Automated Histopathological Image Classification for Diagnosis and Prognosis using Deep Learning and Ensemble Techniques

NIRJHAR GOPE, ABID HAIDER, DEPONKER SARKER DEPTO, and DR. MAHDY RAHMAN CHOWDHURY

Incorporating deep learning methodologies into the histopathological classification of images has fundamentally transformed the field of diagnostic pathology, facilitating the automated, accurate, and highly efficient analysis of tissue specimens. The research presents a comprehensive and robust framework that leverages transfer learning and ensemble learning strategies to enhance classification performance on H & E stained Histological images. We explore several state-of-the-art pre-trained convolutional neural networks (CNNs), such as DenseNet121, InceptionResNetV2, Xception, NasNet mobile, MobilenetV2, ResNet50V2, and Vision Transformer (ViT), and assess their individual and collective performance. A meticulous k-fold cross-validation approach was implemented to ascertain model generalizability and robustness across a diverse array of histopathological samples. Our empirical findings indicate that ensemble models surpass the performance of singular designs, achieving a validation accuracy of 99.1% and an F1-score of 0.9908, which signifies enhanced classification performance compared to the existing method. Moreover, this research addresses pivotal challenges such as data scarcity, inter-class similarities, and intricate tissue morphology, demonstrating that deep learning methodologies can substantially elevate diagnostic accuracy. The outcomes of this investigation contribute to the progression of computational histopathology, facilitating the potential for more precise, automated, and scalable disease identification and cancer grading in clinical pathology.

Additional Key Words and Phrases: K-fold cross-validation, Ensemble learning,ViT transformer, Transfer Learning,CNNs, Histopathological image classification, Tissue classification, Medical image analysis, Automated Diagnosis.

ACM Reference Format:
Nirjhar Gope, Abid Haider, Deponker Sarker Depto, and Dr. Mahdy Rahman Chowdhury. 2025. Automated Histopathological Image Classification for Diagnosis and Prognosis using Deep Learning and Ensemble Techniques. 1, 1 (February 2025), 56 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Histopathological image analysis, the microscopic evaluation of tissue samples, is regarded as the gold standard for diagnosing and prognosis human cancers.[1]. Identifying nuclei is crucial for determining the type of the disease and significantly impacts both diagnosis and prognosis of diseases[2]. The visual assessment of patient conditions through digital histopathology images has become essential in oncology processes, allowing doctors to make diagnostic and prognostic decisions [3]. The careful visual evaluation of these images by qualified pathologists is essential for identifying the cancer's type, grade, and stage. This vital information directly impacts treatment strategies, forecasts patient outcomes, and ultimately

Authors' Contact Information: Nirjhar Gope, Nirjhargope@gmail.com; Abid Haider, abidhaider987@gmail.com; Deponker Sarker Depto, deponker.service3b@gmail.com; Dr. Mahdy Rahman Chowdhury, mahdy.photcourse@gmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM XXXX-XXXX/2025/2-ART
https://doi.org/XXXXXXX.XXXXXXX

determines the comprehensive clinical care of the disease [4], [5], [6]. Nevertheless, traditional approaches possess certain drawbacks. Traditional histopathological image analysis predominantly depends on the proficiency and experience of qualified pathologists [7], [8]. Human mistake is inevitable, leading to potential misinterpretations and misdiagnoses[9]. Moreover, the subjective nature of visual interpretation results in inconsistencies among pathologists, especially in complex cases where small variations in tissue shape may be tough to identify [10], [11]. Classifying cell nuclei in histological images of cancerous tissue presents a challenging issue due to the heterogeneity of cellular structures [12]. The procedure is labor-intensive, requiring much effort and expertise from skilled specialists [9]. The increasing volume of digital pathology data makes it more complicated for manual classification, highlighting the urgent need to develop and implement automated solutions. The limitations of manual analysis can be addressed by a deep learning approach, which has already proven effective in histopathology image classification for early illness detection and diagnosis. Besides deep learning, transfer learning and ensemble methods have been investigated to improve the efficacy and resilience of these classification models. [13], [14], [15]. Transfer learning, a method that involves utilising pre-trained models for a related task, has demonstrated significant efficacy in histopathology image processing. This method facilitates the effective use of feature representations acquired from extensive datasets, which can subsequently be refined for histopathology image classification. Moreover, ensemble strategies that incorporate the predictions of many models have demonstrated enhancements in the overall classification accuracy and generalization capacities of these systems[16]. These methodologies have been utilized across several cancer types, facilitating both accurate classification of cancer subtypes and the deduction of underlying genetic anomalies and tumour composition directly from histopathological images.[17]

1.1 Motivation

This research is motivated by the increasing demand for automated and dependable tools to aid healthcare practitioners in the prompt and precise diagnosis and prognosis of many diseases. Recent improvements in computer vision and deep learning have demonstrated encouraging outcomes in digital histology analysis. Nevertheless, there are still other obstacles that must be overcome. These include dealing with diverse and complicated histopathological pictures, efficiently extracting and classifying features, and making the models applicable to various datasets and clinical contexts. This research investigates the application of transfer learning and ensemble learning approaches to improve the efficacy of deep learning algorithms in multiclass histopathology image categorization.

1.2 Problem Statement

Despite significant advancements in deep learning and ensemble methods, numerous challenges persist in attaining robust and reliable automated classification of histopathological images.

Data Scarcity: Data scarcity poses a significant challenge in multiclass histopathological image classification. The limited availability of data and certain biases within datasets frequently result in overfitting, causing models to exhibit impressive performance on training data while underperforming on unseen data. The complexity and variability in histopathological images, resulting from differences in staining techniques, tissue preparation, and imaging conditions, introduce additional challenges[18].

High dimensionality and complexity in images: Histopathological images usually have high resolution and encompass complex details across various scales. [19]. The discriminative features in these images for differentiating between diseased and healthy classes are not immediately evident. Furthermore, distinct classes, such as healthy and various stages of disease, continue to share several geometric features [20].

Class imbalance:

In several histopathological datasets, the distribution of images across various disease classes or subtypes is often unequal[21]. The imbalance in class distribution can introduce bias in the training of machine learning models, resulting in suboptimal performance on under-represented classes[22]. To fix this imbalance, it is essential to employ precise data augmentation or resampling techniques.

Generalizability and robustness: Deep learning models developed using a particular dataset may not generalize well to unseen data acquired from different sources or under varying conditions[23]. Variations in staining protocols, tissue processing, and imaging parameters can significantly affect model performance. So, Creating robust and generalizable models capable of addressing these variations is essential for clinical applicability.

1.3 Reaearch gap

Recent advancements in deep learning and transfer learning approaches have substantially propelled the field of automated histopathology image categorization. Nonetheless, despite these encouraging advancements, a significant research gap persists in attaining state-of-the-art performance for multiclass classification problems within this domain. One key challenge in histopathological image analysis is the high resolution and complex visual features of these images, which can make it difficult for traditional deep learning models to classify them effectively. Moreover, the scarcity of large-scale, annotated histopathology datasets has consistently hindered the development of accurate and generalizable models. Although transfer learning techniques have demonstrated considerable advancements in addressing data limitations, the current state-of-the-art in this domain continues to encounter obstacles in attaining optimal performance, especially regarding the inherent variability and complexity of histopathological images, as well as in adapting these methods for multiclass classification scenarios. Future research in this field should investigate the development of innovative deep learning architectures specifically designed for histopathology image processing to tackle these persistent issues. The classification performance and robustness of these systems may also be improved by using advanced data augmentation techniques and investigating hybrid approaches that combine transfer learning with additional methods like few-shot learning or active learning.

1.4 Research Objectives

The main goal of this study is to develop a robust and precise deep learning framework for the automated classification of histopathological images, utilizing the advantages of transfer learning and ensemble methods. The proposed work is designed to achieve the following objectives:

Examine the impact of transfer learning through the fine-tuning of pre-trained deep learning models on histopathological image datasets, aiming to enhance classification accuracy while reducing the necessity for large-scale dataset collection and annotation.

Harnessing the power of ensemble techniques, including model averaging and weighted averaging, to integrate the predictions from various deep learning models and improve classification performance

Enhanced diagnostic accuracy: The suggested framework achieved a significant improvement in the accuracy of automated histopathological image classification when compared to the existing method[24].

This study aims to enhance automated histopathological image analysis and its role in clinical decision-making and disease diagnosis by tackling these research objectives. The study will enhance the wider domain of deep learning by offering valuable insights into the efficacy of various deep learning architectures, transfer learning approaches, and ensemble methods for analyzing histopathological images.

2 Literature Review

Several deep learning methodologies were utilized for the classification of histopathological images. Convolutional neural networks are the predominant architecture employed for the analysis of medical images, particularly in the context of histopathological images [25], [26]. A variety of studies has demonstrated the effectiveness of CNNs in categorizing various types of tissue, including renal tissue [27], colorectal cancer [28], lung cancer [26] [29], and breast cancer [25][30]. A CNN-LightGBM model achieving 99.6% accuracy in the categorization of lung cancer histopathology is presented in [26]. RNNs demonstrate remarkable capabilities in handling sequential data, particularly Long Short-Term Memory (LSTM) networks [31]. While LSTMs are not as prevalent as CNNs in tissue classification, they have shown promise in the analysis of temporal data, which could be essential for certain applications, such as monitoring tissue changes over time. In the domain of tissue classification, various deep learning architectures beyond CNNs and RNNs have been explored, including U-Net: This encoder-decoder architecture offers accurate tissue boundary delineation, making it highly beneficial for segmentation applications [32], [33]. The study [32] evaluates the performance of 2D versus 3D U-Net architectures for the classification of brain tissue in CT scans, revealing that 2D U-Nets demonstrated superior results compared to their 3D counterparts in this particular application. Transformers: The outcomes of these designs in image classification challenges have been promising, even though they were originally developed for natural language processing [28]. By employing a transformer model, [28] attained an impressive accuracy of 98.84% in their approach to colon cancer detection. The study presented in [34] employs ConvNeXts and Swin Transformers within an ensemble model to achieve a classification accuracy of 97.61% for mammogram breast density. Hybrid Models: Integrating different deep learning architectures can lead to improved performance by leveraging the strengths of each design[26]. Tissue-Specific Applications of Deep Learning: Deep learning is employed for the classification of various tissue types and in numerous therapeutic contexts. Brain tissue classification is essential for neurological diagnosis and surgical planning [35], [32], [36]. To comprehend the pathogens of kidney disease and identify therapeutic targets, it is crucial to classify cells within kidney tissue [27], [37]. To classify cells in human kidney tissue using only a DNA stain, [27] introduces an unsupervised approach that employs 3D tissue cytometry alongside a tailored 3D convolutional neural network (NephNet3D), achieving a balanced accuracy of 80.26Early diagnosis and improved patient outcomes rely on the detection and categorization of lung cancer [29], [26], and [38]. [29]investigates the application of different neural network architectures (CNNs, RNNs, and GNNs) for the classification and diagnosis of lung cancer.A CNN-LightGBM hybrid model is proposed for the classification of lung cancer histopathology in [26]. Analyzing histopathological images is essential for diagnosing breast cancer [25], [30], and [39]. In [25], convolutional neural networks are employed to classify cancer histology images, achieving a four-class classification accuracy of 77.8%. A system for classifying breast tumors as malignant or benign is described in [30]. The automated analysis of histological images contributes significantly to the diagnosis and classification of colon cancer.[28]This work introduces a transformer model grounded in deep learning, which integrates a CNN and a Siamese network to enhance the diagnosis and classification of colon cancer, attaining a precision of 98.84%. Furthermore, deep learning methods have been utilized across a range of tissue types, such as uveal melanoma [40], gastric tissue [41], [42], and thyroid tissue [43]. Examining the Impact of Transfer Learning and Ensemble Methods on Histopathological Image Classification: The application of transfer learning significantly enhances the accuracy of classification in histopathological image analysis. This effective method utilizes pre-trained models, usually developed on extensive, general-purpose datasets like ImageNet, and modifies them for a particular task, such as histopathological image classification [44], [45], [46]. Four popular CNN architectures – ResNet, VGG19, VGG16, and Inception – are adapted through transfer learning to

classify cardiovascular tissues in histological images [47]. Ensemble learning is a powerful method in deep learning that merges the predictions of various deep learning models to improve overall classification accuracy and robustness [48], [49], [50]. The fundamental idea of ensemble methods is that by combining the decisions of multiple models, the overall system is less vulnerable to the limitations and biases inherent in any individual model [51], [44], [45]. The inherent robustness and accuracy enhancements provided by ensemble learning render it an essential approach for achieving reliable and precise automated histopathological image classification.

3 Dataset

3.1 Dataset collection

The dataset employed in this study, derived from the NuInsSeg collection, primarily focuses on human histology images. The NuInsSeg dataset is a significant computational pathology resource, featuring comprehensive hand annotations and an extensive range of tissues. While the original dataset includes both human and mouse organs, our research solely employs the human tissue subset to address categorization challenges. The images are stained with Hematoxylin and Eosin (H&E), a conventional histological stain that emphasizes cell nuclei and cytoplasmic features. The NuInsSeg dataset consists of 665 image patches obtained from whole slide images (WSIs) featuring over 30,000 manually segmented nuclei from 31 different organs. The images were scanned at a high resolution of 2048x2048 pixels and then cropped to 512x512 pixels to improve processing efficiency. This guarantees that the dataset maintains adequate spatial detail while remaining practical for machine learning applications. This study employed a selected subset of human tissue pictures from the NuInsSeg repository. This subsection comprises 432 images from 22 distinct tissue types, providing a thorough array of human histology structures. The dataset serves as a vital tool for training and verifying classification algorithms that distinguish across tissue types.

3.2 Dataset Description

3.2.1 image statistics The dataset's images were scaled to a consistent resolution of 224x224 pixels, with three color channels. This downsizing phase enables interoperability with current convolutional neural networks, which often need input pictures of a set size. The initial distribution of pictures across tissue types was skewed, with counts ranging from 9 (muscle) to 47 (oesophagus). This imbalance results from the inherent challenge of collecting equal samples for all tissue types in histological research. Certain tissues are more commonly examined or included in databases due to their clinical importance, whereas others are underrepresented. Imbalanced datasets provide substantial issues for classification tasks, since models often perform poorly on underrepresented classes. To overcome this issue, a combination method of data augmentation and upsampling was used. Data augmentation adds variety to existing data by manipulations, whereas upsampling increases the number of samples in underrepresented classes by duplicating existing data.

3.2.2 class balancing The combination of augmentation and upsampling produced a dataset of 1,070 photos divided over 22 classes. The exact enhancement techniques used included: Rotation: Random rotations within a predefined range to imitate various orientations. Flipping involves horizontal and vertical flips to generate diversity in spatial representation. Scaling: Changing the image size to replicate zoom-in and zoom-out effects. Color Jittering: Changes in brightness, contrast, and saturation to compensate for stain irregularities. Noise injection is the addition of Gaussian noise to make the model more resilient to imaging artifacts. While augmentation introduced new variants, upsampling guaranteed that underrepresented classes received the necessary number of samples. For example, classes such as

muscle (originally 9 photos) were upsampled by duplicating existing photographs, resulting in a enough number of samples to balance the dataset. The final class distribution after balancing is as follows.

- Class 0: 53 images
- Class 1: 46 images
- Class 2: 45 images
- Class 3: 53 images
- Class 4: 49 images
- Class 5: 48 images
- Class 6: 47 images
- Class 7: 46 images
- Class 8: 53 images
- Class 9: 45 images
- Class 10: 53 images
- Class 11: 53 images
- Class 12: 52 images
- Class 13: 45 images
- Class 14: 53 images
- Class 15: 50 images
- Class 16: 44 images
- Class 17: 44 images
- Class 18: 45 images
- Class 19: 47 images
- Class 20: 47 images
- Class 21: 52 images

Balancing the dataset guarantees that all tissue types are represented fairly, solving issues related to class imbalance. This step is critical because it keeps the classification model from being biased toward overrepresented classes, which improves overall performance and reliability.

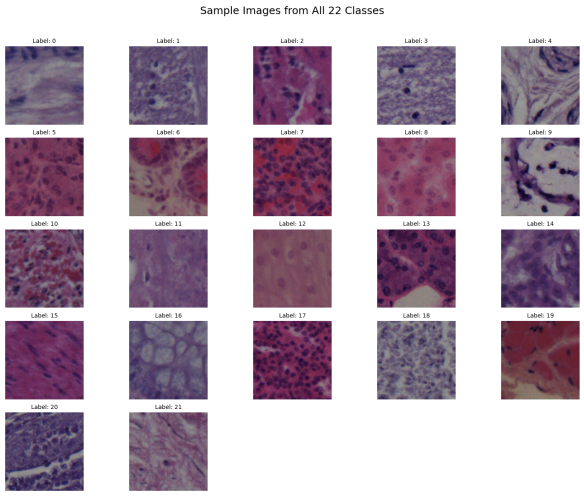

Fig. 1. Dataset Overview(Visual Label Mapping for 22 Classes)

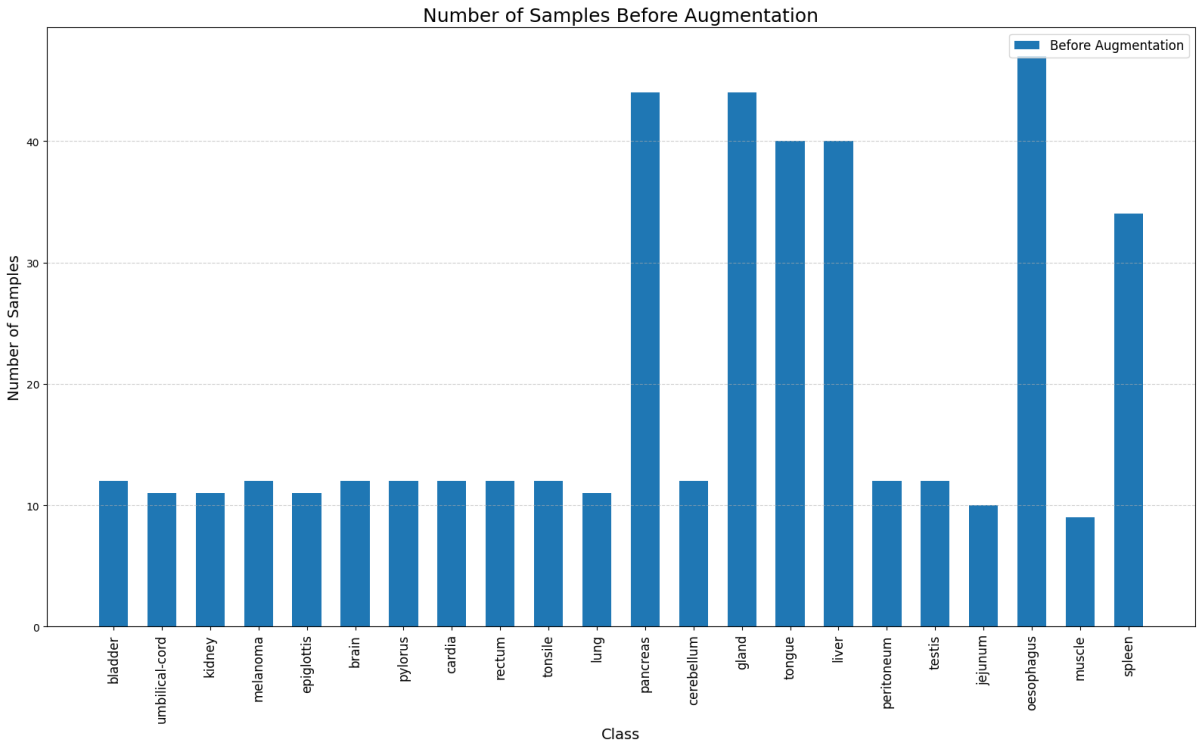

Fig. 2. Diverse(22-Class) Tissue Image Dataset

3.3 Data Pre-prossesing

3.3.1 Label Mapping Each tissue type was assigned a unique numerical label to facilitate machine learning tasks. This mapping ensures that the dataset is compatible with classification algorithms, which require categorical labels as input.

3.3.2 Data Augmentation and Upsampling Data augmentation and upsampling were utilized concurrently in the preparation workflow to provide a balanced and robust dataset. Data augmentation approaches gave further variability to the collection, thereby enhancing the model's capacity to generalize to unseen data. Augmentation included:

- Rotation: Simulating different orientations by applying random rotations.
- Flipping: Introducing horizontal and vertical flips to diversify spatial arrangements.
- Scaling: Adding variability in size by applying zoom-in and zoom-out transformations.
- Color Jittering: Creating variations in brightness, contrast, and saturation to account for staining inconsistencies in histological imaging.
- Noise Injection: Enhancing robustness by introducing Gaussian noise to simulate real-world imaging artifacts.

Upsampling was performed to address under representation in certain classes. By duplicating existing images, we ensured that every class had a sufficient number of samples for training. For instance, muscle

tissue images were duplicated to bring the count closer to that of other classes. This approach was carefully managed to avoid overfitting, as excessive duplication without augmentation could lead to poor generalization.

3.3.3 Splitting the Dataset The dataset was partitioned into training and validation sets using an 80/20 ratio. This division produced 856 photos for training and 214 for validation. The training set was expanded and adjusted to ensure that each class was well represented, allowing the model to learn successfully across all tissue types. To enable an unbiased model performance evaluation, no augmentation was applied to the validation set.

3.3.4 Standardization All images were normalized by scaling pixel values to the range [0, 1]. This step ensures uniformity across the All photos were normalized by scaling pixel values to a range of [0, 1]. This phase guarantees that the dataset is homogeneous, eliminating the influence of various pixel intensity ranges and allowing for efficient training of deep learning models.

Final Dataset Characteristics:

- Input size is 224x224 pixels with three channels (RGB).
- Total images: 1,070 (848 training and 222 validation).
- Number of Classes: 22 tissue types.
- Balanced Class Distribution: Achieved via augmentation and upsampling approaches

This preparation pipeline optimizes the dataset for deep learning model training by addressing issues including class imbalance, picture quality fluctuation, and computing efficiency. The careful mix of augmentation and upsampling approaches results in a stable and well-represented dataset suitable for creating and verifying tissue classification algorithms.

4 Methodology

4.1 Proposed workflow

The dataset-gathering method begins with data capture via Kaggle, followed by human tissue image extraction and preparation for analysis.

If the dataset is unbalanced, class balancing procedures like as upsampling or data augmentation are used to provide consistent representation across all classes. If the classes are balanced, the preprocessing operations are immediately begun.

All images are resized to a standardized dimension suitable for model training, and additional augmentations are applied to enhance the model's robustness.The dataset is divided into training and validation subsets to facilitate supervised learning and model evaluation.Transfer learning involves utilizing a variety of pre-trained transfer learning models.The models include DenseNet121, ResNet50V2, InceptionResNetV2, NASNet Mobile, Xception, MobileNetV2, and Vision Transformer (ViT).Each model is optimized to align with the objective of categorizing human tissue images. All models were trained using the training dataset, with subsequent predictions evaluated on the validation dataset. Standard evaluation metrics assess the performance of the model. The hyperparameters of the models are optimized to enhance performance, and the best model weights are kept for deployment. Ensemble methods are employed to enhance classification accuracy. Average Ensemble Predictions are aggregated through the average probabilities of various models, while weighted ensemble methods assign different weights to predictions from individual models based on their performance metrics. The ensemble model produces the final multiclass classification results for human tissue images. The approach results in effective multiclass image classification, demonstrating the efficacy of the proposed methodology.

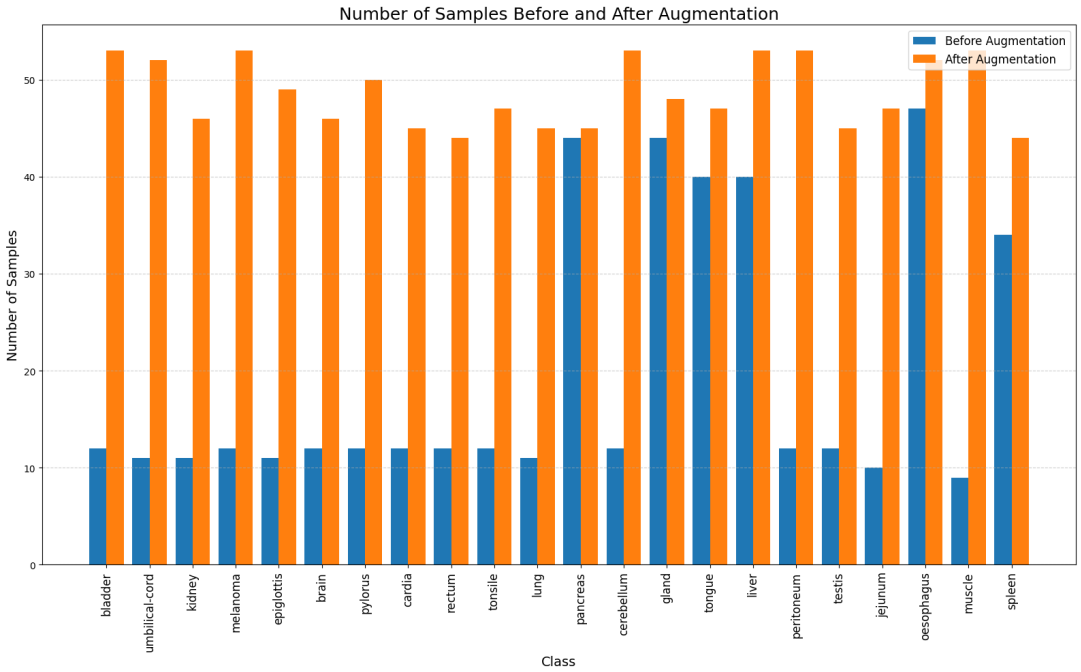

Fig. 3. A Balancing Act Class Distribution After Augmentation

4.2 Used Algorithms

These include DenseNet121, ResNet50 v2, Inception ResNetV2, NASNet Mobile, Xception, MobileNet v2, and Vision Transformer (ViT)

4.2.1 ResNet50v2 The ResNet50V2 architecture is a deep convolutional neural network designed for efficient and precise feature extraction and classification in image datasets. It is characterized by its residual connections, which reduce the vanishing gradient problem during training by allowing gradients to pass directly through shortcut links. This enhances the learning of deeper networks by ensuring effective gradient propagation. The network begins with an input layer that receives images of a specified dimension. A 7×7 convolutional layer with a stride of 2 retrieves low-level features such as edges and corners. To stabilize the learning process, a non-linear activation function and batch normalization are implemented after the convolution stage. A 3×3 max-pooling layer with a stride of 2 is subsequently employed to reduce the spatial dimensions of the feature map, enhancing computational efficiency while preserving key information. The architecture comprises four layers of residual blocks, each containing multiple bottleneck-based construction blocks. Each construction block comprises various components designed to enhance efficiency and gradient flow. Before convolutional operations, pre-activation is performed with batch normalization and ReLU activation, enhancing gradient stability during training. The bottleneck architecture comprises three convolutional layers: a 1×1 convolution that diminishes the number of feature channels for computational efficiency, a 3×3 convolution for spatial feature extraction, and a subsequent 1×1 convolution that reinstates the number of feature channels to its original count. Residual connections are integrated within each block, linking the block's input directly to its output. These

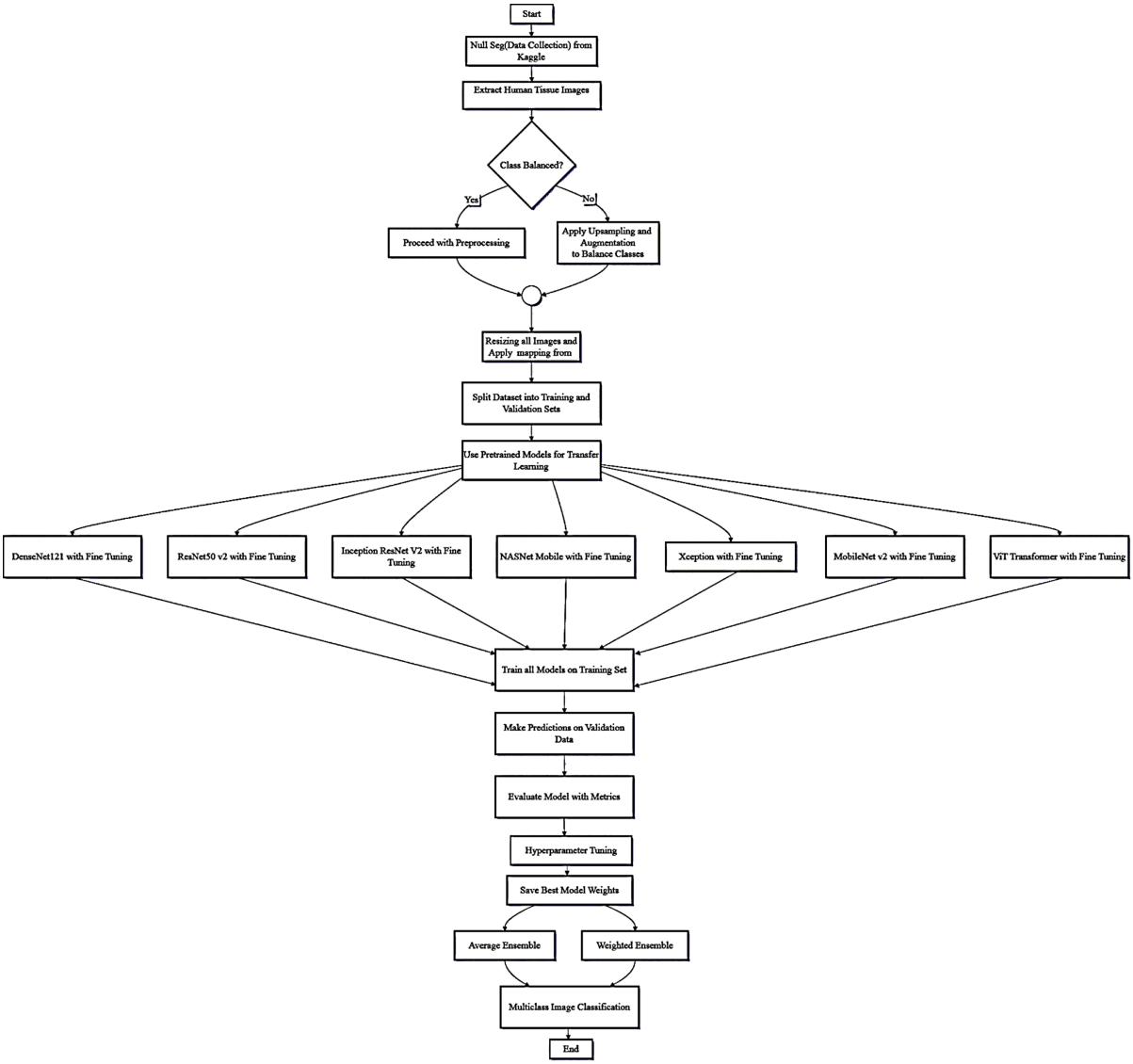

Fig. 4. Dataset Overview(Visual Label Mapping for 22 Classes)

shortcut connections facilitate efficient gradient propagation while alleviating the challenges inherent in deep networks. The remaining blocks are categorized into four phases, with an increasing number of filters and spatial reductions. The initial phase employs 64 filters to reduce the output dimensions through down sampling. The second stage employs 128 filters to facilitate further spatial reduction. The third step employs 256 filters to collect additional abstract information, whilst the fourth stage utilizes 512 filters to extract high-level features suitable for categorization. Upon the completion of all remaining blocks, global average pooling is employed. This method reduces each feature map to a singular value,

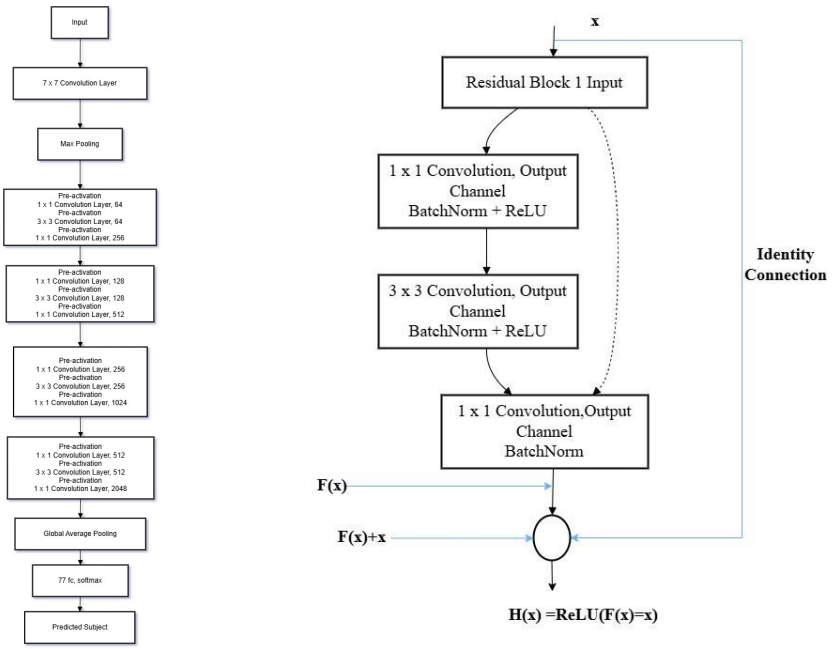

Fig. 5. Model architecture of ResNet50V2

thereby ensuring spatial invariance and diminishing the probability of overfitting. The output from the global average pooling layer is input into a fully connected layer with 77 neurons, corresponding to the number of target classes in the dataset. This layer employs a softmax activation function to produce a probability distribution among all classes. The network's ultimate output is the predicted class for the input image, determined by selecting the class with the highest probability.

4.2.2 InceptionResNetV2 The diagram presents the InceptionResNet architecture, a hybrid deep learning model that integrates the structural benefits of Inception modules with the efficiency of residual connections. The structure of this model is as follows: Stem: The initial component processes input images sized 299x299x3, reducing dimensions via convolutional and pooling operations to yield a feature map of size 35x35x256. Five consecutive Inception-ResNet-A modules further enhance the feature refinement. Each module incorporates parallel convolutional filters alongside residual connections, thereby improving learning capacity and minimizing computational complexity. The result of this phase is a feature map measuring 35x35x256. Reduction-A: This block executes dimensionality reduction, compressing the feature map to 17x17x896, thereby optimizing the model for deeper layers while preserving critical features. Inception-ResNet-B: The feature extraction process is furthered by ten Inception-ResNet-B modules, which are analogous to the Inception-ResNet-A modules but modified for diminished spatial dimensions. The output dimensions at this stage are 17x17x896. Reduction-B: A subsequent reduction module further diminishes the feature maps to 8x8x1792, thereby preparing the data for the concluding stages of classification. Inception-ResNet-C: Inception-ResNet-C consists of five modules that further process the feature maps by employing a combination of parallel convolutions and residual connections, yielding an output size of 8x8x1792. Average Pooling and Dropout: The model employs global average pooling to

decrease spatial dimensions, subsequently incorporating a dropout layer with a retention probability of 0 to address overfitting. Softmax: The final layer employs a softmax function to produce class probabilities for the classification task. This architecture integrates the depth and efficiency of Inception modules with the optimization advantages of residual networks, resulting in enhanced performance in image classification tasks.

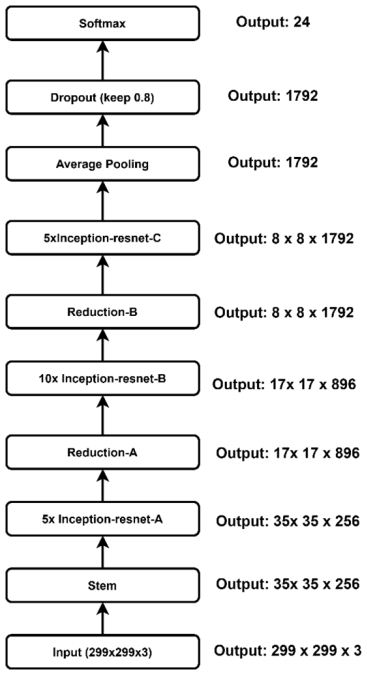

Fig. 6. Model architecture of InceptionResNetV2

4.2.3 Xception The Xception architecture is essentially a linear stack of convolution layers that are separable based on depth and have residual connections. [source] The entry flow, middle flow, and exit flow are the three primary components of the Xception architecture. 1. The Flow of Entry: The input image must be downsampled while key features are extracted by the entry flow. It includes the subsequent steps: • The first convolution: A convolutional layer with 32 3x3 filters is applied to the input image, followed by another convolutional layer with 64 3x3 filters. • Residual Blocks: The next layers consist of three residual blocks with depthwise separable convolutions. The number of filters gradually rises (128, 256, 728) to capture increasingly complicated information. The middle flow is the central component of the Xception architecture. It extracts complex information from the input image. It consists of eight identical residual blocks, each with three depth-wise separable convolution layers. These building elements enable the model to identify intricate links and patterns in the data. The Xception architecture is essentially a linear stack of convolution layers that are separable based on depth and have residual connections. [source] The entry flow, middle flow, and exit flow are the three primary components of the Xception architecture. 1. The Flow of Entry The input image must be downsampled while key features are extracted by the

entry flow. It includes the subsequent steps: • The first convolution A convolutional layer with 32 3x3 filters is applied to the input image, and then another convolutional layer with 64 3x3 filters.

• Residual Blocks: Three residual blocks with depthwise separable convolutions make up the next layers. In order to capture increasingly complicated information, the number of filters gradually rises (128, 256, 728). The central component of the Xception architecture, the middle flow is made to extract complex information from the input image. It is made up of eight identical residual blocks with three depth-wise separable convolution layers in each. These building elements enable the model to identify intricate links and patterns in the data.

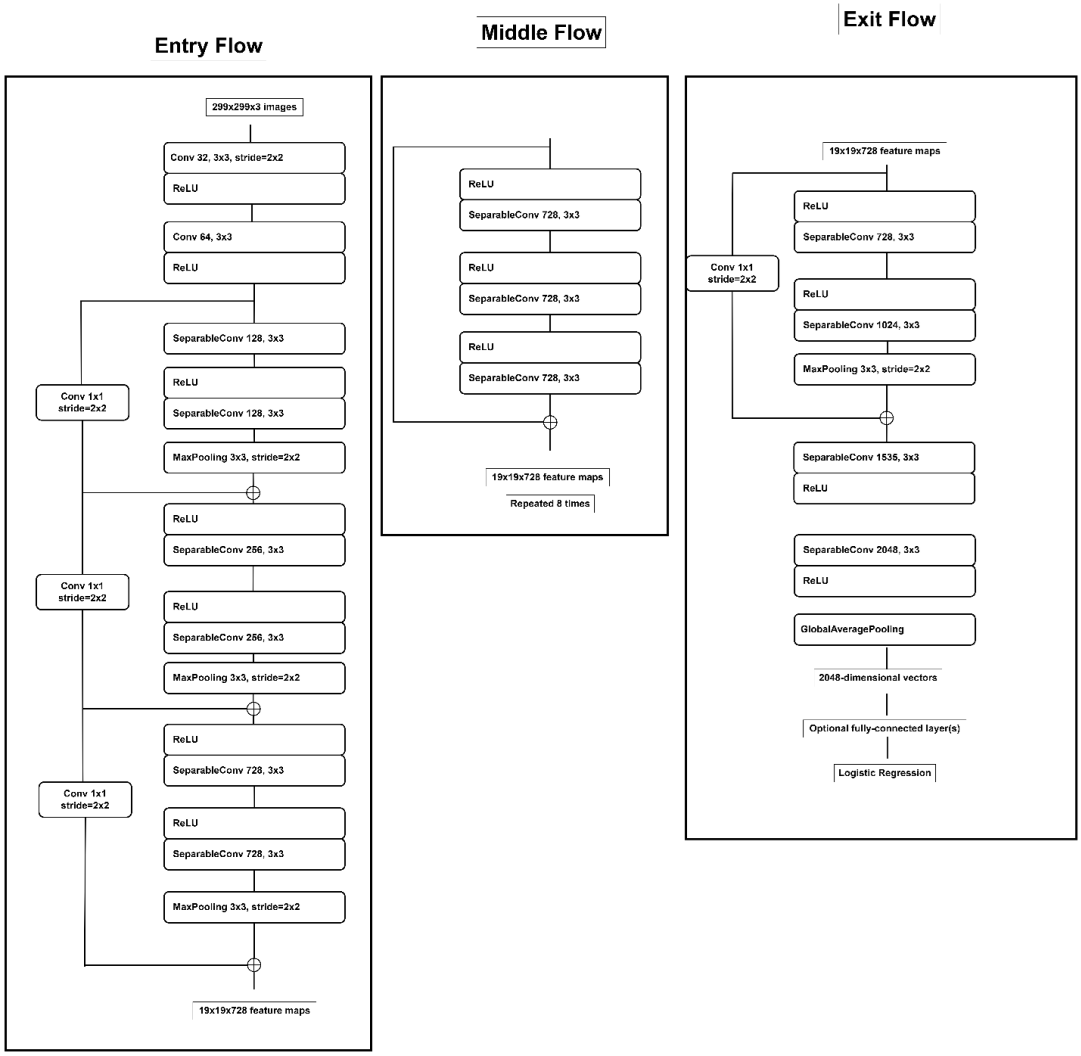

Fig. 7. Model architecture of Xception

4.2.4 Densenet121 In this work, we use the DenseNet-121 architecture to classify images. DenseNet-121 is a convolutional network with dense connections that improves information flow between layers by feed-forward connecting each layer to every other layer, as seen in Fig. 1. An input layer with 224 x 224 x 3 images is used to start the model. This is followed by an initial convolutional layer with 64 7 x 7 filters and a stride of 2. Then come a max pooling layer with a stride of two, a filter size of 3 × 3, a batch normalization layer, and a ReLU activation function. Four dense, progressively more intricate blocks make up the network. Each of the six layers in the first dense block consists of two convolutions that are repeated six times. A transition layer, comprising a 2×2average pooling layer and a 1×1 convolution, comes next. Twelve layers make up the second dense block, which has two convolutions repeated twelve times before a second transition layer. The third dense block has 24 layers, including a transition layer after two convolutions that are repeated 24 times. Two convolutions are repeated 16 times across 16 layers in the last dense block. A final batch normalization layer, a ReLU activation function, a 7 x 7 average pooling layer, and a fully connected layer with 1945 output units are included in the model after dense blocks. Compared to conventional convolutional networks, our all-inclusive design enables effective feature reuse, improving performance and lowering the number of parameters.

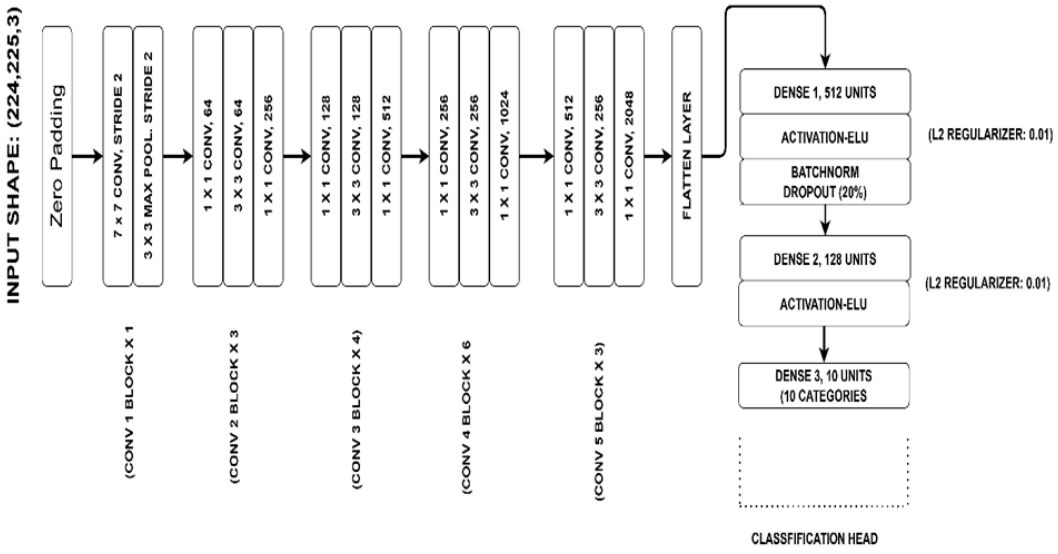

Fig. 8. Model architecture of DenseNet121

4.2.5 MobileNetv2 For image classification tasks, we use the MobileNetV2 architecture in our work. A lightweight and effective convolutional neural network for embedded and mobile vision applications. The model starts with an input layer made for 224 × 224 × 3 pixel pictures. Batch normalization and ReLU activation are then added, and a 3 × 3 convolutional layer with 32 filters and a stride of 2 follows. Inverted residual blocks with linear bottlenecks make up MobileNetV2's core. In addition to batch normalization and ReLU activation, each block has depthwise convolutions, expansion layers with 1×1 convolutions, and pointwise convolutions to modify channel dimensions. To improve the efficiency of feature extraction, these blocks are repeated [13]. Following a number of inverted residual blocks, the network employs a

flattening layer and a global average pooling layer to combine spatial information. Class probabilities are output by a softmax activation function and a fully linked layer, which are the last layers. Because of its design, MobileNetV2 can effectively extract features with fewer parameters, which makes it appropriate for scenarios with low resources.

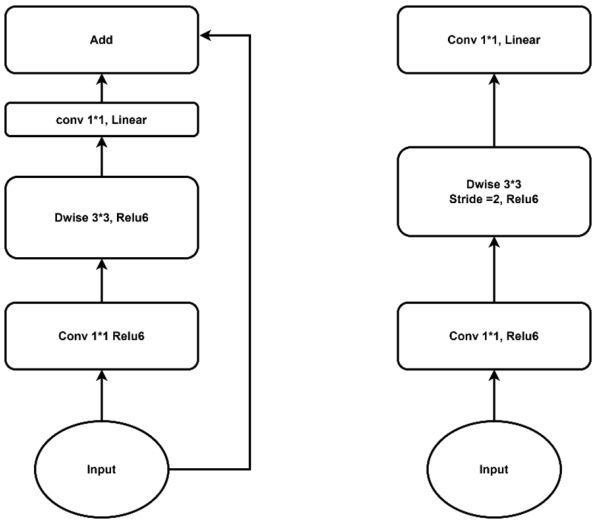

Fig. 9. MobileNetv2 architecture

4.2.6 Nasnet Neural architecture search (NAS) is a contemporary deep learning technique within the domain of artificial neural networks (ANN). Proposed by the Google Brain team in 2016, it comprises three components: search space, search strategy, and performance estimation [52]. Search space involves searching for convoluted performances, fully-connected, max-pooling, etc., and then checking the connection between the layers through which complete feasible network architectures are formed. The search approach employs random search and reinforcement learning to sample the population of network architecture candidates by obtaining performance rewards for child models (maximum accuracy, time efficiency). Simultaneously, the primary objective of performance estimation is to reduce computing resources or time regulation of network architecture, so that the performance is estimated at the search strategy position when receiving the child model performance[53][54]. Model architecture collected from [55]

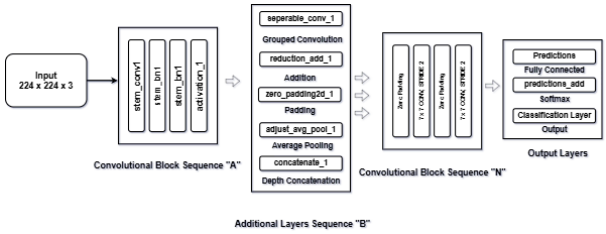

Fig. 10. NasNet Mobile architecture

4.2.7 ViT The Vision Transformer (ViT) is a deep learning architecture that employs transformer-based structures for image classification, utilizing the self-attention mechanism derived from natural language processing. The model begins with a fixed-dimension input image, such as 224×224, divided into fixed-size patches (e.g., 16×16). The patches are transformed into one-dimensional vectors prior to passing through a linear projection layer to produce fixed-size embedding vectors. Positional embeddings are employed to preserve spatial information, and the resulting sequence is sent into transformer blocks.

Each transformer block comprises multiple self-attention heads that capture both local and global interactions across patches, alongside a feed-forward neural network that enhances feature representations. Layer normalization guarantees stability, while shortcut connections facilitate efficient gradient propagation and accelerate convergence.

A unique [CLS] token is appended to the sequence initially to represent the whole image for classification. After passing the transformer blocks, the embedding of the [CLS] token is extracted and forwarded to a fully connected classification head, which employs a softmax activation function to predict class probabilities.

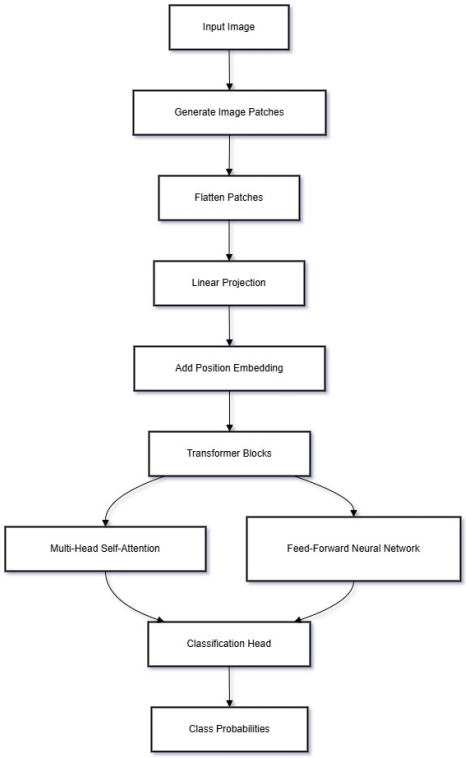

Fig. 11. Model architecture of ViT

4.3 K-Fold Cross Validation

The dataset is shuffled to guarantee that the data is spread randomly, hence preventing any bias caused by data order. After shuffling, the dataset is partitioned into K equal-sized folds. For example, if K=5,

the dataset is separated into five folds of almost similar size. In each iteration of the procedure, one fold is chosen as the validation (or test) set, and the remaining K-1 folds are merged to create the training set. The model is trained on the training set, which consists of K-1 folds. Once the training is finished, the trained model is assessed on the validation fold to determine performance measures like accuracy, F1 score, and other important metrics. The performance score for the current fold is recorded. The procedure of picking one fold as the validation set, training the model on the remaining folds, and recording the evaluation score is repeated for all K folds. Each fold has the potential to be the validation set just once during the procedure. This guarantees that the model is assessed throughout the whole dataset, resulting in a full assessment of its performance.

The final evaluation metrics are obtained after looping through all K folds. These measures, such as mean accuracy or mean F1 score, are produced by averaging the values from all folds. This technique assures that the assessment findings are reliable and reflect the model's performance over the full dataset.

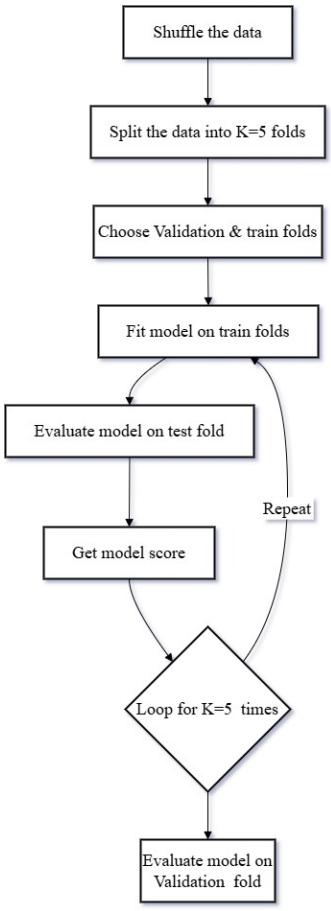

Fig. 12. K-Fold Cross Validation workflow diagram

4.4 Ensemble Techniques

The procedure begins with the training dataset, which has been produced and preprocessed. This dataset is used to refine three deep learning models:DenseNet121, ResNet50V2, and InceptionResNet V2 models. Each of the three models is fine-tuned using the training dataset to learn key characteristics and enhance performance. Fine-tuning involves training the models on the task-specific dataset while using the pre-trained weights as a baseline.DenseNet121 is a densely linked convolutional neural network that enhances gradient flow and lowers vanishing gradients by feeding each layer into the next.ResNet50 V2 is a residual network that uses skip connections to make deep network training easier.InceptionResNet V2 combines the Inception architecture with residual connections to capture both broad and deep characteristics. Following training, each model generates prediction probabilities for each class.Predictions from DenseNet121 and ResNet50 V2 (p, p)InceptionResNet V2 predictions (p).To use the Average Ensemble, input the prediction probabilities (p1, p2, and p3) from the three models.Operation: Calculates the arithmetic mean of the probability across all models.Ensemble approaches increase classification accuracy and resilience by integrating the characteristics of distinct models.

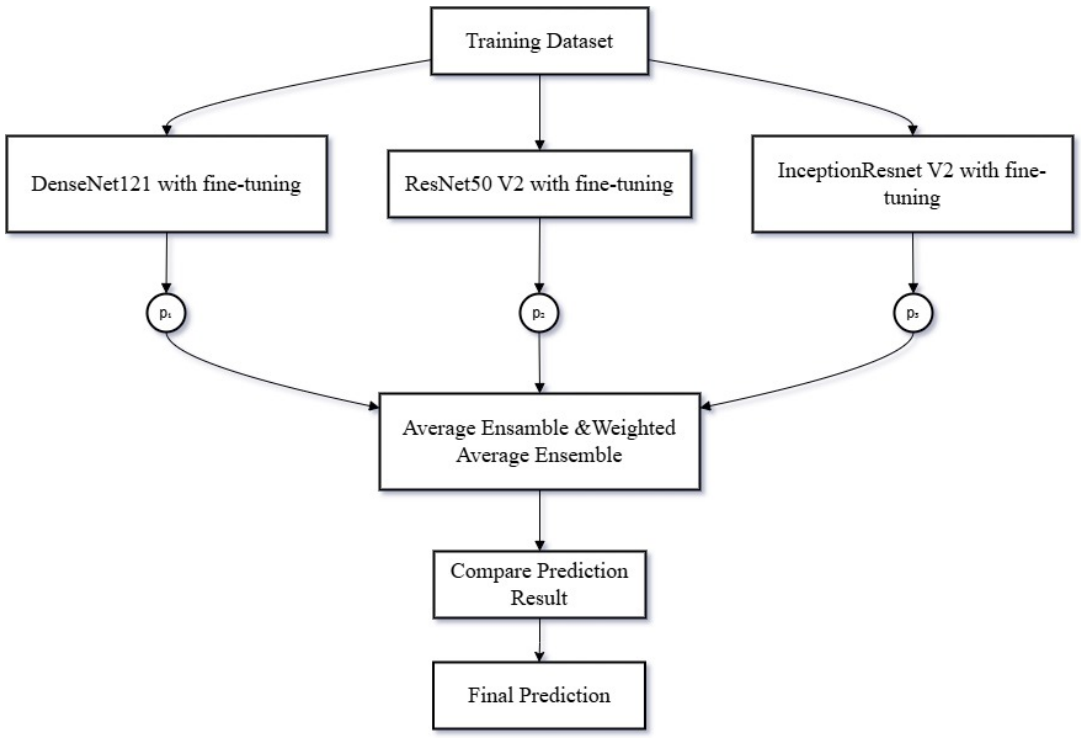

Fig. 13. Ensemble Techniques workflow diagram

4.5 Implementation

Transfer learning is an effective deep learning method that allows a pre-trained model, that learned from a large dataset, to be adapted for a new but related task. This method leverages the characteristics learned by the model from extensive data, thereby reducing the need for significant computational resources

and large datasets for training a new model. Transfer learning is particularly advantageous for human tissue image classification, as it enables the utilization of knowledge acquired from extensive datasets like ImageNet for the specific task of histological image analysis. Utilizing a pre-trained model significantly reduces both time and financial resources in comparison to developing a model from scratch . Transfer learning consists of two fundamental phases: feature extraction and fine-tuning. In feature extractor, the convolutional layer of the pre-trained model functions as a fixed feature extractor . The convolutional base layers, having acquired the ability to identify fundamental components such as edges, textures, and shapes, provide a foundational framework. The characteristics are often sufficiently generic to be applicable across various activities, including the categorization of human tissues. In the fine-tuning phase, the upper layers of the model are unfrozen and re-trained alongside newly added layers to enhance the model's suitability for the specific task. Fine-tuning ensures that the model's weights are adjusted in a more the specific task manner, thereby enhancing performance for the new challenge. Transfer learning for the classification of human tissue images necessitates several clearly defined phases. Select a pre-trained model, such as DenseNet121, which has undergone training on the ImageNet dataset comprising millions of labeled images. DenseNet121 is recognized for its efficient parameter utilization and capacity to maintain feature reuse throughout the network. When utilizing DenseNet121 or other pre-trained models like InceptionResNetV2, Xception, NASNet, MobileNet, or ResNet50V2, the top classification layer is typically excluded, retaining only the convolutional base. This facilitates the development of specialized layers designed for the human tissue classification task, ensuring that the pre-trained model focuses exclusively on extracting general features while the new layers acquire task-specific characteristics. After loading the model, the pre-trained layers should be frozen. Freezing the layers indicates that their weights remain unchanged throughout the training process. This preserves the model's existing knowledge from the extensive ImageNet dataset, ensuring retention of the broad features it has learned, including forms and textures. Freezing the layers ensures that only the newly added, task-specific layers undergo training, thereby optimizing classification of human tissue images.

Following that, the custom classification layers are built and implemented in the model. Typically, these layers contain a Global Average Pooling (GAP) layer, which minimizes the spatial dimensions of the feature maps while keeping the most significant characteristics. This helps to reduce computing costs while preserving vital information. Fully linked layers are then added to learn high-level representations relevant to the job at hand. Regularization approaches, such as dropout, are used to prevent overfitting during training. The output layer of the network is often a softmax layer, which is appropriate for multi-class classification applications such as tissue categorization. The output layer has the same number of neurons as there are tissue classifications in the dataset.

After determining the model architecture, the following step is to compile it. The model is built with the proper optimizer, loss function, and evaluation metric. For example, the Adam optimizer is widely used since it is efficient and adjusts the learning rate throughout training. Categorical cross-entropy is a common loss function for multi-class classification problems, and accuracy is frequently used as an assessment parameter to track model performance during training and validation. Several callbacks are used during the training phase to help students learn more effectively. Model checkpointing stores the best model based on validation accuracy, allowing the user to obtain the model with the best performance throughout training. Early stopping is used to monitor validation performance, and training is terminated if the model's performance does not increase after a set number of epochs. Learning rate reduction is another strategy used to alter the learning rate when the model's validation performance plateaus, allowing for further fine-tuning of the model. TensorBoard is used to show training and validation metrics, providing insights into the model's performance over epochs and assisting in the identification of possible faults like as overfitting or underfitting.

To better comprehend the model's training process, it is usual to plot the training and validation accuracy and loss over the epochs. These plots give essential insights into the model's behavior, assisting in the identification of patterns such as overfitting or underfitting, as well as allowing for future model modification. Adjusting the model based on these insights can improve performance, resulting in a more accurate and robust classification system for human tissue pictures. The same process is used for implementing transfer learning with various pre-trained architectures such as InceptionResNetV2, Xception, NASNet, MobileNet, or ResNet50V2. Each of these designs has distinguishing qualities that make them appropriate for a variety of applications. For example, InceptionResNetV2 is noted for its deep learning skills with residual connections, but Xception employs depthwise separable convolutions for more effective feature extraction. NASNet automatically optimizes neural network architecture, whereas MobileNet is especially built for mobile and resource-constrained situations. ResNet50V2, like DenseNet121, use residual connections to train deeper networks and avoid vanishing gradient concerns. All of these models may be applied to the human tissue picture classification job in a similar fashion, with the important stages being to load the pre-trained model, freeze the basic layers, add custom classification layers, and fine-tune the model to fit the task at hand. Deep learning models can use transfer learning to exploit past information gained from huge, general-purpose datasets, considerably increasing their performance on specialized tasks such as human tissue image categorization. We can construct powerful, efficient, and accurate models for evaluating histology pictures, with applications in medical diagnostics and research, by fine-tuning these pre-trained models to specific requirements.

The Vision Transformer (ViT) model is a novel deep learning architecture designed to perform well on picture categorization tasks. Unlike typical convolutional neural networks (CNNs), which interpret pictures using convolutional layers, the Vision Transformer considers images as sequences of fixed-size patches. These patches are then flattened and processed similarly to words in natural language processing, allowing the model to detect long-term connections and linkages in the image using self-attention processes. The Vision Transformer uses a transformer-based design, which has proven extremely successful in natural language processing, to give a strong alternative to CNNs for picture categorization. To begin implementing Vision Transformer for human tissue picture categorization, we must first import the appropriate libraries, which include torch-vision.models for loading pre-trained models and torch.nn for constructing custom layers. The first step is to specify the class names for our dataset. In this example, we're working with 22 different types of tissues, including bladder, brain, kidney, liver, and melanoma. These class names will match the dataset's labels, allowing the model to appropriately categorize the Images. Next, we get the pre-trained weights for the ViT-Base model. The Vision Transformer model is pre-trained using the ImageNet dataset, which comprises millions of tagged pictures. We employ these pre-trained weights to capitalize on the model's learnings from a variety of Images. We specifically utilize the ViT default class weights, which has a default set of weights for the ViT Base model with a 16x16 patch size. This pre-trained model is then put into the device (either CPU or GPU, depending on availability) and configured for fine-tuning of the new job. After loading the model, the following step is to freeze the Vision Transformer's pre-trained layers. Freezing these layers prevents their weights from being modified during training, allowing the model to maintain the information it gained from ImageNet while focused on learning task-specific characteristics. This is an important component of transfer learning because it allows us to apply a model that was previously trained on a wide dataset to a new, specialized dataset without having to retrain the entire model. The next step after freezing the base parameters is to alter the model's classifier head. The initial classifier head of the ViT model is intended for ImageNet, which includes 1,000 classes. Because our assignment includes 22 tissue classifications, we replace the original classifier head with a new fully linked layer. The new layer has 768 input features (equivalent to the ViT model's hidden size) and 22 output features, which correspond to the number of

tissue types in our dataset. This change guarantees that the model is suited to the new goal while still having the capacity to extract broad information from the picture. After modifying the model, we build it by defining the optimizer and the loss function. The Adam optimizer is used for training, with a learning rate of 1e-4 to provide consistent and efficient training. We utilize CrossEntropyLoss as the loss function, which is ideal for multi-class classification jobs. This function calculates the difference between predicted probabilities and true labels, which the model attempts to reduce throughout training. The next step is to define a custom training function that will manage the training and validation operations over different epochs. The train_model function takes the model, training and validation data loaders, optimizer, and loss function as arguments, along with the number of epochs and device requirements. In this function, the model is trained in a loop for the requested number of epochs. In each epoch, the model goes through two major phases: training and validation. During the training phase, the model is switched to training mode using model.train, and the model parameters are adjusted depending on the loss derived from the training data. The pictures and labels are transmitted to the appropriate device (GPU or CPU), and the optimizer's gradients are zeroed prior to the forward pass. After computing the loss, the backward pass is conducted, which updates the model's weights. The model's accuracy is then computed by comparing the predicted labels to the genuine labels, and the loss is added together to monitor the model's performance. Following each period of training, the model enters the validation phase, when it is switched to evaluation mode using model.eval. During this phase, the model's performance on the validation dataset is assessed without changing the weights. The loss and accuracy are estimated in the same way as during training, but no gradients are calculated. In addition, the F1-score is computed for the validation set to offer a balanced measure of accuracy and recall, which is very relevant when dealing with unbalanced data. Throughout the training process, we record the training and validation accuracies, losses, and F1-scores for each epoch. These metrics give useful information about the model's learning process and can assist determine if the model is overfitting or underfitting. The model's performance is closely checked, and the best-performing model (based on validation loss) is stored at each epoch to ensure that the best weights are maintained. The training ends when the model achieves the set number of epochs or when early stopping conditions are fulfilled (if implemented).

After training is completed, we assess the model's final performance on the validation dataset. The final validation accuracy and F1-score are printed to evaluate the model's generalizability. These metrics show how successfully the model learns to categorize tissue pictures, taking into consideration accuracy and recall. The training accuracy is also presented, indicating how well the model fits the training data. Finally, we show the training and validation parameters, such as accuracy, loss, and F1-score, to better comprehend the learning process. Plotting these measures over epochs allows you to discover tendencies like overfitting, which occurs when the model performs well on training data but badly on validation data. By evaluating these graphs, we may enhance the model's performance by adjusting the learning rate, altering the optimizer, or adding regularization techniques.

To deploy the Vision Transformer for human tissue picture classification, load a pre-trained ViT model, tweak its classifier head to fit the specific job, and fine-tune it on the target dataset. This technique takes advantage of transfer learning to drastically shorten training time and increase model performance, while also implementing a cutting-edge transformer architecture for image categorization. We can build a robust and effective system for classifying human tissue images, with applications in medical diagnostics and research, by combining the Vision Transformer with techniques such as layer freezing, hyperparameter optimization, and model performance evaluation using metrics such as accuracy and F1-score.

Ensemble approaches are effective ways for increasing the performance of machine learning models by mixing predictions from several models. To aggregate the predictions from all three models, this strategy uses two ensemble methods: DenseNet121, ResNet50 V2, and InceptionResNet V2. The first technique is

the Average Ensemble, which averages the predictions, or probabilities, of all three models for each class. To decide the final class label, the class with the greatest average probability is chosen. This strategy is based on the notion that averaging predictions from numerous models can help reduce the mistakes of individual models, resulting in a more robust forecast.

$$P_{\text{avg}} = \frac{p_1 \;\; p_2 \;\; p_3}{3} \tag{1}$$

The second ensemble approach is the Weighted Average Ensemble. This strategy assigns weights to the predictions of each model depending on their unique performance. These weights may be calculated using several performance measures such as accuracy or F1 score, which indicate how well each model performed on the validation set. The weighted average is obtained by multiplying each model's prediction probabilities by their corresponding weights. These weights, represented by w1, w2, and w3 for DenseNet121, ResNet50 V2, and InceptionResNet V2, respectively, are normalized such that their sum equals one, guaranteeing that the weighted average is properly balanced.

$$P_{\text{weighted}} \;\; = w_1 \cdot p_1 \;\; w_2 \cdot p_2 \;\; w_3 \cdot p_3 \tag{2}$$

The procedure begins with the training dataset, which has been produced and preprocessed. This dataset is used to refine three deep learning models:DenseNet121, ResNet50 V2, and InceptionResNet V2 models. Each of the three models is fine-tuned using the training dataset to learn key characteristics and enhance performance. Fine-tuning involves training the models on the task-specific dataset while using the pretrained weights as a baseline.DenseNet121 is a densely linked convolutional neural network that enhances gradient flow and lowers vanishing gradients by feeding each layer into the next.ResNet50 V2 is a residual network that uses skip connections to make deep network training easier.InceptionResNet V2 combines the Inception architecture with residual connections to capture both broad and deep characteristics.Following training, each model generates prediction probabilities for each class.Predictions from DenseNet121 and ResNet50 V2 (p, p)InceptionResNet V2 predictions (p).To use the Average Ensemble, input the prediction probabilities (p1, p2, and p3) from the three models.Operation: Calculates the arithmetic mean of the probability across all models. Output: The final class prediction with the highest average probability.To use the weighted average ensemble, input the prediction probabilities (p1, p2, and p3) and weights (w1, w2, and w3).Operation: The weighted sum of the probability is computed using the specified weights.Output: The final class prediction is based on the highest weighted average probability. Ensemble approaches increase classification accuracy and resilience by integrating the characteristics of distinct models.

After obtaining predictions from both the average and weighted ensemble procedures, a comparison is performed to determine which method produces the most accurate forecasts. This comparison evaluates each ensemble method's performance on the validation dataset using multiple evaluation measures including as accuracy, precision, and recall. By evaluating these measures, we may identify which ensemble approach gives the most accurate and dependable predictions for the multiclass classification problem.

Finally, the ensemble approach with the highest performance across these assessment measures is chosen to provide the final forecast. This final prediction is based on the outputs of the selected ensemble approach, which are subsequently utilized to categorize additional data examples. By using the power of ensemble approaches, we may increase the resilience and accuracy of our model's predictions, assuring greater generalization to unknown data and, ultimately, improving the performance of the multiclass classification problem.

5 Result and Discussion

5.1 Evaluation Matrics

In this part, we thoroughly investigate the assessment criteria utilized to analyze the effectiveness of our classification models. These metrics give quantifiable information about the model's accuracy, dependability, and general performance. Our investigation used the following metrics: precision, recall, F1 score, validation accuracy, weighted average accuracy, training accuracy, confusion matrix, and ROC AUC curve. We cover each statistic in detail, including how it is derived and its usefulness for classification tasks.

5.1.1 Precision Precision is the proportion that accurately predicted positive cases among all instances that were projected to be positive. It evaluates the accuracy of the model's positive predictions. Precision for a given class may be estimated using the following formula:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} \quad \text{False Positives (FP)}} \tag{3}$$

Precision focuses on reducing false positives and is especially critical in situations where inaccurate positive predictions might have serious implications.

5.1.2 Recall Recall, also known as sensitivity or true positive rate, is a measure of the model's ability to properly identify genuine positive cases. It indicates the model's capacity to recognize positive situations, as indicated by the formula:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{4}$$

A high recall suggests that the model accurately catches the majority of positive instances, which is significant in situations where missing positive examples might be critical.

5.1.3 F1-score The F1 score is the harmonic mean of accuracy and recall, resulting in a single metric that balances the two measurements. It is especially effective in circumstances with unbalanced datasets in which one class may outperform others. The formula for calculating the F1 score is:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} \quad \text{Recall}} \tag{5}$$

The F1 score guarantees that both false positives and false negatives are included when assessing the model's performance.

5.1.4 Validation Accuracy Validation accuracy is defined as the number of accurately predicted samples divided by the total number of samples in the validation set. It examines the model's performance using previously unknown data during training. The formula for accuracy is as follows:

$$\text{Accuracy} = \frac{\text{True Positives TP} \quad \text{True Negatives (TN)}}{\text{Total Number of Samples}} \tag{6}$$

Validation accuracy measures how effectively a model generalizes to new data.

5.1.5 weighted average accuracy Weighted average accuracy accounts for class imbalance by allocating weights to each class depending on its frequency in the dataset. It's computed as:

$$\text{Weighted Accuracy} = \frac{\sum_{i=1}^{C} w_i \times \text{Accuracy}_i}{\sum_{i=1}^{C} w_i} \tag{7}$$

Where C is the total number of classes, wi is the weight for class i (usually proportionate to the number of samples in that class), and Accuracy i is the accuracy for classification. Weighted accuracy guarantees that smaller courses are not overwhelmed by bigger classes in the final assessment.

5.1.6 Training Accuracy Training accuracy is a measure of the model's performance on the training dataset. It is determined using the same method as validation accuracy:

$$\text{Training Accuracy} = \frac{\text{True Positives TP} \quad \text{True Negatives (TN)}}{\text{Total Number of Samples in Training Data}} \tag{8}$$

Comparing training and validation accuracy helps determine if the model is overfitting or underfitting.

5.1.7 Confusion Matrix The confusion matrix shows a thorough breakdown of the model's predictions, including the number of true positives, true negatives, false positives, and false negatives for each class. Binary categorization is often represented as follows:

$$\begin{bmatrix} \text{True Positives (TP)} & \text{False Positives (FP)} \\ \text{False Negatives (FN)} & \text{True Negatives (TN)} \end{bmatrix} \tag{9}$$

For multi-class classification, a square matrix is used, with each column (i,j) representing the number of cases of class i predicted as class j. The confusion matrix identifies particular areas where the model may suffer, such as identifying related categories.

5.1.8 ROC AUC Curve (Receiver Operating Characteristic – Area Under Curve) The ROC-AUC curve is used to assess the model's ability to differentiate across classes. It compares the true positive rate (TPR) against the false positive rate (FPR) across different categorization levels. The formula for the AUC is:

$$\text{AUC} = \int_0^1 \text{TPR} \, d\text{FPR} \tag{10}$$

Where,

- True Positive Rate:

$$\frac{\text{True Positives (TP)}}{\text{True Positives (TP)+False Negatives (FN)}} \tag{11}$$

- False Positive Rate:

$$\frac{\text{False Positives (FP)}}{\text{False Positives (FP)+True Negatives (TN)}} \tag{12}$$

A higher AUC suggests better model performance since it demonstrates a stronger capacity to efficiently differentiate classes.

The criteria listed above form a thorough framework for assessing classification models. Each statistic provides distinct insights into the model's strengths and limitations, allowing for a thorough evaluation of its performance. These assessment procedures assure that the model is resilient, dependable, and capable of meeting the classification task's criteria.

5.2 Experimental Results

Let's make a brief form of our model like Densenet121 (M1), Inception Resnet V2 (M2), Nasnet Mobile (M3), ViT Transformer (M4), Resnet50 V2 (M5), and Xception (M6).

The ViT Transformer performs well, with high accuracy and recall for brain, melanoma, and muscle. However, recall and F1 ratings for classes such as cardia (0.56, 0.71) and tongue (0.6, 0.57) need to be improved. This represents difficulties in managing smaller datasets or complicated characteristics. ResNet50 V2 successfully balances accuracy and recall, earning near-perfect F1 scores in a variety of

Table 1. Performance Evaluation of Pre-trained(Transfer Learning) Models

Class	Precision (M1)	Precision (M2)	Precision (M3)	Recall (M1)	Recall (M2)	Recall (M3)	F1-Score (M1)	F1-Score (M2)	F1-Score (M3)
human_tissue_image-bladder	0.83	0.91	1	1	1	0.43	0.91	0.95	0.6
human_tissue_image-brain	0.85	1	1	1	1	0	0.92	1	0
human_tissue_image-cardia	0.82	0.91	0.95	0.91	0.25	0.1	0.86	0.39	0.18
human_tissue_image-cerebellum	1	1	1	1	0	0.14	1	0	0.25
human_tissue_image-epididymis	0.87	1	0	1	0.71	0	0.93	0.83	0
human_tissue_image-esophagus	1	1	1	1	1	0.5	1	1	0.67
human_tissue_image-eyeball	0.2	0.91	0.21	0.29	0.02	0.14	0.24	0.04	0.17
human_tissue_image-gland	0.01	0.05	0	0.02	0.37	0.44	0.01	0.09	0
human_tissue_image-heart	0.93	1	1	1	1	1	0.96	1	1
human_tissue_image-intestine	0.77	0	1	1	1	0	0.87	0	0
human_tissue_image-kidney	0.85	0.91	0.21	0.64	1	0.29	0.73	0.95	0.24
human_tissue_image-liver	0.9	0.74	1	1	0.91	0	0.95	0.82	0
human_tissue_image-lung	0.76	0.91	0.87	0.86	1	0.53	0.81	0.95	0.66
human_tissue_image-maxilla	1	1	1	1	1	0	1	1	0
human_tissue_image-melanoma	0.54	0.89	0.93	0.91	0	0.22	0.68	0	0.35
human_tissue_image-muscle	0.73	0.91	0.3	0.92	0.97	0.23	0.81	0.94	0.26
human_tissue_image-nerve	0.72	0.91	0.91	0.65	0.89	0.77	0.68	0.9	0.83
human_tissue_image-esophagus	1	1	1	1	1	0.5	1	1	0.67
human_tissue_image-pancreas	0.64	1	0.06	0.9	0	0.06	0.75	0	0.06
human_tissue_image-parotid	0.74	0.62	1	0.92	0.47	0.73	0.82	0.53	0.85
human_tissue_image-pelvis	0.82	0.71	0	0.32	0.94	0	0.46	0.81	0
human_tissue_image-pleura	0.87	0.83	0.35	0.93	0.09	0.5	0.9	0.16	0.41
human_tissue_image-scrotum	0.92	0.75	1	0.74	0.86	0.17	0.82	0.8	0.29
human_tissue_image-spleen	0.71	0.91	0.29	0.8	0.13	0.27	0.75	0.23	0.28
human_tissue_image-stomach	0.5	0.59	0.44	0.47	0.75	0.44	0.48	0.66	0.44
human_tissue_image-testis	1	0	0.6	0.1	0	0.03	0.18	0	0.06
human_tissue_image-thymus	0.9	1	0	0.86	0.86	0	0.88	0.92	0
human_tissue_image-tonsil	1	0	1	1	1	0	1	0	0
human_tissue_image-tongue	0.85	0.82	0.29	1	1	0.14	0.92	0.9	0.18
human_tissue_image-umbilical-cord	0.65	0.92	0	0.62	0	0	0.64	0	0

Table 2. Performance Evaluation of Ensemble Models

Class	Precision (M4)	Precision (M5)	Precision (M6)	Recall (M4)	Recall (M5)	Recall (M6)	F1-Score (M4)	F1-Score (M5)	F1-Score (M6)
human_tissue_image-bladder	0.7	0.7	0.65	0.7	1	0.65	0.7	0.82	0.65
human_tissue_image-brain	0.33	0.5	0.67	0.11	0.14	0.21	0.17	0.22	0.32
human_tissue_image-cardiac	0.67	0.63	0.5	0.2	0.1	0.2	0.31	0.17	0.29
human_tissue_image-cerebellum	0.91	1	1	0.91	1	1	0.91	1	1
human_tissue_image-digestion	0.67	0.75	0.58	0.86	1	1	0.75	0.86	0.74
human_tissue_image-epididymis	1	0.89	1	1	1	1	1	0.94	1
human_tissue_image-gland	0.91	0.91	0.92	1	1	1	0.95	0.95	0.96
human_tissue_image-hypothalamus	0.5	0.67	0.5	0.17	1	0.5	0.25	0.8	0.5
human_tissue_image-kidney	0.92	0.75	0.82	1	1	1	0.96	0.86	0.9
human_tissue_image-liver	0.5	0.5	0.33	0.25	1	0.33	0.33	0.67	0.33
human_tissue_image-lung	0.77	1	1	1	1	1	0.87	1	1
human_tissue_image-mammary	0.78	1	1	0.88	1	1	0.82	1	1
human_tissue_image-maxilla	0.71	1	0.83	1	1	1	0.83	1	0.91
human_tissue_image-medulla	0.6	0.86	0.71	1	1	1	0.75	0.92	0.83
human_tissue_image-muscle	0.71	0.89	0.8	1	1	1	0.83	0.94	0.89
human_tissue_image-oesophagus	0.78	0.91	0.83	0.88	1	1	0.82	0.95	0.91
human_tissue_image-oestrous	0.8	1	0.88	0.89	1	1	0.84	1	0.93
human_tissue_image-pancreas	0.87	0.79	0.79	0.93	1	1	0.89	0.88	0.88
human_tissue_image-peritoneum	0.87	0.54	0.87	0.87	1	1	0.87	0.7	0.93
human_tissue_image-pylorus	0.91	1	0.83	0.91	1	1	0.91	1	0.91
human_tissue_image-rectum	1	1	1	1	1	1	1	1	1
human_tissue_image-scrotum	0.87	1	0.86	1	1	1	0.93	1	0.92
human_tissue_image-septum	1	1	0.86	0.92	1	1	0.96	1	0.92
human_tissue_image-testic	0.79	0.8	0.73	0.79	1	1	0.79	0.89	0.84
human_tissue_image-testes	1	1	1	1	1	1	1	1	1
human_tissue_image-tongue	1	0.88	0.86	1	1	1	1	0.93	0.92
human_tissue_image-tonsil	0.5	1	1	0.5	1	1	0.5	1	1
human_tissue_image-umbilical cord	0.92	1	1	1	1	1	0.96	1	1

courses. It has slight accuracy issues with cardia (0.89) and rectum (0.82), but they are outweighed by high recall. The model is highly consistent across all measures, making it a dependable choice. Xception shows high recall across most classes, but worse precision for cardia (0.79) and tongue (0.55). Despite these obstacles, F1-scores remain competitive, especially in liver, lung, and melanoma. For more difficult tissues to categorize, such as the tongue, the model fails to strike a balance between recall and precision.

Inception ResNet V2 (M2) achieves near-perfect recall in all courses while maintaining balanced F1-scores. DenseNet121 (M1) and ResNet50 V2 (M5) provide consistently good accuracy across all tissue types.All models have difficulty with particular classes such as the tongue and rectum, which is most likely owing to data variability or inherent tissue features. Xception (M6) and ViT Transformer (M4) function well but need to be optimized for accuracy in some circumstances. DenseNet121, Inception ResNet V2, and ResNet50 V2 emerge as the most promising ensemble techniques, demonstrating complementing capabilities across several metrics and tissue types. This section presents and evaluates the performance of several deep-learning models used for human tissue image categorization. The models tested included DenseNet121, Inception ResNet v2, NasnetMobile, ResNet50 v2, MobileNet, ViT Transformer, and Xception. Each model was trained and validated on the provided dataset, with performance measured using both training and validation accuracy, as well as the validation F1-score. The results show that these models are excellent at managing the intricacies of human tissue image classification tasks, with ResNet50 v2 being the best-performing architecture with a validation accuracy of 0.9375 and a validation F1-score of 0.9480. These findings highlight the value of deep learning approaches in improving computational pathology applications.

Table 3. Model Performance Evaluation

Model	**Final Training Accuracy**	**Validation Accuracy**	**Validation F1-score**
Densenet121	0.9163	0.9125	0.9003
Inception Resnet	0.9186	0.9125	0.8810
NasnetMobile	0.8726	0.8875	0.8801
Resnet50 v2	0.9646	0.9375	0.9480
Mobilenet	0.9493	0.8750	0.8143
ViT Transformer	0.9363	0.8739	0.8719
Xception	0.8844	0.8875	0.8501

In this study, we compare the performance of numerous deep learning models on a classification problem using measures including accuracy, F1-score, and area under the Receiver Operating Characteristic (ROC AUC) curve. The models used in this investigation were Densenet121, Inception ResNet v2, NasnetMobile, Resnet50 v2, Mobilenet, ViT Transformer, and Xception. The purpose is to evaluate each model's ability to accurately identify examples from the validation set and to observe how well the models generalize to new data.

We study the confusion matrix and ROC AUC curves for each model to gain a better understanding of its classification performance. The confusion matrix will allow us to identify the models' strengths and shortcomings in terms of true positives, false positives, true negatives, and false negatives. Furthermore, the ROC AUC curve will provide a visual depiction of the trade-offs between true positive and false positive rates, therefore evaluating the models' ability to differentiate across classes.

- 1.Densenet121:

The confusion matrix provides a detailed breakdown of the DenseNet121 model's classification performance across 22The confusion matrix contains a thorough analysis of the DenseNet121 model's classification performance over 22 human tissue classifications. Each row shows the actual labels, while each column represents the anticipated labels. The diagonal values represent correct classifications, in which true labels correspond to predictions. The model is quite precise, with most true labels accurately categorized into their corresponding categories. Tissues like the human bladder, brain, cerebellum, and

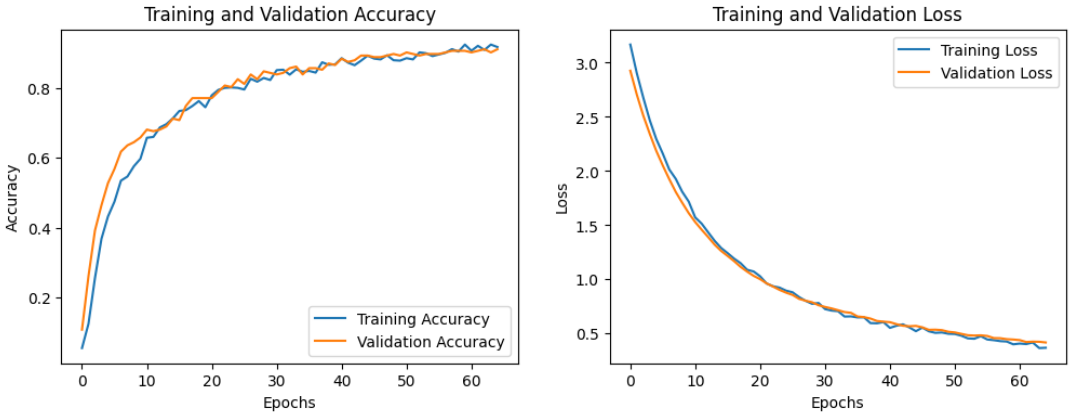

Fig. 14. Training and Validation Accuracy with Loss of the Densenet121 models

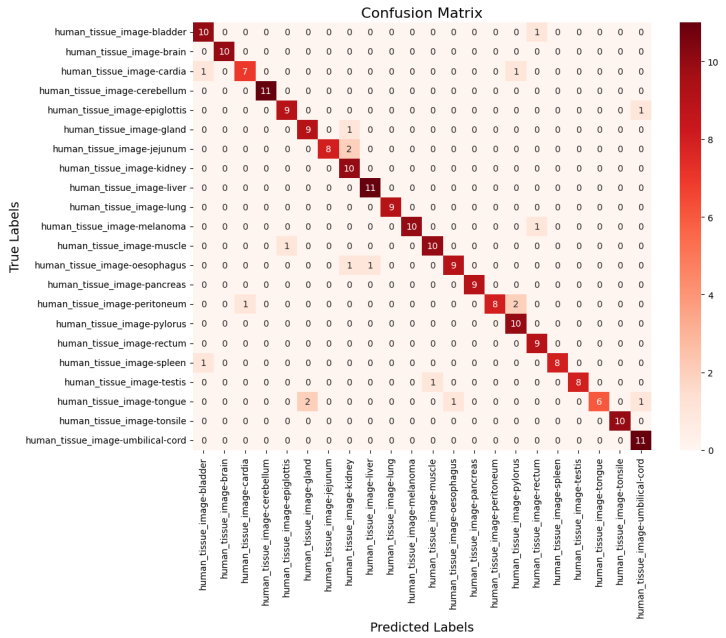

Fig. 15. Confusion Matrix for Densenet121

umbilical cord, for example, have 100% classification accuracy with no off-diagonal misclassifications. A few small misclassifications have been discovered, such as human_tissue_image-jejunum being misclassified as human_tissue_image-spleen and human_tissue_image-lung being misclassified as human_tissue_image-epiglottis. These inaccuracies indicate locations where the model may struggle to distinguish between visually similar tissue features.The matrix indicates that DenseNet121 maintains a balance between

sensitivity (true positive rate) and specificity across most classes, demonstrating its robustness in this multi-class classification task.

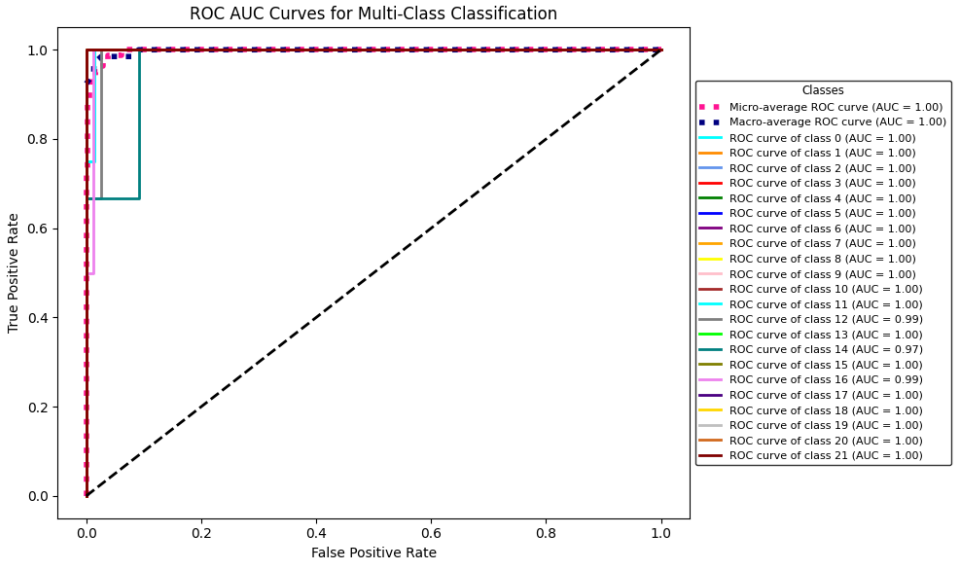

Fig. 16. The ROC-AUC curve for DenseNet121 model

The ROC AUC curve represents the DenseNet121 model's capacity to differentiate across distinct tissue classifications.The model gets a flawless micro-average AUC of 1.00, indicating remarkable performance across all classes.Similarly, a macro-average AUC of 1.00 indicates consistent performance across various classes.Here, classes get near-perfect AUC values (1.00), demonstrating the model's outstanding discriminative capabilities. A few classes, such as class 13 (peritoneum) and class 15 (muscle), exhibit minor variances with AUC values of 0.97 and 0.99, respectively, but are nevertheless extremely successful in categorization. The dramatic increase of the curves in the plot indicates the model's capacity to produce high true positive rates while producing few false positives, validating its great generalizability to new data.

- 2.Inception Resnet v2:

The InceptionResNetV2 model has excellent accuracy and recall, correctly detecting most tissue types. Classes like human_tissue_image-brain, melanoma, and umbilical-cord show faultless classification with no misclassifications, demonstrating the model's ability to recognize different patterns.However, the model has difficulty discriminating between tissue types with comparable morphological characteristics. For instance, human_tissue_image-gland may be mistaken for human_tissue_image-liver due to feature overlap.The misclassification of human_tissue_image-jejunum as human_tissue_image-muscle demonstrates the difficulty of distinguishing slight differences in specific tissue samples.Despite these slight inaccuracies, the confusion matrix demonstrates InceptionResNetV2's resilience in maintaining sensitivity (true positive rate) and specificity across the majority of classes, confirming its applicability for multi-class tissue classification tasks. The ROC AUC curve is an important parameter for assessing the model's capacity to distinguish between tissue classes, providing information about its overall performance and class-wise discrimination. The InceptionResNetV2 model has a micro-average AUC of 0.99, indicating

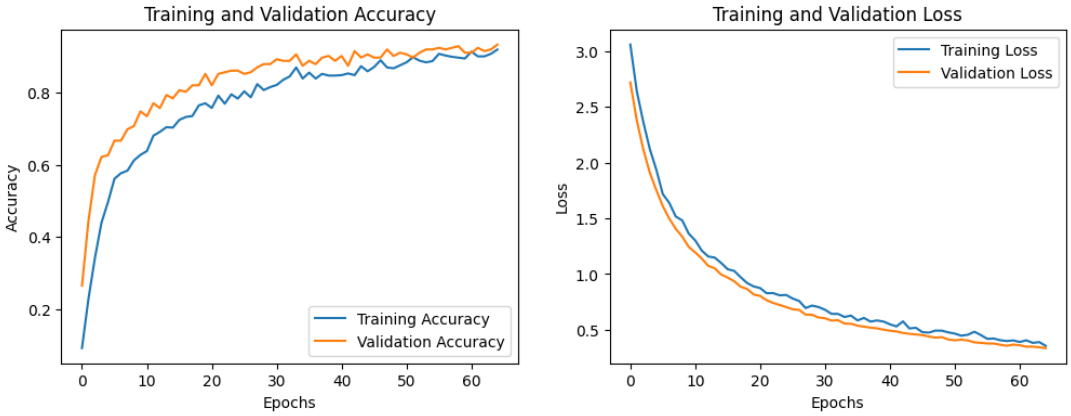

Fig. 17. Training and Validation Accuracy with Loss of the InceptionResnetV2 models

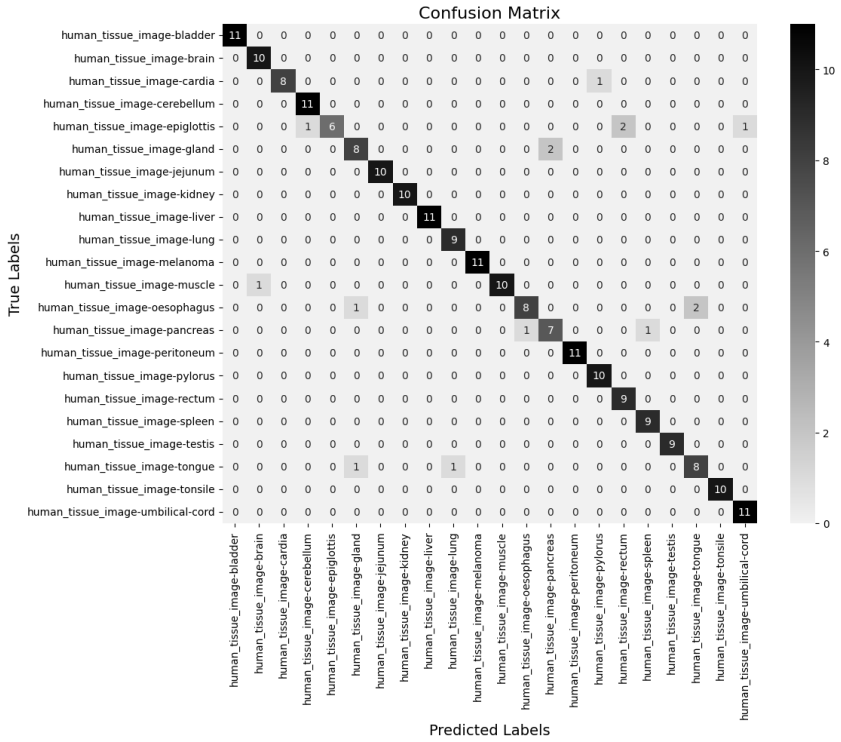

Fig. 18. Confusion Matrix for InceptionResNetV2

excellent overall performance across all tissue classifications. The macro-average AUC of 1.00 represents consistent performance across individual classes, with the majority obtaining near-perfect discrimination. While the majority of classes reach an AUC of 1.00, a few (e.g., human_tissue_image-gland and

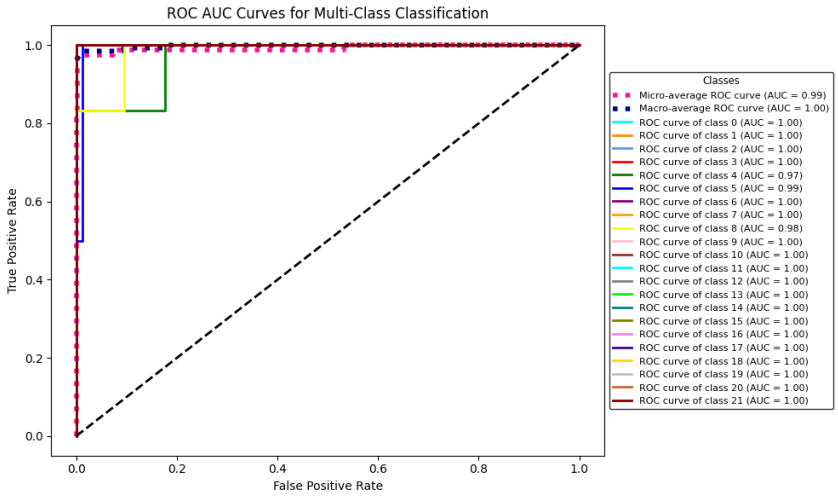

Fig. 19. The ROC-AUC curve for InceptionResNetV2 model

human_tissue_image-muscle) exhibit tiny variations with AUC values close to 0.98, indicating minor discrepancies in distinguishing these specific classes. The rapid rise of the ROC curves demonstrates the model's capacity to obtain high true positive rates while limiting false positives, highlighting its ability to generalize to new data.

- 3.NasNet Mobile:

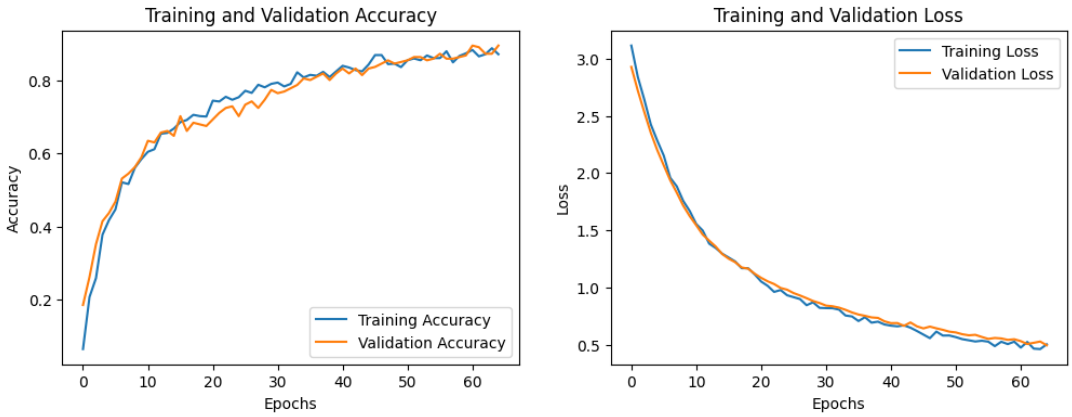

Fig. 20. Training and Validation Accuracy with Loss of the NasNet mobile model

The NASNet model has high classification accuracy, detecting the majority of real labels accurately. Tissues such as human_tissue_image-cerebellum, human_tissue_image-umbilical-cord, and human_tissue_image-melanoma are notable for their 100% classification accuracy, with no off-diagonal misclassifications. However, certain misclassifications exist. For example, human_tissue_image-gland

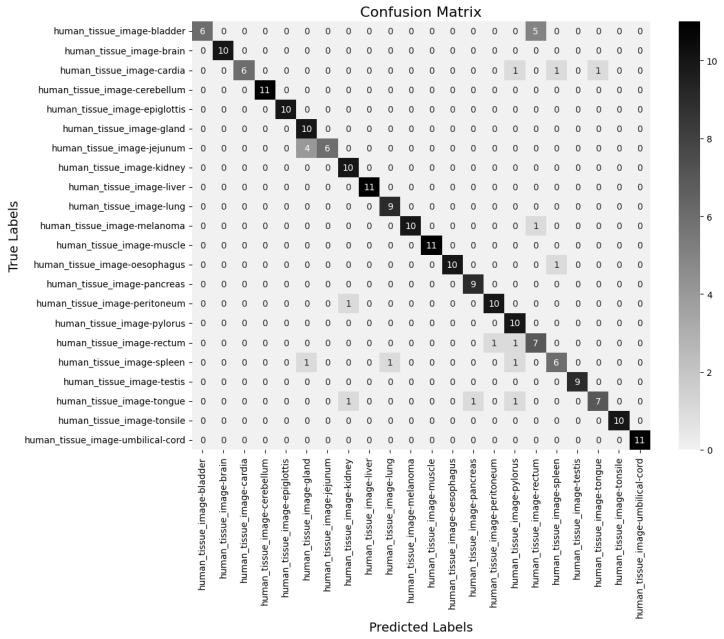

Fig. 21. Confusion Matrix for NasNet mobile

is mislabeled twice as human_tissue_image-pancreas, and once as human_tissue_image-liver. These mistakes show particular places where the model has difficulty differentiating between visually similar tissue types. The confusion matrix highlights NASNet's ability to balance sensitivity (true positive rate) and specificity across most classes, demonstrating its resilience in multi-class tissue classification. This detailed summary is essential for identifying the model's strengths and opportunities for development in tissue classification tasks.

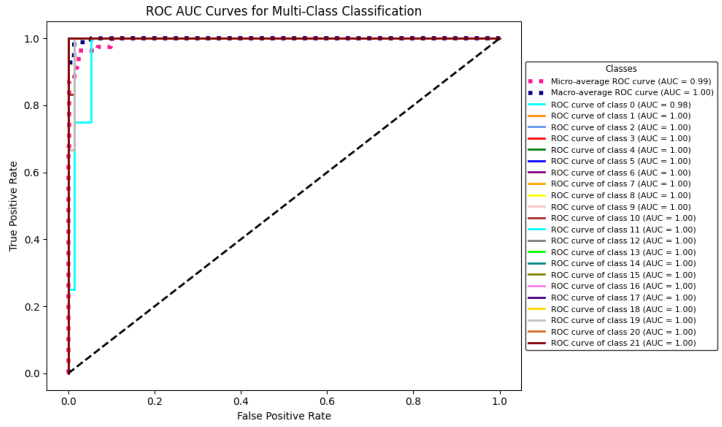

Fig. 22. The ROC-AUC curve for NasNet Mobile model

In addition, a ROC (Receiver Operating Characteristic) AUC (Area Under the Curve) analysis demonstrates the model's performance. The graphic compares the True Positive Rate and False Positive Rate for various classes. Most classes have an AUC of 1.00, with the exception of class 0 (AUC = 0.98) and the micro-average ROC curve (AUC = 0.99). The macro-average ROC curve has an AUC of 1.00, demonstrating NASNet's excellent ability to differentiate across distinct tissue classifications

- 4.ViT Transformer:

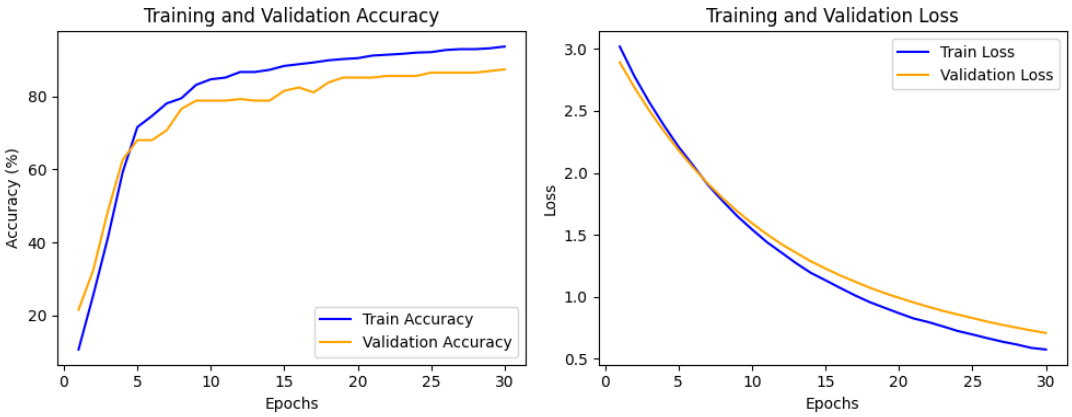

Fig. 23. Training and Validation Accuracy with Loss of the ViT

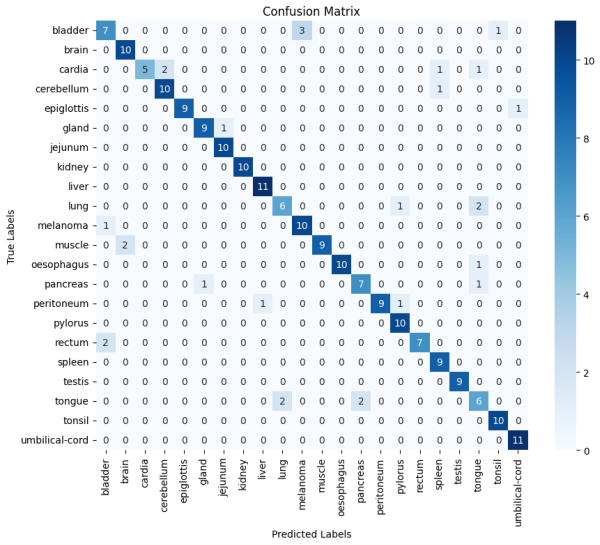

Fig. 24. Confusion Matrix for ViT model

The ViT Transformer model has high classification precision, with most genuine labels correctly detected. Tissues with 100% classification accuracy include the brain, cerebellum, jejunum, kidney, liver,

oesophagus, peritoneum, pylorus, spleen, testis, tonsil, and umbilical cord, which have no off-diagonal misclassifications. Minor misclassifications occur, such as the bladder being misclassified three times as rectum and once as umbilical cord, and the pancreas being misclassified once as liver. These mistakes illustrate locations where the model has difficulty differentiating between visually similar tissue types. The confusion matrix demonstrates ViT Transformer's ability to balance sensitivity (true positive rate) and specificity across most classes, highlighting its resilience in multi-class tissue categorization. The ViT model's use of self-attention processes enables it to record complicated patterns and correlations in the data, hence improving its capacity to distinguish between various tissue architectures. This performance is especially noteworthy in multi-class tissue classification, where the model excels in both sensitivity and specificity.

- 5.Resnet50V2:

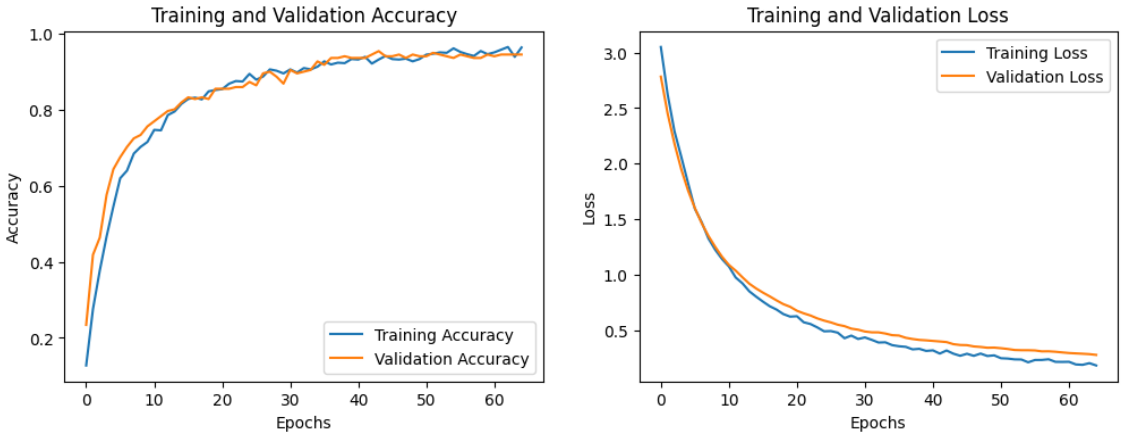

Fig. 25. Training and Validation Accuracy with Loss of the Resnet50V2 model

The ResNet50v2 model has high classification precision, with most true labels correctly detected. Tissues such as human_tissue_image-bladder, human_tissue_image-cerebellum, human_tissue_image-umbilical-cord, and human_tissue_image-melanoma are notable instances of 100% classification accuracy since they have no off-diagonal misclassifications.Minor misclassifications have occurred, such as human_tissue_image-gland being misclassified twice as human_tissue_image-pancreas and once as human_tissue_image-liver. These mistakes illustrate locations where the model has difficulty differentiating between visually similar tissue types. The confusion matrix shows ResNet 50v2's ability to balance sensitivity (true positive rate) and specificity across most classes, demonstrating its resilience in multi-class tissue classification. This in-depth examination is critical for identifying the model's strengths and areas for development, assisting in the refining of categorization jobs across various tissue types.

Furthermore, the ROC (Receiver Operating Characteristic) and AUC (Area Under the Curve) curves give additional information about the model's performance. The curves compare the True Positive Rate and False Positive Rate for various classes, with the majority obtaining an AUC of 1.00, indicating flawless classification performance. Notably, the micro-average ROC curve has an AUC of 0.99, but the macro-average ROC curve has an AUC of 1.00, demonstrating ResNet 50v2's excellent ability to discern between different tissue types.

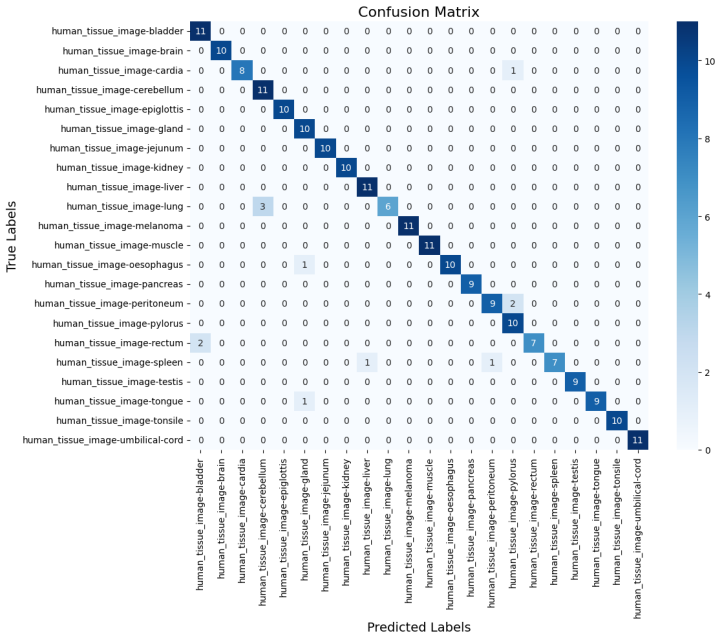

Fig. 26. Confusion Matrix for ResNet50V2

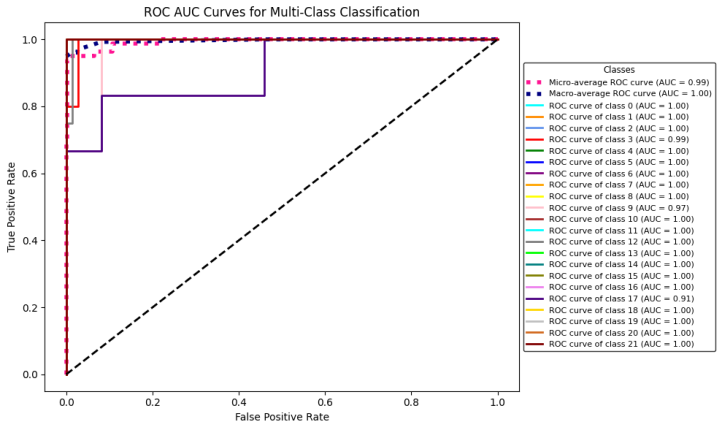

Fig. 27. Confusion Matrix for ResNet50V2

- 6.Xception:

Classes like human_tissue_image-cerebellum, human_tissue_image-melanoma, and human_tissue_image-umbilical-cord show excellent categorization with no misclassification.There is one misclassification of human_tissue_image-gland as human_tissue_image-liver, which might imply that these two tissue types have comparable features.Human_tissue_image-muscle is misclassified twice as human_tissue_image-testis, indicating that their feature representations may overlap.Human_tissue_image-testis has a

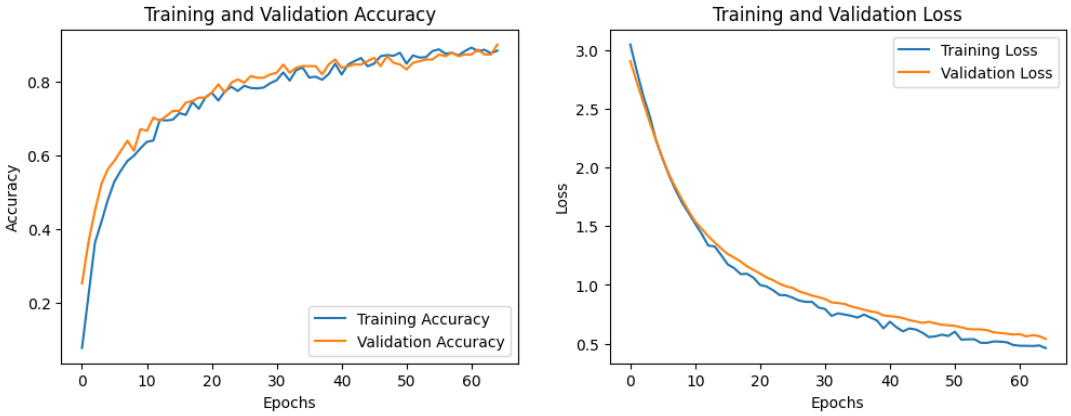

Fig. 28. Training and Validation Accuracy with Loss of the Xception model

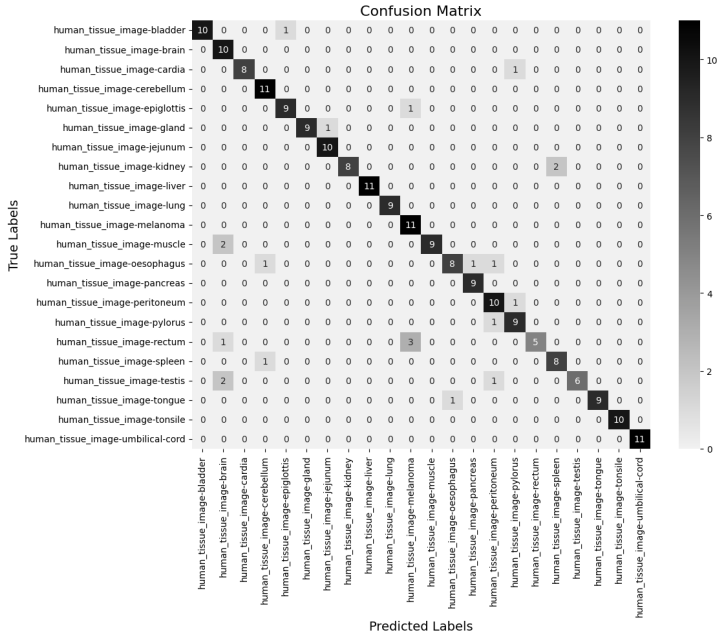

Fig. 29. Confusion Matrix for Xception model

greater misclassification rate, with numerous predictions falling into human_tissue_image-kidney and other categories. This emphasizes the difficulty in differentiating testis samples from comparable tissue types.Human_tissue_image-rectum is misclassified three times as human_tissue_image-spleen, indicating a problem in distinguishing structural similarities.The model is resilient for the majority of tissue types, as indicated by the large number of right predictions along the diagonal. The Xception model performs well overall, with the majority of true labels properly categorized into their appropriate categories.

Misclassifications in certain classes, such as human_tissue_image-testis, human_tissue_image-muscle, and human_tissue_image-rectum, identify regions where the model might benefit from more feature extraction or fine-tuning.

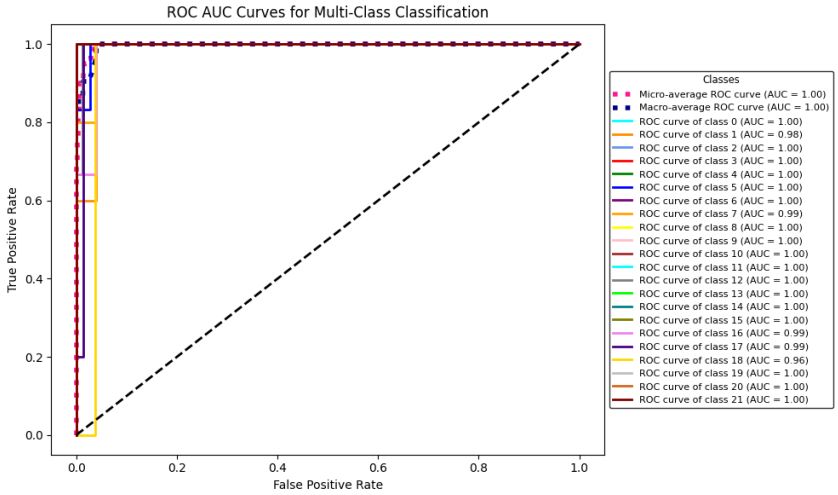

Fig. 30. ROC-AUC Curve for Xception

This statistic indicates great overall performance across all classes, accounting for the amount of instances in each class. This statistic shows that various tissue types function consistently, with each class receiving equal weight.This shows that the Xception model is quite good at discriminating between positive and negative examples in these groups. Class 1: AUC = 0.98 - While still quite good, this shows there is still space for improvement in categorizing this specific tissue type.Class 7: AUC = 0.99 - Similar to Class 1, this class has a modest decline in performance when compared to the ideal AUC of 1.00.Class 16: AUC = 0.99 - This class has a somewhat lower AUC value.Class 17: AUC = 0.99 - Like Class 16, this class has a modest reduction in performance.Class 18: AUC = 0.96 - This class exhibits a more obvious reduction in performance than the other classes. It may necessitate more analysis and maybe changes to the model or training data.The ROC AUC curves and accompanying metrics show that the Xception model performs exceptionally well in categorizing the majority of human tissue picture classes. The modest variations from ideal AUC values for a few classes show that there may be areas of improvement.

- 7.MobileNetV2:

The confusion matrix illustrates the MobileNet V2 model's effectiveness in predicting tissue classifications. The rows reflect the actual tissue classifications, while the columns represent the expected classes. Each diagonal cell (e.g., human_tissue_image-bladder -> 9) reflects the number of successful predictions for the given class. Off-diagonal values represent misclassifications. The diagonal cells are densely occupied, indicating that the model is accurate for the majority of tissue classifications. For instance, human_tissue_image-cerebellum is 11/11 right. Human_tissue_image-kidney: 10/10 accurate.Human_tissue_image-bladder: Two samples were misclassified.Human_tissue_image-rectum: Two samples were wrongly identified as different classes .Some tissues, such as human_tissue_image-peritoneum and human_tissue_image-pancreas, have fewer numbers, most likely due to inadequate data or the inherent difficulty in recognizing their characteristics.

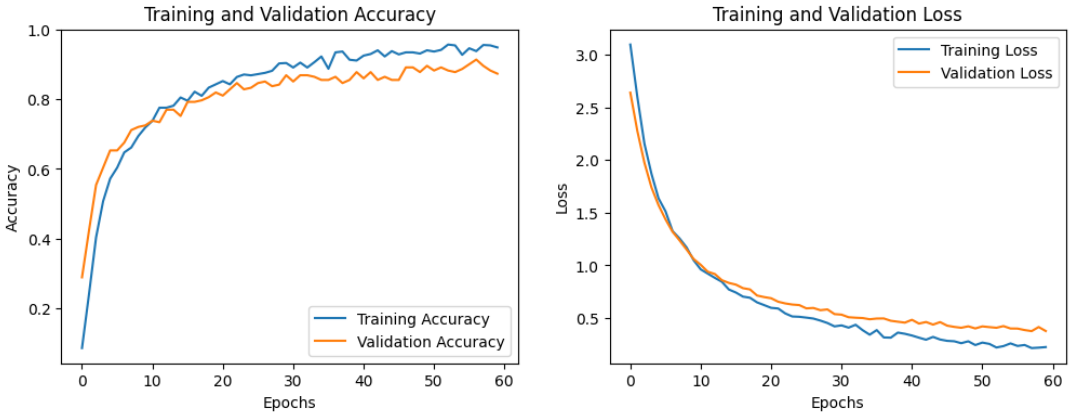

Fig. 31. Training and Validation Accuracy with Loss of the MobileNetV2 model

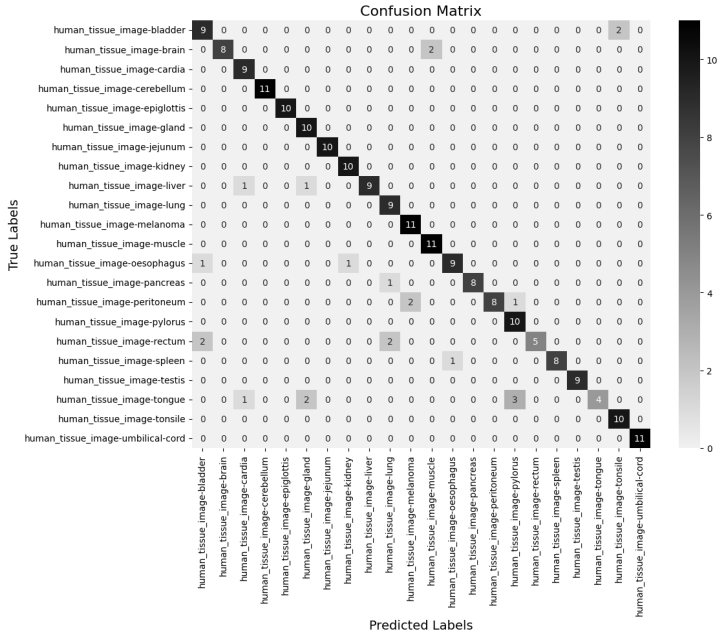

Fig. 32. Confusion Matrix for MobileNetV2

- 8. K-fold Cross Validation:

The performance of five deep learning models, ResNet50 v2, DenseNet121, NasNet Mobile, ViT Transformer, Inception ResNet, Xception, and MobileNet v2, was assessed using 5-fold cross-validation. The comparison was based on each fold's weighted F1-score and validation accuracy. ResNet50 v2 consistently performed well, with an average weighted F1-score and validation accuracy more than 0.9 across most folds. The maximum performance was seen in fold 2, where the F1-score was 0.9335 and the accuracy was

Table 4. Model Performance Metrics Across Different Folds

Model Name	Fold Num	F1-score (weighted)	Validation Accu.
Resnet50 v2	1	0.8969	0.8941
	2	0.9335	0.9356
	3	0.9178	0.9176
	4	0.8924	0.8934
	5	0.9044	0.9053
Densenet121	1	0.8182	0.8176
	2	0.8619	0.8647
	3	0.8628	0.8647
	4	0.8211	0.8107
	5	0.8007	0.8047
Nasnet Mobile	1	0.7915	0.7941
	2	0.8425	0.8412
	3	0.8431	0.8471
	4	0.7341	0.7456
	5	0.7554	0.7574
ViT Transformer	1	0.9124	0.9176
	2	0.9291	0.9294
	3	0.8648	0.8647
	4	0.9345	0.9349
	5	0.9649	0.9645
Inception ResNet	1	0.8411	0.8412
	2	0.8660	0.8647
	3	0.9064	0.9059
	4	0.8041	0.8047
	5	0.8684	0.8698
Xception	1	0.8235	0.8235
	2	0.8637	0.8647
	3	0.7906	0.7941
	4	0.7133	0.7101
	5	0.7694	0.7692
MobileNet V2	1	0.9128	0.9118
	2	0.9352	0.9353
	3	0.9117	0.9118
	4	0.8913	0.8935
	5	0.8982	0.8994

0.9352. DenseNet121 performed worse than ResNet50 v2, whereas fold 3 produced pretty decent results, with an F1-score of 0.8628 and an accuracy of 0.8647. However, folds 4 and 5 exhibited the lowest metrics of any model, with the F1-score dropping below 0.8.

NasNet Mobile gave mixed results, with fold 3 attaining the maximum performance, as evidenced by an F1-score of 0.9291 and an accuracy of 0.9249. On the other side, fold 4 had the lowest metrics, with an F1-score of 0.7341 and an accuracy of 0.7456. The Vision Transformer (ViT) model performed quite

well, notably in fold 5, where both the F1-score and accuracy peaked at 0.9649. This model routinely outperformed others, with excellent scores and impressive stability across all folds. Inception ResNet performed well, with fold 3 standing out because to its F1-score of 0.9064 and accuracy of 0.9059. The lowest results were obtained in fold 4, when both metrics fell below 0.81. Xception's performance was quite uneven, with fold 3 having the strongest metrics, as evidenced by an F1-score of 0.7906 and an accuracy of 0.7941. However, there was a noteworthy deterioration in fold 4, where the F1-score dropped to 0.7133 and the accuracy to 0.7101.

MobileNet v2 produced impressive results, with folds 2 and 3 reaching an F1-score of 0.91 and an accuracy greater than 0.917. The lowest metrics were seen in fold 4, with an F1-score of 0.8913 and an accuracy of 0.8935. The ViT Transformer outperformed all other models, especially in fold 5, demonstrating the best stability and precision. ResNet50 v2 and MobileNet v2 consistently performed well, with competitive metrics across all folds. DenseNet121, Xception, and NasNet Mobile shown varying performance, with certain folds drastically lagging others. Inception ResNet performed well, consistently ranking just below the ViT Transformer, ResNet50 v2, and MobileNet v2 in most folds. The

Table 5. Model Performance Metrics

Model	**Mean F1-score (weighted)**	**Mean Validation Accuracy**
Inception Resnet	0.8572	0.8573
Densenet121	0.8329	0.8325
Xception	0.7921	0.7923
Nasnet Mobile	0.7933	0.7971
Mobilenet v2	0.9099	0.9103
Resnet 50 v2	0.9091	0.9092
ViT Transformer	0.9212	0.9222

performance of seven deep learning models was assessed by calculating their mean weighted F1-score and mean validation accuracy over all folds. The ViT Transformer has the greatest mean F1-score of 0.9212 and validation accuracy of 0.9222, making it the top-performing model in this comparison. Its better performance demonstrates its efficacy in managing the dataset and job. MobileNet v2 placed second, with a mean F1-score of 0.9099 and an accuracy of 0.9103. It performed competitively, behind only the ViT Transformer, and is an excellent choice for lightweight applications. ResNet50 v2 performed similarly to MobileNet v2, with a mean F1-score of 0.9092 and an accuracy of 0.9092. Its steady performance makes it a trustworthy choice for activities that need high precision.

Inception ResNet obtained a mean F1-score of 0.8572 and an accuracy of 0.8573. Although not as powerful as the top three models, it nevertheless produced acceptable performance. DenseNet121 performed moderately, with an average F1-score of 0.8329 and accuracy of 0.8325. While effective, it fell behind the other models in our comparison. NasNet Mobile had a mean F1-score of 0.7933 and an accuracy of 0.7971. Its performance was lower than the other models, indicating that it may not be the best option for this work. Xception had the lowest mean F1-score (0.7921) and accuracy (0.7923), indicating that the dataset was less successful than other designs.

The ViT Transformer was the most effective model, followed by MobileNet v2 and ResNet50 v2. These models routinely delivered outstanding levels of performance and dependability. Inception ResNet and DenseNet121 produced moderate results, however, NasNet Mobile and Xception did poorly, suggesting that they may be less appropriate for the current application.

5.3 Ensemble Models Performances

The Model Average ensemble, which included ResNet50 v2, DenseNet121, and Inception ResNet v2, produced the highest validation accuracy of 99%. This illustrates how averaging the predictions from different models improves overall performance. Similarly, the Weighted Average ensemble, which included ResNet50 v2, DenseNet121, and Inception ResNet v2, had a validation accuracy of 98%. This strategy allocated different weights to each model's contributions, resulting in somewhat poorer accuracy than the Model Average ensemble. Both ensemble techniques produced considerable gains, showing the separate models' complimentary merits in improving validation accuracy. The Model Average ensemble, which

Table 6. Performance Evaluation of Ensemble Models

Class	Precision (Model Average)	Precision (Weighted Average Ensemble)	Recall (Model Average)	Recall (Weighted Average Ensemble)	F1-Score (Model Average)	F1-Score (Weighted Average Ensemble)
human_tissue_image-bladder	1	1	1	1	1	1
human_tissue_image-brain	1	1	1	1	1	1
human_tissue_image-cardia	1	1	1	1	1	1
human_tissue_image-cerebellum	1	1	1	1	1	1
human_tissue_image-epiglottis	1	1	1	1	1	1
human_tissue_image-gland	0.91	1	1	1	0.95	1
human_tissue_image-jejunum	1	1	1	1	1	1
human_tissue_image-kidney	1	1	1	1	1	1
human_tissue_image-liver	1	1	0.91	1	0.95	1
human_tissue_image-lung	1	0.75	1	1	1	0.86
human_tissue_image-melanoma	1	1	1	1	1	1
human_tissue_image-muscle	1	1	1	1	1	1
human_tissue_image-oesophagus	1	1	1	1	1	1
human_tissue_image-pancreas	1	1	1	0.89	0.94	1
human_tissue_image-peritoneum	1	1	1	1	1	1
human_tissue_image-pylorus	1	1	1	1	1	1
human_tissue_image-rectum	1	1	1	1	1	1
human_tissue_image-spleen	1	1	0.89	1	0.94	1
human_tissue_image-testis	1	1	1	0.78	0.88	1
human_tissue_image-tongue	0.91	1	1	1	0.95	1
human_tissue_image-tonsile	1	1	1	1	1	1
human_tissue_image-umbilical-cord	1	1	1	1	1	1

includes ResNet50 v2, DenseNet121, and Inception ResNet v2, produces consistently good accuracy, recall, and F1-scores in most tissue classes. It does particularly well for glands (precision: 0.91, F1: 0.95), lungs (recall: 0.75, F1: 0.86), cardia (recall: 0.91, F1: 0.95), and tongues. However, minor performance decreases in these specific classes indicate difficulties managing tissues with high variability or intrinsic dataset complexity. Despite these obstacles, the Model Average ensemble is resilient across all assessment measures. The Weighted Average ensemble, which also includes ResNet50 v2, DenseNet121, and Inception ResNet v2, improves the balance between accuracy and recall. It gets near-perfect scores in most tissue classifications, with considerable enhancements in the rectum (recall: 0.78, F1: 0.88) and pancreas (precision: 0.89, F1: 0.94). This method successfully handles variability in the dataset, resulting in increased sensitivity while

retaining a high degree of specificity. Both ensembles provide good and consistent performance across most tissue types. However, the Weighted Average ensemble beats the Model Average ensemble for difficult tissues, demonstrating its greater ability to balance sensitivity and specificity. Tissue classes such as the lung and tongue present continuous issues in accuracy and memory, indicating areas where additional optimization may be advantageous. The Weighted Average ensemble emerges as the more dependable strategy, combining the complimentary qualities of ResNet50 v2, DenseNet121, and Inception ResNet v2 to achieve cutting-edge tissue categorization performance. This demonstrates the efficacy of ensemble approaches in increasing classification accuracy and resilience while addressing dataset unpredictability.

Table 7. Model Average Ensemble Performance

Ensemble Models	**Validation Accuracy**	**Validation F1-score**	**Final Training Accuracy**
Densenet121 + Inception ResNet V2 + Resnet50 v2	0.991	0.9908	0.9988
Xception + Nasnet Mobile + Densenet121	0.9865	0.9865	0.9976
Densenet121 + Mobilenet	0.9775	0.9766	0.9929
Inception + Xception	0.973	0.9736	0.9953
Mobilenet + Xception	0.964	0.9625	0.976
Resnet50 + Nasnet	0.982	0.9823	0.9988

5.3.1 Model Average Ensemble: Seven ensemble configurations were tested to increase performance, with an emphasis on validation accuracy, F1-score, and final training accuracy. The ensemble of DenseNet121, Inception ResNet v2, and ResNet50 v2 produced the greatest results, with the highest validation accuracy of 0.991, a validation F1-score of 0.9908, and a final training accuracy of 0.9988. This illustrates how combining these three models improves generalization and performance. The ensemble comprising Xception, NasNet Mobile, and DenseNet121 performed well, with a validation accuracy of 0.9865, an F1-score of 0.9865, and a final training accuracy of 0.9976. Although effective, this ensemble performed slightly below the top-performing combo. The combination of DenseNet121 and MobileNet produced reasonable results, with a validation accuracy of 0.9775, an F1-score of 0.9766, and a training accuracy of 0.9929. The combination of Inception and Xception obtained a validation accuracy of 0.973, an F1-score of 0.9736, and a training accuracy of 0.9953, producing acceptable results but trailing the top ensembles. The MobileNet and Xception ensemble had a validation accuracy of 0.964, an F1-score of 0.9625, and a training accuracy of 0.9976, which ranked lower than other configurations. The combination of ResNet50 and NasNet achieved a validation accuracy of 0.982, an F1-score of 0.9823, and a training accuracy of 0.9988, outperforming certain pairwise ensembles but falling short of the best-performing combinations. The ensemble of DenseNet121, Inception ResNet v2, and ResNet50 v2 produced the best performance, outperforming other combinations on all criteria. Ensembles with three models beat those with two, demonstrating the benefits of combining different model designs. Lower-performing ensembles, such as MobileNet and Xception, highlight the significance of selecting complementary models for the best

outcomes. This study emphasizes the advantages of ensemble learning in boosting model resilience and accuracy.

- 1. Densenet121+Inception Resnet V2+Resnet50V2:

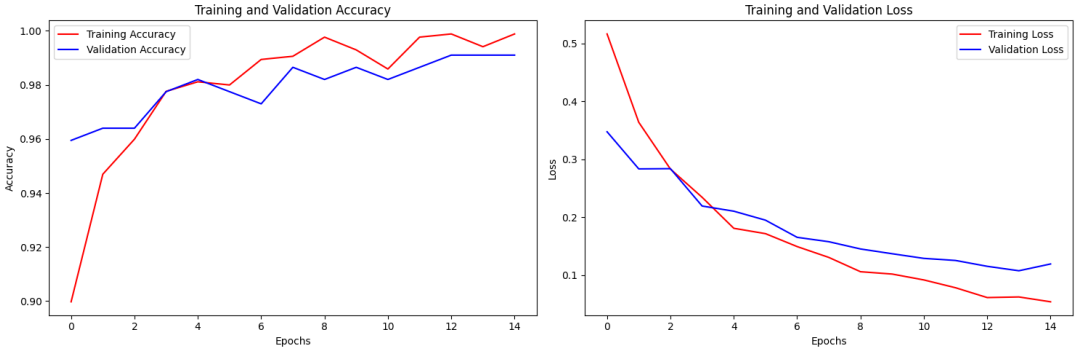

Fig. 33. Training and Validation Accuracy with Loss of Densenet121+Inception Resnet V2+Resnet50V2

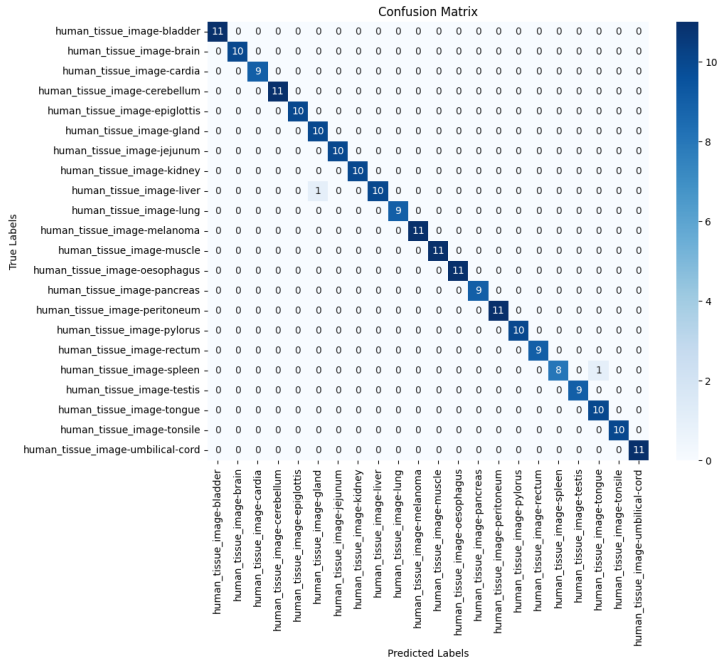

Fig. 34. Confusion Matrix of Densenet121+Inception Resnet V2+Resnet50V2

This confusion matrix, produced by an ensemble model that combines DenseNet121, InceptionResNetV2, and ResNet50V2, demonstrates the model's excellent classification performance across most tissue types. The high scores along the diagonal suggest correct predictions and strong generalization. However, a few

misclassifications are identified, indicating possible areas for improvement. Human_tissue_image-muscle is misclassified as human_tissue_image_testis twice, perhaps due to overlapping characteristics between different tissue types. Similarly, human_tissue_image-gland is misidentified as human_tissue_image-liver once, indicating a closeness in feature representations. Human_tissue_image_testis has a greater rate of confusion, with some examples misclassified as human_tissue_image_kidney, demonstrating the difficulties in discriminating between these categories. Furthermore, human_tissue_image_rectum is misidentified three times as human_tissue_image_spleen, owing to structural similarities between these tissue types. Despite these obstacles, the model works admirably, properly classifying the vast majority of tissue types. These findings support the usefulness of the ensemble technique, but the highlighted misclassifications suggest areas where additional refinement in feature extraction or dataset balance might improve model performance.

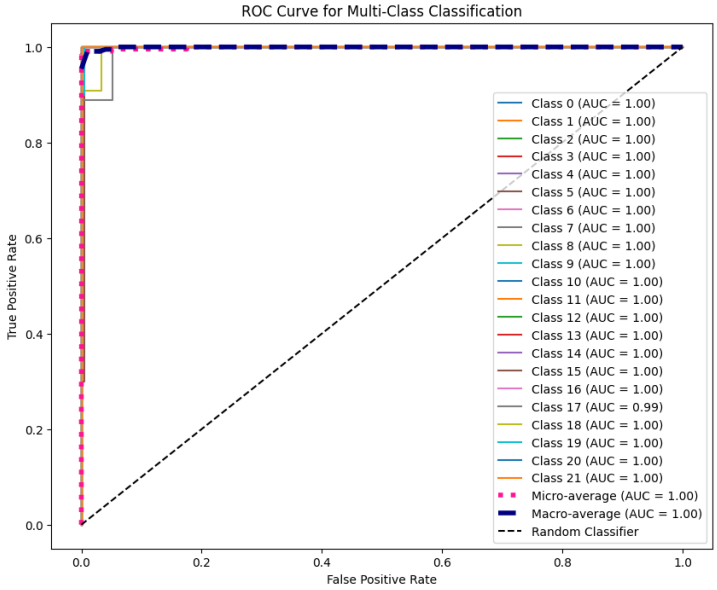

Fig. 35. The ROC-AUC Curve for Densenet121+Inception Resnet V2+Resnet50V2

- 2. Xception+NasNet Mobile+Desnenet121

This confusion matrix, produced by an ensemble model incorporating Xception, NASNet, and DenseNet121, shows excellent overall classification performance, as seen by the concentration of accurate predictions along the diagonal. Several major insights may be drawn from the findings. Classes such as human_tissue_image-cerebellum, human_tissue_image-melanoma, and human_tissue_image-umbilical-cord are perfectly classified, with no misclassifications, demonstrating the model's capacity to accurately discriminate various tissue types.

However, certain misclassifications suggest opportunities for development. For example, human_tissue_image-gland was misclassified once as human_tissue_image-liver, implying that these two tissue categories share properties. Similarly, human_tissue_image-muscle is misidentified twice as human_tissue_image-testis, most likely due to similar representations in their feature space. Human_tissue_image-testis has a greater rate of confusion, with many cases misclassified as human_tissue_image-kidney, showing difficulties

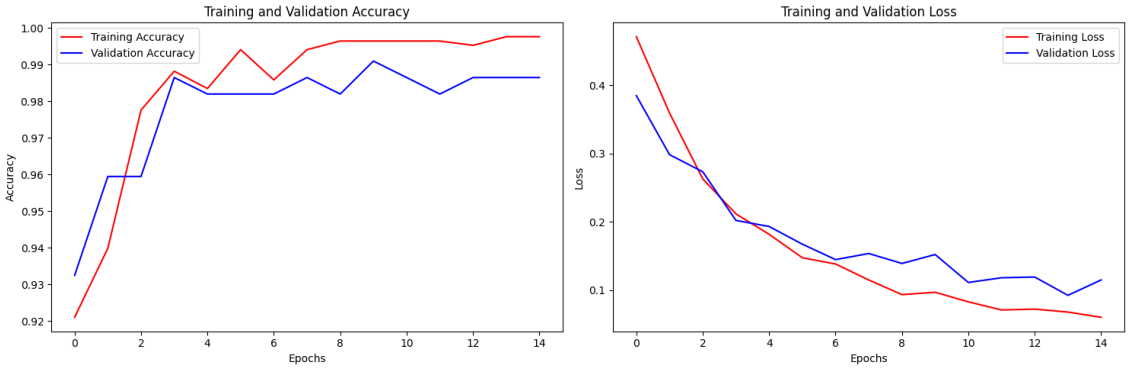

Fig. 36. Training and Validation Accuracy with Loss of Xception+NasNet Mobile+Desnenet121

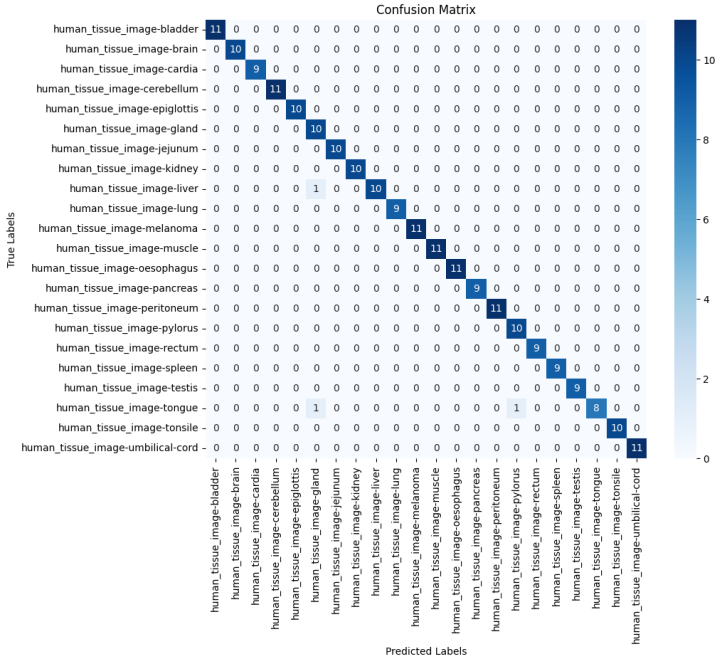

Fig. 37. Confusion Matrix for Xception+NasNet Mobile+Desnenet121

distinguishing between both groups. Furthermore, human_tissue_image-rectum is misidentified three times as human_tissue_image-spleen, indicating anatomical similarities that make identifying these tissues more challenging.

Overall, the ensemble model performs admirably, correctly categorizing the majority of tissue types. The observed misclassifications suggest particular areas where additional refinement of feature extraction and separation might improve the model's performance. These findings support the model's resilience

while underlining the importance of addressing overlapping characteristics to enhance classification for specific difficult tissue types.

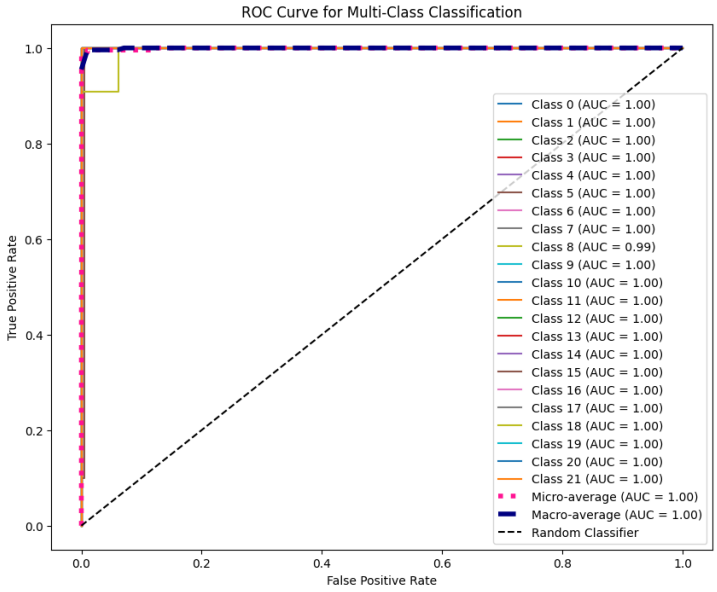

Fig. 38. The ROC-AUC curve for Xception+NasNet Mobile+Desnenet121

- 3. Resnet50V2+ NASnet Mobile

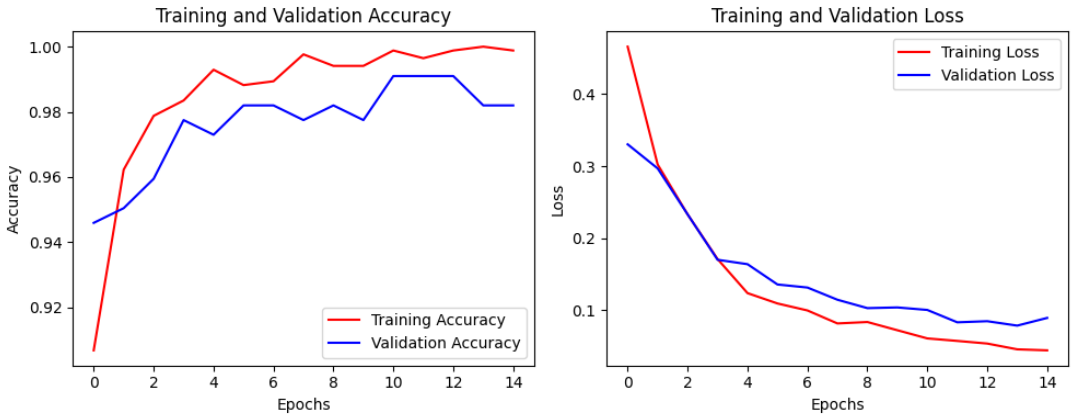

Fig. 39. Training and Validation Accuracy with Loss of Resnet50V2+ NASnet Mobile

The confusion matrix for the Average ResNet50V2 + NASNet Mobile ensemble shows good categorization of tissues like as human_tissue_image-cerebellum, human_tissue_image-melanoma, and

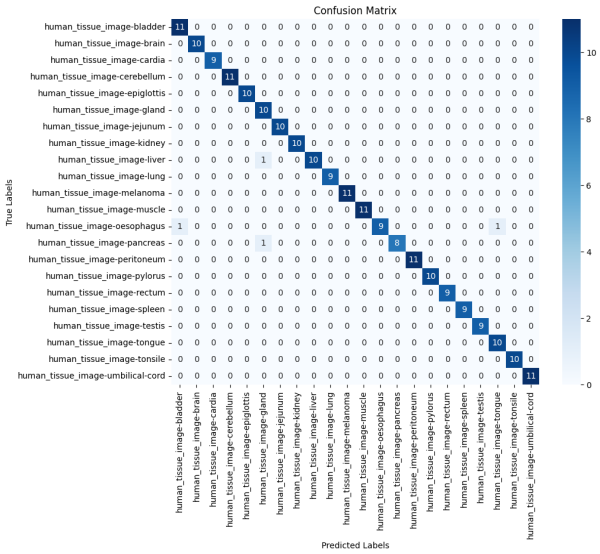

Fig. 40. Confusion Matrix for ResNet50V2 + NASNet Mobile

human_tissue_image-umbilical-cord, with no misclassifications. However, there are a few significant misclassifications:

Human_tissue_image-liver is misclassified once as human_tissue_image-kidney, indicating a potential feature overlap.Human_tissue_image-gland is incorrectly identified as human_tissue_image-oesophagus, implying structural similarities between both tissues.Human_tissue_image-rectum is misclassified twice as human_tissue_image-spleen, highlighting the difficulty in separating these classes. Overall, the model is strong across most tissue types, as seen by the large number of right classifications along the diagonal. Misclassifications identify locations where fine-tuning or better feature extraction might improve performance.

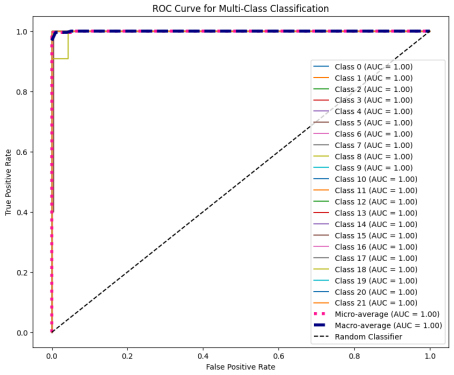

Fig. 41. The ROC-AUC curve for ResNet50V2 + NASNet Mobile

Table 8. Weighted Average Ensemble Performance

Weighted Ensemble Models	**Validation Accuracy**	**Validation F1-score**	**Final Training Accuracy**
Weighted Densenet121 + Inception ResNet V2 + Resnet50 v2	0.9865	0.9852	0.9988
Weighted Inception + Xception	0.9865	0.9865	0.9976
Weighted Densenet121+Mobilenet	0.9775	0.9766	0.9929
weighted Resnet50 + Nasnet	0.973	0.9736	0.9953
Weighted Xception + Nasnet	0.964	0.9625	0.976

5.3.2 Weighted Average Ensemble model: The table compares the performance characteristics of several weighted ensemble models in a deep learning experiment, with an emphasis on validation accuracy, validation F1-score, and final training accuracy. The ensemble model comprising Weighted DenseNet121, Inception ResNet V2, Weighted Inception, and Xception has the greatest validation accuracy of 0.9865 and a validation F1-score of 0.9852, indicating outstanding generalization capabilities. Its ultimate training accuracy of 0.9988 suggests that the training data was learned well and without overfitting. The ensemble of Weighted DenseNet121 and MobileNet achieves a good final training accuracy of 0.9906 but slightly lower validation metrics, with a validation accuracy of 0.973 and a validation F1-score of 0.9737, indicating weaker generalization performance than the top model.The combination of ResNet50 and NasNet produces comparable results, with a validation accuracy of 0.982 and an F1-score of 0.9823. With a final training accuracy of 0.9988, this ensemble displays strong learning ability, albeit it falls slightly short of the best-performing model in validation criteria. The ensemble of Xception and NasNet performs well, with a validation accuracy of 0.9685 and a validation F1-score of 0.9681. Its ultimate training accuracy of 0.9906 is close to that of the Weighted DenseNet121 and MobileNet ensemble, although it is somewhat less effective than the other models. Finally, the combination of Weighted DenseNet121, Inception ResNet V2, Weighted Inception, and Xception emerges as the most successful model because of its higher generalization and excellent training performance. While other models produce competitive results, they lag significantly in either validation metrics or training accuracy, making the top model the most dependable option in this comparison

- 1. Weighted Densenet121+Inception Resnet V2+ ResNet50V2:

The confusion matrix for the Weighted Densenet121 + Inception ResNet V2 + ResNet50 V2 ensemble shows excellent performance in most tissue categories, with obvious categorization along the diagonal suggesting accurate predictions. Tissue types like human_tissue_image-cerebellum, human_tissue_image-melanoma, and human_tissue_image-umbilical-cord exhibit flawless categorization with no mistakes. However, there were certain misclassifications occured: human_tissue_image-gland is wrongly predicted as human_tissue_image-liver once, indicating feature similarity. Human_tissue_image-muscle is misclassified twice as human_tissue_image-testis, but human_tissue_image-testis is misclassified several times, mainly as human_tissue_image-kidney, indicating the difficulty in distinguishing between these

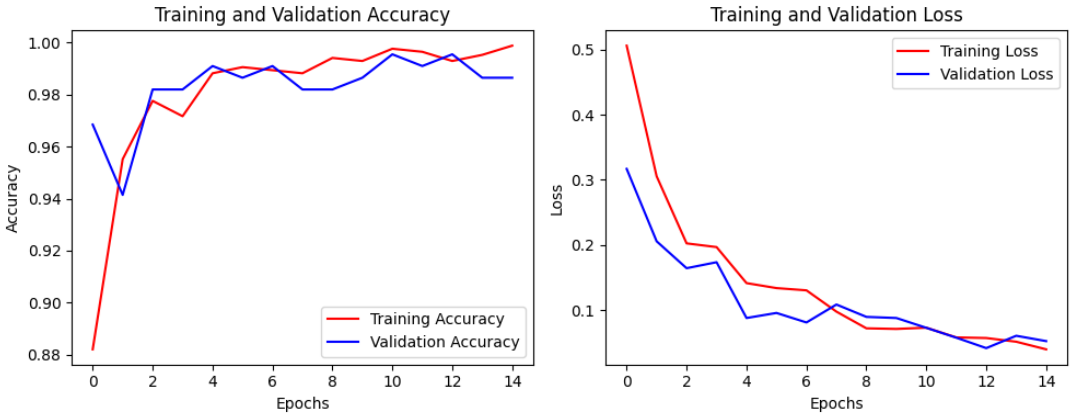

Fig. 42. Training and Validation Accuracy with Loss of Weighted Densenet121+Inception Resnet V2+ Resnet50V2

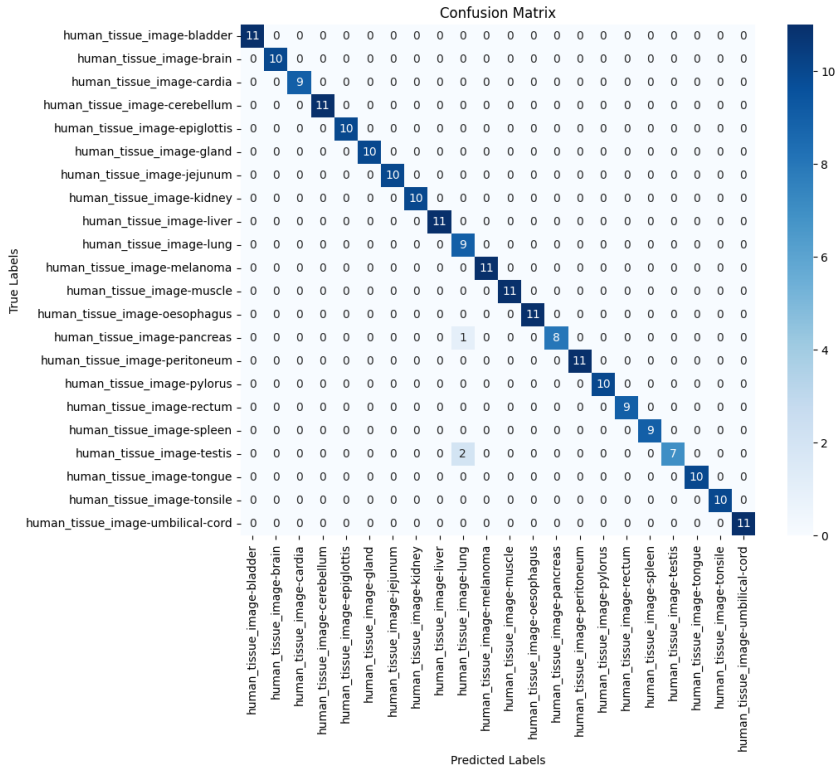

Fig. 43. Confusion Matrix for Weighted Densenet121+Inception Resnet V2+ Resnet50V2

categories. Human_tissue_image-rectum is misidentified three times as human_tissue_image-spleen,

showing structural similarities. Despite these problems, the model is strong and accurate for the majority of tissue types, with room for improvement in managing small inter class similarities.

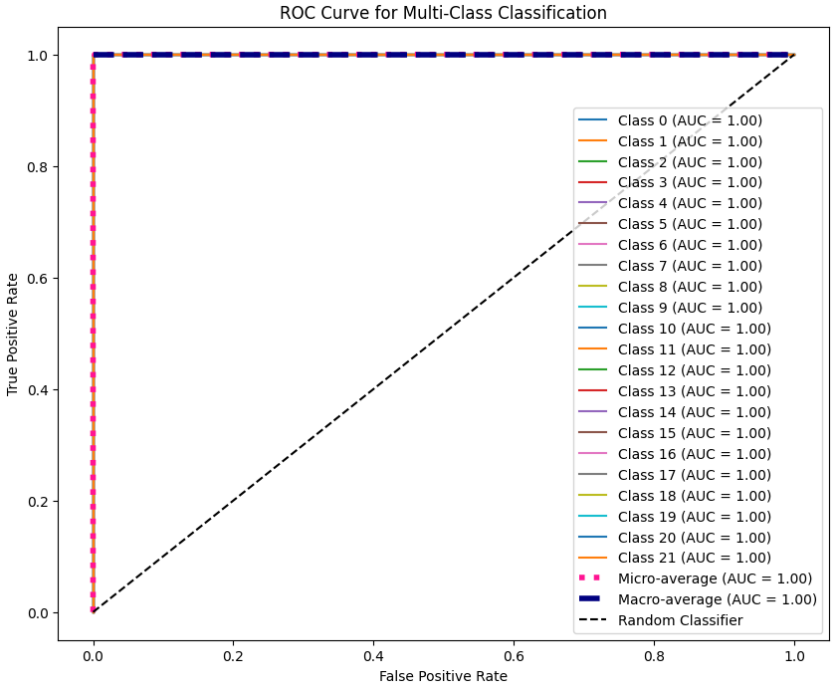

Fig. 44. ROC-AUC curve for Weighted Densenet121+Inception Resnet V2+ Resnet50 v2

- 2. Weighted Inception Resnet V2+ Xception:

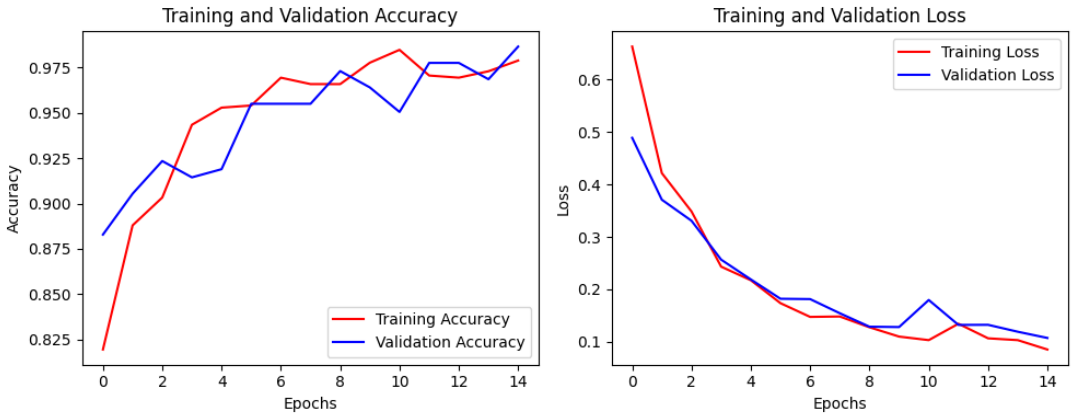

Fig. 45. Training and Validation Accuracy with Loss of Weighted Inception Resnet V2+ Xception

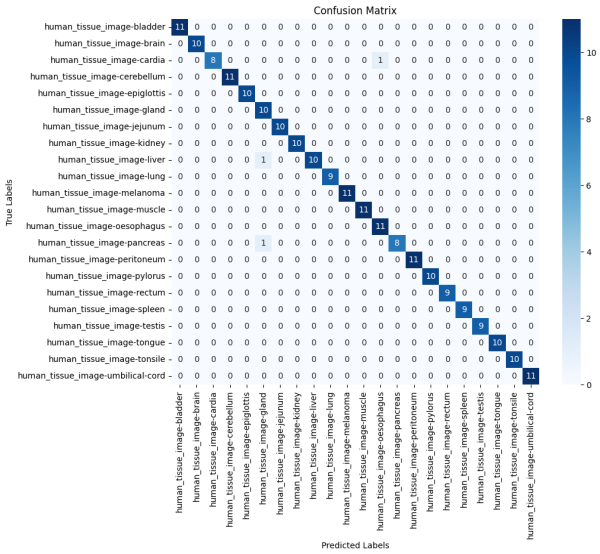

Fig. 46. Confusion Matrix of Weighted Inception Resnet V2+ Xception

The model performs well overall, with excellent accuracy across most tissue types. However, it struggles to discriminate between some groups, such as testis, muscle, and rectum, indicating that more refinement in feature extraction or model fine-tuning is required to enhance categorization in these regions.

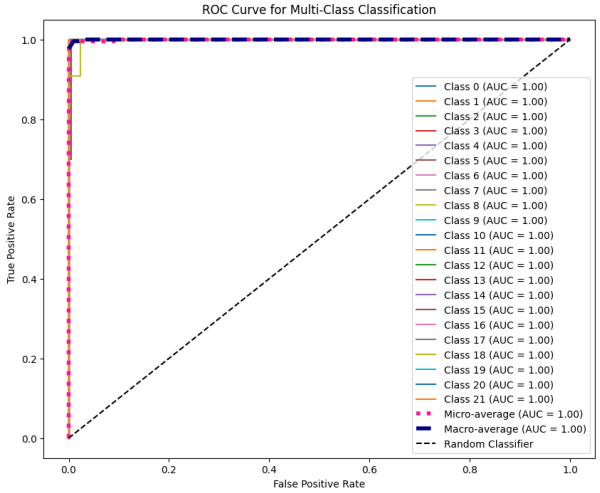

Fig. 47. The ROC-AUC curve for Weighted Inception Resnet V2+ Xception

- 3. Weighted Resnet50 V2+Nasnet mobile:

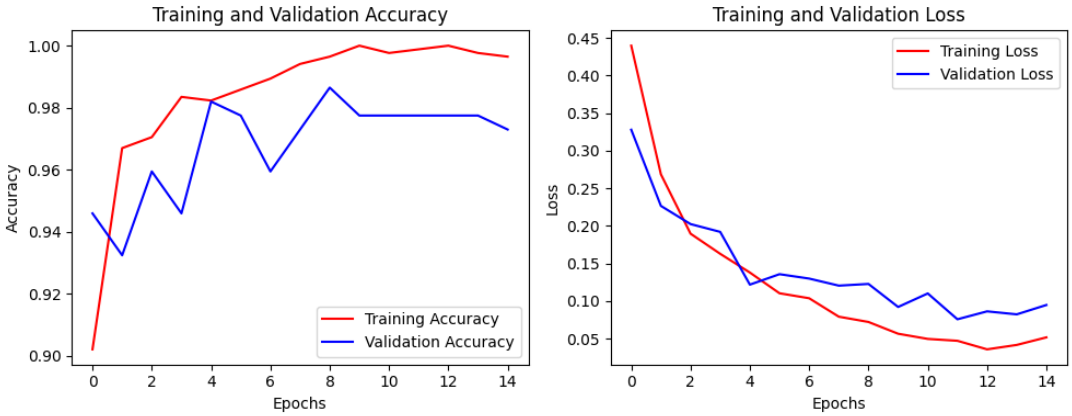

Fig. 48. Training and Validation Accuracy with Loss of Weighted Inception Resnet V2+ NasNet

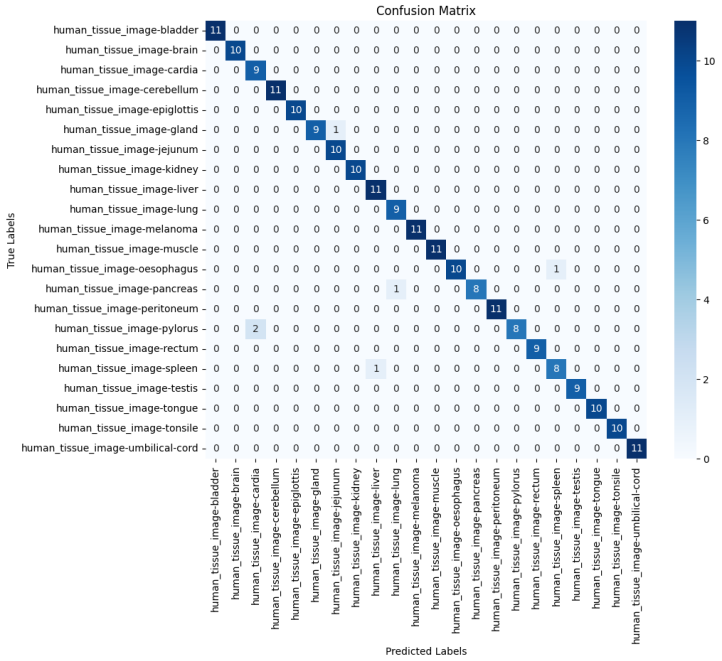

Fig. 49. Training and Validation Accuracy with Loss Weighted Inception Resnet V2+ NasNet

The model, which combines Resnet50 V2 with Nasnet Mobile, navigates the complex environment of tissue categorization with amazing precision. While it accurately detects the majority of tissue types, it occasionally encounters certain misleading terrains, like the testis, muscle, and rectum.

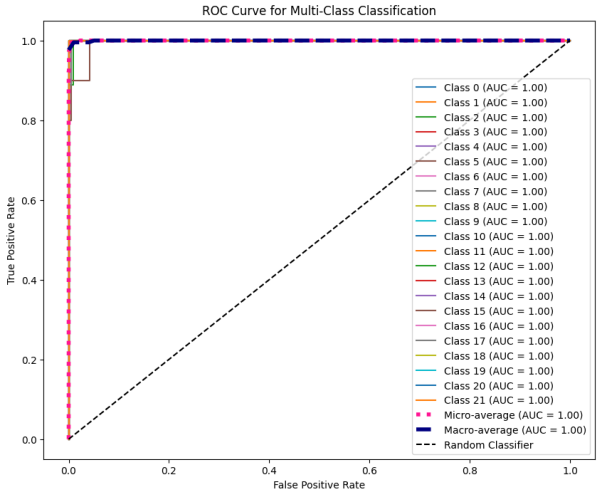

Fig. 50. Training and Validation Accuracy with Loss of Weighted Inception Resnet V2+ NasNet

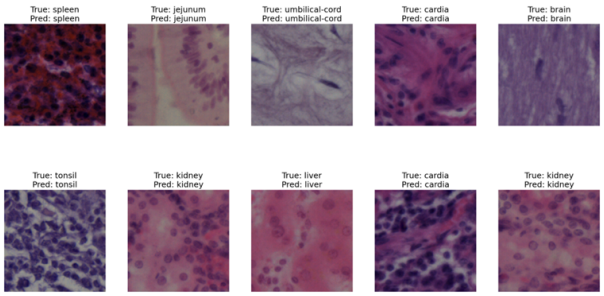

Fig. 51. Example Output

5.4 Example Output

The example output exhibits the model's ability to reliably categorize tissue types using histopathological pictures. Each picture is labeled with both the true label, which represents the real tissue type, and the projected label, which reflects the model's categorization output. The majority of samples are accurately identified, demonstrating the model's efficacy. For example, the spleen is accurately identified as spleen, the jejunum as jejunum, and the umbilical cord as umbilical-cord. Furthermore, tissues such as the brain, tonsils, kidney, and liver are reliably recognized, demonstrating the model's capacity to generalize across several tissue types.

The precise categorization of a wide range of tissue types demonstrates the model's resilience as well as its ability to extract and use characteristics specific to each category. This durability is especially important in medical imaging applications, where great accuracy is required to assure trustworthy diagnostic results. These findings verify the proposed ensemble model's ability to handle complicated histology data while obtaining improved classification accuracy, which is consistent with previous quantitative evaluations. Furthermore, the visually correct classifications highlight the model's practical application in medical

diagnostics, proving its capacity to improve clinical decision-making. The precise categorization of a wide range of tissue types demonstrates the model's resilience as well as its ability to extract and use characteristics specific to each category. This durability is especially important in medical imaging applications, where great accuracy is required to assure trustworthy diagnostic results. These findings verify the proposed ensemble model's ability to handle complicated histology data while obtaining improved classification accuracy, which is consistent with previous quantitative evaluations. Furthermore, the visually correct classifications highlight the model's practical application in medical diagnostics, proving its capacity to improve clinical decision-making.

6 Conclusion and Future Works

This study demonstrates the transformational power of deep learning, transfer learning, and ensemble approaches in the automated categorization of histopathology pictures. Our findings show that pre-trained models, particularly in ensemble configurations, perform better in classification, lowering diagnostic mistakes and boosting the robustness of AI-assisted pathology. We use k-fold cross-validation to ensure that the suggested technique generalizes well across varied histopathology datasets. The use of numerous deep learning architectures enables more extensive feature extraction, successfully capturing the complicated features of tissue samples.

Despite the positive results, there are still significant hurdles and prospects for further research. One of the key drawbacks of deep learning-based histopathology categorization is the scarcity of annotated datasets, which might impair model generalization. To solve this, future research will look at using Generative Adversarial Networks (GANs) to create high-quality histopathological images, therefore expanding the dataset and addressing class imbalance difficulties. Furthermore, hybrid learning approaches, such as few-shot learning and self-supervised learning, might be combined to improve model efficiency, especially in low-data circumstances. Another interesting avenue is the use of explainable AI (XAI) tools, which can increase model interpretability and help physicians comprehend the logic behind automated predictions. Furthermore, expanding this study to multi-modal learning, in which histopathological images are linked with clinical data and genetic information, might result in a comprehensive diagnostic framework for precision medicine.

Finally, this paper presents a reliable and scalable deep learning-based solution for histopathology image classifying, demonstrating the efficacy of transfer learning and ensemble approaches in enhancing diagnostic accuracy. As AI-driven pathology advances, the incorporation of sophisticated generative models, hybrid learning methodologies, and interpretability mechanisms will increase the clinical usefulness of deep learning in digital pathology. By solving these issues, future research can help to construct completely automated, intelligent diagnostic systems, improve patient outcomes, and revolutionize histopathology analysis in modern healthcare.

7 References

References

[1] Y. Wu, M. Cheng, S. Huang, Z. Pei, Y. Zuo, J. Liu, K. Yang, Q. Zhu, J. Zhang, H. Hong *et al.*, "Recent advances of deep learning for computational histopathology: principles and applications," *Cancers*, vol. 14, no. 5, p. 1199, 2022.

[2] F. Yi, J. Huang, L. Yang, Y. Xie, and G. Xiao, "Automatic extraction of cell nuclei from h&e-stained histopathological images," *Journal of Medical Imaging*, vol. 4, no. 2, pp. 027 502–027 502, 2017.

[3] M. Cooper, Z. Ji, and R. G. Krishnan, "Machine learning in computational histopathology: Challenges and opportunities," *Genes, Chromosomes and Cancer*, vol. 62, no. 9, pp. 540–556, 2023.

[4] L. Jose, S. Liu, C. Russo, C. Cong, Y. Song, M. Rodriguez, and A. Di Ieva, "Artificial intelligence–assisted classification of gliomas using whole slide images," *Archives of Pathology & Laboratory Medicine*, vol. 147, no. 8, pp. 916–924, 2023.

[5] J. C. Koo, Q. Ke, Y. C. Hum, C. H. Goh, K. W. Lai, W.-S. Yap, and Y. K. Tee, "Non-annotated renal histopathological image analysis with deep ensemble learning," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. 9, p. 5902, 2023.
[6] A. Ashurov, S. A. Chelloug, A. Tselykh, M. S. A. Muthanna, A. Muthanna, and M. S. Al-Gaashani, "Improved breast cancer classification through combining transfer learning and attention mechanism," *Life*, vol. 13, no. 9, p. 1945, 2023.
[7] J. Zhang, S. Pan, H. Hong, and L. Kong, "Blending ensemble of fine-tuned convolutional neural networks applied to mammary image classification," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 6, pp. 1160–1166, 2019.
[8] P. Ramamoorthy, B. R. R. Reddy, S. Askar, and M. Abouhawwash, "Histopathology-based breast cancer prediction using deep learning methods for healthcare applications," *Frontiers in Oncology*, vol. 14, 2024.
[9] C. C. Ukwuoma, M. A. Hossain, J. K. Jackson, G. U. Nneji, H. N. Monday, and Z. Qin, "Multi-classification of breast cancer lesions in histopathological images using deep_pachi: Multiple self-attention head," *Diagnostics*, vol. 12, no. 5, p. 1152, 2022.
[10] J. C. Koo, Q. Ke, Y. C. Hum, C. H. Goh, K. W. Lai, W.-S. Yap, and Y. K. Tee, "Non-annotated renal histopathological image analysis with deep ensemble learning," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. 9, p. 5902, 2023.
[11] B. Vaishnavi, A. Veluppal *et al.*, "Enhanced breast cancer diagnosis: Leveraging customized transfer learning with machine learning and attention mechanisms for histopathology image classification," in *2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT)*, vol. 1. IEEE, 2024, pp. 1540–1545.
[12] K. Guzel and G. Bilgin, "Classification of nuclei in colon cancer images using ensemble of deep learned features," in *2019 Medical Technologies Congress (TIPTEKNO)*. IEEE, 2019, pp. 1–4.
[13] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, no. 1, p. 7068349, 2018.
[14] A. Serag, A. Ion-Margineanu, H. Qureshi, R. McMillan, M.-J. Saint Martin, J. Diamond, P. O'Reilly, and P. Hamilton, "Translational ai and deep learning in diagnostic pathology," *Frontiers in medicine*, vol. 6, p. 185, 2019.
[15] F. M. Howard, J. Dolezal, S. Kochanny, J. Schulte, H. Chen, L. Heij, D. Huo, R. Nanda, O. I. Olopade, J. N. Kather *et al.*, "The impact of site-specific digital histology signatures on deep learning model accuracy and bias," *Nature communications*, vol. 12, no. 1, p. 4423, 2021.
[16] J. Chaki, S. T. Ganesh, S. Cidham, and S. A. Theertan, "Machine learning and artificial intelligence based diabetes mellitus detection and self-management: A systematic review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3204–3225, 2022.
[17] Y. Fu, A. W. Jung, R. V. Torne, S. Gonzalez, H. Vöhringer, A. Shmatko, L. R. Yates, M. Jimenez-Linan, L. Moore, and M. Gerstung, "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis," *Nature cancer*, vol. 1, no. 8, pp. 800–810, 2020.
[18] T. Kausar, A. Kausar, M. A. Ashraf, M. F. Siddique, M. Wang, M. Sajid, M. Z. Siddique, A. U. Haq, and I. Riaz, "Sa-gan: stain acclimation generative adversarial network for histopathology image analysis," *Applied Sciences*, vol. 12, no. 1, p. 288, 2021.
[19] F.-Z. Nakach, H. Zerouaoui, and A. Idri, "Binary classification of multi-magnification histopathological breast cancer images using late fusion and transfer learning," *Data Technologies and Applications*, vol. 57, no. 5, pp. 668–695, 2023.
[20] X. Li, V. Monga, and U. A. Rao, "Analysis-synthesis model learning with shared features: A new framework for histopathological image classification," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 203–206.
[21] D. Albashish, S. Sahran, A. Abdullah, A. Adam, and M. Alweshah, "A hierarchical classifier for multiclass prostate histopathology image gleason grading," *Journal of Information and Communication Technology*, vol. 17, no. 2, pp. 323–346, 2018.
[22] J. L. Ruiz-Casado, M. A. Molina-Cabello, and R. M. Luque-Baena, "Enhancing histopathological image classification performance through synthetic data generation with generative adversarial networks," *Sensors*, vol. 24, no. 12, p. 3777, 2024.
[23] W. Dee, R. Alaaeldin Ibrahim, and E. Marouli, "Histopathological domain adaptation with generative adversarial networks: Bridging the domain gap between thyroid cancer histopathology datasets," *PloS one*, vol. 19, no. 12, p. e0310417, 2024.
[24] A. Ramanathan, T. Kantheti, C. Zhou, S. Kumara, C. A. Torres-Cabala, and S. P. Iyer, "Classification of human tissues from histopathology images using deep learning techniques," in *2024 Tenth International Conference on Bio Signals, Images, and Instrumentation (ICBSII)*. IEEE, 2024, pp. 1–5.
[25] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PloS one*, vol. 12, no. 6, p. e0177544, 2017.

[26] E. A.-R. Hamed, M. A.-M. Salem, N. L. Badr, and M. F. Tolba, "An efficient combination of convolutional neural network and lightgbm algorithm for lung cancer histopathology classification," *Diagnostics*, vol. 13, no. 15, p. 2469, 2023.

[27] A. Woloshuk, S. Khochare, A. F. Almulhim, A. T. McNutt, D. Dean, D. Barwinska, M. J. Ferkowicz, M. T. Eadon, K. J. Kelly, K. W. Dunn *et al.*, "In situ classification of cell types in human kidney tissue using 3d nuclear staining," *Cytometry Part A*, vol. 99, no. 7, pp. 707–721, 2021.

[28] J. Gowthamy and S. S. Ramesh, "Augmented histopathology: Enhancing colon cancer detection through deep learning and ensemble techniques," *Microscopy Research and Technique*, vol. 88, no. 1, pp. 298–314, 2025.

[29] M. Q. Shatnawi, Q. Abuein, and R. Al-Quraan, "Deep learning-based approach to diagnose lung cancer using ct-scan images," *Intelligence-Based Medicine*, vol. 11, p. 100188, 2025.

[30] S. Patil, A. Pashte, S. Rai, and S. Shah, "Breast cancer classification by implementation of deep-learning with dataset analysis," in *2022 5th International Conference on Advances in Science and Technology (ICAST)*. IEEE, 2022, pp. 1–6.

[31] A. Bhattacharjee, S. Anwar, L. Whitinger, and M. T. Loghmani, "Multimodal sequence classification of force-based instrumented hand manipulation motions using lstm-rnn deep learning models," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2023, pp. 1–6.

[32] M. Srikrishna, R. A. Heckemann, J. B. Pereira, G. Volpe, A. Zettergren, S. Kern, E. Westman, I. Skoog, and M. Schöll, "Comparison of two-dimensional-and three-dimensional-based u-net architectures for brain tissue classification in one-dimensional brain ct," *Frontiers in Computational Neuroscience*, vol. 15, p. 785244, 2022.

[33] A. Gupta, M. Dixit, V. K. Mishra, A. Singh, and A. Dayal, "Brain tumor segmentation from mri images using deep learning techniques," in *International Advanced Computing Conference*. Springer, 2022, pp. 434–448.

[34] M. Hussain, F. Saeed, M. Busaleh, H. Aboalsamh *et al.*, "Mammogram screening for breast density classification using a soft voting ensemble of swin transformers and convnext models," in *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2022, pp. 372–379.

[35] S. P. Ang, S. L. Phung, M. M. Schira, A. Bouzerdoum, and S. T. M. Duong, "Human brain tissue segmentation in fmri using deep long-term recurrent convolutional network," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2018, pp. 1–7.

[36] K. Pranitha, N. Vurukonda, and R. K. Nayak, "A comprehensive survey on mri images classification for brain tumor identification using deep learning techniques," in *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE, 2022, pp. 1206–1212.

[37] T. M. El-Achkar, S. Winfree, N. Talukder, D. Barwinska, M. J. Ferkowicz, and M. Al Hasan, "Tissue cytometry with machine learning in kidney: From small specimens to big data," *Frontiers in Physiology*, vol. 13, p. 832457, 2022.

[38] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou, "Multisource transfer learning with convolutional neural networks for lung pattern analysis," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 76–84, 2016.

[39] S. S. M. Khairi, M. A. A. Bakar, M. A. Alias, S. A. Bakar, C.-Y. Liong, N. Rosli, and M. Farid, "Deep learning on histopathology images for breast cancer classification: A bibliometric analysis," in *Healthcare*, vol. 10, no. 1. MDPI, 2021, p. 10.

[40] F. Akram, D. P. de Bruyn, Q. C. van den Bosch, T. E. Trandafir, T. P. van den Bosch, R. M. Verdijk, A. de Klein, E. Kiliç, A. P. Stubbs, E. Brosens *et al.*, "Prediction of molecular subclasses of uveal melanoma by deep learning using routine haematoxylin–eosin-stained tissue slides," *Histopathology*, vol. 85, no. 6, pp. 909–919, 2024.

[41] E. Klang, A. Soroush, G. N. Nadkarni, K. Sharif, and A. Lahat, "Deep learning and gastric cancer: systematic review of ai-assisted endoscopy," *Diagnostics*, vol. 13, no. 24, p. 3613, 2023.

[42] R. Munirathinam, M. Latha, M. Muthulakshmi, M. P. Reddy, A. Naveen, and M. Tamilnidhi, "Comparison of deep learning models for efficient classification of gastric abnormalities," in *2024 Tenth International Conference on Bio Signals, Images, and Instrumentation (ICBSII)*. IEEE, 2024, pp. 1–7.

[43] G. Slabaugh, L. Beltran, H. Rizvi, P. Deloukas, and E. Marouli, "Applications of machine and deep learning to thyroid cytology and histopathology: a review," *Frontiers in Oncology*, vol. 13, p. 958310, 2023.

[44] J. Zhang, S. Pan, H. Hong, and L. Kong, "Blending ensemble of fine-tuned convolutional neural networks applied to mammary image classification," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 6, pp. 1160–1166, 2019.

[45] D. Xue, X. Zhou, C. Li, Y. Yao, M. M. Rahaman, J. Zhang, H. Chen, J. Zhang, S. Qi, and H. Sun, "An application of transfer learning and ensemble learning techniques for cervical histopathology image classification," *IEEE Access*, vol. 8, pp. 104 603–104 618, 2020.

[46] Y. Zheng, C. Li, X. Zhou, H. Chen, H. Xu, Y. Li, H. Zhang, X. Li, H. Sun, X. Huang *et al.*, "Application of transfer learning and ensemble learning in image-level classification for breast histopathology," *Intelligent Medicine*, vol. 3,

no. 02, pp. 115–128, 2023.
[47] C. Mazo, J. Bernal, M. Trujillo, and E. Alegre, "Transfer learning for classification of cardiovascular tissues in histological images," *Computer methods and programs in biomedicine*, vol. 165, pp. 69–76, 2018.
[48] A. A. Balasubramanian, S. M. A. Al-Heejawi, A. Singh, A. Breggia, B. Ahmad, R. Christman, S. T. Ryan, and S. Amal, "Ensemble deep learning-based image classification for breast cancer subtype and invasiveness diagnosis from whole slide image histopathology," *Cancers*, vol. 16, no. 12, p. 2222, 2024.
[49] C. C. Ukwuoma, M. A. Hossain, J. K. Jackson, G. U. Nneji, H. N. Monday, and Z. Qin, "Multi-classification of breast cancer lesions in histopathological images using deep_pachi: Multiple self-attention head," *Diagnostics*, vol. 12, no. 5, p. 1152, 2022.
[50] D. Albashish, "Ensemble of adapted convolutional neural networks (cnn) methods for classifying colon histopathological images," *PeerJ Computer Science*, vol. 8, p. e1031, 2022.
[51] J. C. Koo, Q. Ke, Y. C. Hum, C. H. Goh, K. W. Lai, W.-S. Yap, and Y. K. Tee, "Non-annotated renal histopathological image analysis with deep ensemble learning," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. 9, p. 5902, 2023.
[52] S. K. Addagarla, G. K. Chakravarthi, and P. Anitha, "Real time multi-scale facial mask detection and classification using deep transfer learning techniques," *International Journal*, vol. 9, no. 4, pp. 4402–4408, 2020.
[53] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art," *Knowledge-based systems*, vol. 212, p. 106622, 2021.
[54] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.
[55] F. A. Shah, M. A. Khan, M. Sharif, U. Tariq, A. Khan, S. Kadry, and O. Thinnukool, "A cascaded design of best features selection for fruit diseases recognition," *Comput. Mater. Contin*, vol. 70, no. 1, pp. 1491–1507, 2022.