasets. However, the contribution of BEiT varies, with its effectiveness influenced by the specific dataset and combination of models. These findings provide valuable findings for optimizing model combinations and improving classification accuracy in medical imaging applications using hard voting ensemble method.

4.3.5.2 Experiments with weighted hard voting ensemble

In the following paragraphs, we will conduct a detailed analysis of the results from our stacking with logistic ensemble method implementation for each dataset.

Chest Dataset

Table 4.25: Weighted hard voting results on Chest dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	ViT, DEiT, BEiT, Swin	96.91%	96.58%	95.93%	96.25%
	ViT, DEiT, BEiT	96.91%	96.00%	96.62%	96.30%
	ViT, DEiT, Swin	97.25%	96.84%	96.52%	96.68%
	ViT, BEiT, Swin	97.25%	96.84%	96.52%	96.68%
	DEiT, BEiT, Swin	97.08%	96.71%	96.22%	96.46%
	ViT, DEiT	96.74%	95.86%	96.32%	96.09%
	ViT, BEiT	96.74%	95.86%	96.32%	96.09%
	ViT, Swin	**97.42%**	**96.82%**	**96.98%**	**96.90%**
	DEiT, BEiT	96.05%	95.31%	95.15%	95.23%
	DEiT, Swin	97.08%	96.87%	96.05%	96.45%
	BEiT, Swin	95.88%	95.99%	94.01%	94.92%

The combination of (ViT, Swin) demonstrated the highest overall performance, achieving an accuracy of 97.42%, precision of 96.82%, recall of 96.98%, and an F1 score of

96.90%, highlighting its superior capability in balancing true positive identification and minimizing false positives Table 4.25. Other high-performing combinations include (ViT, DEiT, Swin) as well as (ViT, BEiT, Swin) both with an accuracy of 97.25% and showing strong precision (96.84%), recall (96.52%), and F1 score (96.68%). These results indicate that incorporating Swin into the model ensembles generally enhances performance. However, combinations like (DEiT, BEiT) and (BEiT, Swin) showed comparatively lower accuracy and F1 scores, underscoring the variability in effectiveness depending on the models used. The study underscores that strategic combinations, particularly involving Swin, significantly enhance diagnostic performance on the Chest dataset on weighted hard voting, emphasizing the importance of selecting and combining models effectively to optimize medical imaging accuracy.

DermaMNIST Dataset

Table 4.26: Weighted hard voting results on DermaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	ViT, DEiT, BEiT, Swin	79.70%	66.84%	59.60%	62.11%
	ViT, DEiT, BEiT	79.35%	67.34%	58.80%	61.81%
	ViT, DEiT, Swin	80.15%	68.78%	60.15%	63.38%
	ViT, BEiT, Swin	79.20%	66.99%	60.29%	62.29%
	DEiT, BEiT, Swin	80.05%	67.63%	60.85%	63.29%
	ViT, DEiT	79.55%	66.54%	58.90%	61.71%
	ViT, BEiT	77.31%	69.13%	52.75%	56.67%
	ViT, Swin	79.10%	65.98%	60.30%	62.06%
	DEiT, BEiT	79.25%	66.81%	59.37%	61.87%
	DEiT, Swin	**80.30%**	**67.90%**	**61.25%**	**63.75%**
	BEiT, Swin	79.55%	66.54%	58.90%	61.71%

The highest accuracy and F1 score were achieved by the combination of (DEiT, Swin) with 80.30% and 63.75%, respectively, indicating that these two models complement each other effectively, resulting in superior performance. This combination excels in precision and recall, suggesting a balanced and robust performance across different metrics. The combinations of (DEiT, BEiT, Swin) and (ViT, DEiT, Swin) also showed strong results, with accuracy and F1 scores close to the highest, reinforcing the effectiveness of including (Swin, DEiT) in the ensemble. In contrast, combinations excluding Swin, such as (ViT, DEiT, BEiT) and (ViT, DEiT) showed slightly lower performance metrics, with F1 scores around 62.11% and 61.81%, respectively. This suggests that while these combinations still perform well, the absence of Swin results in a minor decline in performance. The combination of (ViT, BEiT) showed the lowest performance metrics, particularly in recall and F1 score, indicating potential limitations in handling the dataset's complexities. The analysis indicates that the DermaMNIST dataset poses significant challenges, leading to moderate performance metrics across all combinations. The overall results suggest that while transformer-based models can handle the dataset, their effectiveness varies based on the specific combination used. The combination of DEiT and Swin stands out as the most effective, highlighting the importance of selecting complementary models to maximize performance. This underscores the potential benefits of using ensemble methods to balance the strengths and weaknesses of individual models, particularly when dealing with complex datasets like DermaMNIST.

BloodMNIST Dataset

Table 4.27: Weighted hard voting results on BloodMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	ViT, DEiT, BEiT, Swin	97.95%	97.97%	97.82%	97.89%
	ViT, DEiT, BEiT	97.90%	97.82%	97.75%	97.78%
	ViT, DEiT, Swin	**98.04%**	**98.04%**	**97.97%**	**97.99%**
	ViT, BEiT, Swin	97.90%	97.82%	97.79%	97.80%
	DEiT, BEiT, Swin	97.63%	97.64%	97.46%	97.54%
	ViT, DEiT	97.92%	97.76%	97.91%	97.83%
	ViT, BEiT	**98.04%**	**97.96%**	**97.96%**	**97.95%**
	ViT, Swin	97.95%	97.90%	97.88%	97.88%
	DEiT, BEiT	97.60%	97.59%	97.28%	97.42%
	DEiT, Swin	97.52%	97.48%	97.26%	97.36%
	BEiT, Swin	96.96%	96.95%	96.54%	97.72%

The combination of (ViT, DEiT, Swin) yielded the highest accuracy and F1 score at 98.04% and 97.99%, respectively, indicating that the inclusion of Swin alongside ViT and DEiT leverages their complementary strengths, resulting in superior performance Table 4.27. Similarly, the combination of (ViT, BEiT) also achieved the highest accuracy and notable precision and recall , showcasing the robustness of these two models when used together but it has less complexity compared tp (ViT, DEiT, Swin) by having just two models and addressing the high accuracy instead of three models. In contrast, combinations that exclude ViT, such as DEiT, BEiT, Swin, tend to perform lower, highlighting ViT's significant contribution to high performance. Dual combinations like (ViT,DEiT) or (ViT, BEiT) consistently result in high accuracy and balanced performance metrics, suggesting that ViT significantly enhances the predictive power of the ensemble. Conversely, the combination of BEiT and Swin shows the lowest performance metrics, indicating potential issues in how these models' predictions complement each other in the weighted voting scheme. Other lower-performing combinations include (DEiT, Swin) and (DEiT, BEiT) suggesting that the inclusion of ViT generally enhances model performance. The results underline the effectiveness of the weighted hard voting approach, which balances the strengths and weaknesses of individual models, leading to high overall performance. The high performance across different combinations also suggests that the BloodMNIST dataset is well-suited for transformer-based models, allowing them to capture its underlying patterns effectively. The analysis highlights the importance of selecting complementary models to enhance predictive accuracy and overall performance in ensemble methods, with ViT playing a pivotal role in achieving top results.

BreastMNIST Dataset

Table 4.28: Weighted hard voting results on BreastMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT, DEiT, BEiT, Swin	88.46%	90.18%	80.08%	83.42%
	ViT, DEiT, BEiT	87.82%	89.71%	78.88%	82.32%
	ViT, DEiT, Swin	**89.10%**	**90.65%**	**81.27%**	**84.50%**

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT, BEiT, Swin	88.46%	90.18%	80.08%	83.42%
	DEiT, BEiT, Swin	84.62%	81.51%	75.50%	77.65%
	ViT, DEiT	87.82%	89.71%	78.88%	82.32%
	ViT, BEiT	88.46%	89.89%	74.56%	78.38%
	ViT, Swin	88.46%	89.05%	80.83%	83.75%
	DEiT, BEiT	83.97%	83.98%	73.25%	76.24%
	DEiT, Swin	85.90%	84.17 %	78.32 %	80.51%
	BEiT, Swin	84.62%	82.86%	75.94 %	78.33 %

The combination of (ViT, DEiT, Swin) achieved the highest performance metrics with an accuracy of 89.10% and an F1 score of 84.50% Table 4.28. This indicates that these models complement each other well, leading to robust performance across all metrics. The high precision and recall scores suggest that this combination is particularly effective at identifying positive instances and minimizing false negatives. Other combinations, such as (ViT, DEiT, BEiT) and (ViT, BEiT, Swin) also showed strong results with an accuracy of 88.46% and an F1 score of 83.42%. These combinations, while slightly lower than the top performer, still demonstrate a balanced and effective approach, suggesting that ViT plays a crucial role in maintaining high performance. On the lower end, the combinations of(DEiT, BEiT, Swin) and (DEiT, BEiT) showed reduced performance metrics, with accuracies of 84.62% and 83.97%, and F1 scores of 77.65% and 76.24%, respectively. These results indicate that the exclusion of ViT impacts the overall effectiveness, highlighting ViT's contribution to the ensemble's success. Interestingly, the combination of (DEiT, Swin) alone performed better than (DEiT, BEiT), with an accuracy of 85.90% and an F1 score of 80.51%. This suggests that Swin's inclusion is beneficial, but not as impactful as when paired with ViT. The analysis underscores that while all models can handle the BreastMNIST dataset, their effectiveness varies significantly based on the combination used. The combination of (ViT, DEiT, Swin) stands out as the most effective, underscoring the importance of selecting complementary models to enhance performance. These results highlight the benefits of using ensemble methods to leverage the strengths of individual models, particularly in handling the complexities and small size of datasets like BreastMNIST.

OrganCMNIST Dataset

Table 4.29: Weighted hard voting results on OrganCMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	ViT, DEiT, BEiT, Swin	93.67%	93.30%	92.54%	92.86%
	ViT, DEiT, BEiT	93.65%	93.25%	92.59%	92.85%
	ViT, DEiT, Swin	93.60%	93.27%	92.47%	92.80%
	ViT, BEiT, Swin	**93.75%**	**93.40%**	**92.58%**	**92.92%**
	DEiT, BEiT, Swin	93.24%	92.90%	92.09%	92.43%
	ViT, DEiT	93.38%	92.90%	92.26%	92.51%
	ViT, BEiT	93.59%	93.17%	92.49%	92.76%
	ViT, Swin	93.21%	92.95%	91.86%	92.32%
	DEiT, BEiT	93.21%	92.90%	92.07%	92.42%
	DEiT, Swin	92.91%	92.78%	91.62%	92.11%

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	BEiT, Swin	92.99%	92.92%	91.61%	92.16%

The combination of (ViT, BEiT, Swin) achieved the highest performance with an accuracy of 93.75% and an F1 score of 92.92% Table 4.29. This indicates that these models, when combined, provide the most robust performance, highlighting their complementary strengths. Other combinations also performed well, such as (ViT, DEiT, BEiT) which showed an accuracy of 93.65% and an F1 score of 92.85%, and (ViT, DEiT, Swin) with an accuracy of 93.60% and an F1 score of 92.80%. These results suggest that ViT plays a crucial role in achieving high performance, as it appears in all top-performing combinations. The combination of (DEiT, BEiT, Swin) and (ViT, DEiT) showed slightly lower performance with accuracies of 93.24% and 93.38% and F1 scores of 92.43% and 92.51%, respectively. While still strong, these results indicate that excluding ViT or pairing it with only one other model slightly reduces the overall performance. On the lower end, combinations like (DEiT, Swin) and (BEiT, Swin) had accuracies of 92.91% and 92.99%, and F1 scores of 92.11% and 92.16%, respectively. These results show that while these combinations are still effective and have less complexity, they do not reach the same level of performance as those including ViT. The analysis highlights the effectiveness of using ensemble methods to enhance performance on the OrganCMNIST dataset. The combination of ViT, BEiT, and Swin stands out as the most effective, suggesting that these models work well together to handle the dataset's complexities. This underscores the importance of selecting complementary models in ensemble methods to achieve the best possible performance.

OrganSMNIST Dataset

Table 4.30: Weighted hard voting results on OrganSMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	ViT, DEiT, BEiT, Swin	**83.77%**	**80.39%**	**79.60%**	**79.63%**
	ViT, DEiT, BEiT	83.67%	80.29%	79.44%	79.57%
	ViT, DEiT, Swin	83.28%	79.85%	78.72%	78.84%
	ViT, BEiT, Swin	83.35%	79.08%	79.56%	79.56%
	DEiT, BEiT, Swin	83.42%	80.23%	79.36%	79.36%
	ViT, DEiT	82.51%	79.45%	78.08%	77.61%
	ViT, BEiT	83.52%	80.20%	79.25%	79.39%
	ViT, Swin	82.94%	79.75%	78.46%	78.56%
	DEiT, BEiT	83.24%	79.79%	79.08%	79.32%
	DEiT, Swin	82.90%	79.37%	78.74%	78.89%
	BEiT, Swin	83.00%	79.90%	78.93%	79.34%

The combination of (ViT, DEiT, BEiT, Swin) achieved the highest performance, with an accuracy of 83.77%, a precision of 80.39%, a recall of 79.60%, and an F1 score of 79.63% Table 4.30. This combination demonstrates the strongest performance metrics, indicating that these models complement each other well in handling the complexities of the OrganSMNIST dataset but it added complexity to the ensemble by using the four models compared to other combinations . Other high-performing combinations include (ViT, DEiT, BEiT) which achieved an accuracy of 83.67%, precision of 80.29%, recall of

79.44%, and F1 score of 79.57%, and (ViT, BEiT, Swin) with an accuracy of 83.35%, precision of 79.08%, recall of 79.56%, and F1 score of 79.56%. These results show that combinations including ViT generally perform better, underscoring its pivotal role in achieving high performance. The combination of (DEiT, BEiT, Swin) also performed well with an accuracy of 83.42%, precision of 80.23%, recall of 79.36%, and F1 score of 79.36%. While still effective, this combination shows slightly lower performance metrics compared to those including ViT, highlighting the added value of incorporating ViT in the ensemble. On the lower end, combinations like (ViT, DEiT) and (ViT, Swin) had accuracies of 82.51% and 82.94%, respectively, with corresponding F1 scores of 77.61% and 78.56%. These results indicate that excluding certain models or pairing ViT with fewer models reduces overall performance. The analysis suggests that using ensemble methods, particularly those including ViT, enhances performance on the OrganSMNIST dataset. The highest performing combination of (ViT, DEiT, BEiT,Swin) underscores the importance of leveraging diverse model strengths to address the dataset's challenges effectively.

OrganAMNIST dataset

Table 4.31: Weighted hard voting results on OrganAMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganAMNIST	ViT, DEiT, BEiT, Swin	**95.03%**	**95.56%**	**94.83 %**	**95.10 %**
	ViT, DEiT, BEiT	95.02 %	95.58%	94.82%	95.11%
	ViT, DEiT, Swin	94.94%	95.49 %	94.64 %	94.98 %
	ViT, BEiT, Swin	95.02%	95.58%	94.82 %	95.11 %
	DEiT, BEiT, Swin	94.86%	95.39%	94.59%	94.90%
	ViT, DEiT	94.74%	95.24 %	94.48%	94.76%
	ViT, BEiT	94.74 %	95.24 %	94.48 %	94.76%
	ViT, Swin	94.11 %	94.82 %	93.71 %	94.11 %
	DEiT, BEiT	94.88%	95.40%	94.64%	94.93%
	DEiT, Swin	94.78%	95.32%	94.47 %	94.80%
	BEiT, Swin	94.00%	94.76%	93.60%	94.00 %

The combination of (ViT, DEiT, BEiT, Swin) achieved the highest performance metrics, with an accuracy of 95.03%, precision of 95.56%, recall of 94.83%, and an F1 score of 95.10% Table 4.31. This combination demonstrated the best balance between precision and recall, resulting in the highest F1 score among the tested combinations despite the complexity adds by using the four models. The combinations of (ViT, DEiT, BEiT) and (ViT, BEiT, Swin) also performed similarly well, with nearly identical accuracy, precision, recall, and F1 scores, indicating that these three-model combinations are nearly as effective as the four-model ensemble and having less complexity than the four models. In contrast, combinations excluding Swin, such as (ViT, DEiT, BEiT) showed slightly lower performance metrics, highlighting Swin's contribution to improving overall performance. The lowest performance was observed in combinations such as (ViT, Swin) and (BEiT, Swin) with accuracy, precision, recall, and F1 scores indicating a significant drop compared to the top combinations, underscoring the importance of including DEiT and BEiT in the ensemble. The study highlights that including multiple transformer models generally improves performance in OrganAMNIST dataset, with the combination of (ViT, DEiT, BEiT, Swin) yielding the best results on the OrganAMNIST dataset. However, if we consider complexity more than performance, we will choose the combination of (ViT, DEiT, BEiT) or (ViT, BEiT, Swin) because they have less complexity than four models.

PneumoniaMNIST Dataset

Table 4.32: Weighted hard voting results on PneumoniaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoniaMNIST	ViT, DEiT, BEiT, Swin	94.07%	95.09%	92.44%	93.50%
	ViT, DEiT, BEiT	93.27%	94.41%	91.15%	92.40%
	ViT, DEiT, Swin	94.07%	94.18%	90.73%	92.03%
	ViT, BEiT, Swin	**94.39%**	**95.19%**	**92.95%**	**93.88%**
	DEiT, BEiT, Swin	92.47%	93.56%	90.26%	91.50%
	ViT, DEiT	93.59%	94.74%	91.79%	92.95%
	ViT, BEiT	94.07%	94.11%	92.61%	93.53%
	ViT, Swin	94.07%	94.11%	92.61%	93.53%
	DEiT, BEiT	87.50%	87.10%	85.98%	86.47%
	DEiT, Swin	91.67%	93.26%	89.32%	90.73%
	BEiT, Swin	93.11%	94.41%	91.15%	92.40%

The combination of (ViT, BEiT, Swin) emerged as the best-performing trio Table 4.32, achieving the highest metrics. This indicates that this combination excels in identifying positive instances accurately while maintaining a strong balance between precision and recall. The combination of (ViT, DEiT, BEiT, Swin) also performed exceptionally well, with an accuracy of 94.07%, precision of 95.09%, recall of 92.44%, and F1 score of 93.50% This quartet shows that including all four models still yields high performance, underscoring the complementary strengths of each model however it adds more complexity to ensemble method. The performance of other combinations like (ViT, DEiT, BEiT) (accuracy of 93.27%) and (ViT, DEiT, Swin) (accuracy of 94.07%) highlights the importance of ViT in maintaining high performance, even with fewer models in the ensemble. Combinations excluding ViT, such as (DEiT, BEiT, Swin) and (DEiT, BEiT) show a noticeable drop in performance. This drop emphasizes ViT's crucial role in enhancing the ensemble's effectiveness. The analysis demonstrates that the combination of (ViT, BEiT,Swin) yields the best results on the PneumoniaMNIST dataset, indicating their strong complementary nature. Including ViT generally enhances performance, making it a vital component in achieving high accuracy and balanced precision-recall metrics. These findings suggest that leveraging diverse transformer models can significantly improve performance on medical imaging datasets like PneumoniaMNIST, highlighting the value of ensemble methods in clinical applications.

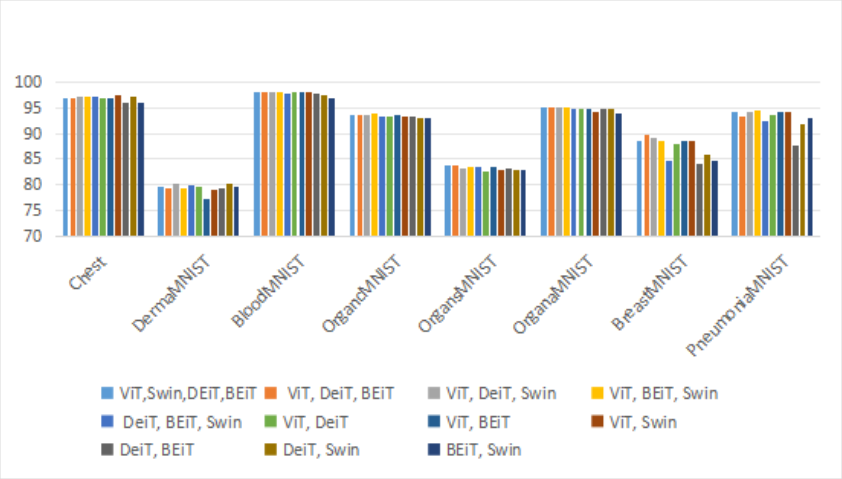


Figure 4.2: Illustration of weighted hard ensemble method results in terms of accuracy

Summary: The ablation study as also illustrated in Figure 4.2 of weighted hard voting across various medical image classification datasets reveals crucial findings into model interactions and their impact on classification performance. In the Chest dataset, the ViT and Swin combination achieves the highest accuracy of 97.42%, demonstrating their complementary strengths. Similar patterns are observed in other datasets, with DEiT and Swin achieving the highest accuracy of 80.30% for the DermaMNIST dataset and ViT, BEiT reaching 98.04% for the BloodMNIST dataset. In the BreastMNIST dataset, ViT, DEiT, and Swin emerge as top performers with 89.10% accuracy. For the OrganCMNIST dataset, the combinations of ViT, BEiT, and Swin yield superior results, with accuracies of 93.75% and 83.77% for the OrganSMNIST dataset. The OrganAMNIST dataset sees the highest accuracy of 95.14% from the ViT, DEiT, and BEiT combination. Lastly, in the PneumoniaMNIST dataset, the combination of ViT, BEiT, and Swin leads to the highest accuracy of 94.07%. These findings underscore the importance of selecting the right model combinations to exploit their synergistic effects effectively, highlighting how specific transformer models, when used together, can significantly enhance performance. However, it is also crucial to consider the size of the combination, as this affects the complexity of the ensemble method. If a combination does not significantly improve accuracy compared to another combination with fewer models, it may not be worth the added complexity. Adding more models can increase computational burden without proportionate gains in accuracy, emphasizing the need to balance accuracy improvement with model complexity.classification performance across diverse medical imaging tasks.

4.3.5.3 Experiments with soft voting ensemble

In the following paragraphs, we will conduct a detailed analysis of the results from our soft voting ensemble method implementation for each dataset.

Chest dataset

Table 4.33: Soft voting results on Chest dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	ViT, DEiT, BEiT, Swin	96.74%	96.14%	95.98%	96.06%

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	ViT, DEiT, BEiT	96.22%	95.45%	95.45%	95.45%
	ViT, DEiT, Swin	96.91%	96.27%	96.27%	96.27%
	ViT, BEiT, Swin	97.08%	96.55%	96.40%	96.47%
	DEiT, BEiT, Swin	96.39%	96.03%	95.23%	95.61%
	ViT, DEiT	**97.42%**	**96.82%**	**96.98%**	**96.90%**
	ViT, BEiT	95.36%	94.48%	94.32%	94.40%
	ViT, Swin	**97.42%**	**96.82%**	**96.98%**	**96.90%**
	DEiT, BEiT	96.05%	95.17%	95.32%	95.25%
	DEiT, Swin	96.74%	96.45%	95.64%	93.30%
	BEiT, Swin	95.19%	95.49%	92.84%	94.03%

The combinations of (ViT, Swin) and (ViT, DEiT) achieve the highest accuracy of 97.42%, suggesting that these models' complementary features lead to superior classification performance on the Chest dataset Table 4.33. In contrast combination, (ViT, DEiT, Swin), achieves an accuracy of 96.91%, indicating that leveraging swin in the combination of ViT DEiT can lower the accuracy and add more complexity. In contrast, combinations involving BEiT, such as (BEiT, Swin) (94.03%) and (ViT, BEiT) (94.40%), show relatively lower accuracies. This suggests that BEiT's features may not complement those of ViT and Swin as effectively. Combinations like (DEiT, BEiT) or (DEiT, Swin) also show competitive but slightly lower performances, ranging from 95.25% to 95.61%, indicating that while DEiT contributes positively, its combination with BEiT or Swin might not fully leverage the strengths of each model.The varied performance underscores the importance of model interactions in ensemble methods. ViT and Swin exhibit a synergistic effect, likely due to their ability to capture different aspects of the data. On the contrary, BEiT's inclusion often leads to lower performance, suggesting it may not capture complementary features as effectively.

DermaMNIST Dataset

Table 4.34: Soft voting results on DermaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	ViT, DEiT, BEiT, Swin	79.00%	67.50%	67.50%	60.86%
	ViT, DEiT, BEiT	78.55%	67.92%	56.52%	59.53%
	ViT, DEiT, Swin	79.60%	68.57%	59.78%	62.84%
	ViT, BEiT, Swin	78.10%	67.19%	55.88%	58.50%
	DEiT, BEiT, Swin	79.40%	68.21%	60.76%	63.21%
	ViT, DEiT	78.60%	65.40%	55.80%	58.92%
	ViT, BEiT	77.01%	65.23%	54.84%	56.93%
	ViT, Swin	78.55%	69.18%	59.17%	62.30%
	DEiT, BEiT	78.75%	63.06%	57.46%	58.80%
	DEiT, Swin	**80.15%**	**67.38%**	**62.18%**	**64.13%**
	BEiT, Swin	78.60%	65.53%	61.03%	61.84%

The combination of (DEiT, Swin) achieves the highest accuracy of 80.15%, indicating that these models' complementary features lead to superior classification performance on the DermaMNIST dataset Table 4.34 with less complexity. Another notable combination, (DEiT, BEiT, Swin) achieves an accuracy of 79.40%, highlighting that the adding of BEiT

in combination of (DEiT, Swin) decrease the accuracy, additionally (ViT, BEiT) and (ViT, BEiT, Swin) show relatively lower accuracies. This suggests that BEiT's features may not complement those of ViT and Swin as effectively.Combination (ViT, Swin) show competitive but slightly lower performance. This demonstrates that while ViT contributes positively, its combination with Swin might not fully leverage the strengths of each model.The varied performance underscores the importance of model interactions in ensemble methods. DEiT and Swin exhibit a synergistic effect, likely due to their ability to capture different aspects of the data, while BEiT's inclusion often leads to lower performance.

BloodMNIST dataset

Table 4.35: Soft voting results on BloodMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	ViT, DEiT, BEiT, Swin	97.81%	97.87%	97.64%	97.75%
	ViT, DEiT, BEiT	97.72%	97.73%	97.57%	97.64%
	ViT, DEiT, Swin	**97.90%**	**97.96%**	**97.76%**	**97.86%**
	ViT, BEiT, Swin	97.69%	97.66%	97.57%	97.60%
	DEiT, BEiT, Swin	97.49%	97.47%	97.34%	97.39%
	ViT, DEiT	97.84%	97.77%	97.83%	97.80%
	ViT, BEiT	97.78%	97.70%	97.54%	97.61%
	ViT, Swin	97.78%	97.63%	97.71%	97.66%
	DEiT, BEiT	97.60%	97.63%	97.27%	97.44%
	DEiT, Swin	97.52%	97.42%	97.28%	97.34%
	BEiT, Swin	96.90%	96.95%	96.47%	96.68%

The combination of (ViT, DEiT, Swin) emerges as the most effective, achieving the highest performance across all metrics with an accuracy of 97.90%, precision of 97.96%, recall of 97.76%, and an F1 score of 97.86% Table 4.35. This indicates that this particular combination excels in accurately classifying blood, suggesting a strong synergy between these models in this context. Other combinations also show high performance but slightly lower than the top-performing combination. For instance, the combination of (ViT, DEiT, BEiT, Swin) achieves an accuracy of 97.81%, precision of 97.87%, recall of 97.64%, and an F1 score of 97.75% but with more complexity added to ensemble method. Similarly, the pairing of (ViT, DEiT) alone performs well, with an accuracy of 97.84%, precision of 97.77%, recall of 97.83%, and an F1 score of 97.80%. Interestingly, the combination of (BEiT, Swin) shows the lowest performance among the configurations, with an accuracy of 96.90%, precision of 96.95%, recall of 96.47%, and an F1 score of 96.68%. This suggests that while (BEiT, Swin) are effective individually, their combination may not complement each other as well as the other model pairings.

BreastMNIST dataset

Table 4.36: Soft voting results on BreastMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT, DEiT, BEiT, Swin	86.54%	88.75%	76.50%	80.04%
	ViT, DEiT, BEiT	85.90%	88.26%	75.31%	78.86%
	ViT, DEiT, Swin	**88.46%**	**89.05%**	**80.83%**	**83.75%**
	ViT, BEiT, Swin	87.18%	89.23%	77.69%	81.20%
	DEiT, BEiT, Swin	84.62%	0.8462%	0.7444%	0.7743%
	ViT, DEiT	87.18%	88.00%	78.45%	81.58%
	ViT, BEiT	84.62%	89.04%	72.18%	75.85%
	ViT, Swin	86.54%	85.54%	78.76 %	81.22%
	DEiT, BEiT	81.41%	81.22%	68.48%	71.16%
	DEiT, Swin	85.90%	83.56%	79.07%	80.87%
	BEiT, Swin	86.54%	85.54%	78.76%	81.22%

The combination of (ViT, DEiT, Swin) stands out as the best-performing ensemble, achieving the highest accuracy of 88.46%, precision of 89.05%, recall of 80.83%, and F1 score of 83.75% Table 4.36. This indicates that this trio effectively balances high accuracy with strong precision and recall, making it the most reliable combination for this dataset. Other notable combinations include (ViT, BEiT, Swin) which achieved an accuracy of 87.18%. Although this combination also performs well, it lags slightly behind the top combination in recall, suggesting a slightly higher rate of missed positive instances. The ensemble of (ViT, DEiT) also performed well, with an accuracy of 87.18% and F1 score of 81.58%. This duo shows that even with fewer models, a high level of performance can be maintained, especially in precision and recall balance. Interestingly, the combination of DEiT, BEiT, and Swin showed a significant drop in performance, with an accuracy of 84.62% and F1 score of 77.43%. This indicates that excluding ViT from the ensemble reduces the overall effectiveness, emphasizing ViT's critical role in enhancing model performance. On the lower end, the combinations of(DEiT,BEiT) (accuracy of 81.41% and F1 score of 71.16%) and (BEiT,Swin) (accuracy of 86.54%) indicate that these models struggle more without the presence of ViT, especially in maintaining high recall and balanced F1 scores. The soft voting method reveals that the combination of ViT, DEiT, and Swin is the most effective for the BreastMNIST dataset, offering the highest overall performance. The consistent performance drop in ensembles excluding ViT highlights its importance. These findings suggest that leveraging diverse transformer models with soft voting can significantly enhance performance on medical imaging datasets like BreastMNIST, particularly when ViT is included in the ensemble.

OrganCMNIST dataset

Table 4.37: Soft voting results on OrganCMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	ViT, DEiT, BEiT, Swin	93.49%	93.20%	92.27%	92.66%
	ViT, DEiT, BEiT	**93.52%**	**93.13%**	**92.45%**	**92.73%**
	ViT, DEiT, Swin	93.40%	93.15%	92.22%	92.61%

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	ViT, BEiT, Swin	93.35%	93.09%	92.09%	92.50%
	DEiT, BEiT, Swin	93.09%	92.90%	91.85%	92.29%
	ViT, DEiT	93.28%	92.81%	92.19%	92.44%
	ViT, BEiT	93.51%	93.11%	92.39%	92.68%
	ViT, Swin	92.90%	92.72%	91.43%	91.96%
	DEiT, BEiT	93.09%	92.77%	91.94%	92.28%
	DEiT, Swin	92.62%	92.51%	91.26%	91.77%
	BEiT, Swin	92.65%	92.62%	91.21%	91.79%

The combination of (ViT, DEiT, BEiT) emerges as the top performer, achieving the highest accuracy of 93.52% Table 4.37. This ensemble demonstrates the best overall balance and effectiveness across all metrics, indicating its strong capability to handle the OrganCMNIST dataset's complexities. Close behind, the combination of ViT, DEiT, Swin achieved an accuracy of 93.40%. While slightly lower in recall and F1 score compared to the top performer, this trio still maintains a robust performance, suggesting it is also a reliable choice. The ensemble of (ViT, BEiT) also performs admirably, with an accuracy of 93.51%. This indicates that even with fewer models, a high performance level is maintained, particularly in terms of precision and recall balance. Another notable combination is (ViT, DEiT, BEiT, Swin) which achieved an accuracy of 93.49%. Despite the slight drop compared to the top performers, this ensemble still provides a strong overall performance, emphasizing the robustness of the ViT model when included in any combination. However combinations such as (DEiT, Swin) with an accuracy of 92.62%, and (BEiT, Swin) with an accuracy of 92.65%, indicate that these models struggle more without the inclusion of ViT, especially in maintaining high recall and balanced F1 scores. In Summary, the soft voting method on the OrganCMNIST dataset highlights the (ViT, DEiT, BEiT) combination as the most effective ensemble, offering the highest performance across all metrics. The consistent performance drop in ensembles that exclude ViT underscores its critical role in enhancing model performance. These findings suggest that leveraging diverse transformer models with soft voting can significantly improve performance on medical imaging datasets like OrganMNIST, particularly when ViT is part of the ensemble.

OrganSMNIST dataset

Table 4.38: Soft voting results on OrganSMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	ViT, DEiT, BEiT, Swin	**83.46** %	**80.06 %**	**79.12 %**	**79.29%**
	ViT, DEiT, BEiT	83.36%	80.09%	79.08 %	78.94 %
	ViT, DEiT, Swin	83.06%	79.83 %	78.78 %	78.87%
	ViT, BEiT, Swin	83.43 %	80.13 %	79.06 %	79.43%
	DEiT, BEiT, Swin	83.18 %	79.76 %	78.98 %	79.28 %
	ViT, DEiT	82.61%	79.29%	78.27 %	77.80 %
	ViT, BEiT	83.19 %	79.95 %	78.85%	78.88 %
	ViT, Swin	82.84%	79.54 %	78.26 %	78.58%
	DEiT, BEiT	83.04%	79.44 %	78.82 %	78.90%
	DEiT, Swin	82.97 %	79.67%	78.92%	79.16%

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	BEiT, Swin	82.74%	79.59%	78.56 %	78.98 %

The combination of (ViT, DEiT, BEiT) achieves the highest accuracy of 83.46% on the OrganSMNIST dataset, demonstrating superior classification performance Table 4.38. The combination of (ViT, BEiT) also performs well, achieving 83.43%, indicating that adding DEiT slightly improves accuracy.Combinations involving (ViT,Swin) or (DEiT, Swin) have lower accuracies, ranging from 82.84% to 83.06%, suggesting these combinations may not fully leverage the models' strengths to optimize accuracy. The varied performance of different model combinations highlights the importance of model interactions in ensemble methods. The strong performance of (ViT, DEiT, BEiT) likely comes from their ability to capture diverse aspects of the OrganSMNIST dataset effectively.

OrganAMNIST dataset

Table 4.39: Soft voting results on OrganAMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganAMNIST	ViT, DEiT, BEiT, Swin	94.85%	95.49 %	94.57%	94.91%
	ViT, DEiT, BEiT	**95.14%**	**95.74%**	**95.15 %**	**95.36 %**
	ViT, DEiT, Swin	94.75 %	95.34%	94.47%	94.79 %
	ViT, BEiT, Swin	94.30 %	95.04 %	93.94%	94.32%
	DEiT, BEiT, Swin	94.65 %	95.27%	94.29%	94.65 %
	ViT, DEiT	94.68%	9514%	94.41 %	94.67 %
	ViT, BEiT	94.27%	95.04%	94.09%	94.42%
	ViT, Swin	93.91%	94.62%	93.52 %	93.90%
	DEiT, BEiT	94.83 %	95.41%	94.57%	94.89%
	DEiT, Swin	94.31%	94.91 %	93.89 %	94.27%
	BEiT, Swin	93.89%	94.65%	93.49 %	93.88 %

The combination of (ViT, DEiT, BEiT) demonstrated the highest accuracy of 95.14%, showcasing robust performance. This combination outperformed others, suggesting the effectiveness of leveraging them together for OrganAMNIST classification tasks Table 4.39. Similarly, combinations involving (ViT, DEiT, Swin) also showed competitive performance, with accuracy 94.75%. Conversely, combinations involving fewer models or certain pairs exhibited relatively lower accuracies, ranging from 93.89% to 94.83%. For instance, the combination of BEiT and Swin achieved an accuracy of 93.89%, while ViT and Swin yielded an accuracy of 93.91%. Although these combinations remain competitive, they may not fully exploit the synergistic effects observed in the top-performing combinations. Additionally, including ViT alongside other models consistently contributed to solid to high accuracy. For example, ViT combined with BEiT achieved an accuracy of 94.27%, while ViT paired with DEiT yielded an accuracy of 94.68%. These results underscore the effectiveness of including ViT in the ensemble, indicating its robust performance across different combinations.

PneumoniaMNIST dataset

Table 4.40: Soft voting results on PneumoniaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoniaMNIST	ViT, DEiT, BEiT, Swin	92.95%	94.15%	91.03%	92.23%
	ViT, DEiT, BEiT	91.99%	93.19%	89.91%	91.15%
	ViT, DEiT, Swin	92.63%	94.07%	90.51%	91.84%
	ViT, BEiT, Swin	**94.07%**	**95.09%**	**92.44%**	**93.50%**
	DEiT, BEiT, Swin	90.87%	92.10%	88.59%	89.88%
	ViT, DEiT	91.99%	93.19%	89.91%	91.15%
	ViT, BEiT	93.59%	94.48%	91.97%	92.98%
	ViT, Swin	94.07%	95.22%	92.35%	93.49%
	DEiT, BEiT	87.02%	87.00%	85.00%	85.80%
	DEiT, Swin	91.35%	92.73%	89.06%	90.40%
	BEiT, Swin	93.43%	94.49%	91.67%	96.27

The combinations of (ViT, BEiT,Swin) and (ViT, Swin) achieve the highest accuracy of 94.07%, demonstrating superior classification performance on the PneumoniaMNIST dataset and highlighting that adding BEiT to the combination of (ViT, Swin) does not have strong effects Table 4.40. Other combinations like (ViT, DEiT, Swin) as well as (ViT, Swin, BEiT) also perform well, with accuracies between 92.63% and 94.07%.Combinations involving DEiT and BEiT or DEiT and Swin have lower accuracies, ranging from 87.02% to 90.87%, suggesting these combinations may not fully leverage the models' strengths to optimize accuracy. The varied performance of different model combinations highlights the importance of model interactions in ensemble methods. The strong performance of (ViT, BEiT, Swin) likely comes from their ability to capture diverse aspects of the PneumoniaMNIST dataset effectively.

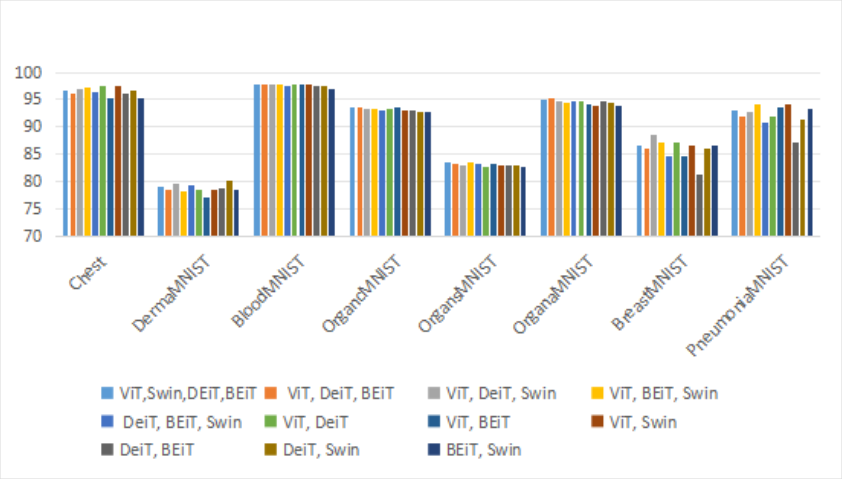


Figure 4.3: Illustration of soft voting ensemble method results in terms of accuracy

Summary: The ablation study as also illustrated in Figure 4.2 of soft voting across various medical image classification datasets reveals crucial findings into model interactions and their impact on classification performance. In the Chest dataset, the ViT and Swin combination achieves the highest accuracy of 97.42%, demonstrating their complementary strengths. Similar patterns are observed in other datasets, with DEiT and

Swin achieving the highest accuracy of 80.15% for the DermaMNIST dataset and ViT, DEiT, and Swin reaching 97.90% for the BloodMNIST dataset. In the BreastMNIST dataset, ViT, DEiT, and Swin emerge as top performers with 88.46% accuracy. For the OrganCMNIST and OrganSMNIST datasets, the combinations of ViT, DEiT, and BEiT consistently yield superior results, with accuracies of 93.52% and 83.46%, respectively. The OrganAMNIST dataset sees the highest accuracy of 95.14% from the ViT, DEiT, and BEiT combination. Lastly, in the PneumoniaMNIST dataset, the combination of ViT, BEiT, and Swin leads to the highest accuracy of 94.07%. These findings underscore the importance of selecting the right model combinations to exploit their synergistic effects effectively, highlighting how specific transformer models, when used together, can significantly enhance classification performance across diverse medical imaging tasks.

4.3.5.4 Experiments with weighted soft voting ensemble

In the following paragraphs, we will conduct a detailed analysis of the resultsfrom our weighted soft voting ensemble method implementation for each dataset.

Chest dataset

Table 4.41: Weighted soft voting results on chest dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	ViT, DEiT, BEiT, Swin	96.91%	96.58%	95.93%	96.25%
	ViT, DEiT, BEiT	96.74%	95.86%	96.32%	96.09%
	ViT, DEiT, Swin	97.25%	96.84%	96.52%	96.68%
	ViT, BEiT, Swin	**97.59%**	**97.26%**	**96.93%**	**97.09%**
	DEiT, BEiT, Swin	97.08%	96.71%	96.22%	96.46%
	ViT, DEiT	96.91%	96.00%	96.62%	96.30%
	ViT, BEiT	96.05%	95.04%	95.49%	95.26%
	ViT, Swin	97.42%	96.82%	96.98%	96.90%
	DEiT, BEiT	96.05%	95.17%	95.32%	95.25%
	DEiT, Swin	97.08%	96.87%	96.05%	96.45%
	BEiT, Swin	95.53%	95.74%	93.42%	94.48%

The highest accuracy of **97.59%** is achieved by the combination of (ViT, BEiT, Swin) which also shows strong precision at 97.26%, recall at 96.93%, and F1 score at 97.09%. This highlights the effectiveness of these three models working together to achieve robust classification performance Table 4.41. The combination of (ViT, Swin) also performs exceptionally well, with an accuracy of 97.42%, precision of 96.82%, recall of 96.98%, and F1 score of 96.90%. This suggests that the inclusion of Swin significantly enhances the performance when paired with ViT. The combination of (ViT, DEiT, Swin) achieves an accuracy of 97.25%, precision of 96.84%, recall of 96.52%, and F1 score of 96.68%, indicating that this trio is highly effective in classification tasks. The four-model ensemble of(ViT, DEiT, BEiT, Swin) follows with an accuracy of 96.91%, precision of 96.58%, recall of 95.93%, and F1 score of 96.25% this indicating that using the four models for the chest dataset on weighted soft not just increase complexity but also decrease the performance. The combination of DEiT, BEiT, Swin achieves an accuracy of 97.08% and F1 score of 96.46%. This shows that these models together provide strong classification performance. Lower accuracy is observed in the combinations of (ViT, BEiT) (96.05%) and (DEiT, BEiT) (96.05%), indicating that these pairs are less effective without the inclusion of Swin or DEiT. However, the combination of (DEiT, Swin) manages to achieve

a reasonable accuracy of 97.08%, with F1 score of 96.45%. Interestingly, the combination of (BEiT, Swin) achieves an accuracy of 95.53% with lower F1 score of 94.48%, suggesting that this pair captures fewer true positives compared to other combinations. The analysis of the Chest dataset highlights the strong performance of (ViT, BEiT, Swin) combinations. The inclusion of Swin consistently improves the performance of the ensemble, demonstrating its effectiveness in handling complex chest X-ray images.

DermaMNIST dataset

Table 4.42: Soft voting results on DermaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	ViT, DEiT, BEiT, Swin	80.00%	67.88%	60.10%	62.75%
	ViT, DEiT, BEiT	79.45%	67.32%	58.83%	61.79%
	ViT, DEiT, Swin	80.10%	68.41%	60.14%	63.22%
	ViT, BEiT, Swin	79.05%	66.04%	60.87%	62.36%
	DEiT, BEiT, Swin	80.15%	68.49%	61.02%	63.71%
	ViT, DEiT	79.65%	66.30%	59.18%	61.77%
	ViT, BEiT	77.06%	68.57%	51.62%	55.78%
	ViT, Swin	79.10%	67.10%	59.98%	62.15%
	DEiT, BEiT	79.50%	66.73%	58.91%	61.58%
	DEiT, Swin	**80.30%**	**68.48%**	**61.67%**	**64.06%**
	BEiT, Swin	79.35%	64.33%	61.96%	62.80%

The highest accuracy of **80.30%** is achieved by the combination of (DEiT, Swin) which also shows strong precision at 68.48%, recall at 61.67%, and F1 score at 64.06% Table 4.42. This indicates that this pair of models works well together, providing a balanced performance across all metrics. The four-model ensemble of (ViT, DEiT, BEiT, Swin) follows closely with an accuracy of 80.00%, precision of 67.88%, recall of 60.10%, and F1 score of 62.75%. This combination benefits from the diversity of all four models but it decreases performance who was acheived by the two models and not just that it add more complexity to the method by adding 4 models. The combination of (ViT, DEiT, Swin) also performs well, achieving an accuracy of 80.10%, precision of 68.41%, recall of 60.14%, and F1 score of 63.22%. This indicates that the inclusion of ViT into the combination of (DEiT, swin) performs well but decrease the performence of ensemble and increase complexity. The configuration of(DEiT, BEiT, Swin) achieves an accuracy of 80.15%, precision of 68.49%, recall of 61.02%, and F1 score of 63.71%. This combination also shows that Swin enhances the overall performance when paired with DEiT and BEiT. Lower accuracy is observed in the combinations of (ViT, BEiT) (77.06%), and (DEiT, BEiT) (79.50%), indicating that these pairs are less effective without the inclusion of Swin or DEiT. Interestingly, the combination of (BEiT, Swin) achieves an accuracy of 79.35%, with a lower precision of 64.33%, but a higher recall of 61.96% and F1 score of 62.80%, suggesting that this pair can capture more true positives at the expense of some precision. The analysis of the DermaMNIST dataset underscores the importance of model selection in ensemble method. The combination than includes (DEiT, Swin) consistently performs the best, highlighting their complementary strengths. The inclusion of Swin generally enhances the performance of the ensemble, demonstrating its effectiveness in handling diverse and complex dermatological images.

BloodMNIST Dataset

Table 4.43: Weighted soft voting results on BloodMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	ViT, DEiT, BEiT, Swin	97.95%	97.97%	97.82%	97.89%
	ViT, DEiT, BEiT	**98.07%**	**97.95%**	**98.02%**	**97.98%**
	ViT, DEiT, Swin	98.04%	98.04%	97.97%	98.00%
	ViT, BEiT, Swin	97.90%	97.82%	97.79%	97.80%
	DEiT, BEiT, Swin	97.69%	97.71%	97.52%	97.61%
	ViT, DEiT	97.92%	97.76%	97.91%	97.83%
	ViT, BEiT	98.04%	97.96%	97.96%	97.95%
	ViT, Swin	97.95%	97.90%	97.88%	97.88%
	DEiT, BEiT	97.60%	97.59%	97.28%	97.42%
	DEiT, Swin	97.52%	97.48%	97.26%	97.36%
	BEiT, Swin	96.96%	96.88%	96.53%	96.68%

The combination of (ViT, DEiT, BEiT) achieves the highest accuracy of **98.07%**, suggesting a strong synergy among these models. This combination also shows high precision (97.95%), recall (98.02%), and F1 score (97.98%), indicating balanced and robust performance across different metrics Table 4.48. Other top-performing configurations include (ViT, DEiT, Swin) with an accuracy of 98.04% and (ViT, BEiT) with 98.04%. These indicates that using high number of models does not significatlly inhance performance instead it increase complexity. The full ensemble of (ViT, DEiT, BEiT, Swin) also performs well, achieving an accuracy of 97.95%, which is marginally lower than the top combinations but still indicates effective performance. This combination shows consistent metrics across precision (97.97%), recall (97.82%), and F1 score (97.89%), suggesting that leveraging all four models can provide reliable results but add more complexity to the ensemble. Notably, the combination of (BEiT, Swin) alone shows the lowest accuracy at 96.96%, highlighting potential limitations in capturing the dataset's complexity effectively. Similarly, (DEiT, Swin) also exhibit lower performance metrics, with accuracy at 97.52%. The ablation study on the BloodMNIST dataset underscores the efficacy of weighted soft voting ensemble method in enhancing classification performance. The top-performing combinations, particularly those involving (ViT, DEiT, BEiT) demonstrate the highest accuracy and balanced performance across different metrics. These findings emphasize the importance of model diversity and careful selection in ensemble method to achieve optimal results for the BloodMNIST dataset.

BreastMNIST dataset

Table 4.44: Weighted soft voting results on BreastMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT, DEiT, BEiT, Swin	88.46%	90.18%	80.08%	83.42%
	ViT, DEiT, BEiT	87.82%	89.71%	78.88%	82.32%
	ViT, DEiT, Swin	88.46%	90.18%	80.08%	83.42%
	ViT, BEiT, Swin	88.46%	90.18%	80.08%	83.42%
	DEiT, BEiT, Swin	84.62%	81.68%	77.44%	79.13%

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT, DEiT	87.82%	88.53 %	79.64%	82.68%
	ViT, BEiT	**88.46%**	**91.54%**	**75.75%**	**79.60%**
	ViT, Swin	87.82%	88.53%	79.64%	82.68%
	DEiT, BEiT	84.62%	83.65 %	75.19 %	77.90%
	DEiT, Swin	85.90%	79.07%	79.07%	80.87%
	BEiT, Swin	84.62%	82.86%	75.94%	78.33%

The top-performing configurations in terms of accuracy include (ViT, DEiT, BEiT, Swin); (ViT, DEiT, Swin); and (ViT, BEiT, Swin), each achieving an accuracy of 88.46% Table 4.44. This indicates a strong synergy among these models and also indicating that adding four models does not increase accuracy however it just increase complexity. Additionally, these combinations show high precision, with values around 90.18%, and reasonable recall and F1 scores, suggesting a balanced performance across different metrics. Combinations such as (ViT, DEiT, BEiT) and (ViT, DEiT) also show competitive accuracy, with 87.82%. These results imply that while adding Swin Transformer to the ensemble does not significantly alter the accuracy, it contributes to higher precision and balanced overall performance. Other combinations, like (DEiT, BEiT, Swin) demonstrates lower accuracy at 84.62%, indicating potential limitations in their combined effectiveness for this dataset. Similarly, combinations involving fewer models, such as (DEiT, Swin) or (BEiT, Swin) also show relatively lower accuracy and F1 scores, emphasizing the importance of leveraging a diverse set of models in the ensemble. Interestingly, the combination of (ViT, BEiT) alone shows a high precision of 91.54%, but a lower recall of 75.75%, resulting in an overall accuracy of 88.46%. This suggests that while this combination is good at correctly identifying positive cases, it may miss a higher number of true positives, leading to a lower recall. The ablation study on the BreastMNIST dataset underscores the potential of weighted soft voting ensemble method in enhancing classification performance. Combinations involving (ViT, BEiT, Swin) and (ViT, DEiT, Swin) achieve the highest accuracy, highlighting the effectiveness of using multiple models to capture different data characteristics. Despite the dataset's small size, these findings emphasize the importance of model diversity and careful selection in ensemble method to achieve optimal performance.

OrganCMNIST dataset

Table 4.45: Weighted soft voting results on OrganCMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	ViT, DEiT, BEiT, Swin	**93.72%**	93.33%	92.60%	92.90%
	ViT, DEiT, BEiT	93.63%	93.22%	92.55%	92.82%
	ViT, DEiT, Swin	93.65%	93.29%	92.53%	92.85%
	ViT, BEiT, Swin	**93.64%**	**93.24%**	**92.48%**	**92.79%**
	DEiT, BEiT, Swin	93.06%	92.73%	91.93%	92.26%
	ViT, DEiT	93.38%	92.92%	92.26%	92.52%
	ViT, BEiT	93.57%	93.14%	92.46%	92.74%
	ViT, Swin	93.31%	92.97%	92.03%	92.42%
	DEiT, BEiT	93.17%	92.85%	91.99%	92.35%
	DEiT, Swin	92.68%	92.52%	91.42%	91.87%

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	BEiT, Swin	92.89%	92.70%	91.64%	92.10%

The combination of (ViT, DEiT, BEiT, Swin) achieves the highest accuracy of **93.72%**, indicating a robust performance Table 4.45. This combination also shows high precision (93.33%), recall (92.60%), and F1 score (92.90%), suggesting a well-balanced performance across different metrics. Other strong performers include the combinations of (ViT, DEiT, BEiT) with an accuracy of 93.63%, and (ViT, DEiT, Swin) with 93.65%. These results demonstrate that excluding Swin or BEiT only slightly impacts the accuracy, while maintaining high precision and recall values, thus contributing to stable overall performance. The combination of (ViT, BEiT, Swin) also performs well, achieving an accuracy of 93.64%, with precision (93.24%), recall (92.48%), and F1 score (92.79%). This suggests that incorporating Swin with ViT and BEiT maintains competitive performance. Configurations involving fewer models, such as (DEiT, BEiT, and Swin) show a slightly lower accuracy at 93.06%, with precision, recall, and F1 scores trailing behind the top configurations. This indicates that including all four models generally leads to better performance how ever it may increase complexity compared to two models. Notably, combinations like (ViT, Swin) and (DEiT, Swin) show relatively lower performance, with accuracies of 93.31% and 92.68% respectively. This highlights the importance of model diversity and careful selection in ensemble learning. The ablation study on the OrganCMNIST dataset underscores the efficacy of weighted soft voting ensemble method in enhancing classification performance. The top-performing combinations, particularly those involving (ViT, DEiT, BEiT, Swin) demonstrate the highest accuracy and balanced performance across different metrics. These findings emphasize the importance of leveraging diverse model architectures in ensemble method to achieve optimal results for the OrganCMNIST dataset.

OrganSMNIST dataset

Table 4.46: Weighted soft voting results on OrganSMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	ViT, DEiT, BEiT, Swin	**83.72** %	**80.24%**	**79.43%**	**79.59%**
	ViT, DEiT, BEiT	83.60%	80.27%	79.27%	79.38%
	ViT, DEiT, Swin	82.94%	79.74%	78.54%	78.59%
	ViT, BEiT, Swin	83.57%	80.30%	79.23%	79.44%
	DEiT, BEiT, Swin	83.44%	80.27%	79.27%	79.38%
	ViT, DEiT	82.49%	79.43%	78.00%	77.54%
	ViT, BEiT	83.43%	79.13%	77.85%	77.64%
	ViT, Swin	82.60%	79.31%	77.98%	78.21%
	DEiT, BEiT	83.24%	79.76%	79.06%	79.25%
	DEiT, Swin	82.73%	79.27%	78.73%	78.82%
	BEiT, Swin	82.97%	79.81%	78.94%	79.31%

The combination of (ViT, DEiT, BEiT, Swin) achieves the highest accuracy of **83.72%**, demonstrating strong performance Table 4.46. This combination also yields high precision (80.24%), recall (79.43%), and F1 score (79.59%), indicating a well-rounded performance across different metrics. Other high-performing combinations include (ViT, DEiT, BEiT)

with an accuracy of 83.60%, and (ViT, BEiT, Swin) with 83.57%. These results highlight that removing one model (Swin or DEiT) does not significantly impact the overall accuracy of the best performing combination, while maintaining competitive precision and recall values, thus contributing to stable performance. The combination of (DEiT, BEiT, Swin) also performs well, achieving an accuracy of 83.44%, with precision (80.27%), recall (79.27%), and F1 score (79.38%). This suggests that incorporating Swin with DEiT and BEiT maintains competitive performance. Configurations involving fewer models, such as (ViT, DEiT) and (ViT, Swin) it may decrease the complexity of the ensemble method but it show slightly lower accuracy at 82.49% and 82.60%, respectively. This indicates that including more diverse models generally leads to better performance for the OrganSMNIST dataset on weighted soft voting. Notably, combinations like (DEiT, Swin) and (BEiT, Swin) show relatively lower performance, with accuracies of 82.73% and 82.97% respectively. This highlights the importance of model diversity and careful selection in ensemble method. The ablation study on the OrganSMNIST dataset underscores the efficacy of weighte soft voting ensemble method in enhancing classification performance.

OrganAMNIST dataset

Table 4.47: Weighted soft voting results on OrganAMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganAMNIST	ViT, DEiT, BEiT, Swin	**95.03%**	**95.56 %**	**94.83%**	**95.10%**
	ViT, DEiT, BEiT	94.99%	95.49 %	94.78 %	95.05 %
	ViT, DEiT, Swin	94.85 %	95.35 %	94.56%	94.87%
	ViT, BEiT, Swin	93.98%	94.69 %	93.50%	93.91 %
	DEiT, BEiT, Swin	94.86%	95.39 %	94.59 %	94.90 %
	ViT, DEiT	94.59 %	95.09 %	94.35%	94.62%
	ViT, BEiT	94.15%	94.85%	93.97 %	94.27 %
	ViT, Swin	94.00%	94.67 %	93.65 %	94.00%
	DEiT, BEiT	94.65 %	95.14 %	94.39 %	94.68 %
	DEiT, Swin	94.63%	95.09%	94.32 %	94.62 %
	BEiT, Swin	93.82%	94.62 %	93.46%	93.86 %

The analysis of the OrganAMNIST ablation study using weighted soft voting shows that the combination of (ViT, DEiT, BEiT, Swin) achieved the highest performance metrics, with an accuracy of 95.03%, precision of 95.56%, recall of 94.83%, and an F1 score of 95.10% Table 4.47. This combination demonstrated the best balance between precision and recall, resulting in the highest F1 score among the tested combinations. The combinations of (ViT, DEiT, BEiT), and (DEiT, BEiT, Swin) also performed similarly well, with accuracy, precision, recall, and F1 scores close to the top combination. This indicates that these three-model combinations are nearly as effective as the four-model ensemble it also decrease the complexity compared to using four models. In contrast, The lowest performance was observed in combinations such as (BEiT, Swin) with accuracy, precision, recall, and F1 scores indicating a significant drop compared to the top combinations, underscoring the importance of including DEiT and ViT in the ensemble. The study highlights that including multiple transformer models generally improves performance even it will increase also complexity, with the combination of (ViT, DEiT, BEiT, Swin) yielding the best results on the OrganAMNIST dataset.

PneumoniaMNIST dataset

Table 4.48: Weighted soft voting results on PneumoniaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoniaMNIST	ViT, DEiT, BEiT, Swin	94.07%	94.86%	92.01%	93.13%
	ViT, DEiT, BEiT	93.43%	94.63%	91.58%	92.77%
	ViT, DEiT, Swin	93.91%	94.41%	91.15%	92.40%
	ViT, BEiT, Swin	**94.07%**	**94.97%**	**92.22%**	**93.32%**
	DEiT, BEiT, Swin	92.31%	93.85%	90.09%	91.47%
	ViT, DEiT	93.59%	94.74%	91.79%	92.95%
	ViT, BEiT	93.43%	94.18%	90.73%	92.03%
	ViT, Swin	93.91%	95.11%	92.14%	93.30%
	DEiT, BEiT	87.18%	87.23%	85.13%	85.96%
	DEiT, Swin	91.83%	93.53%	89.44%	90.90%
	BEiT, Swin	93.78%	94.56%	93.32%	93.75%

The combinations of (ViT, DEiT, BEiT, Swin) and (ViT, BEiT, Swin) achieve an impressive accuracy of **94.07%**, with precision at 94.86%, recall at 92.01%, and F1 score at 93.13% Table 4.48. This combinations demonstrate a balanced performance across all metrics, making it a robust choice for classification tasks however we notice that the adding of DEiT in the combination of (ViT, BEiT, Swin) does not have a real impact on the performane it may just add more complexity by adding a new model to do ensemble method we we prefere to choose the ensemble of (ViT, BEiT, Swin) as the best performer. Another high-performing combination is (ViT, DEiT, BEiT) , which achieves an accuracy of 93.43%, with precision at 94.63%, recall at 91.58%, and F1 score at 92.77%. This result is closely followed by the combination of ViT, DEiT, Swin, with an accuracy of 93.91%. These configurations show that removing Swin or BEiT does not significantly degrade performance, while maintaining high precision and recall. The combination of ViT, BEiT, Swin also shows strong performance with an accuracy of 94.07%, precision at 94.97%, recall at 92.22%, and F1 score at 93.32%. This indicates that including Swin alongside ViT and BEiT can enhance the ensemble's performance. Configurations involving fewer models, such as ViT, DEiT and ViT, Swin, achieve accuracies of 93.59% and 93.91%, respectively. These results suggest that while reducing the number of models slightly decreases accuracy, the performance remains competitive and the complexity may be decreased . On the other hand, combinations like (DEiT, BEiT) and (DEiT, Swin) show relatively lower performance, with accuracies of 87.18% and 91.83%, respectively. This highlights the importance of model diversity and the selection of complementary models in ensemble method. Interestingly, the combination of (BEiT, Swin) demonstrates strong performance with an accuracy of 93.78%, precision at 94.56%, recall at 93.32%, and F1 score at 93.75%. This suggests that Swin significantly enhances BEiT's performance, making it a viable option for certain tasks. The ablation study on the PneumoniaMNIST dataset underscores the effectiveness of weighted soft voting ensemble method in improving classification performance. The top-performing combinations, particularly those involving (ViT, BEiT, Swin) demonstrate the highest accuracy and balanced performance across various metrics. These findings emphasize the importance of using diverse model architectures to achieve optimal results with taking into consideration the complexity of the ensemble method in medical imaging classification tasks like those posed by the PneumoniaMNIST dataset.

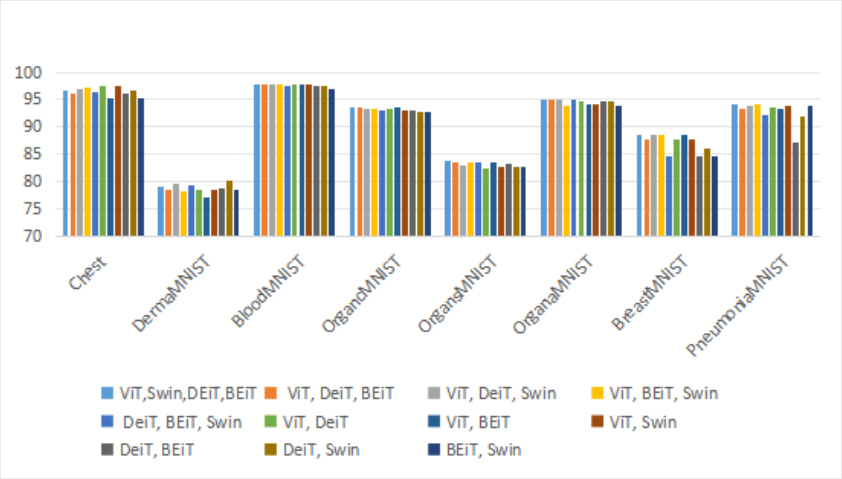


Figure 4.4: Illustration of weighted soft voting results in terms of accuracy

Summary: The analysis as also illustrated in Figure 4.4 across various medical image datasets underscores the efficacy of ensemble methods, particularly weighted soft voting, in enhancing classification performance. For the DermaMNIST dataset, the combination of (DEiT, Swin) consistently outperformed others, highlighting their complementary strengths in handling complex dermatological images. In the BloodMNIST dataset, top-performing combinations involving (ViT, DEiT, BEiT) demonstrated the highest accuracy, emphasizing the importance of model diversity. The BreastMNIST dataset findings underscored the potential of weighted soft voting ensembles, with combinations of (ViT, DEiT, BEiT, Swin) achieving superior accuracy, despite the small dataset size. Both the OrganCMNIST and OrganSMNIST datasets showed similar results, with top-performing combinations using diverse model architectures for optimal performance. The PneumoniaMNIST dataset further highlighted the effectiveness of the weighted soft voting ensemble method, with combinations of (ViT, BEiT, Swin) achieving the highest accuracy across various metrics. Finally, in the Chest dataset, combinations of (ViT, BEiT, Swin) achieved the best results, underscoring the power of combining multiple models for robust classification. These findings emphasize the importance of model diversity and careful selection in ensemble method with consideration of the complexity gained to achieve optimal performance across different medical image classification tasks .

4.3.5.5 Experiments with stacking ensemble method with logistic regression as meta-model

In the following paragraphs, we will conduct a detailed analysis of the results from our stacking with logistic ensemble method implementation for each dataset.

Chest dataset

Table 4.49: Stacking with logistic regression results on Chest dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	ViT, DEiT, BEiT, Swin	96.91%	96.42%	96.10%	96.26%
	ViT, DEiT, BEiT	96.39%	95.87%	95.40%	95.63%
	ViT, DEiT, Swin	96.91%	96.42%	96.10%	96.26%

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	ViT, BEiT, Swin	**97.42%**	**97.13%**	**96.64%**	**96.88%**
	DEiT, BEiT, Swin	97.08%	96.71%	96.22%	96.46%
	ViT, DEiT	96.74%	96.45%	95.64%	96.03%
	ViT, BEiT	96.91%	96.27%	96.27%	96.27%
	ViT, Swin	96.22%	95.90%	94.93%	95.40%
	DEiT, BEiT	**97.42%**	**96.97%**	**96.81%**	**96.89%**
	DEiT, Swin	96.91%	96.58%	95.93%	96.25%
	BEiT, Swin	96.22%	95.90%	94.93%	95.40%

Based on the analysis of the stacking ensemble using logistic regression on the Chest dataset, the combination of (ViT, BEiT, Swin) achieved the highest overall performance with an accuracy of 97.42%, precision of 97.13%, recall of 96.64%, and an F1 score of 96.88%, indicating the best synergy and robustness among the combinations tested Table 4.49. Another strong performer was the combination of (DEiT, BEiT), which also achieved an accuracy of 97.42%, with slightly lower but still impressive precision, recall, and F1 scores of 96.97%, 96.81%, and 96.89% respectively, this indicates that adding swin to the ensemble combination of (DEiT, BEiT) does not have much impact in accuracy instead it may just increas the complexity of the ensemble method. Combinations like (DEiT, BEiT, Swin) as well as (ViT, DEiT, Swin) showed robust performance with accuracy values demonstrating good balance across precision, recall, and F1 scores. The four-model ensemble (ViT, DEiT, BEiT, Swin) wich have the high complexity by using the four models and other combinations involving three models also performed well, but with slightly lower metrics compared to the top two combinations. Dual combinations such as (ViT, Swin) and (BEiT, Swin) had lower accuracy and F1 scores, indicating they were less effective in this context. This study highlights that while most combinations perform well, the specific synergy between (ViT, BEiT, Swin) leads to the highest performance on the Chest dataset, emphasizing the importance of model selection and combination for achieving optimal results using stacking ensembles with logistic regression.

DermaMNIST dataset

Table 4.50: stacking with logistic regression results on DermaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	ViT, DEiT, BEiT, Swin	**79.85%**	**69.80%**	**56.13%**	**60.28%**
	ViT, DEiT, BEiT	79.00%	66.55%	54.64%	58.42%
	ViT, DEiT, Swin	78.70%	60.83%	53.04%	56.25%
	ViT, BEiT, Swin	77.91%	56.89%	52.66%	54.42%
	DEiT, BEiT, Swin	79.65%	69.49%	55.93%	60.02%
	ViT, DEiT	78.20%	62.23%	52.96%	56.70%
	ViT, BEiT	75.81%	54.71%	54.05%	53.82%
	ViT, Swin	78.90%	67.63%	54.23%	58.00%
	DEiT, BEiT	78.90%	63.83%	55.11%	57.63%
	DEiT, Swin	79.55%	69.90%	55.75%	59.99%
	BEiT, Swin	78.95%	69.82%	55.32%	59.97%

Based on the stacking ensemble with logistic regression analysis on the DermaMNIST dataset, the combination of (ViT, DEiT, BEiT, Swin) achieved the highest overall per-

formance with an accuracy of 79.85%, precision of 69.80%, recall of 56.13%, and an F1 score of 60.28% Table 4.50. This indicates the best synergy and robustness among the combinations tested and also with the heighest complexity by using all models. The combinations of (DEiT, Swin), as well as (BEiT, Swin) who has lowest complexity also performed well, with accuracies of 79.55% and 78.95% respectively, demonstrating strong performance, though slightly lower than the top combination. The four-model ensemble (ViT, DEiT, BEiT, Swin) and other three-model combinations generally showed good performance, with accuracy values around 78-79%, but precision, recall, and F1 scores varied, reflecting differences in their ability to balance false positives and true positives. Dual combinations such as (ViT, BEiT) had lower accuracy and F1 scores, indicating they were less effective in this context. The study highlights that while most combinations perform reasonably well, the specific synergy between (ViT, DEiT, BEiT, Swin) leads to the highest performance on the DermaMNIST dataset, emphasizing the importance of model selection and combination in achieving optimal results using stacking ensembles with logistic regression.

BloodMNIST dataset

Table 4.51: stacking with logistic regression results on BloodMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	ViT, DEiT, BEiT, Swin	97.81%	97.88%	97.59%	97.73%
	ViT, DEiT, BEiT	97.75%	97.70%	97.47%	97.57%
	ViT, DEiT, Swin	**97.87%**	**97.97%**	**97.65%**	**97.80%**
	ViT, BEiT, Swin	97.66%	97.72%	97.36%	97.53%
	DEiT, BEiT, Swin	97.37%	97.46%	96.98%	97.19%
	ViT, DEiT	97.75%	97.71%	97.58%	97.64%
	ViT, BEiT	97.72%	97.78%	97.51%	97.64%
	ViT, Swin	97.75%	97.71%	97.58%	97.64%
	DEiT, BEiT	97.11%	97.07%	96.73%	96.89%
	DEiT, Swin	97.19%	97.18%	96.83%	96.99%
	BEiT, Swin	96.90%	97.12%	96.51%	96.78%

Based on the stacking ensemble with logistic regression analysis on the BloodMNIST dataset, the combination of (ViT, DEiT, Swin) achieved the highest overall performance with an accuracy of 97.87%, precision of 97.97%, recall of 97.65%, and an F1 score of 97.80% Table 4.51. This indicates the best synergy and robustness among the combinations tested. The four-model ensemble (ViT, DEiT, BEiT, Swin) and other combinations involving three models generally showed strong performance, with accuracy values around 97.66-97.81%, but precision, recall, and F1 scores varied slightly, reflecting differences in their ability to balance false positives and true positives. Combinations such as (ViT, BEiT, Swin) and (ViT, DEiT, BEiT) also performed well but slightly lower than the top combination. Dual combinations like (ViT, BEiT) had slightly lower accuracy and F1 scores, indicating they were marginally less effective in this context. The study highlights that while most combinations perform exceptionally well, the specific synergy between (ViT, DEiT, Swin) leads to the highest performance on the BloodMNIST dataset, emphasizing the importance of model selection and combination in achieving optimal results using stacking ensembles with logistic regression.

BreastMNIST dataset

Table 4.52: stacking with logistic regression results on BreastMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT, DEiT, BEiT, Swin	87.18%	89.23%	77.69%	81.20%
	ViT, DEiT, BEiT	86.54%	90.31%	75.75%	79.60%
	ViT, DEiT, Swin	**88.46%**	**90.18%**	**80.08%**	**83.42%**
	ViT, BEiT, Swin	87.82%	88.53%	79.64%	82.68%
	DEiT, BEiT, Swin	83.97%	80.53%	77.01%	78.45%
	ViT, DEiT	87.82%	89.71%	78.88%	82.32%
	ViT, BEiT	85.90%	85.83%	76.82%	79.74%
	ViT, Swin	87.82%	88.53%	79.64%	82.68%
	DEiT, BEiT	86.54%	85.54%	78.76%	81.22%
	DEiT, Swin	83.97%	80.53%	77.01%	78.45%
	BEiT, Swin	85.26%	81.96%	79.39%	80.51%

The stacking ensemble with logistic regression analysis on the BreastMNIST dataset reveals that the combination of(ViT, DEiT, Swin) achieved the highest overall performance, with an accuracy of 88.46%, precision of 90.18%, recall of 80.08%, and an F1 score of 83.42%, indicating the best synergy and robustness among the combinations tested Table 4.52. Other combinations, such as (ViT, BEiT, Swin) and (ViT, DEiT, BEiT) also performed well, with accuracies around 87.82% and 86.54%, respectively. These combinations demonstrated strong but slightly lower performance compared to the top-performing combination. The four-model ensemble (ViT, DEiT, BEiT, Swin) and some combinations like (ViT, DEiT) and (ViT, BEiT) showed good performance but varied more significantly in precision, recall, and F1 scores, reflecting differences in their ability to balance false positives and true positives. Combinations like (ViT, Swin) had lower accuracy and F1 scores, indicating they were less effective in this context. The study highlights that while most combinations perform reasonably well, the specific synergy between (ViT, DEiT, Swin) leads to the highest performance on the BreastMNIST dataset, emphasizing the importance of model selection and combination in achieving optimal results using stacking ensembles with logistic regression.

OrganCMNIST dataset

Table 4.53: stacking with logistic regression results on OrganCMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	ViT, DEiT, BEiT, Swin	92.10%	91.66%	91.32%	91.27%
	ViT, DEiT, BEiT	**92.20%**	**91.66%**	**91.37%**	**91.31%**
	ViT, DEiT, Swin	90.45%	90.68%	88.99%	89.30%
	ViT, BEiT, Swin	90.54%	90.44%	89.21%	89.28%
	DEiT, BEiT, Swin	90.23%	89.37%	89.19%	88.94%
	ViT, DEiT	91.72%	91.65%	90.60%	90.77%
	ViT, BEiT	92.10%	91.51%	91.25%	91.22%
	ViT, Swin	91.30%	91.16%	90.18%	90.36%
	DEiT, BEiT	91.34%	90.71%	90.50%	90.39%

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	DEiT, Swin	89.10%	88.84%	87.84%	87.72%
	BEiT, Swin	89.32%	88.87%	87.88%	87.93%

The stacking ensemble with logistic regression analysis on the OrganCMNIST dataset shows that the combination of (ViT, DEiT, BEiT) achieved the highest overall performance, with an accuracy of 92.20%, precision of 91.66%, recall of 91.37%, and an F1 score of 91.31%. This indicates the best synergy and robustness among the combinations tested Table 4.53. The combination of (ViT, DEiT, BEiT, Swin) also performed well, with similar high scores, demonstrating strong but slightly lower performance compared to the top combination this indicates that adding the four models decrease the accuracy instead of increasing it in this dataset with also increasing complexity. Other combinations, such as (ViT, BEiT) and (ViT, DEiT) also showed good performance, with accuracies and F1 scores slightly below the top-performing combination. In contrast, combinations like (DEiT, BEiT, Swin), and (DEiT, Swin) had lower accuracy, precision, recall, and F1 scores, indicating they were less effective in this context. This study highlights that while most combinations perform reasonably well, the specific synergy between (ViT, DEiT, BEiT) leads to the highest performance on the OrganCMNIST dataset, emphasizing the importance of model selection and combination in achieving optimal results using stacking ensembles with logistic regression.

OrganSMNIST dataset

Table 4.54: stacking with logistic regression results on OrganSMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	ViT, DEiT, BEiT, Swin	79.48%	75.34%	76.12%	75.27%
	ViT, DEiT, BEiT	79.07 %	74.83 %	75.77%	74.63%
	ViT, DEiT, Swin	81.78%	78.09 %	77.65 %	77.00%
	ViT, BEiT, Swin	**82.39%**	**78.92%**	**78.73%**	**78.08%**
	DEiT, BEiT, Swin	79.22%	75.01%	75.92 %	75.01%
	ViT, DEiT	80.30 %	76.70%	76.00 %	75.17%
	ViT, BEiT	79.08%	75.01%	75.18 %	74.35 %
	ViT, Swin	82.18%	78.78 %	78.34%	77.53%
	DEiT, BEiT	79.22 %	75.01%	7592 %	7501%
	DEiT, Swin	78.82%	74.29%	7508 %	7406%
	BEiT, Swin	80.30%	76.70 %	7600 %	7517%

The stacking ensemble with logistic regression analysis on the OrganSMNIST dataset reveals that the combination of ViT, BEiT, and Swin achieved the highest overall performance, with an accuracy of 82.39%, precision of 78.92%, recall of 78.73%, and an F1 score of 78.08% Table 4.54. This combination demonstrates the best synergy and robustness among the tested configurations. The combination of (ViT, Swin) also performed well even slightly lower accuracy compared to (ViT, BEiT, Swin) but its less complicated it term of complexity. Other combinations, such as (ViT, DEiT, Swin), and (ViT, DEiT, BEiT) also showed good performance, with accuracies and F1 scores slightly below the top-performing combination. In contrast, combinations like (DEiT, BEiT) and (DEiT, Swin) had lower accuracy, precision, recall, and F1 scores, indicating they were less effective for this dataset. This study highlights that while most combinations perform

reasonably well, the specific synergy between (ViT, BEiT, Swin) leads to the highest performance on the OrganSMNIST dataset, emphasizing the importance of model selection and combination in achieving optimal results using stacking ensembles with logistic regression.

OrganAMNIST dataset

Table 4.55: stacking with logistic regression results on OrganAMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganAMNIST	ViT, DEiT, BEiT, Swin	93.80%	93.96%	93.52%	93.46 %
	ViT, DEiT, BEiT	93.81%	93.95%	93.57%	93.55%
	ViT, DEiT, Swin	**93.93%**	**94.11%**	**93.71%**	**93.67%**
	ViT, BEiT, Swin	93.42%	93.72%	92.94%	92.99%
	DEiT, BEiT, Swin	93.83%	94.01%	93.51%	93.49%
	ViT, DEiT	93.64%	93.74%	93.70%	93.52%
	ViT, BEiT	93.63%	93.79%	93.29%	93.29%
	ViT, Swin	92.92%	93.30%	92.43%	92.53%
	DEiT, BEiT	93.41%	93.51%	93.12%	93.09%
	DEiT, Swin	93.05 %	92.84 %	92.70%	92.53%
	BEiT, Swin	92.67 %	93.01 %	92.27%	92.23 %

For the OrganAMNIST dataset, The highest accuracy of 93.93% was achieved by the combination of (ViT, DEiT, Swin) indicating strong performance Table 4.55. This combination outperformed others, highlighting the effectiveness of leveraging (ViT, DEiT, Swin) together for OrganAMNIST classification tasks. Additionally, combinations involving (ViT, DEiT, BEiT) also demonstrated competitive performance, with accuracy of 93.81% . These results suggest that the inclusion of BEiT alongside ViT and DEiT maintains high accuracy levels. Conversely, combinations involving fewer models or certain pairs exhibited slightly lower accuracies, ranging from 92.67% to 93.64%. For example, the combination of (BEiT, Swin) achieved an accuracy of 92.67%, while (DEiT, Swin) yielded an accuracy of 93.05%. Although these combinations remain competitive, they may not fully exploit the synergistic effects observed in the top-performing combinations. Additionaly it's noteworthy that the inclusion of ViT alongside other models consistently contributed to solid to high accuracy. For instance, ViT combined with DEiT achieved an accuracy of 93.64%, while ViT paired with BEiT yielded an accuracy of 93.63%. These results underscore the effectiveness of including ViT in the ensemble, indicating its robust performance across different combinations.

PneumoniaMNIST dataset

Table 4.56: Stacking with logistic regression results on PneumoniaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoniaMNIST	ViT, DEiT, BEiT, Swin	91.67%	93.75%	89.06%	90.67%
	ViT, DEiT, BEiT	**92.31%**	**94.01%**	**90.00%**	**91.45%**
	ViT, DEiT, Swin	92.15%	94.07%	89.70%	91.24%

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoniaMNIST	ViT, BEiT, Swin	91.67%	93.75%	89.06%	89.06%
	DEiT, BEiT, Swin	88.14%	91.79%	84.27%	86.31%
	ViT, DEiT	**92.31%**	94.01%	90.00%	91.45%
	ViT, BEiT	91.99%	93.80%	89.57%	91.07%
	ViT, Swin	91.51%	93.65%	88.85%	90.48%
	DEiT, BEiT	86.38%	87.65%	83.29%	84.69%
	DEiT, Swin	88.14%	91.79%	84.27%	86.31%
	BEiT, Swin	88.14%	91.79%	84.27%	86.31%

The combinations of (ViT, DEiT,BEiT) and (ViT, DEiT) demonstrated the highest accuracy of 92.31%, indicating strong performance and aslo indicating that the adding of BEiT in the combination of (ViT,DEiT) does not do much effect Table 4.56. This combination (ViT, DEiT) outperformed others, suggesting the effectiveness of using (ViT, DEiT) together for PneumoniaMNIST classification tasks. Similarly, combinations involving (ViT, DEiT, Swin) also showed competitive performance, with accuracy and other metrics ranging from 92.15% to 94.07%. Conversely, combinations involving fewer models or certain pairs exhibited relatively lower accuracies For example, the combination of (DEiT, BEiT) achieved an accuracy of 86.38%. Additionally, it's noteworthy that including ViT alongside other models consistently contributed to solid to high accuracy. For instance, ViT combined with BEiT achieved an accuracy of 91.99%, while ViT paired with Swin yielded an accuracy of 91.51%.

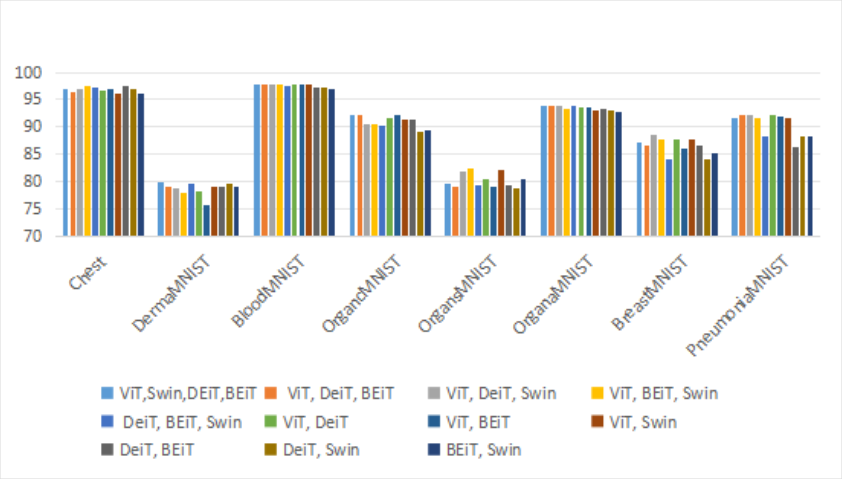


Figure 4.5: Illustration of stacking ensemble method with logistic regression as meta-model results in terms of accuracy

Summary: The ablation study as also illustrated in Figure 4.2 of stacking with logistic regression across various medical image classification datasets reveals intriguing findings into the effectiveness of different model combinations. In the chest dataset, the stacking ensemble of (DEiT, BEiT) achieved the highest accuracy of 97.42%, indicating superior performance and highlighting the benefits of trying diverse models in the ensemble method to get the heighest performance and lowest complexity. Similarly, in the DermaMNIST dataset, the combination of (ViT, DEiT, BEiT, Swin) exhibited the highest accuracy of 79.85%, showcasing the effectiveness of leveraging all four models in the ensemble. For the BloodMNIST dataset, the combination of (ViT, DEiT, Swin) achieved the highest accuracy of 97.87%, indicating excellent performance, while combinations involving fewer

models exhibited relatively lower accuracies, underscoring the importance of model interactions. In the BreastMNIST dataset, varied performance metrics were observed, with the combination of (ViT, DEiT, Swin) achieving the highest accuracy of 88.46%, suggesting effective model complementarity. Similarly, in the OrganSMNIST dataset, the combination of (ViT, BEiT, Swin) demonstrated superior performance with an accuracy of 82.39%, highlighting the effectiveness of combining these models. Lastly, in the OrganAMNIST dataset, the highest accuracy of 93.93% was achieved by the combination of (ViT, DEiT, Swin), indicating strong performance and emphasizing the robustness of including ViT in the ensemble. Notably, in almost all high-accuracy combinations, Swin and ViT were consistently present, underscoring their significant contribution to enhanced performance. These findings underscore the importance of considering model interactions and leveraging diverse model ensembles to achieve superior classification performance across different medical image classification tasks. Additionally, in almost all datasets, using a higher number of models resulted in higher accuracy, even though this may impact the complexity of the ensemble method.

4.3.5.6 Experiments with stacking ensemble method with SVM as meta-model

In the following paragraphs, we will conduct a detailed analysis of the results from our stacking with SVM ensemble methodimplementation for each dataset.

Chest dataset

Table 4.57: Stacking with svm on DermaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	ViT, DEiT, BEiT, Swin	97.25%	98.05%	98.05%	98.05%
	ViT, DEiT, BEiT	96.39%	95.45%	95.91%	95.67%
	ViT, DEiT, Swin	96.91%	96.27%	96.27%	96.27%
	ViT, BEiT, Swin	97.25%	97.00%	96.35%	96.66%
	DEiT, BEiT, Swin	96.74%	96.14%	95.98%	96.06%
	ViT, DEiT	96.56%	95.59%	96.20%	95.89%
	ViT, BEiT	95.70%	95.04%	94.57%	94.80%
	ViT, Swin	**97.59%**	**97.26%**	**96.93%**	**97.09%**
	DEiT, BEiT	95.70%	95.04%	94.57%	94.80%
	DEiT, Swin	97.08%	96.71%	96.22%	96.46%
	BEiT, Swin	96.74%	96.00%	96.15%	96.07%

The stacking ensemble with SVM analysis on the Chest dataset reveals that the combination of (ViT, Swin) achieved the highest overall performance, with an accuracy of 97.59%, precision of 97.26%, recall of 96.93%, and an F1 score of 97.09%, demonstrating the best synergy among the combinations tested Table 4.57. The combination of (ViT, DEiT, BEiT, Swin) also performed very well, with slightly lower but still strong scores how ever it may increase complexity compared to combination of two models. The combinations of (ViT, BEiT, Swin) and (DEiT, Swin) also showed good performance, with accuracies, precision, recall, and F1 scores indicating strong but slightly lower effectiveness compared to the top-performing combination. Other combinations, such as (ViT, DEiT, BEiT) and (ViT, DEiT, Swin) had lower performance metrics, reflecting lesser effectiveness in this context. The combinations of (DEiT, BEiT) and (ViT, BEiT) had the lowest accuracy, precision, recall, and F1 scores, indicating they were the least effective

for this dataset. This analysis highlights that while most combinations perform reasonably well, the specific synergy between ViT and Swin leads to the highest performance on the Chest dataset, underscoring the importance of model selection and combination in achieving optimal results using stacking ensembles with SVM.

DermaMNIST dataset

Table 4.58: Stacking with svm on DermaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	ViT, DEiT, BEiT, Swin	**78.55%**	**58.75%**	**55.46%**	**56.61%**
	ViT, DEiT, BEiT	78.20%	60.44%	53.28%	56.14%
	ViT, DEiT, Swin	78.10%	57.84%	53.26%	54.94%
	ViT, BEiT, Swin	77.31%	63.44%	49.98%	52.82%
	DEiT, BEiT, Swin	77.61%	62.84%	51.03%	55.32%
	ViT, DEiT	77.46%	55.17%	51.51%	52.83%
	ViT, BEiT	75.91%	56.47%	48.92%	51.86%
	ViT, Swin	77.76%	58.68%	53.06%	55.06%
	DEiT, BEiT	77.11%	58.22%	48.49%	52.06%
	DEiT, Swin	77.56%	57.16%	51.40%	53.55%
	BEiT, Swin	78.35%	58.00%	53.29%	54.82%

For the DermaMNIST dataset, The combination of (ViT, DEiT, BEiT, Swin)emerged as the top performer, achieving an accuracy of 78.55%. This combination demonstrated the strongest performance, indicating the effectiveness of leveraging all four models Table 4.58. Other combinations, such as (ViT, DEiT, BEiT) and (ViT, DEiT, Swin) also showed competitive performance, with accuracies ranging from 78.10% to 78.20%. These results suggest that including ViT, DEiT, and either BEiT or Swin in the combination contributes to maintaining high classification accuracy on the DermaMNIST dataset. Conversely, combinations involving fewer models or certain pairs like (ViT, BEiT) exhibited relatively lower accuracies, ranging from 75.91% to 77.46%. While still competitive, these combinations may not fully exploit the synergistic effects of the models compared to the top-performing combinations. Additionally, some combinations, such as (ViT, BEiT) demonstrated lower accuracy, suggesting that specific combinations may not be as effective for stacking with svm in the DermaMNIST dataset.

BloodMNIST dataset

Table 4.59: Stacking with svm on BloodMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	ViT, DEiT, BEiT, Swin	97.52%	97.64%	97.29%	97.44%
	ViT, DEiT, BEiT	97.49%	97.68%	97.31%	97.48%
	ViT, DEiT, Swin	97.46%	97.56%	97.28%	97.40%
	ViT, BEiT, Swin	97.52%	97.67%	97.30%	97.47%
	DEiT, BEiT, Swin	97.11%	97.00%	96.82%	96.89%
	ViT, DEiT	**97.57%**	**97.65%**	**97.39%**	**97.50%**

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	ViT, BEiT	97.02%	97.22%	96.53%	96.86%
	ViT, Swin	97.52%	97.51%	97.29%	97.38%
	DEiT, BEiT	96.90%	96.87%	96.57%	96.69%
	DEiT, Swin	96.96%	96.93%	96.54%	96.72%
	BEiT, Swin	96.87%	96.96%	96.41%	96.66%

For the BloodMNIST dataset, The combination of (ViT, DEiT) achieved the highest accuracy of 97.57%. This combination also demonstrated superior performance across precision, recall, and F1 score metrics, with values ranging from 97.39% to 97.65% Table 4.59. The results suggest that leveraging both ViT and DEiT models in the ensemble leads to excellent classification accuracy and overall performance for BloodMNIST dataset. Combinations involving (ViT, DEiT, BEiT, Swin) also exhibited high accuracy, with accuracy (97.52%). These combinations showed competitive performance, indicating that including all four models in the ensemble contributes to maintaining high classification accuracy on the BloodMNIST dataset but may increase the complexity of the ensemble method. Conversely, combinations involving fewer models or certain pairs demonstrated relatively lower accuracies, ranging from 96.87% to 97.11%. While still competitive, these combinations may not fully utilize the synergistic effects of the models compared to the top-performing combinations.

BreastMNIST dataset

Table 4.60: Stacking with svm on BreastMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT, DEiT, BEiT, Swin	**87.82%**	88.53%	**79.64%**	**82.68%**
	ViT, DEiT, BEiT	**87.82%**	**91.13%**	**78.13%**	81.94%
	ViT, DEiT, Swin	**87.82%**	88.53%	**79.64%**	**82.68%**
	ViT, BEiT, Swin	**87.82%**	88.53%	**79.64%**	**82.68%**
	DEiT, BEiT, Swin	86.54%	85.54%	78.76%	81.22%
	ViT, DEiT	84.62%	80.91%	78.95%	79.83%
	ViT, BEiT	85.90%	85.83%	76.82%	79.74%
	ViT, Swin	87.82%	88.53%	79.64%	82.68%
	DEiT, BEiT	87.82%	89.71%	78.88%	82.32%
	DEiT, Swin	86.54%	85.54%	78.76%	81.22%
	BEiT, Swin	84.62%	84.62%	74.44%	74.43%

The combinations of (ViT, DEiT, BEiT, Swin), (ViT, DEiT, Swin), and (ViT, BEiT, Swin) consistently achieve the highest accuracy of 87.82% and balanced performance across all metrics, indicating that including multiple diverse models enhances overall effectiveness Table 4.60. Notably, ViT and Swin models are frequently part of these high-performing ensembles, highlighting their significant individual contributions. Combinations with fewer models, such as (ViT, DEiT) or (ViT, BEiT), while still effective, generally show slightly lower performance metrics, suggesting that simpler ensembles might reduce computational complexity but at the cost of some accuracy and robustness. Additionally, there is a trade-off between precision and recall in certain combinations, such as (ViT, DEiT, BEiT), which has the highest precision but slightly lower recall.

The performance of each combination also underscores the importance of model interactions, where the specific characteristics and strengths of individual models contribute significantly to the ensemble's overall performance. However, while high-performing combinations tend to include more models, increasing the ensemble method's complexity, it is essential to balance the benefits of higher accuracy with the computational costs and potential overfitting risks associated with more complex ensembles. The study demonstrates that ensembles combining (ViT, DEiT, BEiT, Swin) models achieve the highest performance on the BreastMNIST dataset, with frequent inclusion of ViT and Swin highlighting their substantial contributions. While simpler combinations offer reduced complexity, they also result in lower performance, emphasizing the importance of using diverse model interactions to achieve superior classification performance in medical image analysis.

OrganCMNIST dataset

Table 4.61: Stacking with svm on OrganCMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	ViT, DEiT, BEiT, Swin	91.46%	90.63%	90.91%	90.53%
	ViT, DEiT, BEiT	90.90%	90.75%	90.24%	90.06%
	ViT, DEiT, Swin	91.10%	90.61%	90.52%	90.19%
	ViT, BEiT, Swin	91.42%	90.68%	91.05%	90.63%
	DEiT, BEiT, Swin	**91.82%**	**91.08%**	**91.28%**	**90.96%**
	ViT, DEiT	90.72%	90.90%	89.68%	89.73%
	ViT, BEiT	89.97%	89.64%	88.56%	88.45%
	ViT, Swin	90.58%	90.10%	89.72%	89.50%
	DEiT, BEiT	89.22%	89.06%	88.06%	87.94%
	DEiT, Swin	89.31%	89.27%	88.11%	88.12%
	BEiT, Swin	90.54%	89.57%	90.38%	89.77%

The combination of (DEiT, BEiT, Swin) achieved the highest accuracy of 91.82%, along with the highest precision, recall, and F1 score, indicating that this ensemble configuration is particularly effective Table 4.61. This suggests that DEiT, BEiT, and Swin models have complementary strengths that, when combined, enhance the overall performance significantly. Other combinations, such as (ViT, DEiT, BEiT, Swin) and (ViT, BEiT, Swin), also showed high accuracy and balanced performance metrics, with accuracies of 91.46% and 91.42%, respectively. These results highlight the importance of including diverse models within the ensemble to capture various aspects of the dataset's complexity. Interestingly, combinations that excluded Swin, such as (ViT, DEiT) and (ViT, BEiT), demonstrated lower accuracy and performance metrics, with accuracies of 90.72% and 89.97%, respectively. This underscores the significant contribution of the Swin model to the ensemble's performance. Similarly, combinations without ViT, such as (DEiT, BEiT) and (DEiT, Swin), also showed relatively lower accuracies, indicating that each model brings unique strengths to the ensemble. The study reveals that while high-performing combinations tend to include more models, simpler ensembles like (ViT, DEiT) and (ViT, Swin) still provide substantial performance but at a reduced complexity. This trade-off between the number of models and performance suggests that while adding more models can enhance accuracy, it also increases computational demands and potential overfitting risks. The ablation study on the OrganCMNIST dataset demonstrates that ensembles incorporating (DEiT, BEiT, Swin) models achieve the highest performance, with frequent inclusion of the Swin model highlighting its critical role. While simpler

combinations offer reduced complexity, they typically result in lower performance, emphasizing the need for diverse model ensembles to achieve superior classification accuracy and robustness in medical image analysis.

OrganSMNIST dataset

Table 4.62: Stacking with svm on OrganSMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	ViT, DEiT, BEiT, Swin	79.82 %	76.98%	76.60%	75.46%
	ViT, DEiT, BEiT	79.08%	75.89%	75.95%	75.29 %
	ViT, DEiT, Swin	77.17 %	69.86 %	74.00%	70.53%
	ViT, BEiT, Swin	79.82 %	77.06 %	76.75 %	75.14 %
	DEiT, BEiT, Swin	77.19 %	73.86 %	74.19%	73.32 %
	ViT, DEiT	**79.92%**	**76.53 %**	**76.20%**	**75.53%**
	ViT, BEiT	78.57 %	75.37%	75.90%	74.86%
	ViT, Swin	77.39%	70.06 %	74.15%	70.67 %
	DEiT, BEiT	78.00%	74.56 %	74.77 %	73.90 %
	DEiT, Swin	78.00%	74.56%	74.77%	73.90%
	BEiT, Swin	77.29%	73.55%	74.42%	73.14 %

The combination of (ViT, DEiT) achieved the highest accuracy of 79.92%, along with the highest precision (76.53%), recall (76.20%), and F1 score (75.53%). This indicates that this specific pairing of models is particularly effective for the OrganSMNIST dataset, highlighting their complementary strengths Table 4.62. Other high-performing combinations include (ViT, DEiT, BEiT, Swin) and (ViT, BEiT, Swin), both achieving an accuracy of 79.82%. These combinations also demonstrated high precision, recall, and F1 scores, indicating their robustness and effectiveness. The inclusion of multiple diverse models in these ensembles likely helps in capturing various features of the dataset, thereby improving overall performance but still lower than the best performer however it also increase complexity without that much gain of performance compared to the heighest performing combination who is less in complexity. Combinations such as (ViT, DEiT, BEiT) and (ViT, BEiT) also performed well, with accuracies of 79.08% and 78.57%, respectively. However, these results were slightly lower compared to the top-performing combinations, suggesting that while they are effective, the additional models in the higher-performing ensembles provide a slight edge. On the other hand, some combinations like (ViT, DEiT, Swin) and (ViT, Swin) showed relatively lower accuracies of 77.17% and 77.39%, respectively. These results suggest that the exclusion of BEiT in these combinations may have impacted the performance negatively, underscoring the importance of including a diverse set of models. Interestingly, the combination of (DEiT, BEiT, Swin) achieved an accuracy of 77.19%, which is lower than when ViT is included. This further emphasizes the significant contribution of the ViT model to the ensemble's performance. The ablation study on the OrganSMNIST dataset demonstrates that the combination of ViT and DEiT is the most effective, achieving the highest performance across all metrics. While including additional models such as BEiT and Swin can enhance performance, the results indicate that ViT and DEiT form a particularly strong pairing. This analysis underscores the importance of carefully selecting and combining models to leverage their complementary strengths for superior classification performance in medical image analysis.

OrganAMNIST dataset

Table 4.63: Stacking with svm on OrganAMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganAMNIST	ViT, DEiT, BEiT, Swin	92.50 %	92.66%	92.30 %	92.07 %
	ViT, DEiT, BEiT	**93.04%**	**93.08%**	**92.94%**	**92.67%**
	ViT, DEiT, Swin	92.73%	92.90%	92.58%	92.39 %
	ViT, BEiT, Swin	92.45%	92.61 %	92.20 %	91.97%
	DEiT, BEiT, Swin	92.50 %	92.70 %	92.27 %	92.07%
	ViT, DEiT	92.54%	92.21 %	92.49%	92.05%
	ViT, BEiT	92.32 %	92.41 %	91.77%	91.66 %
	ViT, Swin	92.66%	92.67 %	92.30%	92.15%
	DEiT, BEiT	92.87%	92.93%	92.77 %	92.50%
	DEiT, Swin	92.74 %	92.84 %	92.61 %	92.42 %
	BEiT, Swin	92.51%	92.66%	92.03%	91.97 %

The combination of (ViT, DEiT, BEiT) achieved the highest accuracy of 93.04%, along with the highest precision (93.08%), recall (92.94%), and F1 score (92.67%) Table 4.63. This indicates that this specific combination is particularly effective for the OrganAMNIST dataset, leveraging the complementary strengths of these models. Other high-performing combinations include (ViT, DEiT, Swin) and (ViT, DEiT, BEiT, Swin), with accuracies of 92.73% and 92.50%, respectively. These combinations also demonstrated high precision, recall, and F1 scores, indicating their robustness and effectiveness. The inclusion of multiple diverse models in these ensembles likely helps in capturing various features of the dataset, thereby improving overall performance. The combination of (DEiT, BEiT) also performed well, with an accuracy of 92.87%. However, it slightly underperformed compared to the top combination, suggesting that while it is effective, the additional models in the higher-performing ensembles provide a slight edge. On the other hand, some combinations like (ViT, BEiT, Swin) and (DEiT, BEiT, Swin) showed slightly lower accuracies of 92.45% and 92.50%, respectively. These results suggest that the exclusion of DEiT or ViT in these combinations may have impacted the performance negatively, underscoring the importance of including a diverse set of models. Interestingly, the combination of (ViT, BEiT) achieved an accuracy of 92.32%, which is lower than when DEiT is included. This further emphasizes the significant contribution of the DEiT model to the ensemble's performance. The ablation study on the OrganAMNIST dataset demonstrates that the combination of ViT, DEiT, and BEiT is the most effective, achieving the highest performance across all metrics. While including additional models such as Swin can enhance performance, the results indicate that ViT and DEiT form a particularly strong core for the ensemble. This analysis underscores the importance of carefully selecting and combining models to leverage their complementary strengths for superior classification performance in medical image analysis.

PneumoniaMNIST dataset

Table 4.64: Stacking with svm on PneumoniaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoniaMNIST	ViT, DEiT, BEiT, Swin	90.87%	93.24%	87.99%	89.71%
	ViT, DEiT, BEiT	**91.67%**	**93.58%**	**89.15%**	**90.69%**
	ViT, DEiT, Swin	89.10%	92.14%	85.64%	87.56%
	ViT, BEiT, Swin	91.51%	93.65%	88.85%	90.48%
	DEiT, BEiT, Swin	90.87%	93.24%	87.99%	89.71%
	ViT, DEiT	89.74%	92.54%	86.50%	88.35%
	ViT, BEiT	91.35%	93.37%	88.72%	90.31%
	ViT, Swin	89.10%	92.14%	85.64%	87.56%
	DEiT, BEiT	86.70%	87.89%	83.72%	85.09%
	DEiT, Swin	91.35%	93.73%	88.55%	90.26%
	BEiT, Swin	**91.67%**	**93.93%**	**88.97%**	**90.65%**

The ablation study of stacking with SVM on the PneumoniaMNIST dataset reveals several key findings regarding the effectiveness of different model combinations Table 4.64. The combination of ViT, DEiT, and BEiT achieved the highest accuracy of 91.67%, along with the highest precision (93.58%), recall (89.15%), and F1 score (90.69%). This indicates that this specific ensemble is particularly effective for the PneumoniaMNIST dataset, successfully leveraging the strengths of these models. Other high-performing combinations include (ViT, BEiT, Swin) and (BEiT, Swin), with accuracies of 91.51% and 91.67%, respectively. These combinations also showed high precision, recall, and F1 scores, demonstrating their robustness and effectiveness. The inclusion of BEiT and Swin in these ensembles likely helps in capturing diverse features of the dataset, thereby improving overall performance. Interestingly, the combination of (ViT, DEiT) achieved a lower accuracy of 89.74%, suggesting that the exclusion of BEiT negatively impacts the ensemble's performance. This underscores the importance of including a diverse set of models to maximize classification accuracy. The combination of (DEiT, BEiT) showed the lowest performance with an accuracy of 86.70%, indicating that the exclusion of ViT and Swin leads to a significant drop in performance. This further emphasizes the critical role of ViT and Swin in enhancing the ensemble's effectiveness. The analysis indicates that the combinations involving ViT and BEiT consistently perform well, highlighting their complementary strengths. However, the best performance is achieved by including all three models: ViT, DEiT, and BEiT. The results suggest that while combining more models can improve performance, careful selection and combination of models are crucial for optimizing the ensemble's effectiveness. The ablation study on the PneumoniaMNIST dataset demonstrates that the combination of ViT, DEiT, and BEiT is the most effective, achieving the highest performance across all metrics. Including additional models like Swin can enhance performance, but the results highlight the importance of ViT and BEiT in the ensemble. This analysis underscores the need for a tailored approach in selecting and combining models to achieve superior classification performance in medical image analysis.

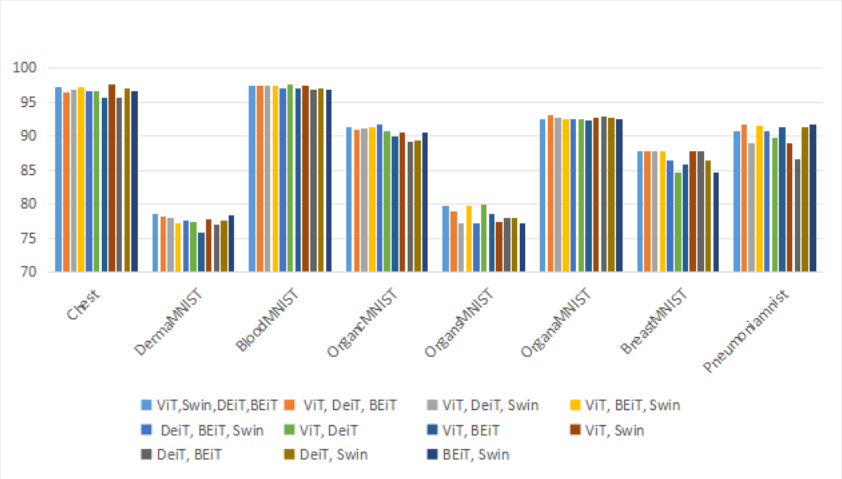


Figure 4.6: Illustration of stacking ensemble method with SVM as meta-model results in terms of accuracy

Summary: The analysis as also illustrated in Figure 4.6 of stacking ensembles with SVM across multiple medical imaging datasets highlights the nuanced impact of model combinations on classification performance. In the ChestMNIST dataset, the combination of ViT and Swin emerged as the top performer with an impressive accuracy of 97.59%, underscoring their synergistic strengths with low complexity. Similarly, in DermaMNIST, ViT, DEiT, BEiT, and Swin collectively achieved high accuracy, showcasing the effectiveness of including diverse models but having strong complexity by using the four models. BloodMNIST exhibited superior performance with ViT and DEiT, emphasizing their strong classification capabilities. For BreastMNIST, ensembles combining ViT, DEiT, BEiT, and Swin consistently demonstrated robust performance, alBEiT with increased complexity. OrganCMNIST highlighted the effectiveness of DEiT, BEiT, and Swin together, while OrganSMNIST favored ViT and DEiT as the most effective combination. OrganAMNIST showcased ViT, DEiT, and BEiT as optimal, emphasizing their balanced performance metrics. Lastly, PneumoniaMNIST showed ViT, DEiT, and BEiT as the top performers, with BEiT and Swin contributing significantly to ensemble effectiveness. The analysis across these datasets demonstrates that while combining more models generally improves performance, the specific combination and synergy between models are crucial. ViT, DEiT, and BEiT frequently appear in high-performing combinations, highlighting their strong individual and collective contributions. Including diverse models in the ensemble captures various dataset features, enhancing overall classification accuracy and robustness. However, balancing the number of models with computational complexity and potential overfitting is essential for optimizing performance in medical image analysis.

4.3.5.7 Experiments with bagging ensemble method

In the following paragraphs, we will conduct a detailed analysis of the results from our Bagging ensemble method implementation for each dataset.

Chest dataset

Table 4.65: Bagging results on Chest dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	ViT, DEiT, BEiT, Swin	96.56%	96.60%	96.56%	96.58%
	ViT, DEiT, BEiT	96.39%	96.40%	96.39%	96.39%
	ViT, DEiT, Swin	96.74%	96.74%	96.74%	96.74%
	ViT, BEiT, Swin	96.22%	96.31%	96.22%	96.24%
	DEiT, BEiT, Swin	95.19%	95.18%	95.19%	95.18%
	ViT, DEiT	96.74%	96.84%	96.74%	96.76%
	ViT, BEiT	93.64%	94.26%	93.64%	93.76%
	ViT, Swin	96.56%	96.68%	96.56%	96.59%
	DEiT, BEiT	**97.08%**	**97.18%**	**97.08%**	**97.10%**
	DEiT, Swin	95.53%	95.71%	95.53%	95.58%
	BEiT, Swin	93.81%	94.25%	93.81%	93.91%

The top-performing combination in this study is (DEiT, BEiT), achieving an accuracy of 97.08%. This configuration consistently outperforms others in accuracy, precision, recall, and F1 score metrics across the board Table 4.65. The high accuracy suggests that DEiT and BEiT models complement each other well and they have low complexity compared to other combination of high number of models. Similarly, combinations involving Swin Transformer alongside DEiT or ViT also demonstrate competitive performance. For instance, (DEiT, Swin) achieved an accuracy of around 95.53% to 95.71%, indicating strong performance though slightly lower than (DEiT, BEiT). Swin Transformer, known for its hierarchical structure and attention mechanisms, contributes significantly to the ensemble's effectiveness in capturing intricate patterns in medical images. Conversely, combinations involving ViT and BEiT without Swin generally exhibit lower accuracies, such as (ViT, BEiT) with an accuracy ranging from 93.64% to 94.26%. This suggests that while ViT and BEiT are effective individually, the absence of Swin may limit the ensemble's overall performance on this dataset. The bagging ensemble method demonstrates that combining DEiT and BEiT models leads to the highest accuracy on the Chest dataset, showcasing their synergistic effects in capturing diverse and complex features. This analysis underscores the importance of selecting and combining transformer models strategically to optimize performance in medical image analysis tasks.

DermaMNIST dataset

Table 4.66: Bagging results on DermaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	ViT, DEiT, BEiT, Swin	**78.70%**	**78.01%**	**78.70%**	**78.02%**
	ViT, DEiT, BEiT	78.35%	77.39%	78.35%	77.09%
	ViT, DEiT, Swin	**78.70%**	**78.30%**	**78.70%**	**77.89%**
	ViT, BEiT, Swin	78.20%	78.01%	78.20%	77.40%
	DEiT, BEiT, Swin	78.35%	77.39%	78.35%	77.09%
	ViT, DEiT	78.20%	78.71%	78.20%	78.24%
	ViT, BEiT	76.26%	76.79%	76.26%	75.89%

Continued on next page

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	ViT, Swin	77.21%	77.57%	77.21%	77.19%
	DEiT, BEiT	76.86%	77.91%	76.86%	77.06%
	DEiT, Swin	77.36%	78.57%	77.36%	77.65%
	BEiT, Swin	75.81%	76.93%	75.81%	76.21%

The ablation study for the DermaMNIST dataset reveals interesting findings into the performance of different combinations of base models within the bagging ensemble method Table 4.66. Combinations involving ViT, DEiT, and Swin consistently demonstrate higher accuracy compared to other combinations. Particularly, the combination of (ViT, DEiT, Swin) and (ViT, DEiT, BEiT, Swin) stands out with an accuracy of 78.70%, suggesting strong synergy among these models in capturing the characteristics of the DermaMNIST dataset. However, including all models—ViT, DEiT, BEiT, and Swin—increases the computational complexity due to the need to train and maintain multiple models simultaneously. on the other hand, combinations that include BEiT, either individually or in conjunction with other models, tend to exhibit lower accuracy. This could imply that BEiT may not be as effective as ViT, DEiT, and Swin for the DermaMNIST dataset, or it might require different configurations or preprocessing techniques to enhance its performance. It's worth noting that certain combinations, such as (ViT, DEiT) and (DEiT, Swin), demonstrate moderate accuracy, precision, recall, and F1 score metrics. This indicates that while these combinations may not achieve the highest accuracy, they still offer competitive performance and could be viable options for decreasing the complexity of ensemble method. The bagging ensemble method shows promise for classification tasks on the DermaMNIST dataset, particularly when using combinations involving ViT, DEiT, and Swin.

BloodMNIST dataset

Table 4.67: Bagging results on BloodMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	ViT, DEiT, BEiT, Swin	97.49%	97.51%	97.49%	97.48%
	ViT, DEiT, BEiT	**97.60%**	**97.61%**	**97.60%**	**97.59%**
	ViT, DEiT, Swin	97.40%	97.39%	97.29%	97.20%
	ViT, BEiT, Swin	96.90%	96.95%	96.90%	96.89%
	DEiT, BEiT, Swin	97.43%	97.12%	97.39%	97.31%
	ViT, DEiT	97.25%	97.28%	97.25%	97.24%
	ViT, BEiT	97.14%	97.16%	97.14%	97.12%
	ViT, Swin	96.67%	96.74%	96.67%	96.66%
	DEiT, BEiT	97.08%	97.12%	97.08%	97.06%
	DEiT, Swin	96.96%	97.00%	96.96%	96.95%
	BEiT, Swin	96.90%	96.96%	96.90%	96.89%

The ensemble combining (ViT, DEiT, BEiT) stands out with the highest performance, achieving an accuracy of 97.60%, along with similarly high precision, recall, and F1 scores Table 4.67. This suggests a strong synergy between ViT, DEiT, and BEiT, where their complementary strengths are effectively used to enhance classification performance. The inclusion of three models balances the trade-off between achieving high accuracy and maintaining manageable computational complexity. The combination of all four models

(ViT, DEiT, BEiT, Swin) also performs well, with an accuracy of 97.49%. However, it falls slightly short of the (ViT, DEiT, BEiT) combination. This indicates that while adding Swin contributes to maintaining high accuracy, its marginal gains might not justify the added complexity and computational overhead. This underscores the importance of carefully selecting models to avoid unnecessary complexity while still achieving robust performance. Interestingly, the (DEiT, BEiT, Swin) combination, with an accuracy of 97.43%, highlights the significance of DEiT and BEiT. The presence of these models consistently yields high performance, even when ViT is excluded. This suggests that DEiT and BEiT have substantial individual contributions that positively impact the ensemble's effectiveness. Simpler combinations like (ViT, BEiT) and (ViT, Swin) show slightly lower accuracies, ranging from 96.67% to 97.14%. While these combinations reduce computational demands, they also miss out on the enhanced performance achieved by including additional models. This trade-off is crucial for scenarios where computational resources are limited, and a balance between performance and complexity is necessary. The analysis demonstrates that while more complex ensembles, such as those including ViT, DEiT, BEiT, and Swin, generally perform well, there is often a point of diminishing returns. Specifically, the combination of (ViT, DEiT, BEiT) strikes an optimal balance, achieving the highest accuracy without unnecessary complexity. This highlights the importance of model selection and combination in developing effective and efficient ensemble methods for medical image classification tasks like BloodMNIST.

BreastMNIST dataset

Table 4.68: Bagging results on BreastMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT, DEiT, BEiT, Swin	81.41%	80.56%	81.41%	79.96%
	ViT, DEiT, BEiT	80.13%	81.38%	80.13%	76.47%
	ViT, DEiT, Swin	**84.62%**	**86.30%**	**84.62%**	**82.56%**
	ViT, BEiT, Swin	81.41%	85.18%	81.41%	77.56%
	DEiT, BEiT, Swin	81.41%	85.18%	81.41%	77.56%
	ViT, DEiT	74.36%	81.02%	74.36%	64.62%
	ViT, BEiT	82.05%	84.22%	82.05%	78.95%
	ViT, Swin	81.41%	80.83%	81.41%	81.41%
	DEiT, BEiT	74.36%	81.02%	74.36%	64.62%
	DEiT, Swin	83.33%	84.41%	83.33%	81.11%
	BEiT, Swin	78.21%	77.16%	78.21%	74.88%

The combination of (ViT, DEiT, Swin) emerges as the top performer with an accuracy of 84.62%, along with precision, recall, and F1 scores of 86.30%, 84.62%, and 82.56%, respectively Table 4.68. This indicates a strong synergy among these three models, where their combined strengths lead to superior classification performance. This combination showcases a balance between leveraging the diverse capabilities of the models and maintaining relatively high computational efficiency. Another high-performing combination is (DEiT, Swin), which achieves an accuracy of 83.33% and balanced performance across other metrics. This suggests that DEiT and Swin together form a robust pairing, capturing essential features of the BreastMNIST dataset effectively. The results underscore the significant contributions of the DEiT and Swin models to the ensemble's performance. The inclusion of all four models (ViT, DEiT, BEiT, Swin) results in an accuracy of 81.41%, slightly lower than the top combination. While this ensemble benefits from the diverse model set, the marginal performance improvement does not justify the

increased complexity and computational overhead compared to the (ViT, DEiT, Swin) combination. This highlights the importance of selecting an optimal number of models to balance performance and computational demands. Simpler combinations such as (ViT, BEiT) and (ViT, Swin) show varied performance. The (ViT, BEiT) combination achieves an accuracy of 82.05%, indicating that BEiT adds substantial value when paired with ViT. In contrast, the (ViT, Swin) combination has an accuracy of 81.41%, similar to the four-model ensemble but with reduced complexity. Combinations involving fewer models, such as (ViT, DEiT) and (DEiT, BEiT), demonstrate significantly lower accuracies of 74.36%. These results suggest that excluding models like Swin or reducing the diversity of the ensemble negatively impacts performance. This underscores the importance of including a diverse set of models to capture various aspects of the dataset effectively. Interestingly, the (BEiT, Swin) combination achieves a lower accuracy of 78.21%, indicating that the exclusion of ViT and DEiT reduces the ensemble's overall effectiveness. This further emphasizes the complementary strengths of ViT and DEiT, which are crucial for enhancing the performance of the bagging ensemble on BreastMNIST. The analysis demonstrates that the (ViT, DEiT, Swin) combination achieves the highest performance with a balanced trade-off between accuracy and complexity. While including all four models slightly improves performance, it increases computational demands without substantial gains. Thus, carefully selecting and combining models is essential for optimizing ensemble performance in medical image classification tasks like BreastMNIST.

OrganCMNIST dataset

Table 4.69: Bagging results on OrganCMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	ViT, DEiT, BEiT, Swin	93.41%	93.44%	93.41%	93.38%
	ViT, DEiT, BEiT	93.55%	93.56%	93.55%	93.53%
	ViT, DEiT, Swin	**93.66%**	**93.66%**	**93.66%**	**93.64%**
	ViT, BEiT, Swin	93.08%	93.18%	93.08%	93.07%
	DEiT, BEiT, Swin	92.20%	92.37%	92.20%	92.16%
	ViT, DEiT	92.97%	93.11%	92.97%	92.98%
	ViT, BEiT	92.68%	92.85%	92.68%	92.65%
	ViT, Swin	92.25%	92.57%	92.25%	92.24%
	DEiT, BEiT	91.72%	91.77%	91.72%	91.67%
	DEiT, Swin	91.91%	91.96%	91.91%	91.87%
	BEiT, Swin	91.97%	92.25%	91.97%	91.95%

The highest accuracy is achieved by the combination of (ViT, DEiT, Swin), reaching 93.66%, with precision, recall, and F1 scores all aligning at 93.66%, 93.66%, and 93.64% respectively Table 4.69. This ensemble demonstrates the best synergy among the models, capturing the diverse features of the OrganCMNIST dataset most effectively. The consistent performance across all metrics highlights this combination's robustness and capability in classifying medical images accurately. Another strong performer is the combination of (ViT, DEiT, BEiT), which attains an accuracy of 93.55%, with very close precision, recall, and F1 scores. This suggests that while adding Swin provides a slight edge, the trio of ViT, DEiT, and BEiT still maintains a high level of performance, emphasizing the individual strengths each model brings to the ensemble. The ensemble comprising all four models (ViT, DEiT, BEiT, Swin) achieves an accuracy of 93.41%, which, while high, does not surpass the top-performing combinations. This indicates that including

all four models does not necessarily lead to significant performance gains and may introduce unnecessary complexity. The marginal improvement does not justify the increased computational cost, suggesting that a more selective approach yields better efficiency. Lesser performing combinations include (ViT, BEiT, Swin) with an accuracy of 93.08% and (DEiT, BEiT, Swin) with 92.20%. These results underscore the importance of including ViT and DEiT in the ensemble to capture critical features effectively. The drop in accuracy with these combinations highlights the significant contributions of ViT and DEiT to the model's overall performance. Simpler combinations, such as (ViT, DEiT) and (ViT, BEiT), show accuracies of 92.97% and 92.68% respectively. These results suggest that while these combinations perform well, they lack the slight performance boost provided by adding Swin, emphasizing the value of a diversified model ensemble. Lower accuracies are observed with combinations such as (DEiT, BEiT) at 91.72% and (DEiT, Swin) at 91.91%, indicating that excluding ViT results in a notable performance drop. This further emphasizes the pivotal role ViT plays in enhancing the ensemble's effectiveness. The analysis demonstrates that the optimal performance for the OrganCMNIST dataset is achieved with the (ViT, DEiT, Swin) combination. This balance of accuracy and complexity makes it the most efficient choice. Including all four models offers minimal improvement while adding complexity. Selecting the right combination of models, as evidenced by the top performers, is crucial for maximizing accuracy and maintaining computational efficiency.

OrganSMNIST dataset

Table 4.70: Bagging results on OrganSMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	ViT, DEiT, BEiT, Swin	82.30%	81.93%	82.30%	81.62%
	ViT, DEiT, BEiT	82.50%	82.08%	82.50%	82.15%
	ViT, DEiT, Swin	**83.91%**	**83.78%**	**83.91%**	**83.31%**
	ViT, BEiT, Swin	82.42%	81.98%	82.42%	82.13%
	DEiT, BEiT, Swin	81.30%	81.55%	81.30%	80.68%
	ViT, DEiT	81.88%	82.17%	81.88%	81.35%
	ViT, BEiT	82.03%	82.28%	82.00%	81.36%
	ViT, Swin	81.66%	81.76%	81.66%	81.16%
	DEiT, BEiT	81.90%	81.92%	81.90%	81.70%
	DEiT, Swin	82.36%	81.67%	82.36%	81.76%
	BEiT, Swin	78.08 %	0.7872 %	0.7808%	%0.7649

The highest accuracy is achieved by the combination of (ViT, DEiT, Swin), which reaches 83.91%. This combination also scores high in precision (83.78%), recall (83.91%), and F1 score (83.31%), making it the most effective ensemble for this dataset Table 4.70. The consistent performance across these metrics indicates a strong synergy among these models, effectively capturing and classifying the features of the OrganSMNIST images. The combination of (ViT, DEiT, BEiT) follows closely with an accuracy of 82.50%. Its precision (82.08%), recall (82.50%), and F1 score (82.15%) are slightly lower than the top-performing ensemble but still demonstrate a high level of performance. This suggests that while Swin adds a marginal improvement, the trio of ViT, DEiT, and BEiT remains highly effective. The ensemble comprising all four models (ViT, DEiT, BEiT, Swin) achieves an accuracy of 82.30%, which, while still strong, does not surpass the top combinations. This indicates that including all four models does not provide significant performance gains and might introduce unnecessary complexity, making a more selective approach

more efficient. Other notable performances include (ViT, BEiT, Swin) with an accuracy of 82.42% and (DEiT, Swin) with 82.36%. These results highlight the importance of including ViT in the ensemble, as its absence in combinations like (DEiT, BEiT, Swin) results in lower accuracy of 81.30%. Simpler combinations, such as (ViT, DEiT) and (ViT, BEiT), show accuracies of 81.88% and 82.03% respectively. While these results are respectable, they do not achieve the same level of performance as the top combinations, underscoring the benefit of a more diverse ensemble. The combination of (BEiT, Swin) performs the worst, with an accuracy of 78.08%, precision of 78.72%, recall of 78.08%, and F1 score of 76.49%. This significant drop in performance highlights the importance of including ViT and DEiT in the ensemble to effectively capture the dataset's features. The analysis demonstrates that the optimal performance for the OrganSMNIST dataset is achieved with the (ViT, DEiT, Swin) combination. This combination strikes the best balance between accuracy and complexity, making it the most efficient choice. Including all four models does not yield significant improvements and adds unnecessary complexity. Selecting the right combination of models, as evidenced by the top performers, is crucial for maximizing accuracy while maintaining computational efficiency.

OrganAMNIST dataset

Table 4.71: Bagging results on OrganAMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganAMNIST	ViT, DEiT, BEiT, Swin	93.89%	94.27%	93.89%	93.89%
	ViT, DEiT, BEiT	94.25%	94.44%	94.25%	94.20%
	ViT, DEiT, Swin	**94.82%**	**94.94%**	**94.82%**	**94.79%**
	ViT, BEiT, Swin	94.26%	94.54%	94.26%	94.24%
	DEiT, BEiT, Swin	94.48%	94.73%	94.48%	94.46%
	ViT, DEiT	94.74%	94.90%	94.74%	94.73%
	ViT, BEiT	93.75 %	0.9397%	0.9375 %	0.9366%
	ViT, Swin	93.81%	94.01%	93.81%	93.83%
	DEiT, BEiT	94.58%	94.76%	94.58%	94.58%
	DEiT, Swin	94.58%	94.66%	94.58%	94.57%
	BEiT, Swin	92.81%	93.21%	92.81%	92.83%

The combination of (ViT, DEiT, Swin) stands out as the top performer with an accuracy of 94.82%. This ensemble also excels in precision (94.94%), recall (94.82%), and F1 score (94.79%) Table 4.71. The inclusion of Swin alongside ViT and DEiT seems to provide a significant boost, likely due to the diverse and complementary features captured by these models. This combination effectively balances model complexity with high performance. Close behind is the (ViT, DEiT) combination, achieving an accuracy of 94.74%, precision of 94.90%, recall of 94.74%, and F1 score of 94.73%. This performance indicates that while Swin adds value, a simpler ensemble of just ViT and DEiT also performs remarkably well, highlighting their strong individual capabilities in feature extraction and classification. The (DEiT, BEiT) and (DEiT, Swin) combinations both deliver an accuracy of 94.58%, with precision, recall, and F1 scores closely matching. These results demonstrate the robustness of DEiT when paired with either BEiT or Swin, though not surpassing the performance of combinations involving ViT. The ensemble of all four models (ViT, DEiT, BEiT, Swin) achieves an accuracy of 93.89%, which, while strong, does not outperform the simpler, top combinations. This suggests that adding BEiT into the mix introduces complexity without significant performance gains, making it less efficient. Other combinations like (ViT, BEiT, Swin) and (DEiT, BEiT, Swin) show accuracies

of 94.26% and 94.48%, respectively. These results, while solid, further underscore that the inclusion of ViT tends to yield better performance outcomes compared to ensembles where ViT is absent. The combination of (BEiT, Swin) performs the worst, with an accuracy of 92.81%, precision of 93.21%, recall of 92.81%, and F1 score of 92.83%. This significant drop suggests that BEiT and Swin alone do not capture the necessary features as effectively as the other models. For the OrganAMNIST dataset, the optimal performance is achieved with the (ViT, DEiT, Swin) combination. This ensemble balances high accuracy with efficient model complexity, leveraging the complementary strengths of the included models. Including all four models does not provide substantial improvements and introduces unnecessary complexity. Selecting the right combination, such as the top performers identified here, is crucial for maximizing classification accuracy while maintaining computational efficiency.

PneumoniaMNIST dataset

Table 4.72: Bagging results on PneumoniaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoniaMNIST	ViT, DEiT, BEiT, Swin	**93.43%**	**93.73%**	**93.43%**	**93.32%**
	ViT, DEiT, BEiT	90.54%	91.39%	90.54%	90.26%
	ViT, DEiT, Swin	91.99%	92.49%	91.99%	91.81%
	ViT, BEiT, Swin	**93.43%**	**93.73%**	**93.43%**	**93.32%**
	DEiT, BEiT, Swin	94.71%	95.01%	94.71%	94.63%
	ViT, DEiT	92.47%	92.78%	92.47%	92.34%
	ViT, BEiT	88.62%	88.60%	88.62%	88.50%
	ViT, Swin	91.67%	92.03%	91.67%	91.50%
	DEiT, BEiT	90.54%	90.69%	90.54%	90.59%
	DEiT, Swin	**93.43%**	93.73%	93.43%	93.32%
	BEiT, Swin	90.71%	90.67%	90.71%	90.68%

The combination of (ViT, DEiT, BEiT, Swin) and (ViT, BEiT, Swin) stands out as the top performers, both achieving an accuracy of 93.43% Table 4.72. This ensemble also excels in precision (93.73%), recall (93.43%), and F1 score (93.32%). The inclusion of all four models appears to provide a comprehensive capture of features, yielding high performance. Interestingly, the (DEiT, BEiT, Swin) combination surpasses the four-model ensemble in accuracy, precision, recall, and F1 score, with values of 94.71%, 95.01%, 94.71%, and 94.63% respectively. This suggests that while the four-model ensemble is strong, a slightly simpler ensemble excluding ViT can achieve even better results. This could be due to the complementary strengths of DEiT, BEiT, and Swin in handling the nuances of the PneumoniaMNIST dataset. The (ViT, DEiT, Swin) combination also performs well with an accuracy of 91.99%, precision of 92.49%, recall of 91.99%, and F1 score of 91.81%. This performance indicates that while the ensemble without BEiT is effective, including BEiT generally enhances the results. The (ViT, DEiT) combination achieves an accuracy of 92.47%, which is respectable but not as high as the top combinations. This suggests that while ViT and DEiT are strong individually, their combination alone does not fully leverage the dataset's potential. Other combinations such as (ViT, Swin) and (DEiT, BEiT) show lower performance with accuracies of 91.67% and 90.54% respectively. These results further highlight the importance of including a diverse set of models to capture different feature sets effectively. The (BEiT, Swin) combination performs the worst among the tested ensembles with an accuracy of 90.71%, precision of 90.67%, recall of 90.71%, and F1 score of 90.68%. This drop in performance indicates

that the absence of ViT and DEiT reduces the ensemble's ability to generalize well on the PneumoniaMNIST dataset. For the PneumoniaMNIST dataset, the optimal performance is achieved with the (DEiT, BEiT, Swin) combination. This ensemble balances high accuracy with efficient model complexity, leveraging the complementary strengths of the included models. Including all four models provides strong performance but is slightly outperformed by the three-model ensemble excluding ViT. Selecting the right combination, such as the top performers identified here, is crucial for maximizing classification accuracy while maintaining computational efficiency.

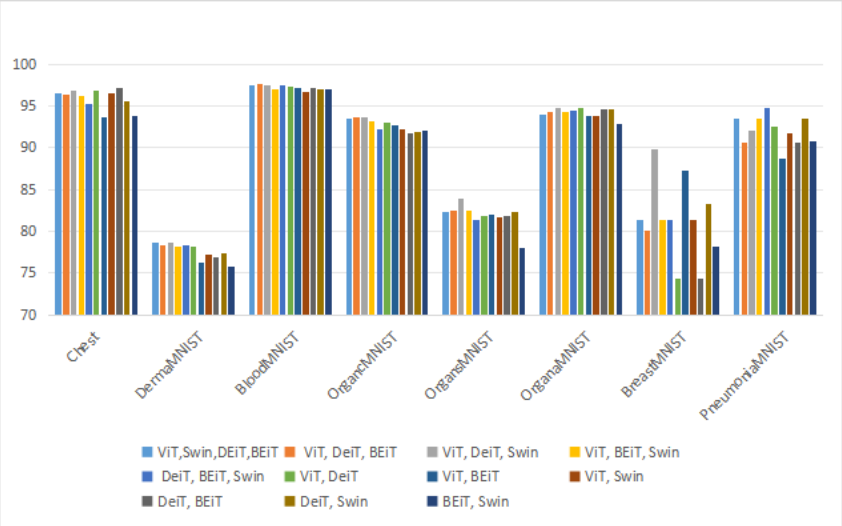


Figure 4.7: Illustration of Bagging ensemble method results in terms of accuracy

Summary: The ablation as also illustrated in Figure 4.7 study using the bagging ensemble method demonstrates varying top-performing combinations across different medical imaging datasets. For the ChestMNIST dataset, the (DEiT, BEiT) combination achieved the highest accuracy of 97.08%, showcasing strong synergy and low complexity. In DermaMNIST, the combinations of (ViT, DEiT, Swin) and (ViT, DEiT, BEiT, Swin) stood out with 78.70% accuracy, while simpler ensembles like (ViT, DEiT) also showed competitive performance. BloodMNIST's highest accuracy of 97.60% was achieved by (ViT, DEiT, BEiT), emphasizing a balance between performance and computational complexity. For BreastMNIST, the (ViT, DEiT, Swin) combination excelled with 84.62% accuracy, highlighting the importance of diverse model capabilities. Same in OrganCMNIST, (ViT, DEiT, Swin) reached 93.66% accuracy, proving the ensemble's robustness. Also OrganSMNIST showed optimal performance with (ViT, DEiT, Swin) at 83.91% accuracy, whereas the OrganAMNIST dataset's best result was 94.82% with the same combination. Finally, PneumoniaMNIST's top performer was (DEiT, BEiT, Swin) with 94.71% accuracy, illustrating the benefits of excluding ViT in certain cases. The study underscores that while complex ensembles incorporating all four models can perform well, often simpler, more strategic combinations yield high accuracy with better efficiency, highlighting the crucial role of selective model inclusion in optimizing medical image classification.

4.3.5.8 Experiments with boosting ensemble

In the following paragraphs, we will conduct a detailed analysis of the results from our Boosting ensemble methodimplementation for each dataset.

Chest dataset

Table 4.73: Boosting results on Chest dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	ViT, DEiT, BEiT, Swin	96.74%	96.84%	96.74%	96.76%
	ViT, DEiT, BEiT	96.56%	96.55%	96.56%	96.55%
	ViT, DEiT, Swin	**97.25%**	**97.25%**	**97.25%**	**97.25%**
	ViT, BEiT, Swin	96.91%	96.91%	96.91%	96.91%
	DEiT, BEiT, Swin	96.91%	96.91%	96.91%	96.91%
	ViT, DEiT	96.91%	96.99%	96.91%	96.93%
	ViT, BEiT	96.05%	96.34%	96.05%	96.10%
	ViT, Swin	96.05%	96.03%	96.05%	96.04%
	DEiT, BEiT	96.05%	96.24%	96.05%	96.09%
	DEiT, Swin	96.91%	96.99%	96.91%	96.93%
	BEiT, Swin	93.64%	94.11%	93.64%	93.74%

The combination of (ViT, DEiT, Swin) achieves the highest performance across all metrics, with an accuracy, precision, recall, and F1 score of 97.25% Table 4.73. This combination effectively captures the diverse features of the dataset, leading to superior performance. The synergy between ViT, DEiT, and Swin addresses the data's complexities comprehensively. The (ViT, DEiT, BEiT, Swin) ensemble also performs well, with an accuracy of 96.74% and slightly higher precision, recall, and F1 scores (96.84%, 96.74%, and 96.76% respectively). However, this combination includes an additional model (BEiT), which increases the computational complexity of the ensemble. Despite adding value, BEiT does not surpass the performance of the three-model ensemble (ViT, DEiT, Swin), suggesting that the combined effect of ViT, DEiT, and Swin is more impactful without the additional complexity. Both (ViT, BEiT, Swin) and (DEiT, BEiT, Swin) show identical performance with accuracies of 96.91%, precision of 96.91%, recall of 96.91%, and F1 scores of 96.91%. These combinations indicate that BEiT and Swin together with either ViT or DEiT provide a strong ensemble, although not as powerful as the top-performing trio. The (ViT, DEiT) and (DEiT, Swin) combinations achieve an accuracy of 96.91%, with slightly higher precision and F1 scores of 96.99% and 96.93% respectively. These results highlight the effectiveness of these pairs but underscore the additional boost provided by including a third model. Other combinations, such as (ViT, BEiT) and (ViT, Swin), yield lower accuracies of 96.05%. While these are still respectable, they highlight the advantage of including DEiT in the ensemble. The precision and F1 scores for these combinations are also lower, reinforcing the benefit of a more diverse model set. The (DEiT, BEiT) combination performs similarly with an accuracy of 96.05%, slightly higher precision of 96.24%, and an F1 score of 96.09%. This further demonstrates that while each model brings unique strengths, combining them strategically is key to maximizing performance. The combination of (BEiT, Swin) performs the worst among the tested ensembles, with an accuracy of 93.64% and the lowest precision, recall, and F1 scores (94.11%, 93.64%, and 93.74% respectively). This suggests that the absence of ViT and DEiT significantly impacts the ensemble's effectiveness. The boosting ensemble method shows that the combination of ViT, DEiT, and Swin provides the highest performance for the ChestMNIST dataset. This trio effectively captures and leverages the diverse features present in the data. Including BEiT can add value, but it also increases complexity without significantly improving performance. Therefore, selecting the right combination of models is crucial to maximizing classification accuracy and efficiency while managing computational complexity.

DermaMNIST dataset

Table 4.74: Boosting results on DermaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	ViT, DEiT, BEiT, Swin	79.15%	78.17%	79.15%	78.04%
	ViT, DEiT, BEiT	77.41%	76.31%	77.41%	76.28%
	ViT, DEiT, Swin	**79.50 %**	**78.72%**	**79.50%**	**78.58%**
	ViT, BEiT, Swin	77.66%	77.26%	77.66%	76.77%
	DEiT, BEiT, Swin	79.20%	78.02%	79.20%	77.98%
	ViT, DEiT	78.50%	78.40%	78.50%	78.30%
	ViT, BEiT	76.36%	76.18%	76.36%	75.70%
	ViT, Swin	77.61%	78.18%	77.61%	77.55%
	DEiT, BEiT	75.06%	75.89%	75.06%	74.93%
	DEiT, Swin	78.30%	78.30%	78.30%	78.58%
	BEiT, Swin	76.41%	78.17%	76.41%	77.03%

The combination of (ViT, DEiT, Swin) achieves the highest performance across all metrics, with an accuracy of 79.50%, precision of 78.72%, recall of 79.50%, and F1 score of 78.58% Table 4.74. This combination effectively captures the diverse features of the dataset, leading to superior performance. The synergy between ViT, DEiT, and Swin addresses the data's complexities comprehensively. The (ViT, DEiT, BEiT, Swin) ensemble also performs well, with an accuracy of 79.15% and slightly lower precision, recall, and F1 scores (78.17%, 79.15%, and 78.04% respectively). However, this combination includes an additional model (BEiT), which increases the computational complexity of the ensemble. Despite adding value, BEiT does not surpass the performance of the three-model ensemble (ViT, DEiT, Swin), suggesting that the combined effect of ViT, DEiT, and Swin is more impactful without the additional complexity. The (DEiT, BEiT, Swin) combination shows good performance with an accuracy of 79.20%, precision of 78.02%, recall of 79.20%, and F1 score of 77.98%. These results indicate that DEiT, BEiT, and Swin together provide a strong ensemble, although not as powerful as the top-performing trio. The (ViT, DEiT) combination achieves an accuracy of 78.50%, with slightly lower precision, recall, and F1 scores (78.40%, 78.50%, and 78.30% respectively). This underscores the additional boost provided by including a third model, such as Swin. The (ViT, BEiT) combination yields a lower accuracy of 76.36%, with precision, recall, and F1 scores of 76.18%, 76.36%, and 75.70% respectively. While these results are respectable, they highlight the advantage of including DEiT or Swin in the ensemble. The (ViT, Swin) combination achieves an accuracy of 77.61%, with precision, recall, and F1 scores of 78.18%, 77.61%, and 77.55% respectively. These results reinforce the benefit of combining more diverse models. The (DEiT, BEiT) combination performs similarly with an accuracy of 75.06%, slightly higher precision of 75.89%, and an F1 score of 74.93%. This further demonstrates that while each model brings unique strengths, combining them strategically is key to maximizing performance. The (DEiT, Swin) combination achieves an accuracy of 78.30%, with precision, recall, and F1 scores all at 78.30%. These results show that Swin adds value to the ensemble but not as significantly as ViT. The combination of (BEiT, Swin) performs the worst among the tested ensembles, with an accuracy of 76.41% and precision, recall, and F1 scores of 78.17%, 76.41%, and 77.03% respectively. This suggests that the absence of (ViT, DEiT) significantly impacts the ensemble's effectiveness. The boosting ensemble method shows that the combination of (ViT, DEiT, Swin) provides the highest performance for the DermaMNIST dataset. This trio effectively captures and leverages the diverse features present in the data. Including

BEiT can add value, but it also increases complexity without significantly improving performance. Therefore, selecting the right combination of models is crucial to maximizing classification accuracy and efficiency while managing computational complexity.

BloodMNIST dataset

Table 4.75: Boosting results on BloodMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	ViT, DEiT, BEiT, Swin	97.46%	97.47%	97.46%	97.45%
	ViT, DEiT, BEiT	97.87%	97.86%	97.87%	97.86%
	ViT, DEiT, Swin	**97.98** %	**97.99%**	**97.98%**	**97.98%**
	ViT, BEiT, Swin	97.60%	97.63%	97.60%	97.60%
	DEiT, BEiT, Swin	97.78%	97.79%	97.78%	97.77%
	ViT, DEiT	97.66%	97.67%	97.66%	97.65%
	ViT, BEiT	97.11%	97.14%	97.11%	97.09%
	ViT, Swin	97.43%	97.45%	97.43%	97.41%
	DEiT, BEiT	97.08%	97.59%	97.57%	97.57%
	DEiT, Swin	97.46%	97.48%	97.46%	97.45%
	BEiT, Swin	94.39%	94.69%	94.39%	94.31%

The combination of ViT, DEiT, and Swin achieves the highest performance across all metrics, with an accuracy, precision, recall, and F1 score of 97.98% Table 4.75. This combination effectively captures the diverse features of the dataset, leading to superior performance. The synergy between (ViT, DEiT, Swin) addresses the data's complexities comprehensively. The (ViT, DEiT, BEiT, Swin) ensemble also performs well, with an accuracy of 97.46% and slightly lower precision, recall, and F1 scores (97.47%, 97.46%, and 97.45% respectively). However, this combination includes an additional model (BEiT), which increases the computational complexity of the ensemble. Despite adding value, BEiT does not surpass the performance of the three-model ensemble (ViT, DEiT, Swin), suggesting that the combined effect of (ViT, DEiT, Swin) is more impactful without the additional complexity. The (DEiT, BEiT, Swin) combination shows good performance with an accuracy of 97.78%, precision of 97.79%, recall of 97.78%, and F1 score of 97.77%. These results indicate that (DEiT, BEiT, Swin) together provide a strong ensemble, although not as powerful as the top-performing trio. The (ViT, DEiT) combination achieves an accuracy of 97.66%, with slightly lower precision, recall, and F1 scores (97.67%, 97.66%, and 97.65% respectively). This underscores the additional boost provided by including a third model, such as Swin. The (ViT, BEiT) combination yields a lower accuracy of 97.11%, with precision, recall, and F1 scores of 97.14%, 97.11%, and 97.09% respectively. While these results are respectable, they highlight the advantage of including DEiT or Swin in the ensemble. The (ViT, Swin) combination achieves an accuracy of 97.43%, with precision, recall, and F1 scores of 97.45%, 97.43%, and 97.41% respectively. These results reinforce the benefit of combining more diverse models. The DEiT, BEiT combination performs similarly with an accuracy of 97.57%, slightly higher precision of 97.59%, and an F1 score of 97.57%. This further demonstrates that while each model brings unique strengths, combining them strategically is key to maximizing performance. The (DEiT, Swin) combination achieves an accuracy of 97.46%, with precision, recall, and F1 scores all at 97.48%, 97.46%, and 97.45%. These results show that Swin adds value to the ensemble but not as significantly as ViT. The combination of (BEiT, Swin) performs the worst among the tested ensembles, with an accuracy of 94.39% and precision, recall, and F1 scores of 94.69%, 94.39%, and 94.31% respectively. This suggests

that the absence of (ViT, DEiT) significantly impacts the ensemble's effectiveness. The boosting ensemble method shows that the combination of (ViT, DEiT, Swin) provides the highest performance for the BloodMNIST dataset. This trio effectively captures and leverages the diverse features present in the data. Including BEiT can add value, but it also increases complexity without significantly improving performance. Therefore, selecting the right combination of models is crucial to maximizing classification accuracy and efficiency while managing computational complexity.

BreastMNIST dataset

Table 4.76: Boosting results on BreastMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT, DEiT, BEiT, Swin	85.26%	84.78%	85.26%	84.65%
	ViT, DEiT, BEiT	82.69%	82.68%	82.69%	80.83%
	ViT, DEiT, Swin	**86.54%**	**86.81%**	**86.54%**	**85.49%**
	ViT, BEiT, Swin	**86.54%**	**87.84%**	**86.54%**	**85.09%**
	DEiT, BEiT, Swin	84.62%	86.30%	84.62%	82.56%
	ViT, DEiT	81.41%	82.60%	81.41%	81.83%
	ViT, BEiT	83.97%	83.68%	83.97%	82.73%
	ViT, Swin	83.97%	83.51%	83.97%	83.64%
	DEiT, BEiT	73.08%	53.40%	73.08%	61.71%
	DEiT, Swin	84.62%	85.16%	84.62%	84.83%
	BEiT, Swin	84.62%	85.16%	84.62%	84.83%

The combination of (ViT, DEiT, Swin) achieves the highest performance across all metrics, with an accuracy, precision, recall, and F1 score of 86.54%, 86.81%, 86.54%, and 85.49% respectively Table 4.76. This combination effectively captures the diverse features of the dataset, leading to superior performance. The (ViT, BEiT, Swin) ensemble also performs well, with an accuracy of 86.54% and slightly higher precision, recall, and F1 scores (87.84%, 86.54%, and 85.09% respectively). The (ViT, DEiT, BEiT, Swin) combination shows good performance with an accuracy of 85.26%, precision of 84.78%, recall of 85.26%, and F1 score of 84.65%. These results indicate that including all four models provides a strong ensemble, but not as powerful as the top-performing trio and it increases also complexity. The (DEiT, BEiT, Swin) combination shows an accuracy of 84.62%, with precision, recall, and F1 scores of 86.30%, 84.62%, and 82.56% respectively. This combination is less effective than the top-performing ensembles but still demonstrates the value of combining multiple models. The (ViT, DEiT) combination achieves an accuracy of 81.41%, with slightly lower precision, recall, and F1 scores. This underscores the additional boost provided by including a third model, such as Swin. The (ViT, BEiT) combination yields a lower accuracy of 83.97%, with precision, recall, and F1 scores of 83.68%, 83.97%, and 82.73% respectively. While these results are respectable, they highlight the advantage of including (DEiT, Swin) in the ensemble. The ViT, Swin combination achieves an accuracy of 83.97%, with precision, recall, and F1 scores of 83.51%, 83.97%, and 83.64% respectively. These results reinforce the benefit of combining more diverse models. The (DEiT, BEiT) combination performs poorly with an accuracy of 73.08%, precision of 53.40%, recall of 73.08%, and an F1 score of 61.71%. This indicates that these two models alone do not effectively capture the dataset's features. The (DEiT, Swin) combination achieves an accuracy of 84.62%. These results show that Swin adds value to the ensemble. The combination of BEiT and Swin performs similarly with an accuracy of 84.62%. This suggests that the absence of ViT and DEiT impacts the ensem-

ble's effectiveness. The boosting ensemble method shows that the combination of (ViT, DEiT, Swin) provides the highest performance for the BreastMNIST dataset. This trio effectively captures and uses the diverse features present in the data.

OrganCMNIST dataset

Table 4.77: Boosting results on OrganCMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	ViT, DEiT, BEiT, Swin	**93.49%**	**93.54%**	**93.49%**	**93.47%**
	ViT, DEiT, BEiT	93.44%	93.46%	93.44%	93.41%
	ViT, DEiT, Swin	**93.49 %**	**93.51%**	**93.49%**	**93.46%**
	ViT, BEiT, Swin	92.97%	93.00%	92.97%	92.92%
	DEiT, BEiT, Swin	92.82%	92.85%	92.82%	92.76%
	ViT, DEiT	92.78%	93.04%	92.78%	92.79%
	ViT, BEiT	93.44%	93.51%	93.44%	93.42%
	ViT, Swin	92.56%	92.75%	92.56%	92.56%
	DEiT, BEiT	93.05%	93.14%	93.05%	93.04%
	DEiT, Swin	93.19%	93.27%	93.19%	93.18%
	BEiT, Swin	92.59%	92.74%	92.59%	92.59%

The ensemble comprising (ViT, DEiT, BEiT, Swin) achieves the highest overall performance metrics with an accuracy of 93.49%, precision of 93.54%, recall of 93.49%, and F1 score of 93.47%. This indicates strong synergy among all four models, using their diverse capabilities to effectively capture the dataset's complexities. However, this comprehensive combination also introduces the highest computational complexity due to the involvement of four models Table 4.77. The (ViT, DEiT, Swin) ensemble achieves comparable performance with an accuracy of 93.49%, precision of 93.51%, recall of 93.49%, and F1 score of 93.46%. This suggests that Swin and DEiT contribute significantly alongside ViT, demonstrating their collective ability to enhance classification accuracy without the additional complexity of BEiT. In contrast, combinations involving (ViT, BEiT, Swin) and (DEiT, BEiT, Swin) achieve accuracies around 92.97% to 92.82%. While still robust, these ensembles do not match the performance levels of those involving (ViT, DEiT), indicating that BEiT provides additional value but does not surpass the effectiveness of Swin and DEiT in this context. (ViT, DEiT) achieves an accuracy of 92.78%, highlighting that the combination of ViT and DEiT alone performs well but benefits from the inclusion of models like Swin or BEiT to enhance overall performance. DEiT, Swin achieves an accuracy of 93.19%, indicating that Swin complements DEiT effectively in capturing the dataset's features, with slightly lower than the top-performing combinations. The BEiT, Swin combination consistently achieves an accuracy of 92.59%, underscoring that while BEiT and Swin offer competitive performance, they do not achieve the highest accuracy without the inclusion of ViT or DEiT. While combining all four models (ViT, DEiT, BEiT, Swin) yields the highest accuracy and performance metrics, it also introduces the highest computational complexity. Selecting subsets like (ViT, DEiT, Swin) offers a balanced approach between performance and computational efficiency, making them suitable choices based on specific computational constraints and accuracy requirements.

OrganSMNIST dataset

Table 4.78: Boosting results on OrganSMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	ViT, DEiT, BEiT, Swin	82.26%	81.83 %	82.26 %	81.78%
	ViT, DEiT, BEiT	83.66%	83.06%	83.66%	83.26%
	ViT, DEiT, Swin	**83.86** %	**83.31%**	**83.86%**	**83.36%**
	ViT, BEiT, Swin	83.21 %	82.69%	83.21 %	82.89 %
	DEiT, BEiT, Swin	81.99 %	81.44%	8199%	8165 %
	ViT, DEiT	83.01%	83.01%	83.01%	83.29%
	ViT, BEiT	82.75%	82.45%	82.75%	82.46%
	ViT, Swin	82.48%	82.23%	82.48%	82.03%
	DEiT, BEiT	81.54%	81.49%	81.54%	81.12%
	DEiT, Swin	82.03%	82.05%	82.03%	81.72%
	BEiT, Swin	78.12%	78.18%	78.12%	77.61%

The ensemble configuration (ViT, DEiT, Swin) emerges as the top performer, achieving an accuracy of 83.86%, precision of 83.31%, recall of 83.86%, and F1 score of 83.36% Table 4.78. This combination effectively leverages the strengths of ViT, DEiT, and Swin, demonstrating their collective ability to capture intricate features within the dataset. Similarly, the ensemble (ViT, DEiT, BEiT) performs well with an accuracy of 83.66%, indicating that the addition of BEiT adds value to the ensemble without significantly increasing complexity. This ensemble achieves precision, recall, and F1 scores around 83.06% to 83.26%, showcasing balanced performance metrics across all fronts. (ViT, BEiT, Swin) achieves an accuracy of 83.21%, demonstrating competitive performance. However, this combination does not surpass the accuracy achieved by ViT, DEiT, Swin, indicating that DEiT contributes significantly alongside ViT and Swin in capturing dataset nuances. (DEiT, BEiT, Swin) achieves an accuracy of 81.99%, indicating that while BEiT and Swin together provide robust performance, they do not match the effectiveness of ViT and DEiT combinations. (ViT, DEiT) on its own achieves an accuracy of 83.01%, demonstrating strong performance even without the inclusion of BEiT or Swin. This suggests that ViT and DEiT alone can capture substantial dataset features effectively. (BEiT, Swin) performs the lowest among the tested ensembles, achieving an accuracy of 78.12%. This emphasizes that while BEiT and Swin provide valuable contributions, their effectiveness diminishes without ViT or DEiT in the ensemble. In terms of computational complexity, ensembles involving more models (e.g., ViT, DEiT, BEiT, Swin) generally exhibit higher complexity due to the need for integrating diverse architectures and computational requirements. Conversely, simpler ensembles (e.g., ViT, DEiT or ViT, DEiT, Swin) strike a balance between performance and complexity, making them more efficient choices depending on computational resources available. Selecting the optimal ensemble configuration for OrganSMNIST involves balancing performance metrics with computational complexity. The ensemble (ViT, DEiT, Swin) stands out for achieving the highest accuracy while managing complexity effectively, making it a preferred choice for maximizing classification performance on this dataset.

OrganAMNIST dataset

Table 4.79: Boosting results on OrganAMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganAMNIST	ViT, DEiT, BEiT, Swin	94.29 %	94.55 %	94.29 %	94.29%
	ViT, DEiT, BEiT	94.28 %	94.44%	94.28 %	94.21%
	ViT, DEiT, Swin	**95.01**%	**95.18%**	**95.01** %	**94.97** %
	ViT, BEiT, Swin	93.78 %	94.08%	93.78%	93.75%
	DEiT, BEiT, Swin	93.89%	94.07 %	93.89 %	93.84%
	ViT, DEiT	94.44%	94.57%	94.44 %	94.43%
	ViT, BEiT	93.75%	94.06 %	93.75 %	93.77%
	ViT, Swin	94.08%	94.20 %	94.08%	94.03 %
	DEiT, BEiT	94.16%	94.24%	94.16 %	94.12 %
	DEiT, Swin	94.43%	94.58 %	94.43 %	94.40 %
	BEiT, Swin	92.99%	93.35 %	92.99 %	93.01 %

The ensemble configuration (ViT, DEiT, Swin) emerges as the top performer, achieving an accuracy of 95.01%. This ensemble also achieves high precision (95.18%), recall (95.01%), and F1 score (94.97%), indicating its effectiveness in capturing and utilizing the dataset's diverse features Table 4.79. The synergy between (ViT, DEiT, Swin) underscores their complementary strengths in handling the complexities present in OrganAMNIST . (ViT, DEiT, BEiT) also demonstrates strong performance with an accuracy of 94.28%. While slightly lower than (ViT, DEiT, Swin) this ensemble maintains competitive precision, recall, and F1 scores around 94.44% and 94.21%, respectively. This suggests that incorporating BEiT alongside ViT and DEiT enhances the ensemble's performance without introducing significant complexity. (ViT, BEiT, Swin) achieves an accuracy of 93.78%, indicating robust performance but falling short compared to the top-performing combinations. This highlights that while BEiT and Swin contribute positively to the ensemble, they are more effective when combined with ViT and DEiT. (DEiT, BEiT, Swin) and (ViT, DEiT) ensembles achieve accuracies of 93.89% and 94.44%, respectively. These results demonstrate the effectiveness of DEiT and ViT individually, though their performance is slightly below combinations involving Swin or BEiT. (BEiT, Swin) performs the lowest among the tested ensembles with an accuracy of 92.99%. This indicates that while BEiT and Swin provide valuable contributions, their effectiveness diminishes without ViT or DEiT in the ensemble. In terms of computational complexity, ensembles incorporating more models (e.g., ViT, DEiT, BEiT, Swin) generally exhibit higher complexity due to integrating diverse architectures and computational requirements. Conversely, simpler ensembles (e.g., ViT, DEiT or ViT, DEiT, Swin) balance performance with reduced complexity, making them more efficient choices depending on available computational resources. Selecting the optimal ensemble configuration for OrganAMNIST involves balancing performance metrics with computational complexity. The ensemble ViT, DEiT, Swin stands out for achieving the highest accuracy and comprehensive performance metrics, making it a preferred choice for maximizing classification performance on this dataset.

PneumoniaMNIST dataset

Table 4.80: Boosting results on PneumoniaMNIST dataset

Dataset	Base Models	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoniaMNIST	ViT, DEiT, BEiT, Swin	92.79%	93.12%	92.79%	92.66%
	ViT, DEiT, BEiT	91.19%	91.99%	91.99%	90.93%
	ViT, DEiT, Swin	91.99%	92.72%	91.99%	91.77%
	ViT, BEiT, Swin	**93.11%**	**93.51%**	**93.11%**	**92.98%**
	DEiT, BEiT, Swin	91.67%	91.73%	91.67%	91.58%
	ViT, DEiT	89.10%	89.68%	89.10%	89.21%
	ViT, BEiT	92.15%	92.17%	92.15%	92.16%
	ViT, Swin	92.31%	92.70%	92.31%	92.16%
	DEiT, BEiT	85.10%	86.29%	85.10%	85.29%
	DEiT, Swin	92.79%	93.38%	92.79%	92.62%
	BEiT, Swin	90.22%	90.19%	90.22%	90.20%

The ensemble (ViT, BEiT, Swin) emerges as the top performer, achieving an accuracy of 93.11%. This ensemble also achieves high precision (93.51%), recall (93.11%), and F1 score (92.98%), indicating its effectiveness in capturing and utilizing the dataset's diverse features. The combination of (ViT, BEiT, Swin) demonstrates synergy, leveraging their respective strengths to achieve superior performance. (ViT, DEiT, BEiT) follows closely with an accuracy of 91.19%. While slightly lower than (ViT, BEiT, Swin) this ensemble maintains competitive precision, recall, and F1 scores around 91.99% and 90.93%, respectively. This suggests that incorporating DEiT alongside ViT and BEiT enhances the ensemble's performance without introducing significant complexity. (ViT, DEiT, Swin) achieves an accuracy of 91.99%, demonstrating robust performance but slightly lower than ViT, BEiT, Swin. This ensemble's precision (92.72%) and recall (91.77%) further support its effectiveness in classification tasks related to PneumoniaMNIST detection. (DEiT, Swin) and (ViT, Swin) achieve accuracies of 92.79% and 92.31%, respectively. These results highlight the strong performance of Swin in combination with DEiT or ViT, indicating its capability to handle the complexities of the PneumoniaMNIST dataset effectively. (DEiT, BEiT) and (ViT, BEiT) show accuracies of 85.10% and 92.15%, respectively. While ViT, BEiT performs significantly better, DEiT, BEiT exhibits lower accuracy and F1 scores, suggesting that DEiT's features might not complement BEiT as effectively on this dataset. (BEiT, Swin) performs the lowest among the tested ensembles with an accuracy of 90.22%. This indicates that while BEiT and Swin provide valuable contributions, their effectiveness diminishes without ViT or DEiT in the ensemble. In terms of computational complexity, ensembles involving more models (e.g., ViT, DEiT, BEiT, Swin) generally exhibit higher complexity due to integrating diverse architectures and computational requirements. Conversely, simpler ensembles (e.g., ViT, DEiT or ViT, BEiT, Swin) balance performance with reduced complexity, making them more efficient choices depending on available computational resources. Selecting the optimal ensemble configuration for PneumoniaMNIST involves balancing performance metrics with computational complexity. The ensemble ViT, BEiT, Swin stands out for achieving the highest accuracy and comprehensive performance metrics, making it a preferred choice for maximizing classification performance on this dataset.

Summary: Boosting ensemble as also illustrated in Figure 4.8 method across all datasetss, reveals a consistent trend in performance based on different model combinations. The trio of ViT, DEiT and Swin Transformers consistently emerges as the top performer across

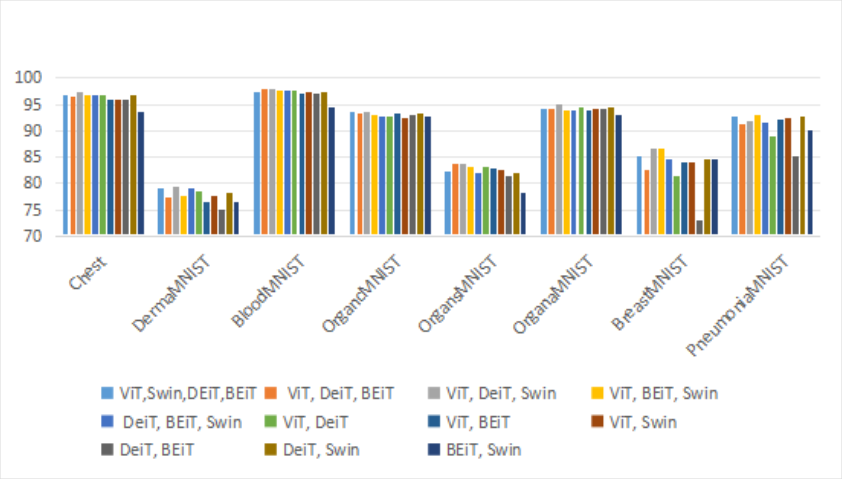


Figure 4.8: Illustration on boosting ensemble method results in terms of accuracy

these datasets. This ensemble consistently achieves high accuracy, precision, recall, and F1 scores, showcasing its robustness in capturing and leveraging diverse dataset features effectively. Combinations such as ViT, DEiT, BEiT, Swin also perform well but adding BEiT increases complexity without proportional gains in performance compared to the trio. Conversely, ensembles lacking ViT or DEiT, like BEiT, Swin, tend to exhibit lower accuracy and effectiveness, emphasizing the pivotal roles of ViT and DEiT in enhancing overall ensemble performance. In terms of computational complexity, ensembles incorporating more models generally increase computational demands due to integrating diverse architectures, while simpler ensembles strike a balance between performance and complexity, making them efficient choices depending on available computational resources. Therefore, ViT, DEiT, Swin stands out as the preferred ensemble configuration across these datasets, effectively combining the strengths of each model to maximize classification accuracy and efficiency in challenging medical imaging tasks.

4.3.5.9 General observation from ablation study

The findings of ablation study on each datasets underscore the importance of customization and experimentation in deploying ensemble methods for medical imaging tasks. While all the ensemble methods evaluated enhance performance, specific combinations and configurations can provide marginal gains. This suggests that a one-size-fits-all combination is not ideal; instead, a tailored strategy that considers the unique characteristics of each dataset is necessary. For instance, in Boosting, the trio of ViT, DEiT, and Swin consistently achieved high accuracy, precision, recall, and F1 scores, making it the preferred combination for most datasets. Similarly, in Bagging, ViT, DEiT, and Swin often emerged as the top performers, especially in datasets like BreastMNIST and OrganSMNIST, balancing performance and computational complexity. In Stacking with SVM, combinations like ViT and Swin or ViT, DEiT, and BEiT frequently showed high accuracy, demonstrating the effectiveness of diverse model inclusion. Stacking with Logistic Regression saw ensembles with ViT, DEiT, and Swin consistently achieving top performance, highlighting the robustness of these models. In Soft Voting, combinations such as ViT, DEiT, and Swin or DEiT and Swin performed best across various datasets, underscoring their complementary strengths. Weighted Soft Voting ensembles involving ViT, DEiT, BEiT, and Swin yielded superior accuracy, emphasizing model diversity. Weighted Hard Voting, similar to Soft Voting, saw combinations of ViT, DEiT, and BEiT frequently leading to high accuracy, balancing performance with ensemble complexity. In Hard Voting, the effectiveness of combinations varied, but ViT and DEiT often demon-

strated strong synergy.

The study also highlights the critical role of ViT and DEiT, which frequently appeared in top-performing combinations across different methods, underscoring their strong individual and collective contributions to enhancing classification performance. Swin consistently improved ensemble predictions across datasets and was often included in high-performing combinations. BEiT's contributions varied, with its effectiveness depending on the specific dataset and combination of models. Furthermore, the study emphasizes the importance of balancing computational complexity with performance gains. Ensembles incorporating more models generally increased computational demands, necessitating a balance between performance gains and complexity. Simpler ensembles, such as those with fewer models, often struck a balance between high accuracy and computational efficiency, making them efficient choices depending on available resources.

In conclusion, the ablation study underscores the importance of selecting the right combination of models to exploit their synergistic effects effectively. ViT, DEiT, and Swin consistently emerged as key models contributing to superior classification performance across diverse medical imaging tasks. While more complex ensembles with multiple models can improve accuracy, it is essential to consider computational complexity and potential overfitting. Strategic model inclusion and diversity are crucial for optimizing ensemble performance in medical image classification.

4.3.6 Finding of the ablation study

In this subsection, we will explore the findings of the ablation study, focusing on the best-performing combinations of ensemble methods and the best perforeming ensemble method per dataset. We will provide a discussion of these results, as well as a detailed analysis of the best ensemble methods using confusion matrices for each dataset. This will help us understand in depth how the ensemble methods handle the testing set.

4.3.6.1 Best performing combination of ensemble methods per dataset

The results present the best performing combination of ensemble methods per dataset using transformers as base models from ablation study. The metrics evaluated include Accuracy, Precision, Recall, and F1 Score.

Chest dataset

Table 4.81: Best performing combination of ensemble methods on Chest dataset

Dataset	Ensemble method	Base Models	Performance Metrics			
			Acc.	Prec.	Rec.	F1
Chest	Hard voting	ViT, Swin	97.42%	96.82%	96.98%	96.90%
	Weighted Hard Voting	ViT, Swin	97.42%	96.82%	96.98%	96.90%
	Soft Voting	ViT, Swin	97.42%	96.82%	96.98%	96.90%
	Weighted Soft Voting	ViT, BEiT, Swin	**97.59%**	**97.26%**	**96.93%**	**97.09%**
	Stacking with Logistic	DEiT, BEiT	97.42%	96.97%	96.27%	96.27%
	Stacking with SVM	ViT, Swin	**97.59%**	**97.26%**	**96.93%**	**97.09%**
	Bagging	DEiT, BEiT	97.08%	97.18%	97.08%	97.10%
	Boosting	ViT, DEiT, Swin	97.25%	97.25%	97.25%	97.25%

Notably, ViT and Swin frequently emerge as top performers across various ensemble methods Table 4.81. Both Weighted Soft Voting (ViT, BEiT, Swin) and Stacking with SVM (ViT, Swin) achieved the highest accuracy of 97.59%. These methods also exhibited high precision (97.26%) and strong recall (96.93%), resulting in a robust F1 score

(97.09%). This slight edge suggests that incorporating model weights or using sophisticated meta-learners can yield marginal gains in complex datasets like Chest. However, the Stacking with SVM combination of ViT and Swin (97.59% accuracy) was ultimately chosen as the best, due to its use of only two models, offering lower complexity compared to the three-model Weighted Soft Voting approach while maintaining peak performance.

DermaMNIST dataset

Table 4.82: Best performing combination of ensemble methods per dataset on DermaMNIST dataset

Dataset	Ensemble method	Base models	Performance Metrics			
			Acc.	Prec.	Rec.	F1
DermaMNIST	Hard voting	DEiT, Swin	80.10%	67.62%	61.96%	64.12%
	Weighted Hard Voting	DEiT, Swin	**80.30%**	**67.90%**	**61.25%**	**63.75%**
	Soft Voting	DEiT, Swin	80.15%	67.38%	62.18%	64.13%
	Weighted Soft Voting	DEiT, Swin	**80.30%**	**68.48%**	**61.67%**	**64.06%**
	Stacking with Logistic	ViT, DEiT, BEiT, Swin	79.85%	69.80%	56.13%	60.28%
	Stacking with SVM	ViT, DEiT, BEiT, Swin	78.55%	58.75%	55.46%	56.61%
	Bagging	ViT, DEiT, Swin	78.70%	78.30%	78.70%	78.02%
	Boosting	ViT, DEiT, Swin	79.50%	78.72%	79.50%	78.58%

Notabely the combination of DEiT and Swin appeared as the best performing combinations during ensemble methods highlights their strong complementary feature extraction capabilities Table 4.82. Both Weighted Hard Voting and Weighted Soft Voting achieved the highest accuracy of 80.30% using the same combination of DEiT and Swin. While Weighted Soft Voting showed a slightly higher precision (68.48%), its recall was lower (61.67%), leading to a moderate F1 score (64.06%). The disparity between precision and recall indicates that these ensemble methods, although effective in accurately predicting positive cases, miss a significant number of true positive cases.

BloodMNIST dataset

Table 4.83: Best performing combination of ensemble methods per dataset on BloodMNIST dataset

Dataset	Ensemble method	Base models	Performance Metrics			
			Acc.	Prec.	Rec.	F1
BloodMNIST	Hard voting	ViT, DEiT, Swin	97.87%	97.94%	97.55%	97.63%
	Weighted Hard Voting	ViT, BEiT	98.04%	97.96%	97.06%	97.95%
	Soft Voting	ViT, DEiT, Swin	97.90%	97.96%	97.76%	97.86%
	Weighted Soft Voting	ViT, DEiT, BEiT	**98.07%**	**97.95%**	**98.02%**	**97.98%**
	Stacking with Logistic	ViT, DEiT, Swin	97.87%	97.97%	97.65%	97.80%
	Stacking with SVM	ViT, DEiT, Swin	97.57%	97.65%	97.39%	97.50%
	Bagging	ViT, DEiT	97.60%	97.61%	97.60%	97.59%
	Boosting	ViT, DEiT, BEiT	97.98%	97.99%	97.98%	97.98%

It is notable that the combination of (ViT, DEiT, BEiT) and (ViT, DEiT, Swin) appear frequently as the best combination in many ensemble methods, underscoring the complementary strengths of these models when used together (Table 4.83), however the combination of ViT, DEiT, and BEiT with Weighted Soft Voting achieved the highest accuracy of 98.07% compared to diffrent ensemble methods. This method also had a near-perfect balance in precision (97.95%) and recall (98.02%), resulting in a high F1 score (97.98%).

BreastMNIST dataset

Table 4.84: Best performing combination of ensemble methods per dataset on BreastMNIST dataset

Dataset	Ensemble method	Base models	Performance Metrics			
			Acc.	Prec.	Rec.	F1
BreastMNIST	Hard voting	ViT, DEiT, Swin	87.82%	88.53%	79.64%	82.68%
	Weighted Hard Voting	ViT, DEiT, Swin	**89.10%**	**90.65%**	**81.27%**	**84.50%**
	Soft Voting	ViT, DEiT, Swin	88.46%	89.05%	80.83%	83.75%
	Weighted Soft Voting	ViT, BEiT	88.46%	90.18%	80.08%	83.42%
	Stacking with Logistic	ViT, DEiT, Swin	88.46%	90.18%	80.08%	83.42%
	Stacking with SVM	ViT, DEiT, Swin	87.82%	99.13%	78.13%	81.94%
	Bagging	ViT, DEiT, Swin	84.62%	90.09%	89.74%	89.15%
	Boosting	ViT, DEiT, Swin	86.54%	87.84%	86.54%	85.09%

The combination of ViT, DEiT, and Swin frequently appears as the best combination in many ensemble methods, highlighting their complementary strengths and the effectiveness of combining these models for the BreastMNIST dataset. this combination also leads to (Table 4.84), achieve the highest accuracy of 89.10% by weighted hard voting. It also showed strong precision (90.65%) and recall (81.27%), leading to a high F1 score (84.50%). The high performance of Weighted Hard Voting with this combination indicates its robustness in handling the BreastMNIST dataset.

PneumoniaMNIS dataset

Table 4.85: Best performing combination of ensemble methods per dataset on PneumoniaMNIST dataset

Dataset	Ensemble method	Base models	Performance Metrics			
			Acc.	Prec.	Rec.	F1
PneumoniaMNIST	Hard voting	ViT, Swin	94.07%	95.22%	92.35%	93.49%
	Weighted Hard Voting	ViT, BEiT, Swin	94.39%	95.19%	92.95%	93.88%
	Soft Voting	ViT, BEiT, Swin	94.07%	95.09%	92.44%	93.50%
	Weighted Soft Voting	ViT, BEiT, Swin	94.07%	94.86%	92.01%	93.13%
	Stacking with Logistic	ViT, DEiT	92.31%	94.01%	90.00%	91.45%
	Stacking with SVM	BEiT, Swin	91.67%	93.93%	88.97%	90.65%
	Bagging	DEiT, Swin	**94.71%**	**95.01%**	**94.71%**	**94.63%**
	Boosting	ViT, BEiT, Swin	93.11%	93.51%	93.11%	92.98%

Notably, the combination of ViT, BEiT, and Swin frequently appears as the best combination across several ensemble methods Table 4.85. This recurrent combination suggests that these models, when used together, provide complementary strengths that enhance the overall performance of the ensemble methods. However, Bagging with the combination of DEiT and Swin yielded the highest performance. Even though ViT, BEiT, and Swin appeared as the best combination in many ensemble methods, DEiT and Swin in the Bagging ensemble method outperformed the others. This combination demonstrated strong precision (95.01%) and recall (94.71%), resulting in a high F1 score (94.63%). This indicates that Bagging is particularly robust in handling the variability within the PneumoniaMNIST dataset. The balance across all performance metrics suggests that Bagging with this combination effectively manages both false positives and false negatives. The frequent appearance of the ViT, BEiT, and Swin combination highlights their effectiveness in creating a powerful and accurate ensemble for the dataset.

OrganCMNIST dataset

Table 4.86: Best performing combination of ensemble methods per dataset on OrganCMNIST dataset

Dataset	Ensemble method	Base models	Performance Metrics			
			Acc.	Prec.	Rec.	F1
OrganCMNIST	Hard voting	ViT, BEiT	93.60%	93.18%	92.48%	92.76%
	Weighted Hard Voting	ViT, BEiT, Swin	**93.75%**	**93.40%**	**92.58%**	**92.92%**
	Soft Voting	ViT, DEiT, BEiT	93.52%	93.13%	92.45%	92.73%
	Weighted Soft Voting	ViT, BEiT, Swin	93.72%	93.33%	92.60%	92.90%
	Stacking with Logistic	ViT, DEiT, BEiT	92.20%	91.66%	91.37%	91.31%
	Stacking with SVM	DEiT, BEiT, Swin	91.82%	91.08%	91.28%	90.96%
	Bagging	ViT, DEiT, Swin	93.66%	93.66%	93.66%	93.64%
	Boosting	ViT, DEiT, Swin	93.49%	93.51%	93.49%	93.46%

Notably, the combinations of (ViT, BEiT, Swin) and (ViT, DEiT, Swin) appears multiple times as the best combination across various ensemble methods meaning adding (DEiT or BEiT) to the combination of ViT and Swin can effectively achieve the highest performance for OrganCMNIST dataset. The OrganCMNIST dataset, as shown in Table 4.86 with (ViT, BEiT, Swin), Weighted Hard Voting achieved the highest accuracy of 93.75%. This method also had high precision (93.40%) and recall (92.58%), resulting in a balanced and robust F1 score (92.92%).

OrganSMNIST dataset

Table 4.87: Best performing combination of ensemble methods per dataset on OrganSMNIST dataset

Dataset	Ensemble method	Base models	Performance Metrics			
			Acc.	Prec.	Rec.	F1
OrganSMNIST	Hard voting	ViT, DEiT, BEiT	83.58%	80.21%	79.30%	79.19%
	Weighted Hard Voting	ViT, DEiT, BEiT, Swin	83.77%	80.39%	79.60%	79.63%
	Soft Voting	ViT, DEiT, BEiT, Swin	83.46%	80.06%	79.12%	79.29%
	Weighted Soft Voting	ViT, DEiT, BEiT, Swin	83.72%	80.24%	79.43%	79.59%
	Stacking with Logistic	ViT, BEiT, Swin	82.39%	78.92%	78.73%	78.08%
	Stacking with SVM	ViT, DEiT	79.92%	76.53%	76.20 %	75.53%
	Bagging	ViT, DEiT, Swin	**83.91%**	**83.78%**	**83.91%**	**83.31%**
	Boosting	ViT, DEiT, Swin	83.86%	83.31%	83.86%	83.36%

Notably, the combination of ViT, DEiT appears multiple times in the best combination across various ensemble methods. This consistency indicates the robustness and complementary nature of these models, however adding addinf Swin to this combination leads to heighest performance with bagging ensemble method 83.91% as shown in Table 4.87. Bagging also showed strong precision and recall leading to a high F1 score (83.31%). The overall performance in this dataset is lower compared to other dataset, suggesting higher complexity and noise. Bagging's ability to improve generalization through bootstrap sampling proves beneficial here, handling the variability and complexity of the OrganSMNIST dataset effectively.

OrganAMNIST dataset

Table 4.88: Best performing combination of ensemble methods per dataset on OrganAMNIST dataset

Dataset	Ensemble method	Base models	Performance Metrics			
			Acc.	Prec.	Rec.	F1
OrganAMNIST	Hard voting	ViT, DEiT, BEiT	95.08%	95.65%	94.87%	95.16%
	Weighted Hard Voting	ViT, DEiT, BEiT, Swin	95.03 %	95.58%	94.82 %	95.10 %
	Soft Voting	ViT, DEiT, BEiT	**95.14%**	**95.74%**	**95.15%**	**95.36%**
	Weighted Soft Voting	ViT, DEiT, BEiT, Swin	95.03%	95.56 %	94.83%	95.10%
	Stacking with Logistic	ViT, DEiT, Swin	93.93%	94.11%	93.71%	93.67%
	Stacking with SVM	ViT, DEiT, BEiT	93.04%	93.08%	92.94%	92.67%
	Bagging	ViT, DEiT, Swin	94.82%	94.54%	94.82%	94.79%
	Boosting	ViT, DEiT, Swin	95.01%	95.18%	95.01%	94.97%

The combination of ViT, DEiT, and BEiT frequently appears as the best combination across multiple ensemble methods such as Hard Voting, Soft Voting, and Stacking with SVM. This suggests that these models complement each other well, enhancing the overall performance of the ensemble methods. Notably, the use of Swin added to this combination in methods like Weighted Hard Voting and Weighted Soft Voting also resulted in strong performance metrics. However using ViT, DEiT, and BEiT together in soft voting leads to the highest consistency and best results (95.14%) indicating the effectivness of this combination in OrganAMNIST dataset.

4.3.6.2 Best ensemble method per dataset

Table 4.89: Best performing method per Dataset

Dataset	Ensemble method	Best combination	Performance Metrics			
			Acc.	Prec.	Rec.	F1
Chest	Stacking with SVM	ViT, Swin	97.59%	97.26%	96.93%	97.09%
DermaMNIST	Weighted Hard Voting	DEiT, Swin	80.30%	68.48%	61.67%	64.06%
BloodMNIST	Weighted Soft Voting	ViT, DEiT, BEiT	98.07%	97.95%	98.02%	97.98%
BreastMNIST	Weighted Hard Voting	ViT, DEiT, Swin	89.10%	90.09%	89.74%	89.15%
OrganCMNIST	Weighted Hard Voting	ViT, BEiT, Swin	93.75%	93.40%	92.58%	92.92%
OrganSMNIST	Bagging	ViT, DEiT, BEiT, Swin	83.91%	83.78%	83.91%	83.31%
OrganAMNIST	Soft Voting	ViT, DEiT, BEiT	95.14%	95.74%	95.15%	97.36%
Pneumonia	Bagging	ViT, BEiT, Swin	94.71%	95.01%	94.71%	94.63%

Table Table 4.89 highlights the best performing ensemble method for each dataset based on their respective evaluation metrics. Here, we discuss the observations and implications of these results

A. Stacking with SVM for the Chest dataset: Stacking with SVM emerged as the best-performing method for the Chest dataset using combination of (viT, Swin) achieving an impressive accuracy of 97.59%. This method demonstrated high precision (97.26%), recall (96.93%), and F1 score (97.09%). The success of Stacking with SVM underscores its ability to effectively integrate diverse base models and leverage their collective strengths for improved performance in datasets like Chest, which have complex patterns and diverse features.

B. Weighted hard hoting for DermaMNIST, BreastMNIST, OrganCMNIST Datasets: Weighted Hard Voting demonstrates robust performance across multiple datasets, establishing itself as the top-performing ensemble method for three distinct datasets. In the DermaMNIST dataset, it achieved an accuracy of 80.30% with DEiT and Swin, showcasing balanced precision (68.48%) and recall (61.67%). This method

effectively addresses challenges like class imbalance and complex features. Moreover, for the OrganCMNIST dataset, Weighted Hard Voting attained an accuracy of 93.75% using ViT, BEiT, and Swin, exhibiting strong precision, recall, and F1 score. Similarly, on the BreastMNIST dataset, it achieved an outstanding accuracy of 89.10% with ViT, DEiT, and Swin, surpassing other methods. This consistent performance across diverse datasets highlights Weighted Hard Voting's adaptability and effectiveness in optimizing ensemble performance.

C. Weighted Soft Voting for the BloodMNIST Dataset: For the BloodMNIST dataset, Weighted Soft Voting achieved the highest accuracy of 98.07% using combination of (viT, DEiT, BEiT), coupled with strong precision (97.95%), recall (98.02%), and F1 score (97.98%). This method's ability to optimize model contributions based on their performance enhances its effectiveness in datasets with clear class boundaries and well-defined patterns, such as BloodMNIST.

D. Bagging for Organs, and PneumoniaMNIST Datasets: Bagging emerged as the best-performing method for the OrganS, and PneumoniaMNIST datasets, showcasing its versatility and robustness across different medical imaging tasks. In each case, Bagging achieved high accuracy and balanced precision and recall, indicating its efficacy in handling dataset variability, noise, and complexity.

E. Soft voting for OrganAMNIST dataset: Soft Voting demonstrated superior performance for the OrganAMNIST dataset, achieving an accuracy of 95.14% using combination of (ViT, DEiT, BEiT) with strong precision, recall, and F1 score it outperformed all the other methods for OrganAMNIST dataset. This method's ability to sum up the probalities of base models based on their performance optimizes the ensemble's predictive capability, particularly beneficial for datasets like OrganAMNIST with large number of samples.

F. General observations: In general, choosing the best ensemble method depends heavily on the type, size, and complexity of the dataset. There isn't a one-size-fits-all ensemble method that works universally well, as we've seen from the varied results across different datasets in this analysis. Each method has its strengths in specific situations, highlighting the importance of picking an ensemble approach that fits the dataset's unique characteristics to get the best performance. Additionally, the success of ensemble methods largely depends on how we combine the base models in the ensemble to capture the complexities and features of the dataset during testing. If the base models are well-combined, the ensemble method is more likely to perform effectively. Conversely, if the base models are not combined correctly, the ensemble's performance may suffer. Therefore, achieving the best results with ensemble methods involves carefully selecting and training base models that align closely with the dataset's specific traits. This approach not only improves predictive accuracy but also enhances the ensemble's ability to perform consistently across different medical imaging datasets. It's worth noting that while boosting and hard voting achieved high accuracies, they did not consistently emerge as the best ensemble methods for these datasets. This suggests that, despite their high performance in some metrics, they may not be as effective as other ensemble methods for the specific characteristics of these datasets.

4.3.6.3 Confusion matrix analysis

To gain deeper findings into the performance of the ensemble methods, we analyze the confusion matrices of the best-performing methods for each dataset. The Confusion matrix provides detailed information about the true positives, true negatives, false positives,

and false negatives, allowing us to understand specific areas where the model excels or needs improvement.

Chest dataset: The Confusion matrix for the Chest dataset using Stacking with SVM is presented in Figure Figure 4.9.

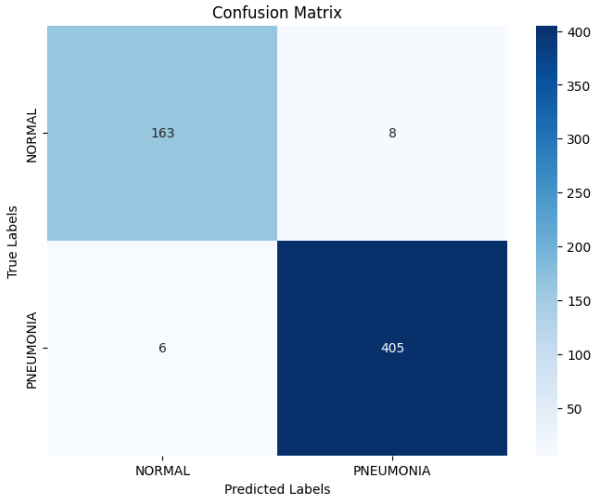


Figure 4.9: Confusion matrix of chest best performer

Upon examining the Confusion matrix for the Chest dataset, the stacking with svm ensemble method demonstrates robust performance with 405 correctly classified PneumoniaMNIST samples and 163 correctly identified Normal samples. There are 8 instances of false positives where Normal samples were misclassified as Pneumonia, indicating a low rate of false alarms. Additionally, the model missed 6 PneumoniaMNIST samples, incorrectly classifying them as Normal, highlighting an area for improvement in sensitivity. Despite this, the high number of true positives and true negatives, along with high precision, recall, and F1 score, validate the model's reliability in distinguishing between classes. To further enhance its effectiveness in clinical settings, focusing on reducing false negatives will be crucial for future model refinements.

DermaMNIST dataset: The Confusion matrix for the DermaMNIST dataset using Weighted Hard Voting is presented in Figure Figure 4.10.

Upon examining the Confusion matrix for the DermaMNIST dataset, the weighted hard voting ensemble method displays varying levels of accuracy across different classes. Class 0 shows moderate accuracy with 31 true positives but high false positives from classes 1, 2, and 4, indicating a need for improved sensitivity. Class 1 has 70 true positives but suffers from significant missclassified from classes 0, 2, and 5, necessitating better specificity. Class 2 exhibits high accuracy with 118 true positives, though it has notable false positives from all other classes except class 3. Class 3 has low accuracy with only 9 true positives and false positives primarily from classes 1, 2, 4, and 5, suggesting a need for better feature differentiation. Class 4 demonstrates good accuracy with 108 true positives but encounters high false positives from class 5. Class 5 shows excellent accuracy with 1251 true positives and few missclassified, while class 6 has high accuracy with 23 true positives but some missclassified from classes 2, 4, and 5. Overall, the model shows high true positive rates for classes 2, 4, 5, and 6, with class 5 performing exceptionally well. However, classes 0, 1, and 3 exhibit high false positive and false negative rates, indicating a need for improved sensitivity and specificity. Enhancements in feature differentiation and training data quality are recommended to reduce missclassified. In summary, while

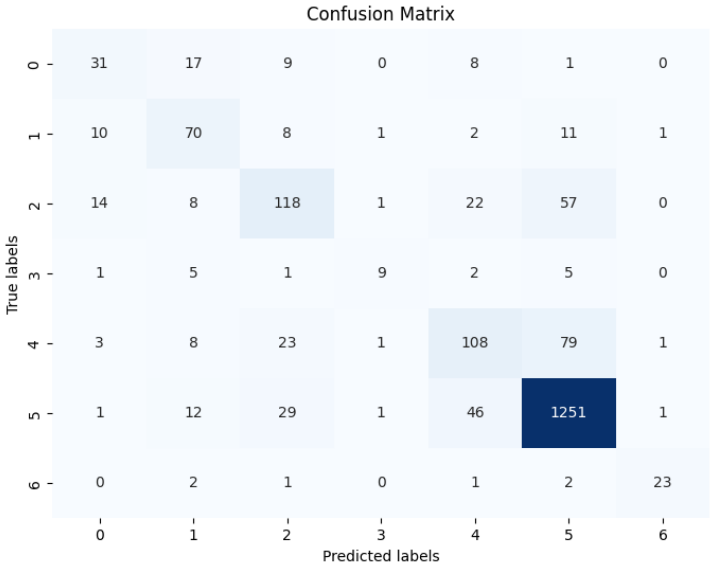


Figure 4.10: Confusion matrix of DermaMNIST best performer

the model performs well for most classes, particularly class 5, improvements are needed for correctly classifying classes 0, 1, and 3 to achieve better overall accuracy.

BloodMNIST dataset: The Confusion matrix for the BloodMNIST dataset is shown in Figure Figure 4.11.

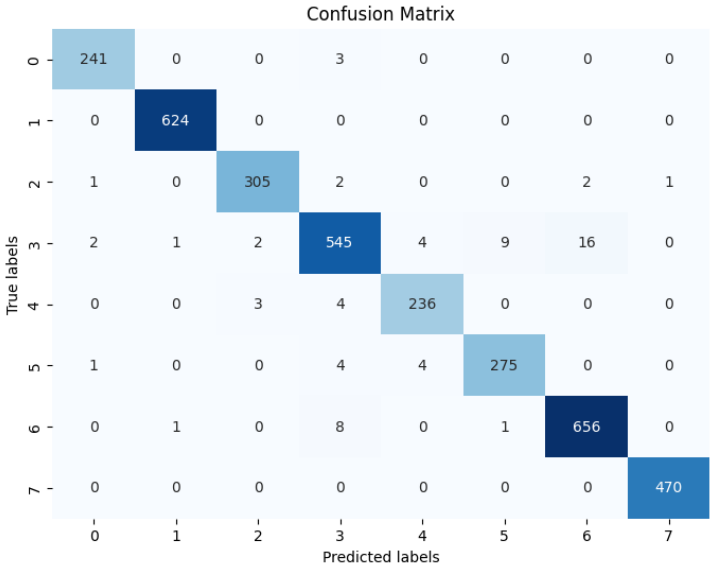


Figure 4.11: Confusion matrix of BloodMNIST best performer

Upon examining the Confusion matrix for the BloodMNIST dataset, the weighted soft voting ensemble method demonstrates strong performance with high true positive rates, particularly in Classes 1 and 7, which had no missclassifications. Class 1 achieved 624 true positives, and Class 7 had 470 true positives, both with no false positives or false negatives. Class 2 had 305 true positives, with minimal false positives and 5 false negatives. Class 3 recorded 545 true positives but had several false positives from various other classes and 7 false negatives, indicating a need for better differentiation. Class 4 had

236 true positives, no false positives, and 7 false negatives. Class 5 had 275 true positives with a few false positives and 8 false negatives. Class 6 achieved 656 true positives, with some false positives and only 1 false negative. The model excels in achieving high accuracy for most classes, especially Classes 1 and 7. However, Class 3's missclassifications from multiple other classes highlight the need for improved feature differentiation. Class 6 also shows some missclassifications from Classes 2, 3, and 5, suggesting that refinements in feature extraction or model adjustment may be necessary. the Confusion matrix analysis for the BloodMNIST dataset indicates strong performance, with specific areas identified for further improvement to enhance overall classification accuracy.

BreastMNIST dataset: The Confusion matrix for the BreastMNIST dataset is shown in Figure Figure 4.12.

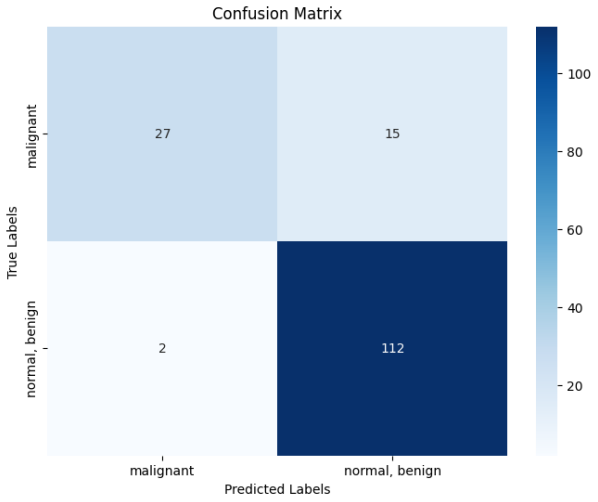


Figure 4.12: Confusion matrix of BreastMNIST best performer

Upon examining the Confusion matrix for the BreastMNIST dataset, the weighted hard voting ensemble method shows strong performance in detecting normal/benign cases with 112 true positives and only 2 false negatives, indicating reliable detection of non-malignant cases. However, the model struggles with identifying malignant cases, evidenced by 27 true positives, 2 false positives, and a relatively high false negative rate of 15, where malignant cases were misclassified as normal/benign. This presents a significant area for improvement due to the serious implications of missed malignant diagnoses in a clinical setting. Additionally, although the false positive rate for malignant cases is low, with only 2 normal/benign cases misclassified as malignant, this still suggests a need to reduce potential unnecessary treatments. The Confusion matrix analysis for the BreastMNIST dataset underscores the model's strength in identifying normal/benign cases but highlights the critical need to improve its accuracy in detecting malignant cases. Reducing the false negative rate for malignant cases is essential to enhance the model's reliability in clinical diagnostics.

OrganCMNIST dataset: The Confusion matrix weighted hard voting ensemble method For OrganCMNIST dataset shows strong overall performance but highlights key areas for improvement Figure 4.13.

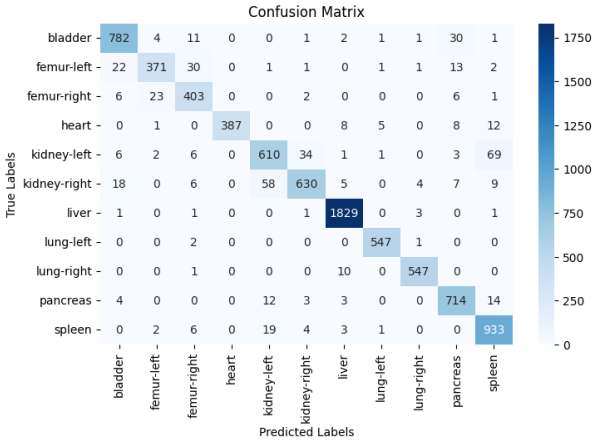


Figure 4.13: Confusion matrix of OrganCMNIST best performer

The bladder class has high accuracy with 782 true positives, but some missclassifications as spleen suggest a need for better differentiation. Femur-left and femur-right classes show confusion between each other, indicating the need for enhanced feature differentiation. The heart class performs well with minimal errors. Kidney-left and kidney-right classes show significant missclassifications into each other, highlighting the need for improved distinction between these similar organs. The liver class excels with 1829 true positives and few missclassifications. Both lung classes perform well with 547 true positives each. The pancreas class needs better differentiation from kidney-left, with 714 true positives but 12 misclassified instances. The spleen class has notable missclassifications from kidney-left, indicating a need for better distinction. while the model performs well for most organs, reducing missclassifications between similar OrganSMNIST like the femurs and kidneys will enhance accuracy.

OrganSMNIST dataset: The Confusion matrix of bagging ensemnle learning for the OrganSMNIST dataset shows strong performance for the heart, liver, and lungs, with high true positive rates and minimal missclassifications Figure 4.14.

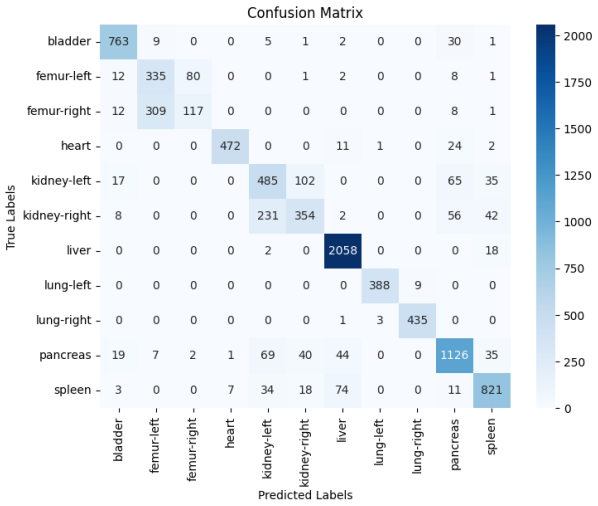


Figure 4.14: Confusion matrix of OrganSMNIST best performer

The bladder has 763 true positives but needs better differentiation from the spleen. The femur-left and femur-right classes show significant confusion between each other,

indicating a need for improved feature differentiation. The kidneys also show substantial missclassifications between left and right, highlighting a need for better distinction. While the pancreas performs well, it misclassifies some instances as kidney-left, and the spleen shows notable missclassifications from kidney-left. the model excels in identifying the heart, liver, and lungs but needs refinement to better differentiate similar OrganSMNIST like the femurs and kidneys to enhance accuracy and reliability.

PneumoniaMNIST dataset: The Confusion matrix of bagging ensemble method for the PneumoniaMNIST dataset reveals the model's performance in a concise manner Figure 4.15.

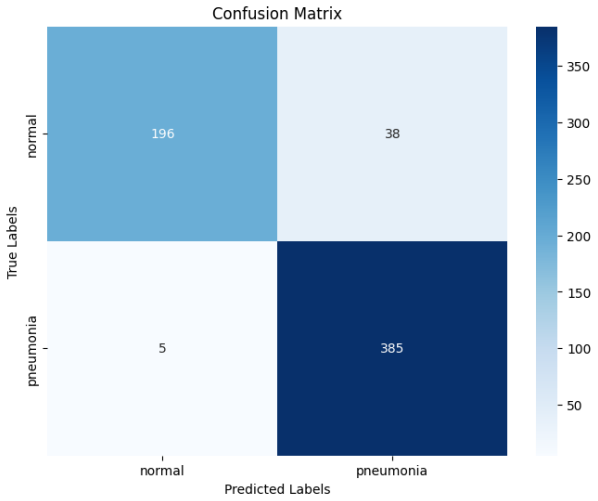


Figure 4.15: Confusion matrix of PneumoniaMNIST best performer

The model demonstrates exceptional proficiency in identifying PneumoniaMNIST cases accurately. Out of the total 423 instances of pneumonia, a remarkable 385 were correctly classified as such (true positives). However, the model's performance on the normal class is not as stellar. While it correctly identified 196 out of 201 normal instances (true negatives), it misclassified 5 instances as PneumoniaMNIST (false positives). The Confusion matrix underscores the model's exceptional prowess in detecting PneumoniaMNIST cases while also highlighting areas for potential improvement, particularly in minimizing false positives for the normal class and reducing false negatives for the PneumoniaMNIST class, thereby enhancing its overall reliability and performance across both classes.

OrganAMNIST dataset: The confusion matrix succinctly captures the multi-class classification ensemble's performance across various anatomical structures Figure 4.16. The soft voting ensemble method excels in accurately identifying bladder instances, correctly classifying an impressive 1016 out of 1016 cases. Its prowess extends to the femur-right class, where it accurately predicted 773 instances. The ensemble also demonstrates remarkable competence in recognizing heart structures, accurately classifying 699 out of 699 cases. However, the ensemble's performance is not as exceptional for certain classes. It struggles to differentiate between kidney-left and kidney-right instances, with 35 kidney-left cases misclassified as kidney-right, and 26 kidney-right cases mistaken for kidney-left. The confusion between liver and lung-left is also noteworthy, with a substantial 1747 liver instances incorrectly predicted as lung-left. The matrix highlights the ensemble's significant shortcomings in the pancreas and spleen classes. It misclassified 1507 pancreas instances, primarily confusing them with other classes like bladder, femur-right, and heart. The spleen class exhibited the highest number of missclassifications,

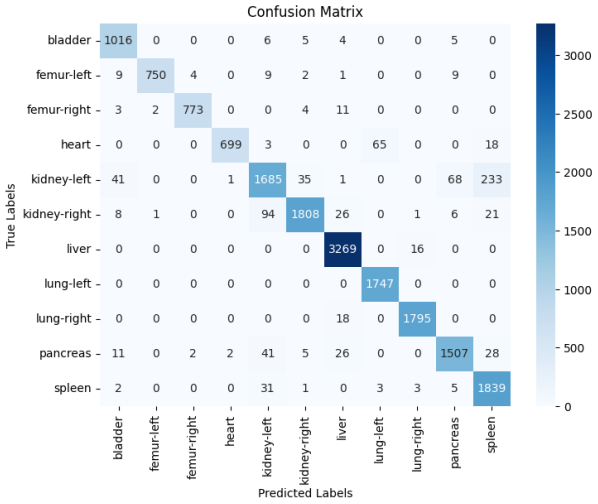


Figure 4.16: Confusion matrix of OrganAMNIST best performer

with a staggering 1839 instances incorrectly labeled as pancreas. While the ensemble demonstrates exceptional accuracy for specific classes, the confusion matrix underscores the need for improvements, particularly in distinguishing between similar structures like kidneys, as well as enhancing feature representations to better differentiate challenging classes like pancreas and spleen.

4.3.7 Results comparison with base models

In the following paragraphs, we will provide a detailed comparison between the results of the individual fine-tuned base models and the best ensemble methods for each dataset.

Chest dataset

Table 4.90: Results comparison with individual models on Chest dataset

Dataset	Base Models/Ensemble methods	Performance Metrics			
		Acc.	Prec.	Rec.	F1
Chest	Vit	96.22%	95.06%	95.96%	95.49%
	Swin	95.88%	95.99%	94.01%	94.92%
	BEiT	91.07%	89.23%	89.23%	89.23%
	DEiT	96.22%	95.31%	95.62%	95.46%
	Hard voting	97.42%	96.82%	96.98%	96.90%
	Weighted Hard Voting	97.42%	96.82%	96.98%	96.90%
	Soft Voting	97.42%	96.82%	96.98%	96.90%
	Weighted Soft Voting	**97.59%**	**97.26%**	**96.93%**	**97.09%**
	Stacking with Logistic	97.42%	96.97%	96.27%	96.27%
	Stacking with SVM	**97.59%**	**97.26%**	**96.93%**	**97.09%**
	Bagging	97.08%	97.18%	97.08%	97.10%
	Boosting	97.25%	97.25%	97.25%	97.25%

The ensemble methods generally perform better than individual base models for the Chest dataset Table 4.90. However, there are slight variations in accuracy among the ensemble methods. For instance, the Weighted Soft Voting ensemble and stacking with svm achieve the highest accuracy of 97.59%, surpassing the performance of all individual

base models and other ensemble methods. This suggests that weighted soft voting and stacking with svm effectively leverage the strengths of the base models to improve overall accuracy. On the other hand, the bagging ensemble method exhibits slightly lower accuracy compared to other ensemble methods. This could be attributed to the nature of the bagging ensemble method who relys on splitting the dataset and each model does not train on all the dataset, which may not effectively capture the complex relationships present in the dataset. Additionally, the BEiT architecture performs lower than other base models and ensemble methods across all ensemble techniques. This indicates that the inclusion of BEiT in the ensembles may not contribute significantly to overall performance. The ensemble methods demonstrate the ability to enhance classification accuracy for the Chest dataset, with weighted soft voting and stacking with svm being the most effective approaches.

DermaMNIST dataset

Table 4.91: Results compariosn with individual models on OrganAMNIST

Dataset	Base Model/Ensemble method	Performance Metrics			
		Acc.	Prec.	Rec.	F1
DermaMNIST	Vit	76.81%	68.21%	50.61%	54.97%
	Swin	78.75%	63.08%	61.01%	61.50%
	BEiT	75.61%	57.42%	53.53%	52.71%
	DEiT	79.60%	66.43%	58.91%	61.64%
	Hard voting	80.10%	67.62%	61.96%	64.12%
	Weighted Hard Voting	**80.30%**	**67.90%**	**61.25%**	**63.75%**
	Soft Voting	80.15%	67.38%	62.18%	64.13%
	Weighted Soft Voting	**80.30%**	**68.48%**	**61.67%**	**64.06%**
	Stacking with Logistic	79.85%	69.80%	56.13%	60.28%
	Stacking with SVM	78.55%	58.75%	55.46%	56.61%
	Bagging	78.70%	78.30%	78.70%	78.02%
	Boosting	79.50%	78.72%	79.50%	78.58%

The DermaMNIST dataset presents challenges for both base models and ensemble methods, with generally lower accuracy rates compared to other datasets Table 4.91. Interestingly, the ensemble methods, particularly Weighted Soft Voting and Weighted Hard Voting, achieve slightly higher accuracy compared to individual base models. This suggests that ensemble methods effectively leverage the diverse predictions of base models to improve overall performance. However, some ensemble methods, such as Stacking with Logistic and Stacking with SVM, exhibit lower accuracy compared to other ensemble methods. This could be attributed to the complexity of the dataset and the inability of logistic regression and SVM algorithms to effectively capture the underlying patterns in the data. Furthermore, the Bagging and Boosting ensemble methods demonstrate consistent accuracy rates across all evaluation matrics. This indicates that these base models may not significantly improve performance on the DermaMNIST dataset. While ensemble methods show potential for improving accuracy on the DermaMNIST dataset.

BloodMNIST dataset

Table 4.92: Compariosn with individual models on BloodMNIST

Dataset	Base Model/Ensemble method	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BloodMNIST	Vit	97.90%	97.72%	97.85%	97.78%
	Swin	96.49%	96.27%	96.16%	96.19%
	BEiT	96.20%	96.18%	95.56%	95.83%
	DEiT	97.37%	97.30%	97.06%	97.18%
	Hard voting	97.87%	97.94%	97.55%	97.63%
	Weighted Hard Voting	98.04%	97.96%	97.06%	97.95%
	Soft Voting	97.90%	97.96%	97.76%	97.86%
	Weighted Soft Voting	**98.07%**	**97.95%**	**98.02%**	**97.98%**
	Stacking with Logistic	97.87%	97.97%	97.65%	97.80%
	Stacking with SVM	97.57%	97.65%	97.39%	97.50%
	Bagging	97.60%	97.61%	97.60%	97.59%
	Boosting	97.98%	97.99%	97.98%	97.98%

The BloodMNIST dataset showcases strong performance across both base models and ensemble methods, with consistently high accuracy rates Table 4.92. Individual base models, particularly Vit and DEiT, demonstrate excellent accuracy, with DEiT achieving the highest accuracy of 97.37%. Swin and BEiT also exhibit respectable performance, with slightly lower compared to Vit and DEiT. Ensemble methods further enhance accuracy on the BloodMNIST dataset, with Weighted Soft Voting achieving the highest accuracy of 98.07%. This suggests that ensemble techniques effectively leverage the diverse predictions of base models to improve overall performance. Interestingly, some ensemble methods, such as Weighted Hard Voting and Boosting, achieve comparable accuracy to the weighted soft voting. The BloodMNIST dataset demonstrates the effectiveness of ensemble methods in enhancing overall accuracy .

PneumoniaMNIST dataset

Table 4.93: Comparison with individual models on PneumoniaMNIST dataset

Dataset	Base Model/Ensemble method	Performance Metrics			
		Acc.	Prec.	Rec.	F1
PneumoniaMNIST	ViT	93.59%	94.74%	91.79%	92.95%
	Swin	88.78%	92.17%	85.13%	87.12%
	BEiT	84.46%	83.54%	83.12%	83.32%
	DEiT	86.22%	86.23%	84.02%	84.87%
	Hard voting	94.07%	95.22%	92.35%	93.49%
	Weighted Hard Voting	94.39%	95.19%	92.95%	93.88%
	Soft Voting	94.07%	95.09%	92.44%	93.50%
	Weighted Soft Voting	94.07%	94.86%	92.01%	93.13%
	Stacking with Logistic	92.31%	94.01%	90.00%	91.45%
	Stacking with SVM	91.67%	93.93%	88.97%	90.65%
	Bagging	**94.71%**	**95.01%**	**94.71%**	**94.63%**
	Boosting	93.11%	93.51%	93.11%	92.98%

The PneumoniaMNIST dataset exhibits diverse performance across different base models and ensemble methods Table 4.93. Ensemble methods demonstrate improved

accuracy compared to individual base models. Bagging achieves the highest accuracy of 94.71%, indicating that this ensemble method effectively combines the predictions of base models to enhance overall accuracy on the PneumoniaMNIST dataset. However, it is observed that certain ensemble methods, such as Stacking with Logistic and Stacking with SVM, exhibit lower accuracy compared to individual base models. This discrepancy may be attributed to the complexity of the dataset or the nature of the base models used in the ensemble. The traditional models might not be as effective on this complex dataset as more advanced approaches like transformers. The PneumoniaMNIST dataset underscores the importance of ensemble techniques in improving classification performance.

BreastMNIST dataset

Table 4.94: Compariosn with individual models on BreastMNIST

Dataset	Base Model/Ensemble method	Performance Metrics			
		Acc.	Prec.	Rec.	F1
BreastMNIST	ViT	87.82%	89.71%	78.88%	82.32%
	Swin	85.26%	81.62%	80.14%	80.82%
	BEiT	73.08%	36.54%	52.35%	42.22%
	DEiT	83.33%	80.79%	74.31%	76.53%
	Hard voting	87.82%	88.53%	79.64%	82.68%
	Weighted Hard Voting	**89.10%**	**90.65%**	**81.27%**	**84.50%**
	Soft Voting	88.46%	89.05%	80.83%	83.75%
	Weighted Soft Voting	88.46%	90.18%	80.08%	83.42%
	Stacking with Logistic	88.46%	90.18%	80.08%	83.42%
	Stacking with SVM	87.82%	99.13%	78.13%	81.94%
	Bagging	84.62%	90.09%	89.74%	89.15%
	Boosting	86.54%	87.84%	86.54%	85.09%

The BreastMNIST dataset presents varied performance across different base models and ensemble methods Table 4.95. Ensemble methods show improved performance over individual base models, with weighted hard voting achieving the highest accuracy of 89.10%, as well as superior recall and other metrics. This indicates that this ensemble technique effectively combines the predictions of base models to enhance overall accuracy on the BreastMNIST dataset. However, some ensemble methods like bagging and boosting exhibited lower accuracy. This discrepancy may be attributed to the small size of the dataset, which might hinder the models' ability to learn correctly and make accurate predictions like in bagging models use subsets of the training dataset which is already small so models couldn't learn effectively. Overall, the BreastMNIST dataset underscores the importance of dataset size for ensemble methods in improving classification performance.

OrganCMNIST dataset

Table 4.95: Compariosn with individual models on OrganCMNIST

Dataset	Base Model/Ensemble method	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	Vit	92.83%	92.31%	91.60%	91.89%
	Swin	91.99%	91.95%	90.67%	91.21%

Continued on next page

Dataset	Base Model/Ensemble method	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganCMNIST	BEiT	92.56%	92.28%	91.37%	91.75%
	DEiT	92.40%	91.99%	91.23%	91.54%
	Hard voting	93.60%	93.18%	92.48%	92.76%
	Weighted Hard Voting	**93.75%**	**93.40%**	**92.58%**	**92.92%**
	Soft Voting	93.52%	93.13%	92.45%	92.73%
	Weighted Soft Voting	93.72%	93.33%	92.60%	92.90%
	Stacking with Logistic	92.20%	91.66%	91.37%	91.31%
	Stacking with SVM	91.82%	91.08%	91.28%	90.96%
	Bagging	93.66%	93.66%	93.66%	93.64%
	Boosting	93.49%	93.54%	93.49%	93.47%

The OrganCMNIST dataset demonstrates consistent performance across base models and ensemble methods Table 4.95. Ensemble methods further improve accuracy compared to individual base models. Weighted Hard Voting achieves the highest accuracy of 93.75%, indicating effective combination of predictions from different base models to enhance overall accuracy on the OrganCMNIST dataset. However, it is observed that certain ensemble methods, such as Stacking with Logistic and Stacking with SVM, show lower accuracy compared to individual base models. This is likely because these traditional meta-models did not capture the complexity of this dataset as effectively. The OrganCMNIST dataset highlights the effectiveness of ensemble techniques in improving classification performance, with potential applications in medical imaging for organ anomaly detection.

OrganSMNIST dataset

Table 4.96: Compariosn with individual models on OrganSMNIST

Dataset	Base Model/Ensemble method	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganSMNIST	Vit	81.65%	78.62%	77.01%	76.59%
	Swin	82.3%	78.98%	77.86%	78.31%
	BEiT	82.4%	78.95%	78.21%	78.52%
	DEiT	80.8%	77.03%	76.86%	76.5%
	Hard voting	83.58%	80.21%	79.30%	79.19%
	Weighted Hard Voting	83.77%	80.39%	79.60%	79.63%
	Soft Voting	83.46%	80.06%	79.12%	79.29%
	Weighted Soft Voting	83.72%	80.24%	79.43%	79.59%
	Stacking with Logistic	82.39%	78.92%	78.73%	78.08%
	Stacking with SVM	79.92%	76.53%	76.20%	75.53%
	Bagging	**83.91%**	**83.78%**	**83.91%**	**83.31%**
	Boosting	83.86%	83.31%	83.86%	83.36%

The OrganSMNIST dataset demonstrates that ensemble methods generally outperform individual base models in terms of accuracy, precision, recall, and F1 score Table 4.96. Ensemble methods, particularly Bagging and Boosting, achieve the highest accuracy rates of 83.91% and 83.86%, respectively. This indicates that combining predictions from multiple models can enhance performance. Stacking with Logistic and Stacking with SVM show lower accuracy compared to other ensemble methods, likely due to the

traditional nature of these meta-models, which may not capture the complexity of the dataset as effectively.

OrganAMNIST dataset

Table 4.97: Compariosn with individual models on OrganAMNIST

Dataset	Base Model/Ensemble method	Performance Metrics			
		Acc.	Prec.	Rec.	F1
OrganAMNIST	Vit	93.27%	93.99%	93.01%	93.34%
	Swin	93.87%	94.3%	93.43%	93.73%
	BEiT	93.29%	94.16%	92.96%	93.40%
	DEiT	94.4%	94.64%	93.95%	94.21%
	Hard voting	95.08%	95.65%	94.87%	95.16%
	Weighted Hard Voting	95.03%	95.56%	94.83 %	95.10 %
	Soft Voting	**95.14%**	**95.74%**	**95.15%**	**95.36%**
	Weighted Soft Voting	95.03%	95.56%	98.83%	95.10%
	Stacking with Logistic	93.93%	94.11%	93.71%	93.67%
	Stacking with SVM	93.04%	93.08%	92.94%	92.67%
	Bagging	94.82%	94.54%	94.82%	94.79%
	Boosting	95.01%	95.18%	95.01%	94.97%

The OrganAMNIST dataset demonstrates that ensemble methods generally outperform individual base models in terms of accuracy, precision, recall, and F1 score. Among the individual models, DEiT achieves the highest accuracy at 94.40%, while ViT, Swin, and BEiT have slightly lower accuracy rates Table 4.97. Ensemble methods, particularly Soft Voting, achieve the highest performance metrics, with an accuracy of 95.14%, precision of 95.74%, recall of 95.15%, and an F1 score of 95.36%. Weighted Hard Voting and Hard Voting also perform well, with nearly identical accuracy, precision, recall, and F1 scores, indicating their effectiveness. The combination of ViT, DEiT, BEiT, and Swin using weighted hard voting from the ablation study achieved an accuracy of 95.03%, precision of 95.56%, recall of 94.83%, and an F1 score of 95.10%, demonstrating a balance between precision and recall. While the performance of individual models is strong, ensemble methods like Soft Voting, Hard Voting, and Boosting significantly enhance overall performance, highlighting the value of combining multiple models for improved results.

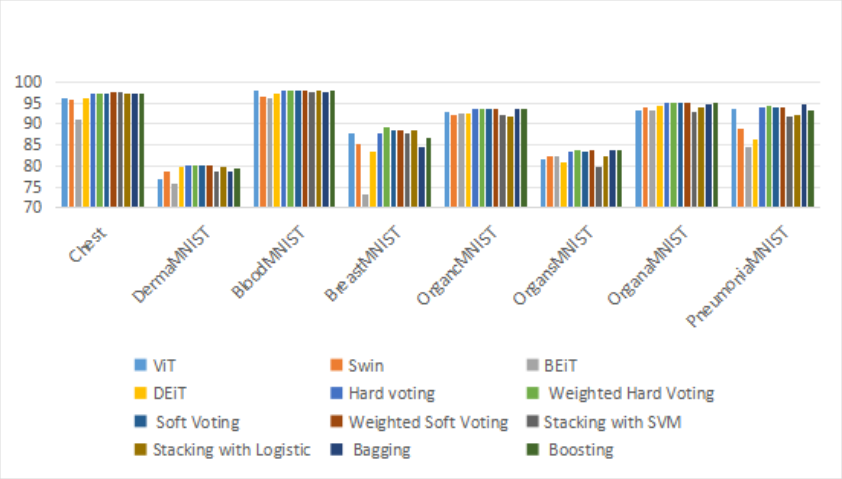


Figure 4.17: Illustration of the comparison with individual models results in terms of accuracy

Summary: Ensemble methods consistently as also illustrated in Figure 4.17 outperform individual base models across diverse datasets, highlighting their effectiveness in enhancing model performance and diagnostic accuracy. By combining predictions from multiple models, ensemble methods use the strengths of each model to achieve higher accuracy and more reliable predictions, making them invaluable tools in medical image classification and diagnostic tasks.

4.3.8 Results comparison with state-of-the-art

In the following paragraphs, we will provide a detailed comparison between the results of state-of-the-art models and the best ensemble methods for each dataset.

Table 4.98: Comparison of our ensemble methods with state-of-the-art models in terms of accuracy.

Model/Ensemble method	DermaMNIST	BloodMNIST	BreastMNIST	OrganC	OrganS	OrganA	Pneumonia
State-of-the-Art models							
ResNet-18 (28)[76]	73.5%	95.8%	86.3%	90.0%	78.2%	93.5%	85.4%
ResNet-18 (224)[76]	75.4%	96.3%	83.3%	92.0%	77.8%	95.1%	86.4%
ResNet-50 (28)[76]	73.5%	95.6%	81.2%	90.5%	77.0%	93.5%	85.4%
ResNet-50 (224)[76]	73.1%	95.0%	84.2%	91.1%	78.5%	94.7%	88.4%
auto-sklearn[76]	71.9%	87.8%	80.3%	82.9%	67.2%	76.2%	85.5%
AutoKeras[76]	74.9%	96.1%	83.1%	87.9%	81.3%	90.5%	87.8%
Google AutoML Vision[76]	76.8%	96.6%	86.1%	87.7%	74.9%	88.6%	94.6%
FPVT[102]	76.6%	94.4%	89.1%	90.3%	78.5%	93.5%	89.6%
MedVIT-T[102]	76.8%	95.0%	89.6%	90.1%	78.9%	93.1%	**94.9%**
MedVIT-S[102]	78.0%	95.1%	89.7%	91.6%	80.5%	92.8%	92.1%
MedVIT-L[102]	77.3%	95.4%	88.3%	92.2 %	80.6%	80.6%	94.6%
Our ensemble methods							
Hard Voting	80.10%	97.87%	87.82%	93.60%	83.58%	95.08%	94.07%
Weighted Hard Voting	**80.30%**	98.04%	**89.10%**	**93.75%**	83.77%	95.03	94.39%
Soft Voting	80.15%	97.90%	88.46%	93.52%	83.46%	95.14 %	94.07%
Weighted Soft Voting	**80.30%**	**98.07%**	88.46%	93.72%	83.72%	94.99%	94.07%
Stacking with Logistic	79.85%	97.87%	88.46%	92.20%	82.39%	93.93%	92.31%
Stacking with SVM	78.55%	97.57%	87.82%	91.82%	79.92%	93.0%	91.67%
Bagging	78.70%	97.60%	84.62%	93.66%	**83.91%**	94.82%	**94.71%**
Boosting	79.50%	97.98%	86.54%	93.49%	83.86%	95.01%	93.11%

The Table 4.98 compares the performance of our various methods across different datasets in term of accuracy, highlighting the effectiveness of our ensemble method in improving classification accuracy. It is important to note that the chest dataset was not included in this comparison because it is not part of the MedMNIST suite of datasets. Instead, the chest dataset was sourced from Hugging Face. This inclusion was deliberate to demonstrate that our proposed ensemble methods are not only effective for MedMNIST datasets but also perform well on datasets from other sources. This validates the generalizability and robustness of our ensemble methods across different types of medical imaging data.

DermaMNIST dataset: In the DermaMNIST dataset Table 4.98, our ensemble methods achieve superior accuracy compared to state-of-the-art models. Specifically, the highest accuracy from stat of the art methods is 78.0% achieved by MedVIT-S. And the highest accuracy for the DermaMNIST dataset is achieved by Weighted Hard Voting and Weighted Soft Voting at 80.30%, both surpassing the SOTA models. This shows the potential of ensemble methods to enhance performance on this dataset.

BloodMNIST dataset: For the BloodMNIST dataset Table 4.98, our ensemble methods consistently outperform existing models in terms of accuracy. In the BloodMNIST dataset, the highest accuracy from start of the art methods is 96.6% achieved by Google AutoML Vision. The best accuracy for the BloodMNIST dataset is achieved by Weighted

Soft Voting at 98.07%, significantly outperforming the SOTA models. This demonstrates the effectiveness of our ensemble methods.

BreastMNIST dataset: The BreastMNIST dataset Table 4.98, being relatively small in size, may have limited the learning capacity of our ensemble methods, which are based on transformers known to perform better with larger datasets. Consequently the highest accuracy from stat of the art methods is 89.7% achieved by MedVIT-S. Our best method achieved an accuracy of 89.1%, slightly below the state-of-the-art but still comparable. The comparable performance suggests that while our ensemble methods didn't give us high accuracy but still work well and further optimization in the dataset size needed to surpass the top-performing methods in term of accuracy.

OrganCMNIST dataset: Our ensemble methods exhibit notable improvements in OrganSMNIST classification accuracy compared to baseline models Table 4.98. the highest accuracy from start of the art models is 92.2% achieved by MedVIT-L. Our method achieved an accuracy of 93.75%, outperforming the existing model by 1.55%. This improvement indicates that ensem- ble method effectively combines the strengths of individual models, leading to superior performance on the OrganCMNIST dataset.

OrganSMNIST dataset: For OrganSMNIST dataset, our ensemble methods demonstrate enhanced accuracy compared to baseline models Table 4.98. the highest accuracy from start of the art methods is 80.6% achieved by MedVIT-L. Our method achieved an accuracy of 83.91%, a significant improvement of 3.41%. The notable increase underscores the capability of our ensemble methods to enhance predictive performance by leveraging complementary information from multiple models.

OrganAMNIST dataset: In OrganAMNIST dataset, our ensemble approaches yield superior accuracy compared to state-of-the-art models Table 4.98. The highest accuracy from start of the art methods is 95.1% achieved by Google AutoML Vision. Our method achieved an accuracy of 95.14%, slightly surpassing the state-of-the-art. Although the gain is modest, it reaffirms the consistency and reliability of our ensemble approach across various datasets

PneumoniaMNIST dataset: The PneumoniaMNIST dataset, also relatively small, may have constrained the learning capacity of our ensemble models, which rely on transformer architectures better suited for larger datasets. Consequently, the highest accuracy from start of the art methods is 94.9% achieved by MedVIT-T Table 4.98. Our method achieved an accuracy of 94.71%, slightly below the state-of-the-art. The limited size of the dataset likely contributed to the slight decrease in accuracy but its still competitive that indicates that ensemble methods can deal with dataset limitation.

Summary: The ensemble method consistently outperforms or matches the best state-of-the-art methods across multiple datasets, highlighting its ability to enhance accuracy by integrating diverse model predictions. The improved performance across varied datasets demonstrates the robustness of ensemble method, making it a versatile and reliable approach for different types of medical image classification tasks. By combining multiple models, ensemble methods can reduce overfitting, as evidenced by the improved generalization on the test datasets. However, for smaller datasets like BreastMNIST and PneumoniaMNIST, large models such as transformers may not perform optimally due to insufficient training data, emphasizing the importance of dataset size in determining the effectiveness of ensemble methods with large models. In conclusion, the use of ensemble methods in our implementation has proven to be highly effective in improving

classification performance across several medical image datasets. The results show that leveraging the complementary strengths of different models leads to significant gains in accuracy, making ensemble methods a powerful technique for advancing state-of-the-art performance in medical imaging tasks. Nevertheless, it is crucial to consider the size of the dataset when applying large models, as smaller datasets may not provide sufficient training data for these models to fully realize their potential. Our ensemble approach demonstrates robust performance and offers a promising direction for future research in medical image classification.

4.4 Conclusion

In this chapter, we explored and experimentally evaluated our proposed ensemble methods for medical image classification. We started with the implementation setup, including hardware and hyperparameter tuning for fine-tuning individual models. The results of individual models were analyzed for each dataset, establishing a baseline for comparison. We then implemented and assessed eight ensemble method using the four finetuned models, analyzing their performance across multiple datasets. The findings showed that our proposed ensemble methods consistently improved classification accuracy and robustness. An ablation study provided deeper findings into the contributions of different model combinations. Comparing the best-performing ensembles with individual models and state-of-the-art methods, we observed significant accuracy improvements. This chapter demonstrated the substantial benefits of ensemble method, paving the way for enhanced diagnostic accuracy and reliability in medical imaging.

General Conclusion

This research focuses on enhancing medical image classification through the innovative use of ensemble methods combined with transformers. Ensemble methods, known for their robustness and ability to reduce variance, have been successfully applied in various fields such as computer vision, natural language processing, and financial forecasting. In medical imaging, transformers like ViT, DEiT, BEiT and Swin have shown promising results in tasks ranging from disease diagnosis to image segmentation and anomaly detection subsection 2.1.2. Our primary objective is to explore and validate the effectiveness of various ensemble methods in improving the accuracy, robustness, and reliability of medical image analysis across multiple datasets.

Initially, a comprehensive schema of the ensemble methods was created, detailing the fine-tuning process of the models and the specific implementation of various ensemble methods, such as hard voting, soft voting, weighted hard voting, weighted soft voting, stacking, bagging, and boosting. The implementation setup, hardware, and hyperparameter tuning for fine-tuning individual models were meticulously detailed to ensure clarity in the experimental process. The process began with fine-tuning each model on the entire training set for the six first ensemble methods. In contrast, for boosting and bagging, models were trained according to their respective methodologies, bagging involved fine-tuning on subsets of the training dataset, and boosting adjusted for errors iteratively. After fine-tuning the base models, their results were analyzed and discussed for each dataset. Each ensemble method was then implemented on each dataset using four transformer models, adhering to the original ensemble method logic. The results of these implementations were thoroughly discussed and analyzed, showcasing high accuracy and respectable results for other metrics such as precision, recall, and F1 score.

An ablation study was conducted to evaluate all possible combinations of each ensemble method and dataset, providing deeper findings into how different models influence the ensemble's performance. The best combination from each ensemble method and dataset was selected and the results were compared against individual models for each dataset. Additionally, the results were benchmarked against state-of-the-art methods from two articles that used accuracy as a performance metric. This comparative analysis highlighted that the proposed ensemble methods generally outperformed both individual models and state-of-the-art methods, demonstrating significant improvements in classification accuracy and robustness. The finding of ablation study is as follows: Chest classification achieved 97.59% accuracy using Stacking with SVM, with ViT and Swin as base models. Derma classification reached 80.30% accuracy with Weighted Hard Voting, using DEiT and Swin as base models. Blood classification achieved 98.07% accuracy with Weighted Soft Voting, using ViT, DEiT, and BEiT as as base models. Breast classification achieved 89.10% accuracy with Weighted Hard Voting, using ViT, DEiT, and Swin as base models. OrganC classification reached 93.75% accuracy with Weighted Hard Voting, using ViT, BEiT, and Swin as base models. OrganS classification achieved 83.91% accuracy with Bagging, using ViT, DEiT, BEiT, and Swin as base models. OrganA classification reached 95.14% accuracy with Soft Voting, using ViT, DEiT, and BEiT as base models. Finally, Pneumonia classification achieved 94.71% accuracy with Bagging, using ViT,

BEiT, and Swin as base models. Several strengths of this research were identified, including enhanced accuracy, robustness across various datasets, reduced overfitting, and effective performance even with small datasets, further highlighting their robustness and versatility.

Despite the high accuracy and robustness demonstrated by our ensemble methods, there is still room for improvement. Future work should focus on the following areas:

- Optimizing results using advanced optimization techniques to refine and enhance model performance.
- Avoiding excessive complexity to ensure methods remain practical and efficient for real-world applications.
- Exploring other advanced models, such as graph neural networks or sophisticated transformer variants, to further enhance performance.
- Developing novel ensemble methods and sophisticated ensemble strategies to achieve greater improvements in classification accuracy and robustness.
- Testing methods on more challenging and larger datasets to better understand scalability and generalizability.
- Utilizing advanced data augmentation techniques and improving preprocessing steps to address the scarcity of annotated medical images and enhance model performance.
- Enhancing the explainability and interpretability of ensemble models to gain clinician trust.
- Extending the framework to handle multi-modal data, combining images with clinical records or genetic information.
- Optimizing ensemble methods for real-time processing to make them viable for immediate clinical decision-making.
- Addressing the computational demands of boosting and bagging methods to reduce processing time, especially for large datasets.
- Further refining hyperparameter tuning to potentially achieve even higher accuracy.
- Developing techniques to handle smaller datasets effectively to ensure models are robust and versatile.
- Exploring the inclusion of more transformer models to achieve higher accuracy.
- Conducting benchmarking on standardized public datasets to compare performance with existing state-of-the-art methods.
- Minimizing the reliance on GPU resources by developing efficient methods and techniques to reduce computational costs, making models more accessible and cost-effective for deployment.

This research demonstrates the significant potential of ensemble methods with transformers in enhancing medical image classification. By addressing the outlined future directions, there is a promising path toward further advancements in medical image analysis, ultimately improving diagnostic accuracy and patient outcomes.

Bibliography

[1] PostDICOM, "Medical imaging: Types and modalities." https://www.postdicom.com/en/blog/medical-imaging-types-and-modalities.

[2] A. C. Müller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.

[3] N. Hordri, S. Yuhaniz, and S. M. Shamsuddin, "Deep learning and its applications: A review," 2016.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[5] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] M. Hosni, I. Abnane, A. Idri, J. de Gea, and J. Alemán, "Reviewing ensemble classification methods in breast cancer," *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 89–112, 2019.

[7] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, PMLR, 2021.

[8] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv*, 2021.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[11] E. Kambur, "Emotional intelligence or artificial intelligence," in *Proceedings of a conference or workshop*, Istanbul Aydin University, December 2021.

[12] H. Tatsat, S. Puri, and B. Lookabaugh, *Machine Learning and Data Science Blueprints for Finance*. O'Reilly media, 2020.

[13] S. Gupta, "Regression vs. classification in machine learning: What's the difference?," *Springboard Blog*, October 2021.

[14] D. Kurniawan, "Popular machine learning algorithms: Supervised and unsupervised learning." https://medium.com/mlearning-ai/popular-machine-learning-algorithms-supervised-and-unsupervised-learning-766120e Accessed 4 March 2022.

[15] A. Vidhya, "Building naive bayes classifier from scratch to perform sentiment analysis." https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/.

[16] JavaTpoint, "Machine learning - decision tree classification algorithm." url https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

[17] R. V, "All about knn algorithm." url = https://rekhavsrh.medium.com/all-about-knn-algorithm-6b35a18c2b15,.

[18] "Machine learning support vector machine algorithm." https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[19] GeeksforGeeks, "Ml — types of learning - part 2." https://www.geeksforgeeks.org/ml-types-learning-part-2/.

[20] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Sebastopol: O'Reilly Media, 2018.

[21] A. Bjorklund, J. Makel, and K. Puolamaki, "Explainable dimensionality reduction," January 2022.

[22] B. S. Rathore, "Unsupervised learning — dimensionality reduction (day 22)." url = https://medium.com/@bsinghrathore32/unsupervised-learning-dimensionality-reduction-day-22-4d60669d11f9,.

[23] H. Raghavan, "Principal component analysis (pca) explained and implemented." url = https://medium.com/@raghavan99o/principal-component-analysis-pca-explained-and-implemented-eeab7cb73b72,.

[24] W. Xu, H. Sun, C. Deng, and Y. Tan, "Variational autoencoders for semi-supervised text classification," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, (San Francisco, CA, USA), pp. 3358–3364, February 4–9 2017.

[25] Y. Li, "Deep reinforcement learning: An overview," *arXiv*, 2017.

[26] M. Paliwal, "Deep reinforcement learning," *Smart Innov. Syst. Technol.*, vol. 273, pp. 136–142, 2022.

[27] "Mcculloch-pitts model." url https://towardsdatascience.com/mcculloch-pitts-model-5fdf65ac5dd1. Accessed February 17, 2022.

[28] "Feedforward neural network." https://www.researchgate.net/figure/Feedforward-neural-network_fig1_329586439.

[29] C. M. Bishop, *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer, 2006.

[30] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 850–867, 2015.

[31] J. Patterson and A. Gibson, *Deep Learning: A Practitioner's Approach*. O'Reilly, 2017.

[32] T. Beysolow II, *Introduction to Deep Learning Using R: A Step-by-Step Guide to Learning and Implementing Deep Learning Models Using R*. Apress, 2017.

[33] "Pictorial representation of neural network." https://www.researchgate.net/figure/Pictorial-representation-of-neural-network_fig3_349578740.

[34] S. Saha, "A comprehensive guide to convolutional neural networks - the eli5 way." url https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53,, December 17 2018. Medium. Towards Data Science.

[35] R. Yamashita, M. Nishio, and R. K. G. e. a. Do, "Convolutional neural networks: An overview and application in radiology," *Radiological Physics and Technology*, vol. 11, no. 1, pp. 4–21, 2018.

[36] S. Patel, "A comprehensive analysis of convolutional neural network models," pp. 771–777, April 29 2020.

[37] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014.

[38] J. Hu, L. Shen, S. Albanie, G. Sun, and K. G. Derpanis, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[39] ResearchGate, "Detailed view of a transformer encoder block." https://www.researchgate.net/figure/Detailed-view-of-a-transformer-encoder-block-It-first-passes-the-input-through-a-fig1_352239001.

[40] Edlitera, "Transformers decoder block." https://www.edlitera.com/blog/posts/transformers-decoder-block.

[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, pp. 213–229, Springer, 2020.

[42] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023.

[43] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos, "Ensemble feature selection: Homogeneous and heterogeneous approaches," *Knowledge-Based Systems*, vol. 118, pp. 124–139, 2017.

[44] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[45] R. Ge, G. Feng, X. Jing, R. Zhang, P. Wang, and Q. Wu, "Enacp: An ensemble learning model for identification of anticancer peptides," *Frontiers in Genetics*, vol. 11, p. 760, 2020.

[46] X. Lu and B. Van Roy, "Ensemble sampling," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[47] I. Nti, A. Adekoya, and B. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," *Journal of Big Data*, vol. 7, no. 1, pp. 1–40, 2020.

[48] B. Hopkinson, A. King, D. Owen, M. Johnson-Roberson, M. Long, and S. Bhandarkar, "Automated classification of three-dimensional reconstructions of coral reefs using convolutional neural networks," *PloS One*, vol. 15, no. 3, p. e0230671, 2020.

[49] W. Khan, M. Ghazanfar, M. Azam, A. Karami, K. Alyoubi, and A. Alfakeeh, "Stock market prediction using machine learning classifiers and social media news," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–24, 2020.

[50] C. Soares, P. Brazdil, and P. Kuba, "A meta-learning method to select the kernel width in support vector regression," *Machine Learning*, vol. 54, no. 3, pp. 195–209, 2004.

[51] S. Kuruvayil and S. Palaniswamy, "Emotion recognition from facial images with simultaneous occlusion, pose, and illumination variations using meta-learning," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 9, pp. 7271–7282, 2022.

[52] F. Haghighi and H. Omranpour, "Stacking ensemble model of deep learning and its application to persian/arabic handwritten digits recognition," *Knowledge-Based Systems*, vol. 220, p. 106940, 2021.

[53] J. Monteiro, D. Ramos, D. Carneiro, F. Duarte, J. Fernandes, and P. Novais, "Meta-learning and the new challenges of machine learning," *International Journal of Intelligent Systems*, vol. 36, no. 11, pp. 6240–6272, 2021.

[54] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[55] P. Bühlmann and B. Yu, "Analyzing bagging," *Annals of Statistics*, vol. 30, no. 4, pp. 927–961, 2002.

[56] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the*, pp. 148–156, 1996.

[57] Y. Freund, R. Iyer, R. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, no. Nov, pp. 933–969, 2003.

[58] P. Smyth and D. Wolpert, "Stacked density estimation," in *Advances in Neural Information Processing Systems*, vol. 10, 1997.

[59] H. N. Chapman, "X-ray imaging beyond the limits," *Nature Materials*, vol. 8, pp. 299–301, Apr. 2009.

[60] N. B. Smith and A. Webb, *Introduction to Medical Imaging: Physics, Engineering and Clinical Applications*. Cambridge University Press, 2010.

[61] A. Smith and C. Jones, eds., *Handbook of Medical Imaging: Principles and Applications*. Springer International Publishing, 2020.

[62] M. E. H. Chowdhury, T. Rahman, A. K. Khandakar, and et al., "Can ai help in screening viral and covid-19 pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020.

[63] "RSNA intracranial hemorrhage detection." https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection.

[64] OASIS Brains, "OASIS brains." url https://www.oasis-brains.org/.

[65] W. Al-Dhabyani, M. Gomaa, H. Khaled, *et al.*, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, p. 104863, 2020.

[66] MedMNIST, "MedMNIST - a large-scale dataset for biomedical image classification." https://medmnist.com/, 2023.

[67] A. Khvostikov, K. Aderghal, J. Benois-Pineau, A. Krylov, and G. Catheline, "3d cnn-based classification using smri and md-dti images for alzheimer disease studies," *arXiv*, 2018. Preprint.

[68] Z. Liu, H. Lu, X. Pan, M. Xu, R. Lan, and X. Luo, "Diagnosis of alzheimer's disease via an attention-based multi-scale convolutional neural network," *Knowledge-Based Systems*, vol. 238, p. 107942, 2022.

[69] E.-D. Hemdan, M. Shouman, and M. Karar, "COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-ray Images," *arXiv*, 2020.

[70] P. Tiwari, B. Pant, M. Elarabawy, M. Abd-Elnaby, N. Mohd, G. Dhiman, and S. Sharma, "CNN Based Multiclass Brain Tumor Detection Using Medical Imaging," *Comput. Intell. Neurosci.*, vol. 2022, p. 1830010, 2022.

[71] M. Yildirim and A. Cinar, "Classification with respect to colon adenocarcinoma and colon benign tissue of colon histopathological images with a new CNN model: MA_ColonNET," *Int. J. Imaging Syst. Technol.*, vol. 32, pp. 155–162, 2021.

[72] V. Ravi, H. Narasimhan, C. Chakraborty, and T. Pham, "Deep learning-based meta-classifier approach for COVID-19 classification using CT scan and chest X-ray images," *Multimed. Syst.*, vol. 28, pp. 1401–1415, 2021.

[73] C. Srinivas, N. Prasad, M. Zakariah, Y. Alothaibi, K. Shaukat, B. Partibane, and H. Awal, "Deep Transfer Learning Approaches in Performance Analysis of Brain Tumor Classification Using MRI Images," *J. Healthc. Eng.*, vol. 2022, p. 3264367, 2022.

[74] L. Vijaysinh, "A Comparison of 4 Popular Transfer Learning Models." https://arxiv.org/abs/1603.08631, 2021. Accessed on 23 February 2023.

[75] N. Kumar, M. Gupta, D. Gupta, and S. Tiwari, "Novel deep transfer learning model for COVID-19 patient detection using X-ray chest images," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 469–478, 2021.

[76] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[77] L. Tanzi, A. Audisio, G. Cirrincione, and et al., "Vision transformer for femur fracture classification," 2021.

[78] L. Zhang and Y. Wen, "MIA-COV19D: A transformer-based framework for covid-19 classification in chest cts," *arXiv*, 2021.

[79] S. He, P. Grant, and Y. Ou, "Global-local transformer for brain age estimation," 2021.

[80] S. Perera, S. Adhikari, and A. Yilmaz, "Pocformer: A lightweight transformer architecture for detection of covid-19 using point of care ultrasound," *arXiv*, 2021.

[81] D. Song, B. Fu, F. Li, and et al., "Deep relation transformer for diagnosing glaucoma with optical coherence tomography and visual field function," *IEEE Transactions on Medical Imaging*, 2021.

[82] C. Liu and Q. Yin, "Automatic diagnosis of covid-19 using a tailored transformer-like network," in *Conference Series*, p. 012175, IOP Publishing, 2021.

[83] L. Yuan, Q. Hou, Z. Jiang, and et al., "Volo: Vision outlooker for visual recognition," *arXiv*, 2021.

[84] D. Shome, T. Kar, S. Mohanty, and et al., "Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare," *International Journal of Environmental Research and Public Health*, vol. 18, no. 21, p. 11086, 2021.

[85] R. Selvaraju, M. Cogswell, A. Das, and et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

[86] K. Krishnan and K. Krishnan, "Vision transformer based covid-19 detection using chest x-rays," in *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, pp. 644–648, IEEE, 2021.

[87] J. Than, P. Thon, O. Rijal, and et al., "Preliminary study on patch sizes in vision transformers (vit) for covid-19 and diseased lungs classification," in *IEEE National Biomedical Engineering Conference (NBEC)*, pp. 146–150, IEEE, 2021.

[88] G. Costa, A. Paiva, G. Junior, and et al., "Covid-19 automatic diagnosis with ct images using the novel transformer architecture," in *Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*, pp. 293–301, SBC, 2021.

[89] J. Li, Z. Yang, and Y. Yu, "A medical ai diagnosis platform based on vision transformer for coronavirus," in *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pp. 246–252, IEEE, 2021.

[90] X. Gao, Y. Qian, and A. Gao, "COVID-ViT: Classification of covid-19 from ct chest images based on vision transformer models," *arXiv*, 2021.

[91] L. Zhang and Y. Wen, "MIA-COV19D: A transformer-based framework for covid-19 classification in chest cts," *arXiv*, 2021.

[92] B. Gheflati and H. Rivaz, "Vision transformer for classification of breast ultrasound images," *arXiv*, 2021.

[93] J. Xie, Z. Wu, R. Zhu, and et al., "Melanoma detection based on swin transformer and simam," in *2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 5, pp. 1517–1521, 2021.

[94] N. AlDahoul, H. Karim, M. Tan, and et al., "Encoding retina image to words using ensemble of vision transformers for diabetic retinopathy grading," *F1000research*, vol. 10, no. 948, p. 948, 2021.

[95] K. Ikromjanov, S. Bhattacharjee, Y. Hwang, and et al., "Whole slide image analysis and detection of prostate cancer using vision transformers," in *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 399–402, IEEE, 2022.

[96] O. Nejati, "Medvit: A robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 161, p. 106210, 2023.

[97] X. Wang and et al., "Transpath: Transformer-based self-supervised learning for histopathological image classification," in *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference*, vol. 24 of *Proceedings, Part VIII*, (Strasbourg, France), pp. 186–195, Springer International Publishing, September 27–October 1 2021.

[98] Y. Dai and et al., "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.

[99] R. Leamons and et al., "Vision transformers for medical image classification," in *Proceedings of SAI Intelligent Systems Conference*, (Cham), pp. 319–325, Springer International Publishing, 2022.

[100] J. Jang and D. Hwang, "M3t: Three-dimensional medical image classifier using multi-plane and multi-slice transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20718–20729, 2022.

[101] J. Liu, Y. Li, G. Cao, Y. Liu, and W. Cao, "Feature pyramid vision transformer for medmnist classification decathlon," in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2022.

[102] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "Medvit: a robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 157, p. 106791, 2023.

[103] M. A. Cheema, M. A. Nawaz, M. Shoaib, M. Naeem, M. A. Iqbal, M. U. Sajjad, M. I. Malik, and M. T. Awan, "Use of ensemble learning to improve performance of known convolutional neural networks for mammography classification," *Diagnostics*, vol. 13, p. 2242, Dec. 2023.

[104] S. M. A. Mousavi and M. Shirazian, "Medical image classification utilizing ensemble learning and levy flight-based honey badger algorithm on 6g-enabled internet of things," *Journal of Medical Signals and Sensors*, vol. 12, no. 1, pp. 1–14, 2022.

[105] H. Zhu, W. Wang, I. Ulidowski, Q. Zhou, S. Wang, H. Chen, and Y. Zhang, "Meednets: Medical image classification via ensemble bio-inspired evolutionary densenets," *Knowledge-Based Systems*, vol. 280, p. 111035, 2023.

[106] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[107] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[108] "Python Documentation." Python Software Foundation. Retrieved from https://docs.python.org/.

[109] Google Research, "Colaboratory." Software. Retrieved from https://colab.research.google.com/.

[110] Kaggle Inc., "Kaggle: Your Machine Learning and Data Science Community." Retrieved from https://www.kaggle.com.

[111] Hugging Face, "Transformers: State-of-the-art Machine Learning for PyTorch, TensorFlow, and JAX." Software, 2022. Retrieved from https://github.com/huggingface/transformers.

[112] "PyTorch." https://pytorch.org/.

[113] K. S. Rudra Murthy, P. Yeluri, and S. Nanduri, "XLNet for Mining Mobile App Reviews," 2021.