ECNNXAI: Ensembled CNNs with eXplainable Artificial Intelligence for Colon Histopathology Image Classification

Juwaria Qadri
Department of Computer Science
Birla Institute of Technology and Science
Pilani, Dubai Campus
Dubai, United Arab Emirates
f20200043@dubai.bits-pilani.ac.in

J. Angel Arul Jothi
Department of Computer Science
Birla Institute of Technology and Science
Pilani, Dubai Campus
Dubai, United Arab Emirates
angeljothi@dubai.bits-pilani.ac.in

ABSTRACT

Colon cancer is ranked as the third most commonly diagnosed cancer and second for causing the most cancer related deaths. Histopathology is a crucial diagnostic tool for cancer since it enables the microscopic analysis of tissue samples to pinpoint abnormal cells, to identify the stage of the cancer and its kind. There is a significant need for precise detection and diagnosis from histopathology images. This research proposes a stacking ensemble model called Ensembled Convolutional Neural Networks with eXplainable Artificial Intelligence (ECNNXAI) for multiclass colon histopathology image classification. Our ensemble model consists of three pre-trained convolutional neural networks (XceptionNet, DenseNet-121 and InceptionNetV3) as base classifiers and the logistic regression as the meta classifier. Gradient-weighted Class Activation Mapping (Grad-CAM) visualization technique is used to interpret and understand the regions focused by the base classifiers to arrive at the final predictions. SHapley Additive exPlanations (SHAP) is used for understanding the predictions made by the ECNNXAI. The proposed model achieves the best overall performance with accuracy of 72.83%, precision of 77.78%, recall of 66.52% and F1 score of 71.71% on the Chaoyang dataset.

Keywords

Histopathology, colon cancer, stacking ensemble, explainable artificial intelligence, Grad-CAM, SHAP.

1 INTRODUCTION

Colon cancer is a malignant growth in the colon or rectum. Globally, colon cancer is ranked as the third most diagnosed cancer and the second leading cause of cancer-related deaths by the World Health Organization, with over 1.9 million new cases and nearly 935,000 deaths annually [WHO23]. Age, family history, genetic conditions and unhealthy lifestyle choices are a few factors that could raise the risk of colorectal cancer. Physical examination, magnetic resonance imaging, abdominal ultrasound, colonoscopy, tissue sample collection for histopathology examination are a few methods to detect colon cancer. Timely detection enhances treatment success, lowers mortality, and enables less invasive interventions, improving patient outcomes.

Accurate histopathology examination is vital in diagnosing cancer. Usually pathologists use a microscope to examine tissues and cells to make diagnoses. It helps identify tumor type, stage, and grade, guiding treatment and predicting outcomes. However, manually analyzing complex-natured histopathology images takes a lot of time, is laborious, challenging and may be error prone [Ham20]. Additionally, there is some subjectivity in the criteria used by different pathologists to identify and classify these images. This subjectivity makes it possible for different pathologists to interpret the same set of images differently. Computer assisted diagnosis can play a significant role in assisting pathologists in examining histopathology images.

Digital pathology (DP) [Sen22] is the process of digitizing histopathology slides to produce high resolution images. Due to the advent of DP, computer assisted diagnosis systems employing deep learning models are recently used to segment images, identify objects and detect diseases from histopathology images. Convolutional Neural Networks (CNNs), are deep learning models that are used in numerous computer vision (CV) applications, specifically those which process image data. They exhibit remarkable automatic feature extraction ability requiring minimal pre-processing. CNNs are a popular choice for medical image analysis

https://www.doi.org/10.24132/CSRN.3401.29

applications such as segmentation, classification and anomaly detection.

Stacking ensemble is a technique used to build a strong classification model by combining multiple individual classifiers called base models. Predictions are obtained by training various base models and then a meta-model is constructed to produce the final output. In a stacked ensemble technique, the meta model is fed with the predictions of the base models from the preceding level [Pav18]. Using an ensemble improves accuracy and reduces generalization error limiting the impact of error causing factors like noise, bias and variance.

It is crucial to understand how a model arrives at its final prediction especially in domains like healthcare. Machine learning models, especially deep architectures, are regarded as black boxes because the way in which these models arrive at their final predictions is not explicit. This is attributed to the complex architecture of these models and the difficulty in understanding their internal working. EXplainable Artificial Intelligence (XAI) is a field of research that adds transparency to the working of the models by explaining, visualizing and interpreting the results. Today, image visualization techniques like Gradient weighted Class Activation Mapping (Grad-CAM) [Sel17] and SHapley Additive exPlanations (SHAP) [Lun17] are used to generate visual explanations for any CNN-based model.

In this study, a novel model called Ensembled CNNs with eXplainable Artificial Intelligence (ECNNXAI) is proposed, which is able to accurately classify colon histopathology images as well as visually explain decisions taken at every level of the stacked ensemble. ECNNXAI contains three pre-trained CNN models as its base classifier and a machine learning model as the meta classifier. Explainable AI (XAI) techniques like Grad-CAM and SHAP are used to provide model interpretability. The following are the contributions of this paper: (1) This study performs multiclass image classification on colon histopathology images. (2) A novel approach called ECNNXAI is put forth with the aim to enhance overall classification accuracy by using stacked ensemble techniques to combine three individually trained CNN models. (3) Visual explanations are produced using XAI techniques like Grad-CAM and SHAP.

The structure of this paper is as follows: Section 2 elaborates on the previous works, Section 3 describes the dataset. Section 4 discusses our proposed model. Section 5 provides the implementation details and evaluation metrics. Section 6 describes the results and discussions. Section 7 provides conclusions and suggests future works.

2 LITERATURE REVIEW

This section elaborates the previous work done on colon histopathology images using various techniques. Zhu et al. [Zhu21] developed an Easy/Hard/Noisy (EHN) image detection model accompanied by a CNN classification model. The EHN model utilized the sample training history to separate the useful hard samples from the detrimental noisy data. It was then incorporated into a self-training algorithm to gradually correct label errors and reduce noise rate. A Noise Suppressing and Hard Enhancing (NSHE) strategy was also suggested to train the noise robust model using the generated almost clean dataset. ResNet-32 was used as the backbone of the classification model for the Chaoyang dataset. Kadian et al. [Kad23] used the pipeline model proposed in [Zhu21] and replaced the backbones with different models like ResNet-34, Cross-Covariance Image Transformer (XCiT), SqueezeNet, and MobileNet. These models were integrated individually with this architecture that incorporated data cleaning. A two-phased architecture was utilized where Phase I generated a dataset that was almost clean through label correction, and the Phase 2 utilized the dataset generated for obtaining a classification model that was robust. It was found that the MobileNet model performed the best on the Chaoyang dataset.

Tepe and Bilgin [Tep22] used Graph Neural Networks (GNNs) to classify the tissue types from the Chaoyang dataset. The construction of a super-pixel graph from an image was the first step in this process that was followed by the application of the GNN models to the constructed graph. The study experimented with the Graph Convolution Network (GCN), Graph Isomorphism Network (GIN), and the Graph Attention Network (GAT) models. Out of the models experimented, the GIN model performed the best.

For computer vision related tasks, the Vision Transformer (ViT) is becoming popular, however pure ViT models do not work well on small datasets. The work by Li et al. [Li22] suggested locality guidance for enhancing the performance of ViT on small datasets. This approach involved using a lightweight ResNet-56 that was trained on the exact same dataset on which the ViT was trained. The local information extracted by the CNN was then combined with the global information extracted from the ViT. This approach enabled the ViT to learn and use both local and global information for the classification of the Chaoyang dataset.

Three primary architectures were used by Nergiz [Ner22] to benchmark the ResNet-18 model for Chaoyang dataset classification: Single Learning (SL), Centralized Learning (CL), and Federated Learning (FL). The traditional FL failed to converge the models on a highly biased dataset to produce good results. As a result, a brand-new Federated Neural Style Transfer

(FNST) technique was put forth that federated the traditional Neural Style Transfer (NST) algorithm and generated synthetic images. The ResNet-18 model was used to test the SL, CL, and FL architectures. The synthetic images produced by the proposed FNST method were also utilized to compare with pure FL findings. The results demonstrated that medical institutions, particularly those that specialize in treating uncommon diseases or medical problems, can effectively apply the FNST algorithm .

Zeid et al. [Zei21] proposed the Compact Convolutional Transformers (CCT) model for ColoRectal Cancer (CRC) tissue classification. The CCT used a convolutional based patching technique that preserves local information and was capable of encoding relationships between the patches. The images were passed to a convolutional layer before passing it to the transformer encoder. This way the CCT was able to combine the advantages of both CNN and transformers. The CCT model outperformed the ViT model. Albashish [Alb22] proposed an ensemble model that was built on four pre-trained CNN models namely DenseNet-121, MobileNetV2, InceptionV3 and the VGG-16 model for classifying colon histopathology images. A blockwise fine-tuning approach was used and additional drop out and dense layers were incorporated to improve the colon image analysis. The ensemble learning methods used were majority voting and product rule. The model that used the product rule achieved better performance as compared to the model that used majority voting.

3 DATASET DESCRIPTION

The Chaoyang dataset introduced by Zhu et al. [Zhu21] is a publicly available dataset which contains images of colon slides scanned at ×20 magnification and collected from the Chaoyang hospital. The images are of size 512×512 and in jpg format. All images in the dataset belong to one of the four categories namely normal, serrated, adenocarcinoma and adenoma. Figure 1 shows images belonging to each of the four classes. Three qualified pathologists collectively assigned labels to the images. The images on which all the three pathologists unanimously agreed on are added to the test set. Rest of the images are added to the training set with labels suggested by one of the pathologists randomly selected. The achieved training set consists of 4021 images and the test set consists of 2139 images in total. Figure 2 depicts the distribution of the four classes in both the training and test dataset.

4 METHODOLOGY

The proposed ECNNXAI is shown in Fig.3. Initially, the images from the dataset are pre-processed. The preprocessed images are then fed to the stacked ensemble model consisting of base and meta classifiers. The XceptionNet, InceptionV3 and DenseNet-121 models are used as the base classifiers and logistic regression model is used as the meta classifier. The meta classifier generates the final predicted class label. By combining predictions from multiple different base models, the strengths of one model can compensate for the weaknesses of another, resulting in a more robust and less biased overall prediction. Grad-CAM is used for visualizing and interpreting the predictions made by the base classifiers. SHAP offers visual explanations for the ensemble model and helps to understand how each base classifier influences the final label predicted.

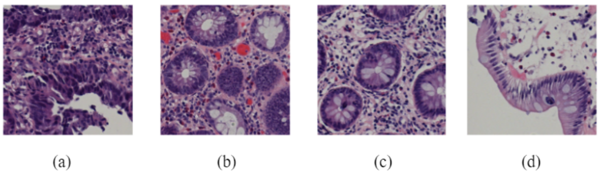

Figure 1: Sample images from the dataset. (a) Adenocarcinoma; (b) Adenoma; (c) Normal; (d) Serrated

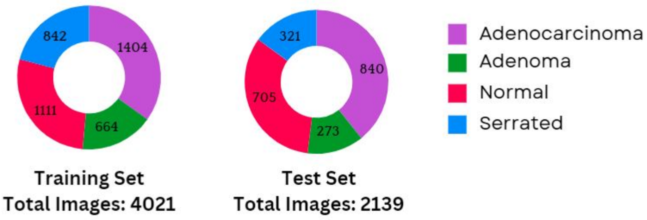

Figure 2: Dataset Distribution

4.1 Data Pre-processing

The training dataset exhibits a visible class imbalance as seen in Fig.2 which can affect the performance of the model. Hence, in order to prevent the model from overfitting and improve its generalization capability the training dataset is enhanced by augmentation techniques such as right-angled rotations and vertical and horizontal flip. These methods are applied to generate new samples for all classes and to increase the number of samples in each class to 1500. All images are then resized to 256×256. Unique numerical identifiers from 0 to 3 are assigned to the class labels adenocarcinoma, adenoma, normal and serrated respectively.

4.2 Base Classifiers

This work uses three pre-trained CNNs namely XceptionNet [Sze16], InceptionV3 [Cho17] and DenseNet-121 [Hua17] as the base classifiers. Pre-trained CNNs are models that are designed and trained for one purpose but can be retrained with little effort on another dataset for a closely related task. CNNs are made by stacking convolution layer, pooling layer, flatten layer, fully connected layer and an output layer. A CNN can

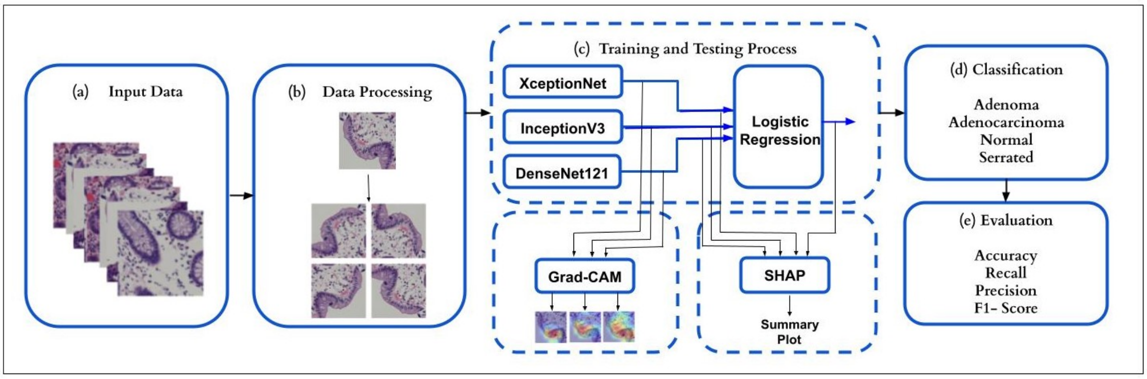

Figure 3: Structure of ECNNXAI

have multiple convolution and pooling layers making the architecture deep. The convolution layers are the key feature extractors of a CNN. The pooling layers are tasked to reduce the number of parameters thereby reducing complexity and improving efficiency. The output of the final pooling layer will be a set of 2-D feature maps. The flatten layer converts the 2-D feature maps into a 1-D feature vector. The fully connected layer is the classification layer that classifies the input image based on the features extracted from the previous layers. The arrangement of the nodes in a fully connected layer is such that all nodes of one layer are connected to each node of the next layer. Finally, the output layer provides the results of classification. The base classifiers in this work are chosen experimentally as explained in Section 6.1.

4.2.1 InceptionV3

Designed by Google, InceptionV3 [Cho17] is a CNN particularly developed for classification of images, and is part of the Inception family. It is 48 layers deep and consists of 23,626,728 parameters. It consists of multiple layers of convolution having various sizes of filters (1×1, 3×3, 5×5) as well as pooling layers for extracting the hierarchical features of the input image. Various techniques are utilized in the InceptionV3 for improving accuracy and efficiency. It enables feature capture at multiple scales using its parallel convolutions having different sizes of filters. The use of 1×1 convolutions for factoring larger convolutions help reduce the cost of computation. Factorized convolutions as well as parallel operations help maintain optimal balance between different cell regions in the colonoscopy images.

4.2.2 XceptionNet

XceptionNet [Sze16], also developed by Google, is an extension of the Inception architecture leveraging separable depth-wise convolutions for performance and computation enhancement. A series of normal convolution layers, depth-wise separable convolutions, pooling layers and residual connections make up its 71 layers deep architecture. Depth-wise separable convolutions enable factorizing of the standard convolution to two separate operations, that are the depth-wise convolutions and pointwise convolutions. The number of parameters are reduced through this separation as well as there is reduction in the complexity of computation, along with allowing more learning of the discriminative features. Despite having almost equal number of parameters as of InceptionV3, the XceptionNet architecture is able to perform better because of the capacity increase due the effective use of the parameters of the model.

4.2.3 DenseNet-121

Densely connected convolutional networks, or DenseNet [Hua17] for short, presents a novel design with the motive of enhancing feature propagation, encouraging feature reuse, and overcoming the vanishing-gradient issue. It is a deep pre-trained CNN model consisting of 121 layers in totals. A unique feature of DenseNet is its dense connectivity where every layer transfers its feature maps to every layer that comes after it and receives feature maps from every layer that comes before it. The network is also made up of numerous layers that are tightly connected within each of its dense blocks. Feature maps from earlier layers are concatenated within a dense block and then passed on to later layers. Reusing features helps improve gradient flow, promote feature propagation, and facilitates the acquisition of more discriminating features. The DenseNet uses batch normalization and ReLU activations as well. Finally, a global average pooling layer is employed followed by a fully connected layer for classification.

4.3 Meta Classifier

The logistic regression (LR) model [Pen02] is used as the meta classifier. It uses a statistical approach. It models the probability of a given input belonging to a certain class through the use of the sigmoid (logistic) function. Logistic regression algorithm is employed often for data having linear relationship between features and the target variable or for datasets that are relatively simple. Results obtained through logistic regression are interpretable, enabling the understanding of the influence of the individual feature on the final prediction. Low cost of computation of logistic regression makes it suitable for cases where the datasets are large and computational resources are few in comparison to models that are complex.

4.4 Explainability

In this work, explainability is incorporated for understanding the working of the base classifiers and the meta classifier using Grad-CAM and SHAP.

4.4.1 Gradient-weighted Class Activation Mapping (Grad-CAM) for base classifiers

The decisions of deep CNNs for the tasks of image classification are visualized and understood through the Grad-CAM. Grad-CAM helps increase the transparency of the model by indicating the features of the image that are most crucial for determining the class label by the model. It works as follows: The input image is passed through the CNN and the target class score gradients are computed using the feature maps of the final convolution layer. Rich spatial information is lost in the fully connected layer, therefore the final convolution layer offers high-level semantics and detailed spatial information. Global average pooling of the gradients is performed to identify the importance of every feature map based on the target class. The result of the pooling operation is a heat map that shows the significance of different regions in the input image for the predicted class. The heat map generated is overlapped onto the original input image to indicate the areas focused by the network while the particular class predictions are made. High intensity regions in the heat map indicate greater significance for the selection of the target class than the lower intensity ones. Fig.4. illustrates the working of Grad-CAM.

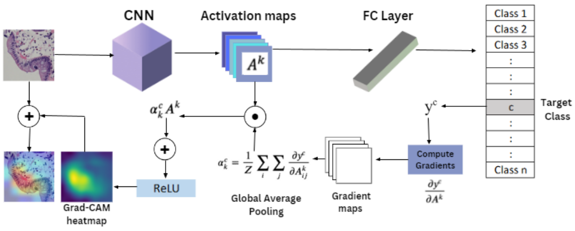

Figure 4: Working of Grad-CAM

4.4.2 SHAP for the level 1 classifiers

Ensemble models are more complex and can be challenging to interpret. SHAP employs cooperative game theory to give each attribute a value that represents how much of an impact it has on the final output. In this work, this property is used to provide explanations on how the contributions from the individual classifiers influence the overall result. This aids in determining the relative significance of every base model in the ensemble.

Following are the steps to calculate the shapley values for the base classifiers in the ensemble: (1) Initially, choose one base classifier as the classifier of interest (*CoI*). (2) List all possible classifier combinations excluding the *CoI*. (3) Compute the ensemble's prediction both with and without the *CoI* for each subset. (4) Determine the marginal contribution (*MC*) of the *CoI* to the final prediction by subtracting the above two values. (5) Compute the average *MC* of the *CoI* (*AvgMC_CoI*) to all possible subsets as the mean of the *MC*s of the *CoI* over all possible subsets. The above steps 1-5 are repeated for each base classifier in the ensemble. The *AvgMC_CoI* is given by Eq. 1 where *N* is the set of all base classifiers, *s* is the subset of base classifiers that does not include the *CoI*, $f(s)$ is the ensemble prediction for the subset *s*, $f(s \cup i)$ is the ensemble prediction for the subset *s* plus *CoI* and |·| denotes the number of elements in a set.

$$AvgMC\_CoI = \sum_{s \subseteq N|i} \frac{|s|!(|N|-|s|-1)!}{|N|!}[f(s \cup i) - f(s)] \tag{1}$$

4.5 Training the ECNNXAI

In this work, the pre-trained CNNs are accessed from the applications module of the Keras library. The top layers of the CNN models are set to false to build our own classification block. This classification block is common for all the models and it consists of a flatten layer followed by two fully connected layers having 512 and 256 neurons respectively and ReLU activation function. This is then followed by an output layer which consists of 4 neurons representing the four target classes in the colon histopathology dataset with softmax activation function.

The training set is further split in the ratio of 80:20 into the train and validation sets. All the base classifiers are trained using the following hyperparameters: 100 epochs, batch size of 64, Adam optimizer and learning rate of 0.001. A stacking classifier is then created using StackingClasssifier from the sklearn.ensemble module with the InceptionV3, XceptionNet and DenseNet-121 as the base classifiers and Logistic regression as the meta classifier.

In this work, pre-processed images from the training set are the inputs for the base classifier's training. Each base classifier predicts the class label of the input images in the training dataset. The final predictions of the base classifiers are formulated into a meta dataset having 4800 rows, 3 independent features (predictions of the base classifiers), and one target attribute. The actual class label of the input images becomes the target attribute. The LR model is then trained on this data. After training the LR model, the test set of the Chaoyang dataset is fed to the ECNNXAI. The images pass through the three base classifiers which provide their predictions. Finally, the predictions from the base classifiers are fed to the LR model which provides the final predictions by combining the predictions of the base models. This helps improve generalization capacity and prediction performance.

5 IMPLEMENTATION AND EVALUATION

This work is implemented in the Jupyter Notebook environment using Python with Tensorflow version 2.1.0. and Keras version 2.3.1. A V100-PCIE-32GB GPU with Ubuntu operating system was used. Let true positives (TP) be the correct predictions of a class of interest, true negatives (TN) denote the correct negative predictions with respect to a class of interest, false positives (FP) denote the number of predictions where samples of other classes are incorrectly predicted as a class of interest, and false negatives (FN) denote the number of predictions where samples of a class of interest are incorrectly predicted as belonging to other classes. All models are evaluated using four evaluation metrics namely accuracy, recall, precision and F1 score as presented in Table 1.

Metric	**Formula**
Accuracy	(TP+TN)/(TP+FP+TN+FN)
Recall	TP/(TP+FN)
Precision	TP/(TP+FP)
F1 score	(2×Precision×Recall)/(Precision+Recall)

Table 1: Evaluation metrics

6 RESULTS AND DISCUSSIONS

6.1 Experiments for choosing the base classifiers

To choose the base classifiers for the ensemble model, six pre-trained models like VGG-16 [Sim14], ResNet-50 [He16], EfficientNetb0 [Tan19], InceptionV3 [Cho17], XceptionNet [Sze16] and DenseNet-121 [Hua17] are trained and tested using the same hyperparameters (epochs: 100, optimizer: Adam, learning rate: 0.001, batch size: 64) with the Chaoyang dataset. According to Table 2, the DenseNet-121 is the best performing pre-trained model achieving an accuracy of 70.1%, precision of 75.67%, recall of 62.97% and F1 score of 68.74%. The XceptionNet is the second best performing model that achieves an accuracy of 68.56%, precision of 73.87%, recall of 59.89% and F1 score of 66.15%. The InceptionV3 is the third best performing model with an accuracy of 67.25%, precision of 72.86%, recall of 55.64% and F1 score of 63.09%. Thus, the three best performing models: XceptionNet, InceptionV3 and DenseNet-121 are then selected to be the base classifiers for our ECNNXAI.

6.2 Ablation study

Experiments are conducted in order to understand the importance of all base classifiers in the ECNNXAI. This is done by forming all possible 2-subsets of the base classifiers and comparing their results against the ECNNXAI which is a 3 classifier combination. The different base classifier combinations that we experimented are: XceptionNet + InceptionV3 + LR (X+I+LR), XceptionNet + DenseNet-121 + LR (X+D+LR), DenseNet-121 + InceptionV3 + LR (D+I+LR). As seen from Table 3 and Table 4, the ECNNXAI outperformed all base classifier combinations.

In order to assess the importance of the logistic regression meta classifier of the ECNNXAI, it is swapped with other popular machine learning algorithms like the naive-Bayes [Wic21] and the decision tree [Cos23] classifiers to create ensemble-NB and ensemble-DT respectively. As seen from Table 4, the ECNNXAI outperformed the ensemble-NB and ensemble-DT ensemble models.

6.3 Performance of the proposed model

It could be inferred from Table 2, Table 3 and Table 4 that the ECNNXAI comprising of a logistic regression model stacked on top of the InceptionV3, XceptionNet and DenseNet-121 outperformed the individual CNNs and all other combinations of the base classifiers. It could be demonstrated from the results that utilizing different base classifiers with varying strengths and weaknesses helped build a stronger classifier model. This is because the different base classifiers perform well in certain areas of the feature space. The overall robustness of ECNNXAI is enhanced and a wider range of patterns are captured helping it to achieve better performance with an accuracy of 72.83%, precision of 77.78%, recall of 66.52% and F1 score of 71.71%.

6.4 Evaluation of explainability

In this work, the use of Grad-CAM highlights how specific regions from the colonoscopy images are used by the base classifiers in order to arrive at the final predictions thereby enhancing the interpretability. This is crucial to understand the classification decisions made by the base classifiers and enables specialists in the field to understand areas focused by the network for predictions. Figure 5 displays the specific regions in sample images belonging to different classes that contribute to the final decisions made by the base classifiers.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Resnet-50	61.87	63.92	52.3	57.52
VGG-16	62.74	66.42	51.90	58.26
EfficientNetb0	65.21	68.28	53.77	60.16
InceptionV3	67.25	72.86	55.64	63.09
XceptionNet	68.56	73.87	59.89	66.15
DenseNet-121	**70.1**	**75.67**	**62.97**	**68.74**

Table 2: Experimental results for choosing the base classifiers

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
X+I+LR	63.19	68.13	50.33	57.81
D+I+LR	63.85	70.49	**53.05**	**60.53**
X+D+LR	**64.24**	**71.43**	51.85	60.08

Table 3: Experimental results for evaluating the importance of the combinations of the base classifiers. X denotes XceptionNet, I denotes InceptionV3, D denotes DenseNet-121 and LR denotes Logistic Regression.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Ensemble-DT	70.79	76.88	62.91	69.2
Ensemble-NB	72.04	77.73	64.17	69.95
ECNNXAI	**72.83**	**77.78**	**66.52**	**71.71**

Table 4: Experimental results to assess the importance of logistic regression (meta classifier)

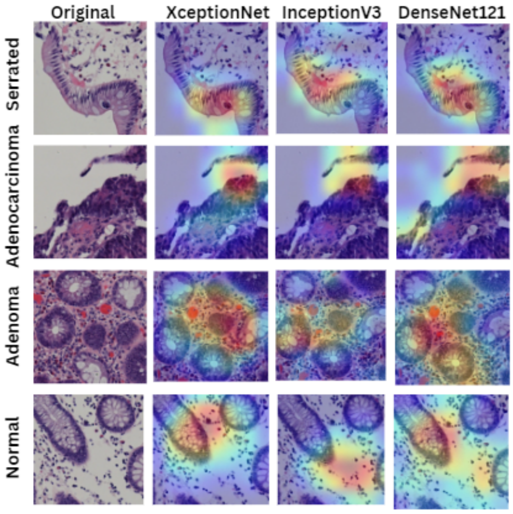

Figure 5: Grad-CAM visualization of sample images belonging to different classes and the regions focused by the base classifiers

Summary plot of SHAP offers a clear and understandable illustration of how each individual base classifier influences the output of the ECNNXAI. It is a horizontal bar chart where every bar represents a base classifier. The impact of each base classifier on the output of the model is shown by the length of the bar. Longer bar in SHAP summary plot indicate that the impact of a classifier is more on the final output generated while shorter bar indicate that the impact of the classifier is lesser. The base classifiers are arranged from highest impact to lowest impact according to how important they are for the model while making the final decision.

Figure 6 illustrates that the DenseNet-121 model has the highest impact on the final prediction of the ECNNXAI, which is followed by the XceptionNet and InceptionV3. From Fig. 6, it is observed that the DenseNet-121 has the highest influence on the predictions of ECNNXAI for class 1 (adenoma) and class 2 (normal) and the XceptionNet has the highest influence on the predictions of ECNNXAI for class 0 (adenocarcinoma) and class 3 (serrated).

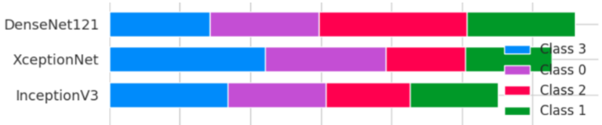

Figure 6: Summary plot showing impact of base classifiers on final predictions of the ECNNXAI

7 CONCLUSION

In this work, we proposed an ensemble network called the ECNNXAI for classifying colon histopathology images into one of the four target classes namely, adenoma, adenocarcinoma, serrated and normal. Three pre-trained CNN models were used as the base classifiers for the stacked ensemble model while the logistic regression model as the meta classifier. Combining the three CNNs using an ensemble model increased the overall performance, the generalization ability to unseen data and the reliability of predictions while reducing the impact of biases as the models errors. Explainable AI techniques like Grad-CAM and SHAP provided interpretation and aided the understanding of the predictions made by the models at various levels. The proposed model can accurately classify colon histopathology images and identify the critical regions in the images that correspond to the cancer types. Future research would examine the optimal number of base classifiers needed in an ensemble model for the classification of images related to colon histopathology and the use of various different deep learning architectures as base classifiers in an ensemble.

8 REFERENCES

[Alb22] Albashish, D. Ensemble of adapted convolutional neural networks (CNN) methods for classifying colon histopathological images. PeerJ Computer Science, 8, p.e1031, 2022.

[Cho17] Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 1251-1258, 2017.

[Cos23] Costa, V.G. and Pedreira, C.E.. Recent advances in decision trees: An updated survey. Artificial Intelligence Review, 56(5), pp.4765-4800, 2023.

[Ham20] Hameed, Z., Zahia, S., Garcia-Zapirain, B., Javier Aguirre, J. and Maria Vanegas, A. Breast cancer histopathology image classification using an ensemble of deep learning models. Sensors, 20(16), p.4373, 2020.

[He16] He, K., Zhang, X., Ren, S. and Sun, J.. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[Hua17] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q.. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708, 2017.

[Jun21] Junayed, M.S., Anjum, N., Noman, A. and Islam, B.. A deep CNN model for skin cancer detection and classification, 2021.

[Kad23] Kadian, V., Singh, A. and Sharma, K.. A Robust Colon Cancer Detection Model Using Deep-Learning. In 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC), pp. 665-670, IEEE, 2023.

[Kad20] Kadhim, M.A. and Abed, M.H.. Convolutional neural network for satellite image classification. Intelligent Information and Database Systems: Recent Developments 11, pp.165-178, 2020.

[Li22] Li, K., Yu, R., Wang, Z., Yuan, L., Song, G. and Chen, J.. Locality guidance for improving vision transformers on tiny datasets. In European Conference on Computer Vision, pp. 110-127, Cham: Springer Nature Switzerland, 2022.

[Lun17] Lundberg, S.M. and Lee, S.I.. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 2017.

[Mhe22] Mhedhbi, M., Mhiri, S. and Ghorbel, F.. A new deep convolutional neural network for 2D contour classification, 2022.

[Ner22] Nergiz, M.. Collaborative Colorectal Cancer Classification on Highly Class Imbalanced Data Setting via Federated Neural Style Transfer Based Data Augmentation. Traitement du Signal, 39(6), 2022.

[Pav18] Pavlyshenko, B., August. Using stacking approaches for machine learning models. IEEE second international conference on data stream mining & processing (DSMP) (pp. 255-258). 2018.

[Pen02] Peng, C.Y.J., Lee, K.L. and Ingersoll, G.M.. An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), pp.3-14, 2002.

[Sel17] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D.. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision pp. 618-626, 2017.

[Sen22] Sengoz, N., Yigit, T., Ozlem, O. and Isik, A.H.. Importance of preprocessing in histopathology image classification using deep convolutional neural network. Advances in Artificial Intelligence Research, 2(1), pp.1-6, 2022.

[Sim14] Simonyan, K. and Zisserman, A.. Very deep convolutional networks for large-scale image recognition, 2014.

[Sze16] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826, 2016,

[Tan19] Tan, M. and Le, Q.. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114), PMLR, 2019.

[Tep22] Tepe, E. and Bilgin, G.. Graph neural networks for colorectal histopathological image classification. In 2022 Medical Technologies Congress (TIPTEKNO), pp. 1-4, IEEE, 2022.

[Wic21] Wickramasinghe, I. and Kalutarage, H.. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing, 25(3), pp.2277-2293, 2021.

[WHO23] World Health Organization. (2023, July 11). Colorectal cancer. Retrieved from https://www.who.int/news-room/factsheets/detail/colorectal-cancer

[Zei21] Zeid, M.A.E., El-Bahnasy, K. and Abo-Youssef, S.E.. Multiclass colorectal cancer histology images classification using vision transformers. In 2021 tenth international conference on intelligent computing and information systems (ICICIS), pp. 224-230, IEEE, 2021.

[Zhu21] Zhu, C., Chen, W., Peng, T., Wang, Y. and Jin, M.. Hard sample aware noise robust learning for histopathology image classification. IEEE transactions on medical imaging, 41(4), pp.881-894, 2021.