Article

Not peer-reviewed version

Digital Pathology and Ensemble Deep Learning for Kidney Cancer Diagnosis: Dartmouth Kidney Cancer Histology Dataset

Muskan Naresh Jain , Salah Mohammed Awad Al-Heejawi , Jamil R. Azzi , Saeed Amal *

Posted Date: 21 November 2024

doi: 10.20944/preprints202411.1615.v1

Keywords: kidney cancer diagnosis; deep learning; convolutional neural network; artificial intelligence; computer vision; histopathology images; foundation models; image processing

Article

Digital Pathology and Ensemble Deep Learning for Kidney Cancer Diagnosis: Dartmouth Kidney Cancer Histology Dataset

Muskan Naresh Jain [1], Salah Mohammed Awad Al-Heejawi [1], Jamil R. Azzi [2] and Saeed Amal [3,*]

[1] College of Engineering, Northeastern University, Boston, MA 02115, USA
[2] Transplantation Research Center, Renal and Engineering Divisions, Brigham and Women's Hospital, Harvard Medical School
[3] The Roux Institute at Northeastern University, Portland, ME, United States; Department of Bioengineering, Northeastern University, Boston, MA, United States
* Correspondence: s.amal@northeastern.edu

Simple Summary: Kidney cancer is a significant health concern, requiring the immediate need for early diagnosis and precise treatment strategies to decrease mortality issues. Identifying the accurate type of kidney cancer is important for determining the best treatment plan to cure the cancer. Although many AI methods are used for classifying various cancers, they are not yet widely adopted in clinical settings. In our paper, we aimed to bridge this gap by developing an ensemble of deep learning algorithms and transformer models including ViT (Vision Transformer), CAiT (Class-Attention in Image Transformers), DeiT (Data Efficient Image Transformers), ResNet to classify the Dartmouth Kidney Cancer Histology Dataset into five types of kidney cancers: benign, chromophobe, clear cell, oncocytoma, and papillary. We organized the patch extraction resulting in 26088 patches. The accuracy results after training the images on ResNet, as well as other state-of-the-art transformer models such as CAiT, ViT, DeiT, Swin and the Ensemble model, are 95.03%, 98.73%, 97.65%, 99.24%, 98.43% and 99.26% respectively which indicates that the Ensemble model, combining Swin and Vision Transformer, has the greatest validation accuracy. By leveraging the strengths of Swin and Vision Transformers, the ensemble model excels at identifying unique features necessary for accurate classification. This paper presents a detailed analysis and comparison of our models and methodologies, demonstrating their potential to improve kidney cancer diagnosis.

Abstract: Kidney cancer has become a major global health issue over time, showing how early detection can play a very important role in mediating the disease. Traditional histological image analysis is recognized as the clinical gold standard for diagnosis although it is highly manual and labor-intensive. Due to this issue, many are interested in computer-aided diagnostics technologies to assist pathologists in their diagnostic. Specifically, deep learning (DL) has become a viable remedy in this field. Nonetheless, the capacity of existing DL models to extract comprehensive visual features for accurate classification is limited. Towards the end, this study proposes using ensemble models that combine the strengths of multiple transformers and deep learning model architectures. By leveraging the collective knowledge of these models, the ensemble enhances classification performance and enables more precise and effective kidney cancer detection. This study compares the performance of these suggested models to previous studies, all of which used the publicly accessible Dartmouth Kidney Cancer Histology Dataset. This study showed that the Vision Transformers, with an average accuracy of over 99%, were able to achieve high detection accuracy across all complete slide picture patches. In particular, the CAiT, DeiT, ViT, and Swin models outperformed ResNet. All things considered, the vision Transformers consistently produced an average accuracy of 98.51% across all five folds 5 folds. These results demonstrated that Vision Transformers might perform well and successfully identify important features from smaller patches. Through utilizing histopathological images, our findings will assist pathologists in diagnosing kidney cancer, resulting in early detection and increased patient survival rates.

Keywords: kidney cancer diagnosis; deep learning; convolutional neural networks; image classification; artificial intelligence; computer vision; histopathology images; foundation models; image processing doi:10.20944/preprints202411.1615.v1

1. Introduction

Patients with renal cancer frequently exhibit symptoms such as anemia and fever which can lower blood cancer levels and lead to poor red blood cell counts. The most common harmful subtype of renal cancer is clear cell renal carcinoma [Figure 1 (a)]. The stroma of these tumors is highly vascular, which often leads to hemorrhagic regions. The characteristic yellow appearance of the tumor surface is due to the lipid composition of the cells, which includes high levels of cholesterol, neutral lipids, and phospholipids. [1,2]. Roughly 10% of renal cell carcinomas are papillary renal cell carcinomas [Figure 1 (b)]. Like clear cell renal cell carcinoma, papillary renal cell carcinoma has an age distribution with a reported mean age at diagnosis typically ranging from 50 to 65 years. Necrosis is a common feature of papillary renal cell carcinomas [3–5]. About 5% of renal cancers are chromophobe renal cell carcinomas. Their prognosis is better than that of clear cell kidney carcinoma. The death rate is under 10%. There have been cases of chromophobe renal carcinoma that have distantly metastasized to the pancreas, liver, and lung. It has been proposed that chromophobe renal tumors have a higher incidence of liver metastasis than other histological subtypes. [6,7] [Figure 1 (c)]. Renal oncocytoma is a benign tumor, and it is believed to be the precursor of eosinophilic chromophobe renal cell carcinoma, which is the malignant variant of this tumor [Figure 1(d)]. Benign kidney cancer can manifest in several forms including renal clear cell, papillary, chromophore, oncocytoma, and benign. [Figure 1]

RCCs (renal cell carcinomas) typically do not cause any symptoms until late in the course of the illness, and more than 50% of tumors are discovered by chance [9]. Merely 10 to 15% of patients can exhibit the "classic triad," which consists of flank fullness, hematuria, and discomfort.

The experience and expertise of the pathologist plays a major role in the accuracy of the histopathological analysis, which leaves the manual method open to human error including improper diagnosis and detection. Additionally, a lack of pathologists causes major delays in the examination of patient cases, which may result in the diagnosis of cancer later than expected [10,11].

The three main components of the proposed CAD system as described by Shehata et al. [13] that preprocessing of grey images to creates 3D segmented objects representing renal tumors; extracting various discriminating features (texture and functional) from segmented objects; and completing a two-stage classification process using various machine learning classifiers to determine the renal tumor's final diagnosis. [13]. Texture analysis is frequently utilized to give multiple quantitative patterns or descriptors that may be obtained by linking the grey values of each pixel in each tumor image or volume during the process of obtaining discriminating features.

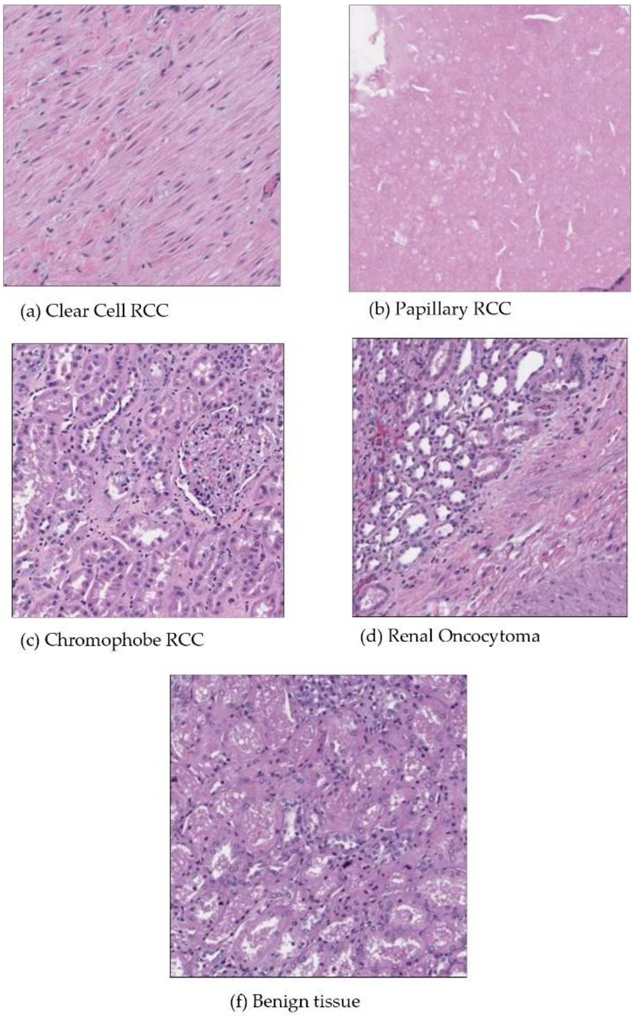


Figure 1. An example of microscopy images for five classes in Dartmouth Kidney Cancer Histology Dataset (on 40 × magnification).

2. Materials and Methods

2.1. Literature Reviews

There are imaging techniques such as CT (Computed Tomography), MRI (Major Research Instrumentation) that can detect tumors or abnormal growths in the kidneys. Blood and urine tests are also conducted to identify the features linked to kidney cancer. The identification and categorization of kidney cancer has made great progress in recent years because of the application of deep learning techniques. By combining radiomic characteristics and clinical data, Liu et al. (2021) developed a multimodal deep-learning system that achieved a 94% accuracy rate in the early diagnosis of kidney cancer disease [23]. To classify kidney cancers in MRI Scans, Chen et al (2020) used transfer learning with pre-trained VGG-16 and ResNet-50 models, obtaining accuracies of 87% and 89%, respectively. Using a modified U-Net architecture, Gao et al. (2019) achieved 90% classification accuracy and 91% segmentation accuracy when they categorized renal cell carcinoma in histopathological pictures. Additionally, Nguyen et al. (2020) obtained a Dice coefficient of 0.89 and a classification accuracy of 92%, demonstrating the efficacy of a combined CNN-RNN strategy for segmenting and categorizing renal masses. These findings demonstrate the revolutionary potential of deep learning to improve the precision and efficacy of kidney cancer diagnosis, opening the door to more dependable and individualized treatment approaches. The paper by Breggie et al. (2023) presents a web app using AI and multimodal data to improve prostate cancer diagnosis. Users valued its summary tabs and high-resolution images. The study suggests improvements while highlighting the app's potential to enhance diagnostic accuracy and efficiency.

Originally, hand-crafted features like color, texture, and morphology were extracted from histopathology pictures and used to identify kidney cancer using classic machine learning (ML) methods including support vector machines (SVM), random forests, and Ada boost. Compared to manual analysis, these techniques significantly improved results, producing more reliable and consistent outcomes. For example, Zhou et al. (2019) achieved a noteworthy accuracy of 88% by using a mix of color and texture information in a random forest classifier to identify between benign and malignant kidney cancers.[12] Deep learning's introduction has completely changed the field of digital pathology. Convolutional neural networks (CNNs) have proven to be incredibly effective at tasks like cancer-related histopathology, image recognition, and segmentation. CNNs are especially useful for medical image analysis since they can automatically learn hierarchical feature representations from raw pixel data. Research has demonstrated that CNNs are as accurate as human pathologists, if not more so, in certain areas, including determining the severity of a tumor and recognizing malignant tissues. Esteva et al. (2017), for instance, showed that a CNN could classify skin cancer with dermatologist-level accuracy [14]. This discovery has now been applied to other cancer types, such as kidney cancer.

A study by Jiang et al. (2024) assessed the suggested models using the recently released Dartmouth Kidney Cancer Histology Dataset [16] to determine their effectiveness. Ivanova et al. [22] review AI models for renal cell carcinoma (RCC) diagnosis using the histological image dataset, highlighting high accuracies in classification and grading tasks. Convolutional Neural Networks (CNNs) and deep learning models often exceed 90% accuracy, with one CNN achieving 99.1% accuracy in RCC tissue identification. Other effective approaches include Bayesian classifiers and Support Vector Machines. These AI techniques show significant potential for improving RCC diagnosis and management in clinical practice.

This study's primary contributions are the creation of transformer models and an efficient deep ensemble learning model that outperforms existing research on the Kidney Cancer Histology dataset for kidney cancer detection. Moreover, the successful identification of kidney histology patches by the ensemble model of Swin and Vision Transformer may lead to a reduction in the number of digital scanners, data storage devices, and computer servers required for histopathology-related tasks. This has the potential to improve patient survival rates and raise the likelihood of renal or kidney cancer being detected early [13,16].

2.2. Dataset Description

The Dartmouth Kidney Cancer Histology Dataset is a large collection of 563 whole-slide images (WSIs) stained with hematoxylin and eosin (H&E) that have been carefully chosen for analysis and kidney cancer. The images provide a broad dataset that is essential for research in digital pathology and machine learning applications in medical diagnostics. The dataset includes a wide range of kidney cancer subtypes, including oncocytoma, chromophobe renal cell carcinoma (chRCC), papillary renal cell carcinoma (pRCC), and clear cell renal cell carcinoma (ccRCC). To properly categorize and diagnose kidney tumors, machine learning models need to be trained with these as labels. The dataset includes metadata including the file name, image class, slide type, and split type (Train, Test, and Val), in addition to various demographic data. Understanding the context of each histopathology image and performing in-depth analysis is facilitated with this information. In particular, the dataset is useful for creating and comparing computer-aided diagnostic (CAD) systems. It offers a wealth of data for deep learning models and other machine learning algorithms to improve their performance in kidney cancer diagnosis and classification. Additionally, it facilitates clinical decision-making by offering a point of reference for the confirmation and comparison of diagnostic results.

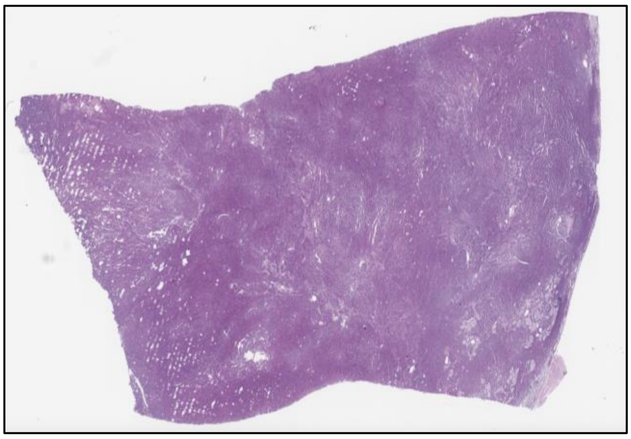
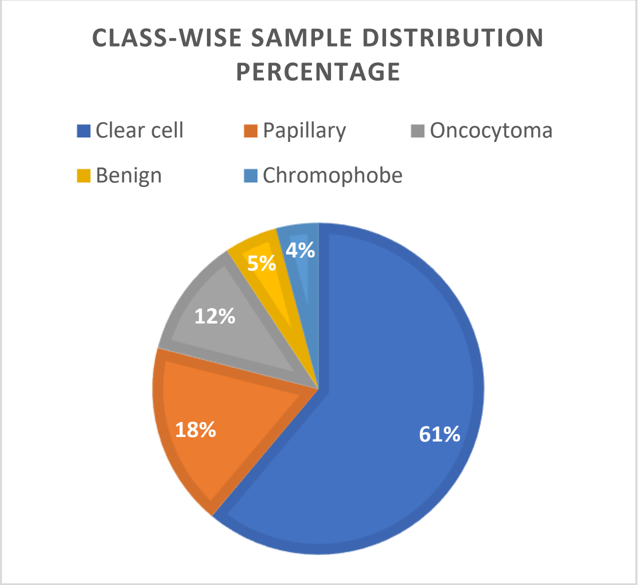
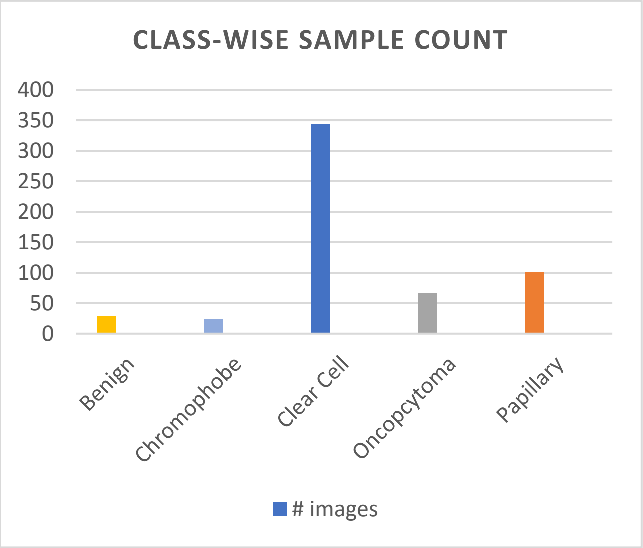


Figure 2. Example of histopathological Kidney Cancer whole slide images.



(a) pie chart representing the total number of images for every kidney cancer subclass



(b) bar chart representing class-wise sample distribution percentage

2.3. Methodology Overview

This study demonstrates the use of Vision Transformers techniques with CNN architectures to identify kidney cancer patches. There are four main steps in the process: First, the dataset is created by removing empty patches and augmenting it. Next, pre-trained networks or base models are tailored. Third, the most successful base models are selected to generate ensemble models. Finally, the models are evaluated and presented using various metrics and the class activation map.

The ensemble model approach is a key feature of this study, combining the strengths of Vision Transformers (ViT) and Swin Transformer architectures. This ensemble strategy leverages the complementary capabilities of both models, with ViT excelling in capturing global image context and Swin Transformer adept at handling multi-scale feature hierarchies. By averaging the outputs of these two powerful models, the ensemble achieves a synergistic effect, enhancing overall classification accuracy and robustness. The ensemble model demonstrated exceptional performance, achieving a remarkable accuracy of 99.26% in classifying kidney cancer histology images.

Data preprocessing was done to improve the model's performance by deleting non-informative empty patches from the dataset. These patches would have biased the training process and compromised the model's performance. Following the elimination of empty patches, data augmentation was used to expand the training dataset.

2.4. Empty Patch Removal Process

This study focuses on the efficient management and processing of whole-slide images (WSIs) for patch extraction using OpenSlide Library. The main goal is to eliminate empty patches, defined as those with over half of pixels having RGB intensity values greater than 230 in all channels. OpenSlide, an open-source C library, is used to read and modify digital pathology images. The implementation involves using OpenSlide to read WSIs and tools from the tools package for tissue detection and patch extraction. The process includes setting up paths, reading metadata, and using a Tissue Detector class with a Gaussian Naive Bayes model for tissue recognition. A Patch Extractor class is employed with specific parameters to extract relevant patches. The workflow is optimized through parallel processing using Python's multiprocessing package, resulting in an efficient and transparent approach for managing WSIs and extracting valuable data for further analysis. With this process, every high-resolution image is broken into different patches depending on the RGB intensity, and the ensemble models are trained on them.

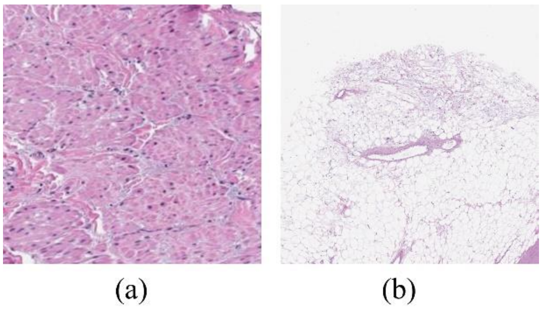


Figure 3. The pictures above exhibit examples of histopathology: (a) tissue patch image and (b) empty patch image.

2.5. Pretrained networks as Base Models

Since the beginning of deep learning, Convolutional Neural Networks (CNNs) have been helpful in many applications because of their constant improvements in strength, efficiency, and adaptability. CNNs, which are specifically designed for computer vision problems and use convolutional layers inspired by natural visual processes, are a great example of this innovation. The accuracy, speed, and overall performance of various CNN structures have improved over time, and they are frequently compared to the ImageNet project —a sizable visual database that fosters advances in computer vision.

In the past, training CNNs from scratch took a lot of time and computer power. By using previously learned information from trained models, transfer learning (TL) offers a useful shortcut that can speed up optimization and possibly increase classification accuracy. TL entails transferring weights from pre-trained models, using insights acquired from varied datasets, and speeding training processes to improve model accuracy, particularly in complicated architectures.

ResNet50 Architecture:

Deeper than ResNet34, ResNet50 is a 50-layer variant of the ResNet architecture. While this increased depth can lead to better performance on some tasks, training with it requires more processing power. By enabling gradients to flow across shortcut connections, ResNet50, a deep convolutional neural network with 50 layers, introduces the idea of residual learning and helps to address the disappearing gradient issue. This design is efficient in several computer vision applications, most notably picture categorization.

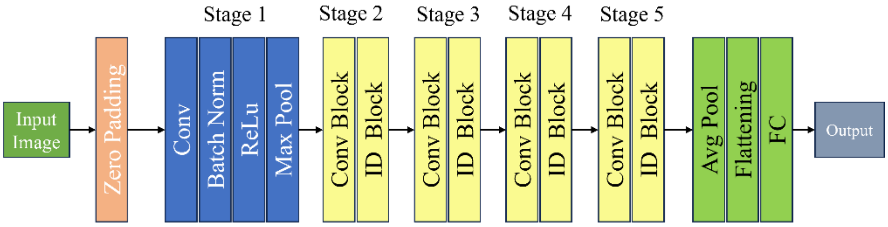


Figure 4. ResNet50 Model Architecture.

Transformers

Transformers are network models that use attention to understand the sequence of information like frames in a movie, words in a sentence, notes in music, or pixels of an image. The transformer networks can capture relationships and dependencies between the elements even if they are far apart from each other. The ability to capture long-range dependencies makes transformers powerful for tasks like language understanding where the meaning of the words depends on words that appear earlier or later in the sentence. The Transformer network consists of two main parts: 1) Encoder; and 2) Decoder.

Encoder and Decoder

The input sequence that we get from positional encoding is passed through the encoder. Each encoder consists of a self-attention mechanism and feed feed-forward neural network to capture the contextual information and dependencies between the words. Multi head attention layer helps the model to figure out which words are important to each other and how they relate to one another. In the self-attention layer, each word will have three jobs such as query, key, and value. A query is a word looking for other words to pay attention to. The key is a word being looked at by other words. The self-attention layer looks at each word and compares it with all other words in the sentence and see how they are related to each other. It calculates the similarity between each word query and all word keys. The words with the higher scores will be prioritized. Add and norm layer is applied after the multiheaded attention layer and feed-forward neural network in each transformer network. It preserves the original information from the previous layer which allows the model to learn and update the new information captured by the sub layer. It assists in addressing the vanishing gradient problems and allows the model to learn more effectively.

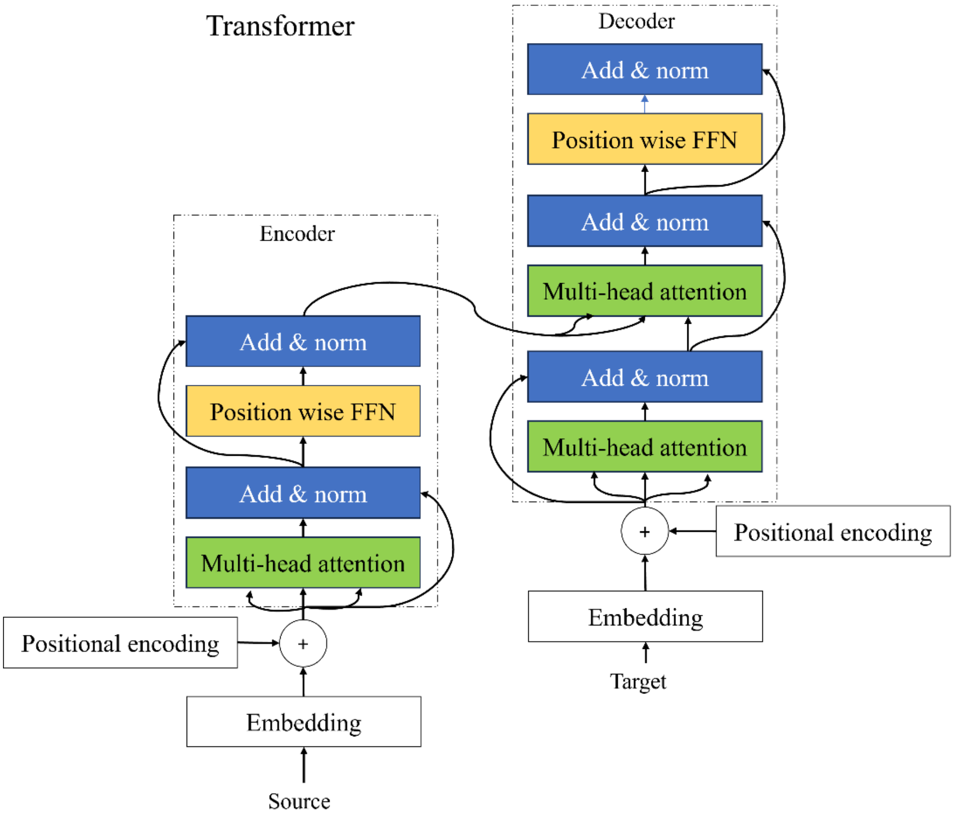


Figure 5. Architecture of Encoder and Decoder of Transformer.

The primary function of the decoder is to transform encoded representations back into the desired output format. In sequence generation tasks like machine translation and text summarization, the decoder predicts the next token in the sequence at each time step, often utilizing techniques like beam search to improve output quality. In data reconstruction applications, such as image or audio reconstruction, the decoder transforms encoded latent representations back into the original data format, a common approach in autoencoders and generative models like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). Additionally, decoders are designed for conditional output production, where the output is conditioned on additional context or input data, and for error correction and denoising, where they reconstruct clean data from noisy inputs. Transformer-based decoders.

The decoder is integral to various neural network architectures, especially in sequence-to-sequence models and data reconstruction tasks. It typically includes an embedding layer, recurrent layers, attention mechanisms, transformer layers, and an output layer. These components collaboratively transform encoded representations into meaningful outputs. The decoder's primary functions include sequence generation, data reconstruction, conditional output production, and error correction. Transformer-based decoders utilize self-attention, cross-attention, and feed-forward networks for enhanced performance. Applications of decoders span natural language processing, computer vision, speech processing, and healthcare, highlighting their versatility and importance in modern neural network models. Understanding their architecture and functions is essential for optimizing data transformation tasks.

CAiT Architecture (Class Attention in Image Transformers)

The CAiT (Class-Attention in Image Transformers) transformer is a novel architecture designed to enhance the performance of vision transformers (ViTs) in image classification tasks. Traditional vision transformers apply self-attention mechanisms uniformly across all patches of an input image, which can sometimes lead to suboptimal learning of class-specific features. CAiT introduces a unique class-attention mechanism that focuses on improving the interaction between class tokens and image patches, leading to better representation learning and classification accuracy. In the CAiT architecture, a class token is appended to the sequence of image patches, and attention is specifically directed towards this class token. This design allows the model to aggregate and emphasize class-specific information more effectively. The class-attention mechanism is integrated at multiple stages of the transformer, enhancing the model's ability to capture and utilize discriminative features necessary for accurate classification. Additionally, CAiT incorporates deeper transformer layers and a progressive learning approach, gradually increasing the model's complexity and capacity. This results in improved convergence and performance on various image recognition benchmarks, making CAiT a powerful architecture for vision tasks.

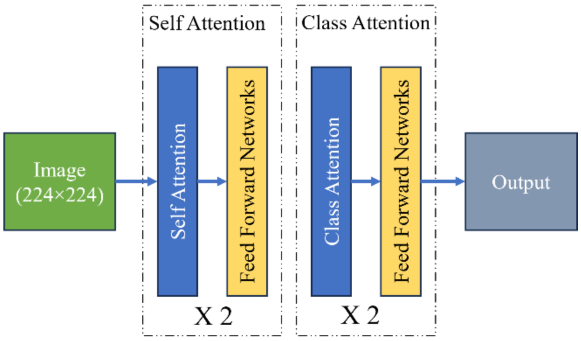


Figure 6. CAiT Transformer Architecture.

VitNet Architecture:

Vision Transformers use self-attention. It allows the model to understand the relationship between different parts of an image by assigning important scores to patches and focusing on the most relevant information. This helps the model make better sense of the image and perform various tasks related to computer vision. It breaks images into smaller patches. [19] The statement used in the paper ‘An image is worth 16x16 words’ means how many pixels the sliding window moves each time. Each patch is treated as a separate input token. There is no decoder in the vision transformer, it is an encoder only transformer. Linear projection works on flattened patches by transforming 1D vector into lower dimensional representation. It preserves the important features.

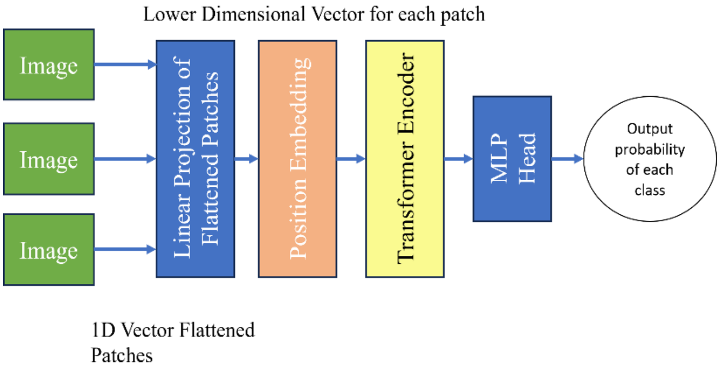

Lower Dimensional Vector for each patch
Image
Image
Image
Linear Projection of Flattened Patches
Position Embedding
Transformer Encoder
MLP Head
Output probability of each class
1D Vector Flattened Patches

Figure 7. VITNet Model Architecture.

DeiT Architecture:

The difference between ViT and DeiT is that originally ViT was trained on a massive dataset having 300M samples of data [20]. DeiT on the other hand, trains on well-known ImageNet Dataset. ViT takes a long time to get trained whereas DeiT trains in 2 or 3 days on a single 8GPU or 4GPU machine. DeiT uses knowledge distillation which means transferring knowledge from one model/network to another. Regularization is used which means the overfitting of a network is being reduced to limited training data so that the model does not learn the noise from the training data but the actual information from the data. Augmentation is when multiple samples are created of the same input with some variations. Suppose there is a model which classifies cats and dogs. We pass the cat image through the model and get the embeddings of the image. The embeddings are passed through self max function to get the probabilities of the dog and cat. Cross entropy loss is compared with the ground truth label and the entire function. With distillation we distill the knowledge from another network called the teacher network, we get the embeddings and pass it through self-max with the temperature parameter to get the output probabilities so that it becomes smoothened.

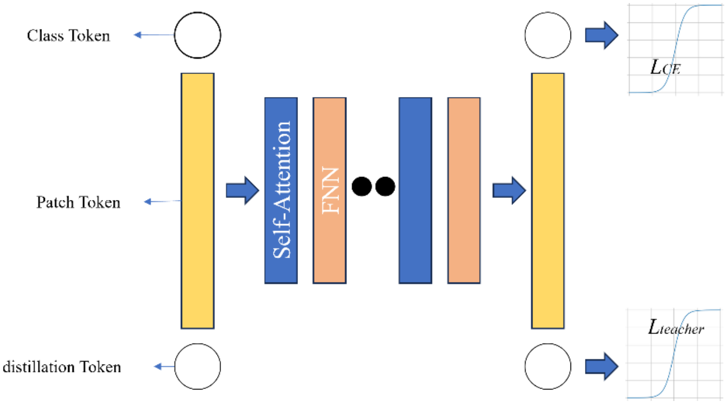


Figure 8. DeiT Model Architecture.

Swin Architecture

Swin Transformers are more accurate than Vision Transformers in some cases due to their capacity to handle large images and high-resolution images with lower computational complexity. The Swin Transformer, or Shifted Window Transformer, enhances traditional vision transformers by targeting their limitations in image processing and the process by which they do it. It is constructed as a hierarchical design with shifted windows, enabling efficient and scalable visual data modeling. Instead of using the whole slide image all at once, it is divided into different sections. The model looks at the relationship between all the features and then analyzes the section. These windows or sections are shifted across all layers so that they can make connections with different features of the image. This method shows that Swin Transformer can detect images with accuracy.

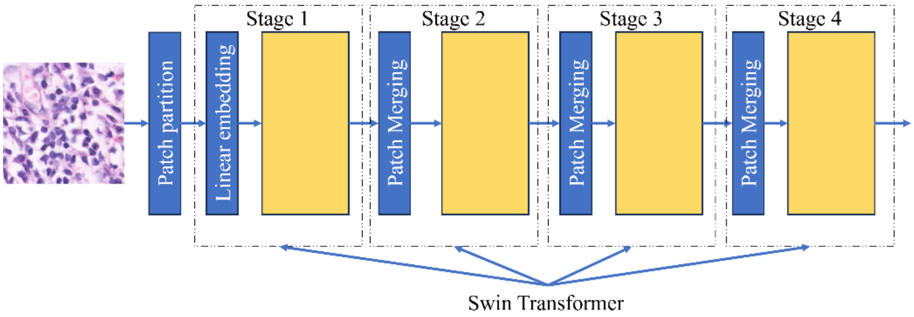


Figure 9. Diagram showing Swin Transformer architecture.

In this research, CNN model architecture and vision transformers were used. Initially, each model was trained independently to determine its unique performance. Then the best-performing epochs for each model are based on validation accuracy.

To improve the robustness and generalizability of the techniques, 5-fold cross-validation is used. During each fold of the cross-validation approach, the research utilized the average calculation of the last epoch of every fold to calculate the best performing validation accuracy.

2.6. Experimental Setting

The data was divided into training and validation sets. Each network was trained for 12 epochs using 5-fold cross-validation to create the model. The weights from the epoch with the best validation accuracy were chosen as the final representations for each model. Various metrics were then employed to assess accuracy, followed by many objective assessment factors to determine overall performance.

3. Results

The performance evaluation criteria used include validation Accuracy and Validation Cohen Cappa Score. Positive samples include abnormal or malignant patches, whereas negative samples contain normal or healthy patches. The phrases true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are used to describe the various prediction results.

1. Train Loss: A machine learning model's fit to the training set of data is shown by its train loss. On the training dataset, it measures the difference between the expected and actual outputs. Reducing this loss is the goal of training to enhance the model's functionality.

2. Validation Loss: A machine Learning model's ability to generalize to previously unknown data is measured by the validation loss. On the validation dataset, it measures the difference between the expected and actual outputs. ￼

3. Validation Accuracy means the ratio of correctly predicted instances out of the total number of instances in the validation dataset. It's computed as:

$$\text{Validation Accuracy} = \frac{(Tp+TN)}{(TN+FP+TP+FN)}$$

4. Validation Cohen Cappa Score is a statistical measure.

A complete view of the model's performance, especially in differentiating between positive and negative data, can be obtained by looking at these metrics.

Performance metrics must be considered while evaluating the efficacy of machine learning models. These metrics offer quantifiable figures that represent a statistical or machine-learning technique's overall performance. Performance metrics assess the model's ability to consistently produce the correct classifications and its ability to classify data points accurately in classification tasks. The table below displays the study's conclusions, which were arrived at by looking at various performance criteria.

Table 1. The effectiveness of the several deep learning models was assessed as displayed below.

Model	Fold	Train Loss	Train Accuracy	Val Accuracy	Val Cohen Cappa Score	Average Val Accuracy
Resnet50	1	0.049504	0.981212	0.935595	0.915757	**0.9503162**
	2	0.028646	0.990174	0.975848	0.970563	
	3	0.091897	0.964484	0.936362	0.906291	
	4	0.116948	0.95624	0.926586	0.889975	
	5	0.013555	0.995399	0.97719	0.967462	
CAiTNet	1	0.002974	0.999617	0.981407	0.973953	**0.9873108**
	2	0.001046	0.999712	0.988499	0.983126	
	3	0.004492	0.998993	0.988116	0.977851	
	4	0.000931	0.999569	0.989266	0.978501	
	5	0.000439	0.999664	0.989266	0.983871	
ViTNet	1	0.000885	0.999808	0.992141	0.989653	**0.9924862**
	2	0.000567	0.999856	0.992141	0.988295	
	3	0.001646	0.999712	0.990416	0.983298	
	4	0.000606	0.999856	0.993483	0.987726	
	5	0.000449	0.999952	0.99425	0.988553	
DeiTNet	1	0.002418	0.999377	0.983899	0.977618	**0.9765**
	2	0.02344	0.993434	0.970098	0.946267	
	3	0.003358	0.999425	0.986007	0.978339	

	4	0.007856	0.996885	0.97834	0.967509	
	5	0.02124	0.993817	0.964156	0.946163	
SwinNet	1	0.005455	0.998083	0.987157	0.98223	
	2	0.005398	0.998322	0.986007	0.979366	
	3	0.011216	0.99583	0.980449	0.96823	**0.9843206**
	4	0.009326	0.997364	0.982749	0.979266	
	5	0.00346	0.998898	0.985241	0.974935	
Ensemble	1	0.001756	0.9788	0.992973	0.953791	
	2	0.00345	0.9779	0.986007	0.979366	
	3	0.004576	0.9867	0.996423	0.96823	**0.99267**
	4	0.00474	0.989967	0.992749	0.979266	
	5	0.00475	0.9930475	0.995241	0.984935	

4. Discussion

In this study, we successfully implemented an ensemble of deep learning and transformer models to classify kidney cancer histopathology images, achieving great validation accuracy rates. Our ensemble, which included ViT and Swin models, demonstrated that this ensemble model is capable of detecting critical features from histopathological images. The ensemble model's approach of processing images as grids of patches facilitates effective diagnosis of images, which was crucial in achieving the highest validation accuracy of 99.26% and less training loss as well. These results highlight the potential of combining Vision Transformers and Swin Transformers in digital pathology, offering a significant improvement relative to traditional convolutional neural network models such as ResNet and Vgg 16.

Moreover, our research demonstrates the potential of these advanced models to enhance diagnostic accuracy and efficiency in clinical settings. The consistent performance of Vision Transformers and Swin Transformers across the five kidney cancer types—benign, chromophobe, clear cell, oncocytoma, and papillary—demonstrates their robustness. This could lead to earlier and more accurate detection diagnosis of kidney cancer, improving patient and doctor report outcomes. By taking into consideration the strengths of both Swin Transformers and Vision Transformers, our ensemble approach not only provides a good diagnostic tool but also paves the way for future research in the application of advanced deep learning models in medical image analysis. The successful implementation and high performance of these models suggest a promising direction for integrating AI-based solutions into routine pathological workflows.

4.1. Future Directions

This study highlights important directions for enhancing kidney cancer diagnosis through digital pathology and deep learning. The main areas for improvement include the integration of medical reports and X-ray scans. This paper and study emphasize the importance of AI tools that can detect the accuracy and the type of the image easily just by seeing the image. By harnessing the full potential of AI-driven digital pathology and web tools to detect the type of cancer in the images by uploading the cancer image, this research paves the way for more accurate, efficient, and reliable diagnostic tools using deep learning models in oncology.

5. Conclusion

This study of kidney cancer diagnosis shows the effectiveness of deep learning and ensemble transformer models in classifying kidney cancer histopathology images, with a focus on comparing their performance on metrics such as validation accuracy, validation cohen-cappa score, training loss, and validation loss. Our analysis reveals that the Ensemble model of Vision Transformer and Swin Transformer and Vision Transformers alone, particularly the ViTNet model, excels in identifying critical features from histopathological images, with the highest validation accuracy of 99.26% achieved by the ensemble of the Swin and Vision transformers. The improvement in accuracy across various models signifies the potential of the Ensemble transformer to outperform convolutional neural network models.

The performance of the models trained along 5 different kidney types shows the robustness of those models in clinical applications such as detecting kidney cancer application. Integration of AI can also be made so that the use cases can be extended in other domains as well. By reducing the errors, this application can be used in many different domains by various kinds of people. Our findings show precise use cases where such a study will be very helpful in clinical domains.

Supplementary Materials: The Kidney Cancer dataset is openly available at this Link: https://bmirds.github.io/KidneyCancer/.

Funding: This research received no funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to express our deepest gratitude to the Roux Institute, the IEAI, and the Alfond Foundation for their invaluable support and contributions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Enrichetta Corgna, Maura Betti, Gemma Gatta, Fausto Roila, Pieter H.M. De Mulder, Renal cancer, Critical Reviews in Oncology/Hematology, Volume 64, Issue 3, 2007, Pages 247-262, ISSN 1040-8428, https://doi.org/10.1016/j.critrevonc.2007.04.007.
2. J. Eble, G. Sauter, J. Epstein, al. et (Eds.), World Health Organization classification of tumors. Pathology and genetics of tumors of the urinary system and male genital organs, IARC Press, Lyon (2004), pp. 23-25.
3. Holger Moch, An overview of renal cell cancer: Pathology and genetics, Seminars in Cancer Biology, Volume 23, Issue 1, 2013, Pages 3-9, ISSN 1044-579X, https://doi.org/10.1016/j.semcancer.2012.06.006.
4. H. Moch, T. Gasser, M.B. Amin, J. Torhorst, G. Sauter, M.J. Mihatsch. Prognostic utility of the recently recommended histologic classification and revised TNM staging system of renal cell carcinoma: a Swiss experience with 588 tumors. Cancer, 89 (2000), pp. 604-614.
5. Amin, Mahul B. M.D.; Corless, Christopher L. M.D., Ph.D.; Renshaw, Andrew A. M.D.; Tickoo, Satish K. M.D.; Kubus, James M.S.; Schultz, Daniel S. M.D.. Papillary (Chromophil) Renal Cell Carcinoma: Histomorphologic Characteristics and Evaluation of Conventional Pathologic Prognostic Parameters in 62 Cases. The American Journal of Surgical Pathology 21(6):p 621-635, June 1997.
6. Thoenes, W., Störkel, S. & Rumpelt, H.J. Human chromophobe cell renal carcinoma. *Virchows Archiv B Cell Pathol* 48, 207–217 (1985). https://doi.org/10.1007/BF02890129
7. W. Thoenes, S. Störkel, H. Rumpelt, R. Moll, H. Baum, S. Werner. Chromophobe cell renal carcinoma and its variants report on 32 cases. Journal of Pathology, 155 (1988), pp. 277-287.
8. Amin, Mahul B.; Crotty, Thomas B.; Tickoo, Satish K.; Farrow, George M.. Renal Oncocytoma: A Reappraisal of Morphologic Features with Clinicopathologic Findings in 80 Cases. The American Journal of Surgical Pathology 21(1):p 1-12, January 1997.
9. Luciani LG, Cestari R, Tallarigo C. Incidental renal cell carcinoma-age and stage characterization and clinical implications: the study of 1092 patients (1982-1997). Urology. 2000 Jul; 56.
10. Chen SC, Kuo PL. Bone Metastasis from Renal Cell Carcinoma. Int J Mol Sci. 2016 Jun 22;17(6) [PMC free article] [PubMed]
11. Pandey J, Syed W. Renal Cancer. [Updated 2023 Aug 8]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK558975/
12. Leilei Zhou, Zuoheng Zhang, Yu-Chen Chen, Zhen-Yu Zhao, Xin-Dao Yin, Hong-Bing Jiang, A Deep Learning-Based Radiomics Model for Differentiating Benign and Malignant Renal Tumors, Translational Oncology, Volume 12, Issue 2, 2019, Pages 292-300, ISSN 1936-5233, https://doi.org/10.1016/j.tranon.2018.10.012.
13. M. Shehata et al., "A New Computer-Aided Diagnostic (Cad) System For Precise Identification Of Renal Tumors," 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 2021, pp. 1378-1381, doi: 10.1109/ISBI48211.2021.9433865. keywords: {Image segmentation;Three-dimensional displays;Sensitivity;Malignant tumors;Tools;Feature extraction;Reliability;Renal Cancer;CE-CT;CAD}.
14. Esteva, A., Kuprel, B., Novoa, R. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017). https://doi.org/10.1038/nature21056

15. Agrawal, R.K., Juneja, A. (2019). Deep Learning Models for Medical Image Analysis: Challenges and Future Directions. In: Madria, S., Fournier-Viger, P., Chaudhary, S., Reddy, P. (eds) Big Data Analytics. BDA 2019. Lecture Notes in Computer Science(), vol 11932. Springer, Cham. https://doi.org/10.1007/978-3-030-37188-3_2
16. Shuai Jiang, Liesbeth Hondelink, Arief A. Suriawinata, Saeed Hassanpour, Masked pre-training of transformers for histology image analysis, Journal of Pathology Informatics, Volume 15, 2024, 100386, ISSN 2153-3539, https://doi.org/10.1016/j.jpi.2024.100386.
17. Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, Mahadev Satyanarayanan, OpenSlide: A vendor-neutral software foundation for digital pathology, Journal of Pathology Informatics, Volume 4, Issue 1, 2013, 27, ISSN 2153-3539,https://doi.org/10.4103/2153-3539.119005.
18. Deep Residual Learning for Image Recognition, Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778
19. K. Han et al., "A Survey on Vision Transformer," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 87-110, 1 Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.keywords: {Transformers;Task analysis;Encoding;Computer vision;Computational modeling;Visualization;Object detection;Computer vision;high-level vision;low-level vision;self-attention;transformer;video}.
20. Liu, Ze & Lin, Yutong & Cao, Yue & Hu, Han & Wei, Yixuan & Zhang, Zheng & Lin, Stephen & Guo, Baining. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 9992-10002. 10.1109/ICCV48922.2021.00986.
21. Singh A, Randive S, Breggia A, Ahmad B, Christman R, Amal S. Enhancing Prostate Cancer Diagnosis with a Novel Artificial Intelligence-Based Web Application: Synergizing Deep Learning Models, Multimodal Data, and Insights from Usability Study with Pathologists. Cancers (Basel). 2023 Nov 30;15(23):5659. doi: 10.3390/cancers15235659. PMID: 38067363; PMCID: PMC10705310.
22. Ivanova E, Fayzullin A, Grinin V, Ermilov D, Arutyunyan A, Timashev P, Shekhter A. Empowering Renal Cancer Management with AI and Digital Pathology: Pathology, Diagnostics and Prognosis. Biomedicines. 2023 Oct 24;11(11):2875. doi: 10.3390/biomedicines11112875. PMID: 38001875; PMCID: PMC10669631.
23. W. Liu, J. -L. Qiu, W. -L. Zheng and B. -L. Lu, "Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition," in IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 2, pp. 715-729, June 2022, doi: 10.1109/TCDS.2021.3071170. keywords: {Emotion recognition;Electroencephalography;Robustness;Deep learning;Correlation;Brain modeling;Computational modeling;Bimodal deep autoencoder (BDAE);deep canonical correlation analysis (DCCA);electroencephalography (EEG);eye movement;multimodal deep learning;multimodal emotion recognition;robustness}.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.