Using SHAP for Deep Convolutional Neural Network Debugging and Transferability

Rebecca Tsekanovskiy[1]

Rensselaer Polytechnic Institute NY 12180, USA tsekar@rpi.edu

Abstract. In recent years, the demand for transparency and interpretability in deep learning models, particularly black-box algorithms, has grown significantly. This paper explores the effectiveness of Explainable Artificial Intelligence techniques, specifically SHapley Additive exPlanations, in providing clear insights into model behavior. Our study evaluates the generalization capabilities of a deep convolutional neural network trained on brain scans when tested on lung scans, without retraining. We employed a deep convolutional neural network trained on thousands of medical images and conducted a comprehensive SHapley Additive exPlanations analysis to elucidate the model's predictions. While the model demonstrated high accuracy and low loss on both training and validation datasets, SHapley Additive exPlanations visualizations revealed challenges in cross-domain generalization, particularly in accurately classifying lung scans. The results underscore the importance of domain-specific training for reliable model performance across different types of medical images. This study highlights the critical role of Explainable Artificial Intelligence techniques like SHapley Additive exPlanations in developing interpretable and trustworthy artificial intelligence models, offering practical insights into understanding complex neural network decisions and their implications for clinical applications.

Keywords: Black Box Algorithms · SHapley Additive exPlanations · Explainable Artifical Intelligence · Deep Convolutional Neural Networks · Cross Domain Model Transferability · Cancer

1 Introduction

1.1 Motivation

The rapid advancements in deep learning have led to its widespread adoption across various industries. However, the inherent complexity of deep learning models, particularly deep convolutional neural networks (DCNNs), poses significant challenges in understanding and interpreting their decision-making processes. This opacity is often referred to as the "black box" problem, where the internal workings of the model remain obscure, even to domain experts.

This paper aims to explore the effectiveness of SHapley Additive exPlanations (SHAP) in debugging deep learning neural networks. By comparing traditional debugging methods with SHAP, we demonstrate how SHAP can improve model

interpretability, identify data quality issues, and ultimately enhance model accuracy and reliability. Our study contributes to the growing body of research on XAI by providing practical insights into the application of SHAP in real-world scenarios, paving the way for more transparent and accountable AI models.

1.2 Research Contribution

Despite the significant advancements in deep learning, there remains a substantial challenge in effectively debugging and interpreting these models. Traditional debugging methods for DCNNs often lack transparency, making it difficult to understand why a model makes certain predictions. This is especially problematic in applications requiring high reliability and trust, such as medical image analysis, where incorrect predictions can have severe consequences. We evaluated a DCNN model trained on brain scans and tested it on lung scans without retraining, finding that the model fails to generalize across domains, as evidenced by SHAP analysis. These findings underscore the need for domain-specific training and robust interpretability tools. Our contributions are as follows:

1. We utilized SHapley Additive exPlanations to analyze the transferability of deep learning models trained on brain scans when applied to lung scans, highlighting the model's inability to generalize effectively without domain-specific retraining.
2. We investigated the challenges of cross-domain generalization in deep convolutional neural networks by training a model on brain scans and applying it to lung scans, providing clear evidence of the model's limitations in adapting to different types of medical data.
3. Our research underscores the critical need for developing domain-specific models to enhance the reliability and interpretability of artificial intelligence in medical applications, ultimately contributing to safer and more effective clinical decision-making.

1.3 Specific Focus and Relevance

This research specifically focuses on the transferability of features learned by a DCNN model who only trained on brain scans, focusing on glioma, pituitary, menginoma, and no-cancer when testing the model on an entirely different dataset, like lung scans with indications of benign and malignant tummors. By using SHAP, we aim to provide a trasparent understanding of how features transfer across diferent types of cancers. This is important because of the following:

1. Understanding features that are specific to a certain type of cancer can help with fine tuning models focusing on a specific disease.
2. Understanding features that are universal to cancer can help with developing general models.
3. Enhancing a models adaptability to different types of data, specifically in healthcare where variability in data occurs.

1.4 Deep learning

Deep learning (DL), a type of machine learning (ML), has changed various industries by causing models to learn and make decisions without the need for humans. This consists of using neural networks, which is made up of multiple layers of interconnected nodes, helping to understand and make sense of the data. [7] This is done in a way that simulates the functionality of neurons, ultimately contributing to the concept of neural networks [4]. These neural networks are created in a way similar to how a human brain process information[4]. The average DL model consists of different layers: input layer, multiple hidden layers, and an output layer. Data will go through each layer and have information extracted from it and passed onto the next layer [7].

DL models can be classified into four categories: deep supervised, unsupervised, reinforcement learning, and hybrid models [7]. Deep supervised learning models show their strength in tasks such as natural language processing, image recognition, and predictive analytics, whereas a regular algorithm falls short due to their inability to handle unstructured and large datasets[16, 7]. In the aspect of our project, we will be focusing on deep convolution neural network, a type of deep learning used for image recognition [17].

Deep Neural Networks (DNNs), a subcategory of deep supervised learning, are only becoming more intricate [7]. With the large numbers of layers, their decision making processes make it difficult to understand how and why specific predictions were made. Despite this, DNNs outperform that of a traditional ML model, yielding results with high accuracy[4].

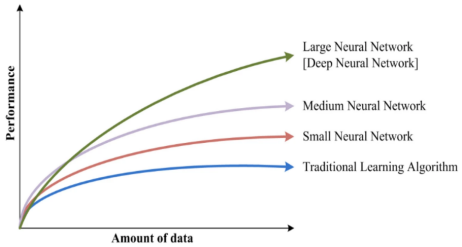


Fig. 1: Amount of data vs. Performance [4]

Black Box Algorithms Deep learning models, specifically DNNs, are sometimes mentioned as a "black box" algorithms because of their complicated and hidden decision making processes [15]. In other words, it is unclear what information causes a model to come to a certain conclusion. With this, many applications view this as a huge negative. This is causing a strain on a majority of users wanting to trust and adopt these models, like in stake areas like healthcare

and finance [7, 15]. In some scenarios, the reasoning behind a certain conclusion may not be critical, but in cases like self-driving cars, understanding it is crucial due to the potential impact on human lives. [15]. Today, these models are being trained with millions of different datasets, causing the model to observe certain patterns that a human may not have been able to identify [15]. Thus, there has been a growing need to understand black box algorithms, creating the demand for explainable Artificial Intelligence (XAI).

Explainable Artificial Intelligence and SHapley Additive exPlanations Explainable Artificial Intelligence techniques have risen to help combat the transparency issues with black box algorithms. The goal of XAI is to ensure there is interpretabilty privacy, robustness, trust, fairness, and transparency. SHAP is a type of XAI technique based off a cooperative game. SHAP helps a user understand the contributions of different features to a model's predictions by quantifying all aspects [6]. SHAP is also possible of both global and local explainability[8]. This means it can provide detailed insights into individual predictions as well as an understanding of the overall behavior of the model across the entire dataset.[8] This dual capability makes SHAP a powerful tool for enhancing model interpretability and making sure AI systems are both transparent and accountable.

By using SHAP to help explain deep learning models, we can better understand the underlying framework of these systems, identify data quality issues, and improve model accuracy. This study explores the effectiveness of SHAP in debugging deep learning neural networks, paving a new path for the future by developing more transparent AI models.

2 Literature Review

This section reviews various methodologies and approaches to interpreting and explaining deep learning models. We focus on sensitivity analysis, layer-wise relevance propagation, Locally-interpretable model-agnostic explanations, and SHapley Additive exPlanations. Additionally, we explore the application and evaluation of SHAP in different domains.

2.1 Interpreting and Explaining Deep Learning Models

There are many different methods in place to help understand a deep learning model. This includes sensitivity analysis, layer-wise relevance propagation, SHapley Additive exPlanations, Locally-interpretable model-agnostic explanation, and more [5, 15].

Sensitivity analysis (SA) refers to explaining a models prediction from the locally evaluated gradient [15]. For each input variable, an image pixel, sensitivity analysis quantifies each input. Here, it is mathematically represented as [15]

$$R_i = ||\frac{\partial}{\partial x_i} f(x)||. \tag{1}$$

The system will first identify the input image and classify it. Then, a heatmap will be able to be visualized, demonstrating each pixels worth towards the prediction [15]. It is identifying the pixels needed to be altered to make the image resemble more of the prediction. A change in these pixels would ultimately impact classification score [15]. With this, there are drawbacks, such as SA in explainability often focuses on irrelevant features, leading to misleading heatmaps and poor quantitative performance in accurately identifying key predictive features [15].

Layer-Wise Relevance Propagation (LRP) can also help explain predictions made by models by showing which specific input is crucial for the prediction[15]. LRP can provide a detailed breakdown on how each part of the input impacts the output. LRP can also be used with different AI models, not specifically just deep neural networks [15]. The effectiveness of LRP can be sensitive to the chosen parameters. Incorrect parameter settings can affect the accuracy and reliability of the relevance scores [15]. Moreso, LRP can be computationally expensive, especially in cases when the model is large [15].

Locally-interpretable model-agnostic explanations (LIME) is used as an XAI technique to help simplify the interpretation process by approximating the model with a linear model [13, 18]. In other words, the goal of LIME is to represent the behavior of a complex model, such as the black box model, in the realm of a specific data instance [13] LIME can be applied to any machine learning model because it is model-agnostic. LIME can provide information on why the model came to a certain prediction, more specifically on the local level[18]. LIME does not try to approximate the black box model on the global level because it may not always be feasible due to the complexity[13].

SHAP has stemmed from a cooperative game theory, used to determine the impact certain features have on a model's prediction, and it is not impacted by complexity or structure of the model. This is because each feature is assessed based off all possible feature combinations. SHAP contains many different great qualities, ultimately contributing to the popularity. SHAP can be applied to decision trees, neural networks, linear models, and more [6]. Since it can evaluate both on the global and local level, unlike most other popular XAI tools, it makes it extremely versatile. The usage of SHAP values have been used for helping transparency and improving model debugging [6].

SHAP calculates the average output of the model when only specific features are identified. This can help understand what the model predicts when information is limited [6]. SHAP also calculates the contribution of an individual feature by focusing on the change of the model's prediction, specifically in cases where a new feature is added[6]. Ultimately, this is demonstrating how much weight a certain feature carries in relation to a combined subset of other features. SHAP will also calculate the value for a feature by averaging the way it is contributing across all possible combinations to order the features, making sure that the SHAP value accounts for every possible variation providing an accurate measurement[6]. Moreso, if a complex model is made from several simpler models, the SHAP value for a feature in the complex model is the sum of the SHAP

values for that feature in each simpler model, adjusted by their weights. It must be noted that SHAP explanations are calculated in #P-hard, meaning it is extremely computationally intense and can not always compute SHAP values in polynomial time [6]. Because of this, the practical application of SHAP can be limited for large models.

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) identify relevant features in an image without any form of supervision and are widely used in areas such as computer vision and facial recognition [2]. Using a CNN architecture instead of standard neural networks is beneficial because it reduces the number of trainable parameters, which helps mitigate overfitting. Additionally, concurrent learning occurs during the classification layer, allowing the model to better understand and learn the extracted features [2]. Furthermore, implementing a CNN is generally easier compared to other types of networks.

CNNs are composed of five main layers: input, convolutional, pooling, activation function, and fully connected layers [12]. As in other model architectures, the input layer consists of the pixel values from the selected images. The convolutional layer is a fundamental part of the architecture, using kernels to learn and understand spatial information in the data [12]. This layer is crucial for generating the output feature map. Following the convolutional operations, the pooling layer reduces the size of the feature map, which is essential for decreasing the number of parameters and preventing overfitting.

The activation layer can employ various functions such as sigmoid, tanh, ReLU, Leaky ReLU, Noisy ReLU, and more. This layer introduces non-linearity to the model, enabling it to learn and understand more complex features [2]. Finally, the fully connected layer, which is the last layer of the architecture, connects each neuron to every neuron in the previous layer. The output of the fully connected layer represents the final prediction of the model. A CNN model ultimately becomes deep when more layers are involved in the model architecture.

DCNNs have become crucial in high performing image classification. Training these models can become quite complicated due to the complexity of the architecture [2]. To help deal with these issues, transfer learning has emerged where prior models that were trained on very large datasets can be fined tune and focused on a specific task[2]. An example of this is ImageNet. Moreso, the accuracy of a DCNN model can be on the lower end if the dataset size is either not large enough or if the complexity of the model is not deep enough. As such, models with less layers tend to have lower accuracy.

2.3 Inception V3

Inception-V3 is a deep convolutional neural network known for its complexity and robust performance, though it presents significant challenges during training due to its extensive architecture, often requiring considerable computational resources and time. To overcome these challenges, Inception-V3 is designed with

computational efficiency in mind, utilizing a factorization approach that breaks down convolutions into smaller operations [14]. This design not only reduces resource demands but also enhances the model's ability to process data more effectively [14].

Additionally, to make Inception-V3 more adaptable to specialized tasks, transfer learning is often employed. By fine-tuning the final layers while retaining the knowledge from earlier layers, researchers can significantly reduce training time and computational requirements, making it feasible to apply Inception-V3 to tasks such as medical image analysis [10]. This method allows the model to leverage its pre-trained features effectively, ensuring strong performance even with limited data, while maintaining high accuracy and reliability in predictions. The model's ability to apply multiple filters to the same input and concatenate the outputs further enhances its capacity to capture both cross-channel and spatial correlations, allowing it to analyze data from different perspectives simultaneously and recognize complex patterns with greater accuracy [14].

2.4 Application and Evaluation of SHAP in Various Domains

SHAP has been implemented in multiple different domains, demonstrating its effectiveness in various aspects.

Healthcare SHAP has been applied in the healthcare sector, ultimately aiding in combining transparency and AI. In one aspect, the authors of [9] implemented eXtreme Gradient Boosting (XGBoost), a type of decision tree, with SHAP to analyze predictions for heart failure stages. Using SHAP, researchers were able to view how clinical features impacted the model's prediction. For instance, the study identified gender, BMI, and blood pressure as significant predictors of heart failure stages, validating the model's logic and highlighting areas for potential refinement [9]. Moreover, it was stated that "in the SHAP interpretation for our prediction model, heart failure patients tend to have EF scores higher than average if the BMI value is high and vice versa" [9]. Using SHAP has helped better debugging, improving the model's overall accuracy and reliability.

By using SHAP, the researchers were able to find information that they would not have been able to recognize by just analyzing the information given by XGBoost [9] . Additionally, the article highlights how SHAP can be used to detect and mitigate biases in deep neural networks. The researchers identified gender differences in the model's predictions, which could indicate potential biases in how the model processes clinical data [9] . The researchers conclude they believe it is possible for future use of machine learning models in the health sector to aid with clinical trials [9] .

3 Methods and Data Overview

We began our study by getting a dataset comprising over 10,000 images of abdomen and breast scans, each with a resolution of 66x66 pixels. The dataset was

divided into training, validation, and test sets in a 70:20:10 ratio. We utilized this dataset to train a DCNN to evaluate its ability to accurately recognize and classify these distinct types of medical images. The dataset was sourced from [3].

This initial phase served as a proof of concept to ensure that the model could effectively distinguish between two vastly different types of scans. Following this, we expanded our study by incorporating a dataset containing 7,022 images from [11], representing four types of brain scans: glioma, meningioma, pituitary tumors, and non-cancerous (normal) tissues. We maintained the same 70:20:10 split for training, validation, and testing.

Finally, to assess the transferability of the model's learned features, we tested it on a completely different dataset consisting of lung cancer scans, obtained from [1]. This dataset had about 1,100 images. This step was critical for evaluating whether the model could generalize and apply the knowledge gained from the initial datasets to an entirely new type of medical imaging data.

Our initial approach involved installing all necessary dependencies and ensuring that the image files were in valid formats (e.g., .jpg, .png). The following libraries and tools were crucial to our workflow:

By integrating these tools, we established a comprehensive and efficient environment for conducting our experiments on chest and stomach scans. This setup enabled effective data preprocessing, model building, and performance evaluation, laying the foundation for the success of our methods and results.

- TensorFlow: Used for building and training deep learning models, with Keras providing a user-friendly API for defining neural networks.
- OS: Utilized for handling file paths and directory structures, ensuring proper organization of the dataset.
- cv2 (OpenCV): Employed for image preprocessing tasks, including resizing, filtering, and augmenting images.
- imghdr: Used to verify the integrity and type of image files, ensuring only valid formats were processed.
- Matplotlib: Used to generate visualizations, such as training curves, confusion matrices, and data distributions.
- NumPy: Provided support for numerical operations on multi-dimensional arrays, crucial for image data manipulation and processing.
- Scikit-learn: Assisted with model evaluation, data splitting, and generating metrics not directly available in TensorFlow.
- Keras: Used for image augmentation, enhancing model generalization by modifying training images on the fly, and creating model structure.
- Seaborn: Employed for more advanced data visualization, especially for statistical analysis.
- TensorBoard: Used for tracking and visualizing model training metrics, providing insights into model performance over time.

3.1 Initial Methods and Results for Chest and Stomach Scans

With these tools in place, we began by confirming the model's ability to differentiate between abdomen and breast scans. Once the model demonstrated a clear ability to distinguish between these two distinct types of medical images, we proceeded to a more challenging dataset where the scans were more similar to each other.

To illustrate the types of images used for training and validation, we present a sample batch from the dataset below:

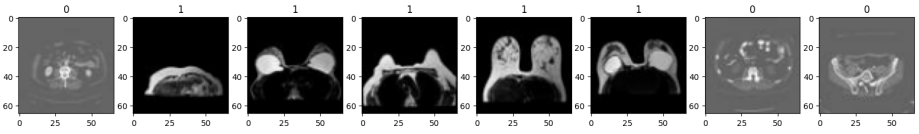


Fig. 2: Sample batch from the dataset, representing the two image classes

The images are labeled as either '0' or '1', corresponding to the two classes the model is trained to distinguish. Each image highlights various features that are characteristic of its respective class. The differences between the classes are visually evident in this figure, showcasing the patterns that the model is learning. The model achieved high accuracy on the test set, showing its effectiveness in distinguishing between the two classes. By the third epoch, accuracy reached 100% and remained consistently at this level. The training loss was relatively higher in the first epoch but dropped significantly during subsequent epochs. By the second epoch, the training loss decreased dramatically, nearing zero by the third epoch. This indicates the model's strong performance in both learning and generalization, as evidenced by the near-perfect accuracy scores. The close alignment of the training and validation curves suggests that the model did not overfit, and the low validation loss confirms that the model is generalizing well to unseen images.

However, these results raise concerns about the possibility of the model memorizing the training data rather than genuinely learning to generalize. The rapid achievement of 100% accuracy could indicate that the model is overfitting or simply memorizing specific features of the dataset, especially since the training and validation sets are relatively straightforward to differentiate. Despite the color differences between the scans—one being a lighter gray and the other a darker gray, the model was tested using grayscale images (i.e., 1-channel). Thus, these color differences should not significantly impact the model's ability to distinguish between the two classes.

It is important to note that this model was developed as a proof of concept. The primary objective was to confirm that the model could easily distinguish between two very different types of scans without any significant issues. This initial step laid the groundwork for more complex and challenging tasks, which involve datasets where the distinctions between classes are much subtler.

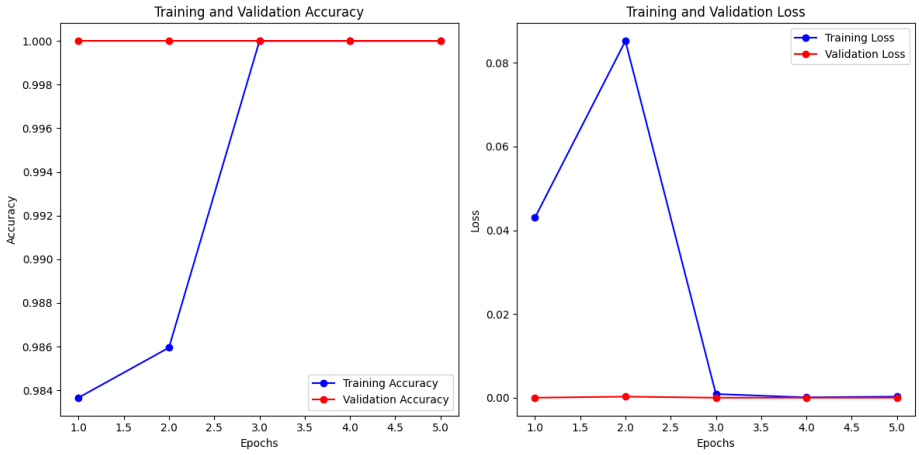


Fig. 3: Validation and accuracy as epoch's increase

3.2 Brain Cancer Scans

For the analysis of brain cancer, we utilized a dataset consisting of MRI scans representing four categories: gliomas, meningiomas, pituitary tumors, and normal brain tissues [11]. These categories were selected due to their visual similarities, which pose a significant challenge for accurate classification. To address the similarities amongst the scans, we employed the InceptionV3 model, pre-trained on ImageNet, as our base architecture. This was because the prior architecture was not complicated enough for such a task. Given the specific nature of our dataset, we fine-tuned the model by unfreezing the last few layers of InceptionV3, allowing these layers to update their weights during training while keeping the rest of the network frozen. The top layers were constructed using a combination of GlobalAveragePooling2D, Dense layers, and Dropout, optimizing the model for this particular classification task.

The fine-tuned model was then trained on the brain cancer images with the objective of accurately distinguishing between the different types of brain conditions present in the dataset. This step was critical in assessing the model's ability to classify amongst the groups.

3.3 Methods Enhanced

To ensure that our model is effectively learning rather than simply memorizing specific patterns, we implemented data augmentation to increase the complexity of the training process. Below is the data augmentation strategy applied during the final training of the brain scan models.

```
## batch size = 16
# image size = (299,299)
## channels = 3
```

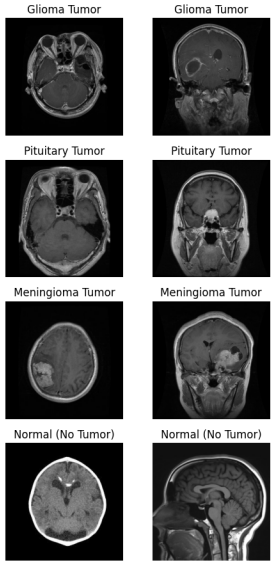


Fig. 4: Brain Cancer Dataset Example Images

```
tr_gen = ImageDataGenerator(
    rotation_range=20,
    height_shift_range=0.2,
    zoom_range=0.09,
    width_shift_range=0.2,
    horizontal_flip=True,
    vertical_flip=False,  ## Not applicable to scans
    brightness_range=[0.8, 1.5],
    rescale=1./255,
    fill_mode='constant',
    channel_shift_range=0.2,
)
```

Listing 1.1: Data Augmentation using ImageDataGenerator

5.

Model Architecture This is the model architecture that provides the final results.

```
base_model = InceptionV3(weights='imagenet',
    include_top=False, input_shape=img_shape)
x = GlobalAveragePooling2D()(x)
x = BatchNormalization()(x)
x = Dense(128, activation='linear',
    kernel_regularizer=regularizers.l2(0.0001))(x)
```

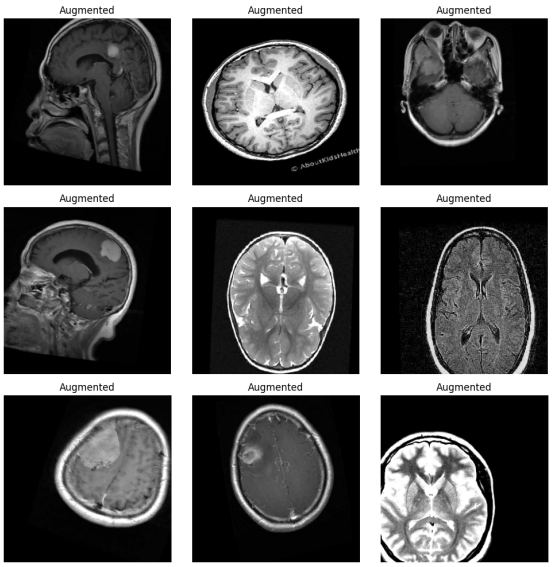


Fig. 5: Data Augmentation Example

```
x = LeakyReLU()(x)
x = Dropout(0.4)(x)
x = Dense(64, activation='linear',
    kernel_regularizer=regularizers.l2(0.0001))(x)
x = LeakyReLU()(x)
x = Dropout(0.3)(x)
x = Dense(4, activation='softmax')(x)

for layer in base_model.layers[:100]:
    layer.trainable = False
for layer in base_model.layers[100:]:
    layer.trainable = True
optimizer = Adam(learning_rate=3e-5, amsgrad=True)

model.compile(optimizer=optimizer,
    loss='categorical_crossentropy', metrics=['accuracy',
    'Precision', 'Recall', 'AUC'])

model_checkpoint = ModelCheckpoint('best_model.keras',
    monitor='val_accuracy', save_best_only=True, verbose=1)
reduce_lr = ReduceLROnPlateau(monitor='val_loss',
    factor=0.5, patience=3, min_lr=1e-7, verbose=1)
early_stopping = EarlyStopping(monitor='val_loss',
    patience=8, restore_best_weights=True, verbose=1) # To
    prevent overfitting

```

```
history = model.fit(train_gen, epochs=200,
    validation_data=valid_gen, callbacks=[early_stopping,
    reduce_lr, model_checkpoint])
```

Listing 1.2: Model Definition and Training Configuration

Incorporating these strategies, the model showed promising results, with its ability to distinguish between these visually similar categories being a strong indicator of its potential utility in real-world medical diagnosis scenarios. This analysis serves as a crucial test of the model's robustness and its applicability to more challenging classification tasks in medical imaging.

3.4 Lung Cancer Scans

To evaluate the transferability of the model, we tested its performance on an entirely different domain—lung cancer scans. This step aimed to assess the model's ability to generalize knowledge gained from training on brain scans to make accurate predictions on lung scans. The InceptionV3 model, trained solely on brain cancer scans, was applied to lung cancer histopathological images without any additional fine-tuning. This experiment sought to observe how well the features learned from brain scans could translate to recognizing patterns in lung scans. To further evaluate the model's robustness and its ability to generalize to completely different types of medical images, we tested it on a dataset of lung cancer scans [1]. This dataset was chosen not only because of its clinical significance but also to assess the model's transferability of learned features from brain cancer scans to lung cancer images—two vastly different domains in medical imaging.

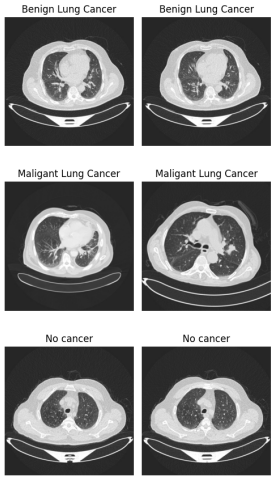


Fig. 6: Lung Cancer Dataset Example Images

The lung cancer scans differ significantly from the brain MRI scans in terms of texture, shape, and overall visual characteristics. This presented a considerable challenge, as the model was required to apply the features it learned from the brain cancer dataset to a new and visually distinct set of images. The objective here was not only to measure the model's accuracy but also to understand its ability to generalize knowledge across different types of cancerous images.

3.5 SHAP Plot Generation

Below is the code used to generate the plots for brain and lung scans.

```
class_names = ["glioma", "meningioma", "no tumor",
    "pituitary"]
def explain_image_with_shap(image_path):
    # Preprocess the image
    def preprocess_image(image_path):
        image = tf.io.read_file(image_path)
        image = tf.image.decode_jpeg(image, channels=3)
        image = tf.image.resize(image, [299, 299])  # Resize
            image to (200, 200)
        image = tf.cast(image, tf.float32) / 255.0   #
            Normalize pixel values
        return image.numpy()

    image = preprocess_image(image_path)

    # Define the model prediction function for SHAP
    def f(x):
        return model.predict(x)

    # Create an Image masker for SHAP
    masker_blur = shap.maskers.Image("blur(20,20)", (299,
        299, 3))

    explainer_blur = shap.Explainer(f, masker_blur)


    predictions = model.predict(image[np.newaxis, :, :, :])
    print("Predictions:", predictions)
    # Identify the top class index and corresponding class
        name
    top_class_index = np.argmax(predictions[0])
    top_class_name = class_names[top_class_index]
    top_class_probability = predictions[0][top_class_index]

    print(f"Top class: {top_class_name} with probability
        {top_class_probability}")


```

```
    top_4_indices = np.argsort(predictions[0])[::-1][:4]

    print(f"Top 4 class indices: {top_4_indices}")
    print(f"Top 4 class names: {[class_names[idx] for idx in
        top_4_indices]}")
    print("Name of image ",image_path)

    # Calculate SHAP values for the top 4 classes
    shap_values_fine = explainer_blur(image[np.newaxis, :,
        :, :], max_evals=1000, outputs=top_4_indices)
    ## Note the max_evals as it ranged from 6,000-10,000
        during final results for more precision
    # Plot the SHAP values for the top 4 classes with labels
    for i, class_index in enumerate(top_4_indices):
        shap.image_plot([shap_values_fine.values[0][:, :, :,
            i]], image, show=False)
        plt.title(f"SHAP Explanation for Class:
            {class_names[class_index]}")
        plt.show()
```

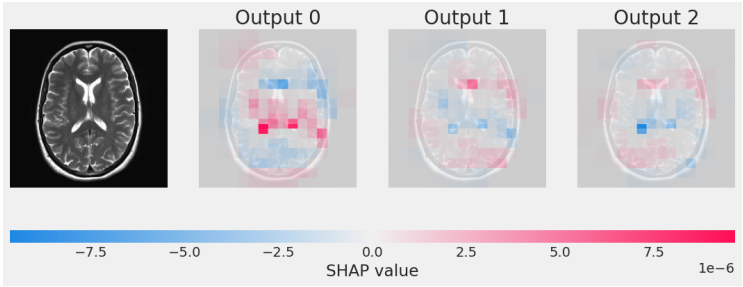
Listing 1.3: SHAP Plot Generation

4 Results

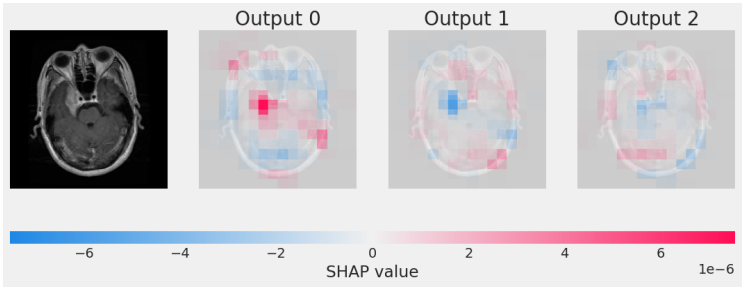
This section presents the results of our deep learning model's performance on both brain and lung cancer scans. We evaluate the model's ability to accurately classify images into their respective categories and assess its generalization capabilities across different types of medical images. The results are presented in two main parts: the initial results of the model on brain scans and the subsequent analysis of its performance on lung cancer scans.

4.1 Initial Results on Brain Cancer Scans

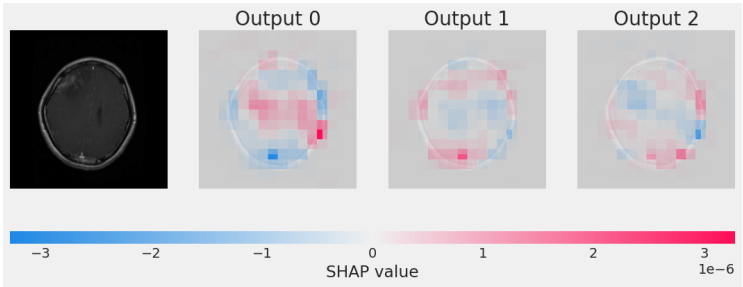
The figures 7a, 7b,7c demonstrate that the mdoel is not performing as well as expected and is struggling a little bit with generalizing correctly. We are able to tell this by the extremely small shapley values. The model printed out the first 3 classes it resembled to (output 0, output 1, output 2). These probabilities were consistently changing. For 7a, it was not completely confident that it was a normal brain scan as it corresponded to output 1. In other words, it was not considered the top class. For 7b and 7c, again the shapley values are extremely small indicating a smaller confidence. In these cases, the model correctly predicted the classification, but due to the extremely low confidence and unlocalized pixel blur, it is clear that the model is making predictions on irrelevant features and not locating the tumor properly. This is a clear indication of a weak model as it is not properly understanding, generalizing, and finding hidden patterns in the images as it is supposed to do.



(a) Normal Brain Scan



(b) Meningioma Brain Scan



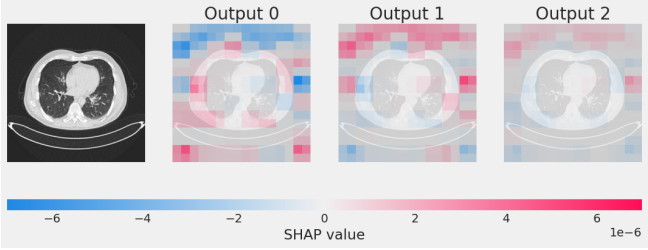
(c) Glioma Brain Scan

Fig. 7: Brain scans with SHAP values applied: Normal, Meningioma, and Glioma.

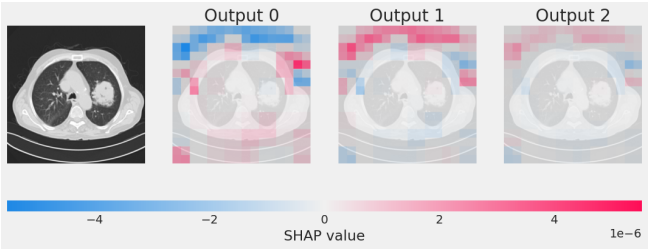
4.2 Initial Results on Lung Cancer Scans

The SHAP values provide a comprehensive picture of how different parts of the image influence the model's predictions across all classes. Even if the scan is normal, the SHAP values for other classes can be significant, indicating the model's decision-making process involves ruling out other possibilities by highlighting what it does not see (evidence for benign or malignant) as much as what it does see (evidence for normal). In this evaluation we are mapping confidence to the first 3 classes corresponding to glioma, menigionoma, pituitary, and no tumor. By plotting the scans, it is clear that the model is not focusing on the correct

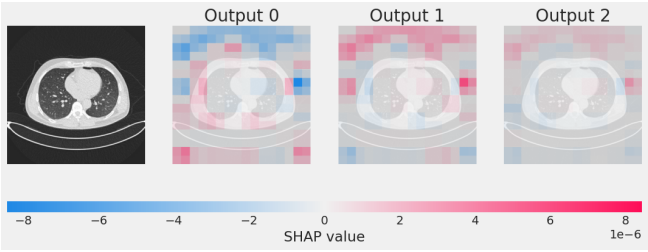
areas of the lung scans and making the predictions by focusing on irrelevant features outside of the lung scan.



(a) Using a benign lung scan to assess transferability



(b) Using a malignant lung scan to assess transferability



(c) Using a normal lung scan to assess transferability

Fig. 8: Assessing transferability using different types of lung scans.

4.3 Initial Results Analysis

Our study aimed to evaluate the generalization capabilities of a deep learning model trained on brain scans when tested on lung scans, without retraining. Additionally, we assessed the model's performance on brain scans to identify potential issues in generalization within its own domain.

When applying the brain-trained model to lung scans, the SHAP analysis revealed several key observations:

1. The regions highlighted by SHAP values in the lung scans did not correspond to clinically relevant features for lung pathology. Instead, the model appeared to focus on irrelevant areas, indicating a misalignment in feature extraction.

These results suggest that the features learned from brain scans are not directly transferable to lung scans without domain-specific retraining. This finding demonstrates the necessity of training or fine-tuning models on data specific to the target application.

Initial Results Concluded Our initial results demonstrate that the model, when trained on brain scans, does not generalize effectively to lung scans, highlighting the need for domain-specific training. Additionally, the SHAP analysis within brain scans suggests areas for improvement in the model's feature extraction and generalization processes. Future work will focus on refining the model to enhance its robustness and clinical applicability across different types of medical imaging data. Another approach could involve using pre-trained models on similar tasks and fine-tuning them on lung scans to help in learning more generalizable features.

4.4 Final Results on Brain Scans

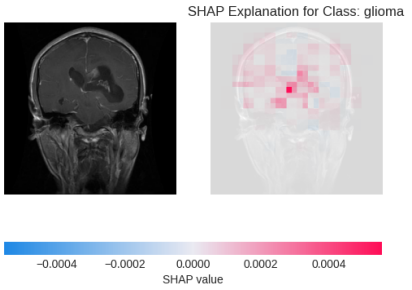
The primary objective of this phase was to evaluate the performance of our fine-tuned InceptionV3 model on a diverse set of brain and lung cancer scans. The aim was to assess the model's ability to accurately classify images into categories such as No Tumor, Meningioma, Pituitary Tumors, and Glioma, and to explore its generalization capabilities across different types of medical images. subcaption

In this round of testing and evaluation, the model was correctly predicting specifically to what class the scans belong to. The shapley values are much bigger indicating a higher confidence. The pixels are localized onto the corresponding tumor area, demonstrating the model is focusing on the correct pathology parts to make the predictions. Moreso, the scans above were ran with more evaluations ranging from 6,000 to 10,000 compared to 8c compared to that of 1,000, which helps to focus the pixels more and make them more precise. Due to the computational complexity, max evaluations were not run on every singular image.

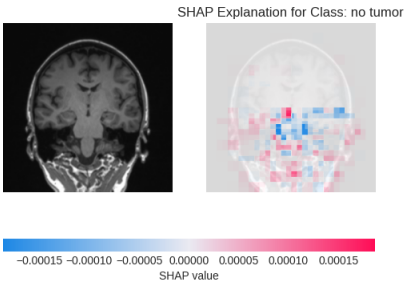
4.5 Final Results on Lung Cancer tests

The results of our model's predictions are summarized in Table 1. This table provides a comprehensive view of the model's performance across different image categories, including the predicted probabilities for each class (No Tumor, Meningioma, Pituitary, Glioma), the correct labels, and the model's prediction accuracy. Additionally, the table includes the sum of tumor-related probabilities, which indicates the model's certainty in its predictions.

The model was tested on lung cancer scans to assess its generalization capabilities across different types of medical images. In many cases, the model

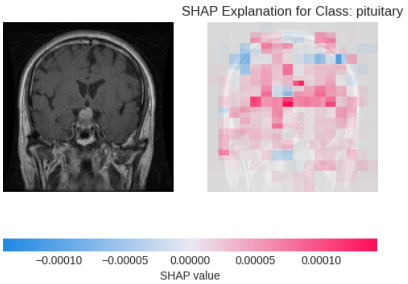


(a) Precise glioma scan applied with SHAP values



(b) No tumor scan applied with SHAP values

(c) Meningioma scan applied with SHAP values



(d) Pituitary scan applied with SHAP values

Fig. 9: Application of SHAP values to different types of scans: glioma, no tumor, meningioma, and pituitary.

assigned higher probabilities to tumor categories, indicating that it recognized patterns in lung scans that it had associated with brain tumors. However, there were instances where lung cancer scans were incorrectly classified as resembling normal brain tissue.

The shapley values of the above scans are much larger than the first round indicating a higher confidence from the model. This is ultimately because we changed the model structure significantly, causing a creation of a very well developed model.

Summary of Key Findings The results demonstrate that our fine-tuned InceptionV3 model is capable of accurately classifying brain cancer scans and shows potential in generalizing to lung cancer scans. However, the analysis also reveals certain limitations, particularly in the model's ability to differentiate between cancerous and non-cancerous lung scans. The presence of misclassifications and variable confidence margins suggests that while the model performs well on brain

Actual Class	Glioma Prob	Meningioma Prob	No Tumor Prob	Pituitary Prob	Combined Tumor Prob	Highest Prob Class
normal	0.001618	0.000207	0.998106	0.000068	0.001894	no tumor
normal	0.009308	0.005801	0.982685	0.002206	0.017315	no tumor
normal	0.002369	0.002954	0.994548	0.000129	0.005452	no tumor
normal	0.088900	0.058281	0.817691	0.035128	0.182309	no tumor
normal	0.054912	0.272921	0.566007	0.106160	0.433993	no tumor
benign	0.029764	0.669881	0.179649	0.120705	0.820351	meningioma
benign	0.002314	0.773614	0.223873	0.000199	0.776128	meningioma
benign	0.002289	0.744679	0.252881	0.000151	0.747119	meningioma
benign	0.009986	0.730734	0.256931	0.002349	0.743069	meningioma
benign	0.009862	0.663708	0.321695	0.004735	0.678305	meningioma
malignant	0.000002	0.999935	0.000062	0.000001	0.999938	meningioma
malignant	0.000002	0.999935	0.000063	0.000001	0.999937	meningioma
malignant	0.000006	0.999894	0.000099	0.000001	0.999901	meningioma
malignant	0.000003	0.999884	0.000110	0.000003	0.999890	meningioma
malignant	0.000193	0.999528	0.000223	0.000056	0.999777	meningioma

Table 1: Top probabilities for each class with the actual class and highest probability class.

Actual Class	Glioma	Meningioma	No Tumor	Pituitary	Tumor Total	Max Probability Class
benign	0.002886	0.001316	0.994657	0.001140	0.005343	no tumor
benign	0.000345	0.004903	0.994573	0.000179	0.005427	no tumor
benign	0.004296	0.074044	0.872770	0.048890	0.127230	no tumor
benign	0.062278	0.001284	0.936270	0.000168	0.063730	no tumor
benign	0.000498	0.002880	0.996582	0.000039	0.003418	no tumor
malignant	0.269130	0.279730	0.342372	0.108768	0.657628	no tumor
malignant	0.030440	0.665805	0.296193	0.007562	0.703807	meningioma
malignant	0.000327	0.936217	0.063391	0.000065	0.936609	meningioma
malignant	0.010783	0.093766	0.894730	0.000722	0.105271	no tumor
malignant	0.008686	0.813801	0.176570	0.000943	0.823430	meningioma

Table 2: Probabilities for the benign and malignant class only showing less promising results.

scans, its generalization capabilities could be further enhanced with additional training on more diverse datasets.

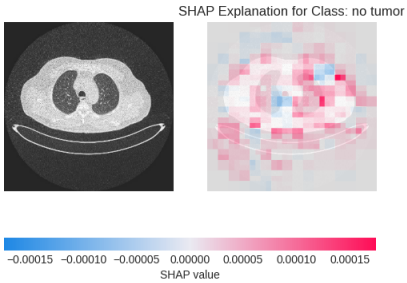
5 Metrics of Evaluation

In this section, we present the evaluation metrics used to assess the performance of our model. These metrics provide a comprehensive overview of the model's accuracy, precision, recall, and overall performance

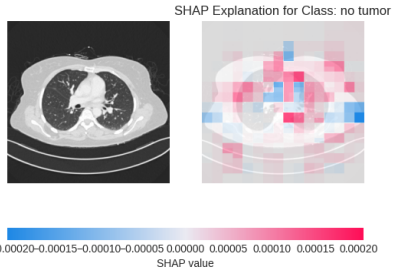
Overall, the metrics presented in the figures 11 and 12 above demonstrate the model's strong performance across various evaluation criteria. The high precision and recall values, coupled with a well-balanced accuracy, indicate that the model is proficient in correctly identifying different tumor types. The Confusion Matrix provides additional insight into how the model handles each class, with a majority of predictions correctly aligning with the true labels.

6 Discussion

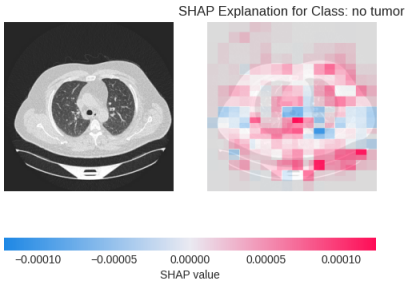
Our study demonstrates that while SHAP is a powerful tool for enhancing the interpretability of deep convolutional neural networks, the model's performance



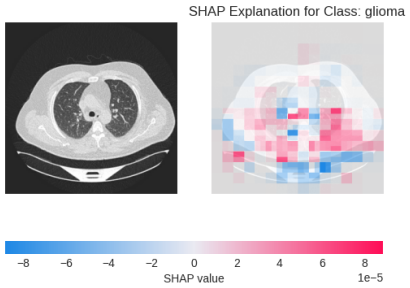
(a) Using a normal lung scan to see what part the model is focusing on to make the correct prediction. The model identified this scan as a normal brain scan.



(b) Using a malignant lung scan to see what part the model is focusing on to make the correct prediction. The model identified this scan as a normal scan, with much lower probabilities than that of a normal brain scan.



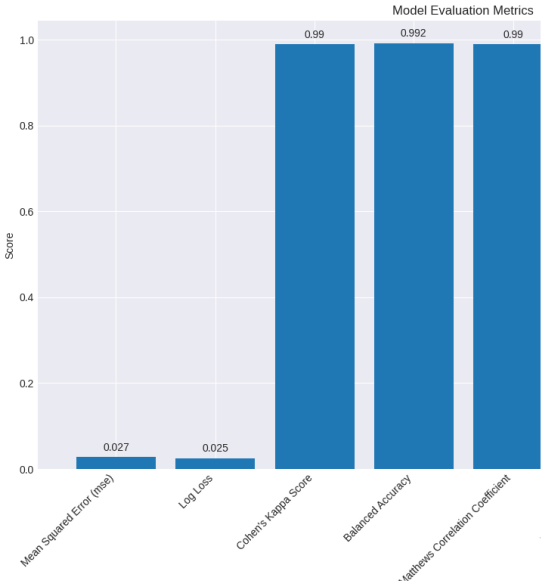
(c) Using a benign lung scan to see what part the model is focusing on to make the correct prediction. The model identified this scan as a normal scan, with much lower probabilities than that of a normal brain scan.



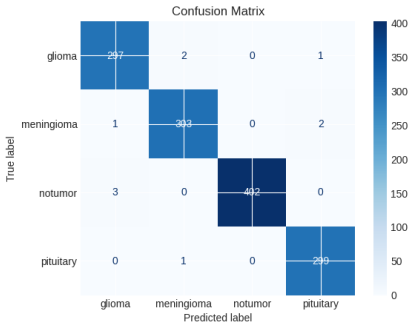
(d) Using the same benign scan to see what part the model is focusing on to make the correct prediction. The model identified the second probability to be glioma.

Fig. 10: Model focus points using different types of lung scans to assess prediction accuracy.

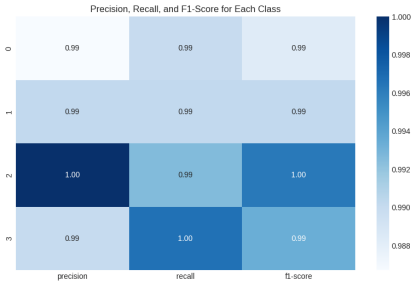
and generalizability are heavily dependent on the domain specificity of the training data. The application of SHAP revealed that the model, when trained on brain scans, struggled to generalize to lung scans, frequently misidentifying relevant features and focusing on non-significant areas of the images. The inability of the model to generalize effectively from brain to lung scans highlights the importance of domain-specific training in medical imaging. SHAP analysis revealed that the model's focus on irrelevant features in lung scans could be attributed to the stark differences in texture and shape between brain and lung tissues. This finding show the necessity for models to be trained on data that closely resembles the target application domain, particularly in medical contexts where precision



(a) Metrics of evaluation include Mean Squared Error, Log Loss, Cohen's Kappa Score, Balanced Accuracy, and more.



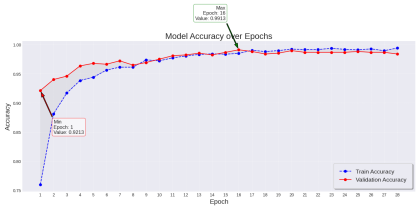
(b) Confusion Matrix on the test set.



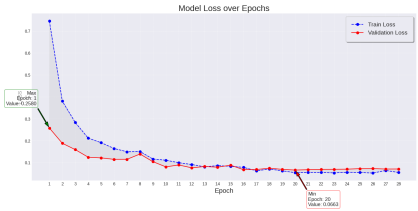
(c) Class 0 represents glioma, class 1 represents meningioma, class 2 represents no tumor, and class 3 represents pituitary.

Fig. 11: Evaluation metrics, confusion matrix, and class-specific metrics for the test set.

is critical. The implications of our study for clinical practice are significant. The ability to interpret model decisions through SHAP is crucial for gaining trust in AI systems, particularly in sensitive fields like healthcare. However, our findings indicate that clinicians should be cautious when applying models trained on one type of medical data to another without appropriate retraining. This study high-



(a) Accuracy over all epochs during model training.



(b) Loss over all epochs during model training.

Fig. 12: Training metrics over all epochs: Accuracy and Loss.

Dataset	Loss	Accuracy	Precision	Recall	AUC
Training Set	0.0341	0.9996	0.9996	0.9992	1.0000
Validation Set	0.0699	0.9926	0.9933	0.9926	0.9979
Test Set	0.0621	0.9900	0.9918	0.9900	0.9982

Note: The values displayed above are derived from the evaluation of the model using TensorFlow/Keras. Discrepancies between reported and printed values (e.g., precision, recall, and AUC) might occur due to differences in rounding, batch averaging, and metric calculation timing.

lights the need for more extensive validation and retraining processes to ensure that AI models can be safely and effectively used in diverse clinical scenarios.

7 Limitations

This study encountered several limitations that are important to acknowledge.

7.1 GPU Limitations

The computational resources available for this project were constrained, particularly with regard to GPU availability. This limitation impacted the ability to experiment with more complex models or conduct more extensive hyperparameter tuning. Despite switching over to Google Colab's L4 GPU, the computational speed was still not fast enough. Future work could benefit from access to more computational infrastructure to fully explore the potential of the models.

7.2 Accuracy of Interpretability Methods

While SHAP values were used to interpret the model's decisions, the accuracy and comprehensiveness of these interpretability methods are limited. SHAP provides valuable insights but may not capture the full complexity of the model's decision-making process. Additional interpretability techniques or improvements in SHAP itself could enhance the understanding of how the model makes predictions.

7.3 Clinical Applicability

The models developed in this study, though promising, are not yet ready for clinical application. The limited dataset size, the specificity of the data, and the need for further validation in diverse clinical settings all present challenges to immediate clinical implementation. More extensive testing and validation, potentially through clinical trials, are necessary before these models can be reliably used in real-world healthcare settings.

7.4 Time Constraints

The project was completed within a strict 12-week timeframe, which included the entire process from selecting a topic to finalizing the results. This time constraint limited the depth of exploration and the ability to address certain challenges more thoroughly. Future research with a longer timeline could allow for a more comprehensive investigation and refinement of the methods used.

7.5 Learning Curve

At the outset of this project, I had no prior experience in data science or machine learning, and I had to teach myself these skills throughout the course of the research. This steep learning curve may have affected the efficiency and effectiveness of some of the methodologies applied. With further experience and education, future work could build on this foundation to achieve more refined and optimized results.

8 Future Work

For future work, several key directions will be pursued to enhance the robustness of the model. First, we plan to expand the dataset by incorporating additional types of cancer, such as breast, prostate, and skin cancers. This broader dataset will allow us to evaluate whether the model's generalization improves when exposed to a more diverse set of tumor characteristics. The expansion will not only test the model's ability to classify different cancers but also explore its transferability across various cancer types.

Additionally, we will train the model to predict not just the tumor type but also the specific stage of cancer. By including cancer staging in the prediction task, the model could become a more powerful tool for clinical decision-making, potentially aiding in both diagnosis and treatment planning. We will assess the model's performance on new datasets, ensuring that it can accurately differentiate between early and late-stage cancers, which is crucial for effective patient management.

Moreover, we will explore other XAI methods to gain new insights into the model's decision-making process. While SHAP has been key in understanding feature importance, alternative XAI approaches such as LIME or Grad-CAM

could offer different perspectives and help identify any potential biases or weaknesses in the model. Testing multiple XAI methods will contribute to a more comprehensive understanding the model, ultimately improving transparency and trust in the model's predictions.

By pursuing these other methods, we aim to not only enhance the model's accuracy and generalization but also ensure that it remains interpretable and applicable in real-world clinical settings.

9 Acknowledgements

I would like to express gratitude to Professor Neha Keshan, who not only organized this course but also provided clear guidance and support throughout this project. As someone who came in with no prior research experience, her assistance was invaluable in helping me stay organized and focused. I am also thankful to my peers in CSCI 4960, whose constant feedback during peer review sessions significantly contributed to the development of my research. My sincere thanks go to the TAs of the class for their constructive feedback on improving my work.

A special thank you goes to my friends, Nicole Bienasz and Shoshana Sugerman, for their unwavering support and encouragement everyday. Their willingness to listen to every singular thought I had at any hour of the day about this research meant the world to me. I would also like to express my deepest appreciation to my father, Vladislav Tsekanovskiy, who has always been my greatest supporter.

Finally, I would like to acknowledge the resources and environment provided by Rensselaer Polytechnic Institute, where this research was conducted.

References

1. Al-Yasriy, H.F.: The IQ-OTH/NCCD lung cancer dataset (May 2020), https://www.kaggle.com/ds/672399
2. Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L.: Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. Journal of Big Data **8** (03 2021). https://doi.org/https://doi.org/10.1186/s40537-021-00444-8, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8
3. Apolanco3225: Medical mnist classification. https://github.com/apolanco3225/Medical-MNIST-Classification (2017), gitHub repository
4. Bhatt, C., Kumar, I., Vijayakumar, V., Singh, K.U., Kumar, A.: The state of the art of deep learning models in medical science and their challenges. Multimedia Systems (09 2020). https://doi.org/10.1007/s00530-020-00694-1
5. Bozorgpanah, A., Torra, V.: Explainable machine learning models with privacy. Progress in artificial intelligence **13**, 31–50 (03 2024). https://doi.org/10.1007/s13748-024-00315-2

6. Van den Broeck, G., Lykov, A., Schleich, M., Suciu, D.: On the tractability of shap explanations. Journal of Artificial Intelligence Research **74**, 851–886 (06 2022). https://doi.org/10.1613/jair.1.13283
7. Khoei, T.T., Slimane, H.O., Kaabouch, N.: Deep learning: systematic review, models, challenges, and research directions. Neural Computing and Applications (09 2023). https://doi.org/10.1007/s00521-023-08957-4
8. Kostopoulos, N., Kalogeras, D., Pantazatos, D., Grammatikou, M., Maglaris, V.: Shap interpretations of tree and neural network dns classifiers for analyzing dga family characteristics. IEEE Access **11**, 61144–61160 (2023). https://doi.org/10.1109/ACCESS.2023.3286313
9. Lu, S., Chen, R., Wei, W., Belovsky, M., Lu, X.: Understanding heart failure patients ehr clinical features via shap interpretation of tree-based machine learning model predictions. AMIA ... Annual Symposium proceedings. AMIA Symposium **2021**, 813–822 (2021), https://pubmed.ncbi.nlm.nih.gov/35308970/
10. Mujahid, M., Rustam, F., Álvarez, R., Luis Vidal Mazón, J., Díez, I.d.l.T., Ashraf, I.: Pneumonia classification from x-ray images with inception-v3 and convolutional neural network. Diagnostics **12**, 1280 (05 2022). https://doi.org/10.3390/diagnostics12051280
11. Nickparvar, M.: Brain tumor mri dataset (2021). https://doi.org/10.34740/kaggle/dsv/2645886
12. O'Shea, K., Nash, R.: An introduction to convolutional neural networks, https://ar5iv.labs.arxiv.org/html/1511.08458
13. Ribeiro, M., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning (2016), https://arxiv.org/pdf/1606.05386
14. S, Y.K., Jeya, J.J., R, M.T., Khan, S.B., Alzahrani, S., Alojail, M.: Explainable lung cancer classification with ensemble transfer learning of vgg16, resnet50 and inceptionv3 using grad-cam. BMC medical imaging **24** (07 2024). https://doi.org/10.1186/s12880-024-01345-x
15. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models (08 2018), https://arxiv.org/pdf/1708.08296
16. Shiri, F.M., Perumal, T., Mustapha, N., Mohamed, R.: A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru (09 2023)
17. Traore, B.B., Kamsu-Foguem, B., Tangara, F.: Deep convolution neural network for image recognition. Ecological Informatics **48**, 257–268 (11 2018). https://doi.org/10.1016/j.ecoinf.2018.10.002, https://www.sciencedirect.com/science/article/pii/S1574954118302140
18. Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M.B., Kang, B.H.: Survey on explainable ai: From approaches, limitations and applications aspects. Human-centric intelligent systems **3**, 161–188 (08 2023). https://doi.org/10.1007/s44230-023-00038-y