School of Information Technology and Engineering at the ADA University

School of Engineering and Applied Science at the George Washington University

REFINING NEURAL NETWORK INTERPRETABILITY THROUGH ACTIVATION MODIFICATION TECHNIQUES: AN EXPLORATION OF THRESHOLD-BASED APPROACHES

A Thesis
Presented to the Graduate Program of Computer Science and Data Analytics
of the School of Information Technology and Engineering
ADA University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Nigar Mammadova

April, 2025

THESIS ACCEPTANCE

This Thesis by: Nigar Mammadova
Entitled: *Refining Neural Network Interpretability through Activation Modification Techniques: An Exploration of Threshold-Based Approaches*

has been approved as meeting the requirement for the Degree of Master of Science in Computer Science and Data Analytics of the School of Information Technology and Engineering, ADA University.

Approved:

(Adviser) Abzatdin Adamov	08.08.2025 (Date)
(Program Director) Abzatdin Adamov	(Date)
(Dean)	(Date)

ABSTRACT

Interpretability in deep learning models has recently emerged as a major and growing concern, especially in high-stakes settings such as medical diagnostics. This thesis focuses on the problem of how to design real post-hoc modifiable Deep Neural Networks (DNNs) that can achieve or exceed state-of-the-art performance while also providing increased transparency that can help in understanding how predictions made by DNNs were reached. Existing techniques for interpretability are mostly concentrated on inspecting neuron activations as is. Here, we study controlled neuron activation adjustments during inference and examine whether these adjustments can help improve the explainability and generalization of Fully Connected Neural Networks (FCNNs) without retraining.

The dataset utilized in our research is a publicly available benchmark brain tumor classification dataset, which has been divided into four classes: glioma, meningioma, pituitary tumor, and no tumor. Both a baseline Fully Connected Neural Network (FCNN) was constructed and assessed, and an interpretable framework was created where activation patterns along the layers were visualized. Finally, we further studied the activation dynamics through experiments with partial network connections, underfitting, and overfitting, in order to investigate the relationship among sparsity, generalization, and interpretability of the network.

Based on these results, the study introduces three activation method adaptation strategies: a thresholding method based on qualitative analysis of heatmaps, a robust Linear moments (L-moments), and a probabilistic Gaussian Mixture Model (GMM) threshold determination method. All of them introduce a systematic adjustment of neuron activations according to individual activation magnitude, which tends to make the latent feature representation more significant in the inference phase. Experimental results show that the improvement of classification accuracies can be significant on misclassified samples as well as on overall model performance, achieving up to 14% improvements without retraining.

The proposed approaches provide realistic post-deployment methodologies to enhance model performance without significant computational overhead or regulatory liability. Furthermore, the improvements in interpretability achieved by activation visualization and correction provide useful information regarding how deep neural networks arrive at their decisions, building trust with users. The thesis observes this “post-hoc” activation manipulation as a promising, scalable path to improving the interpretability and the usability of Deep Learning (DL) models, especially in sensitive domains where not only performance but also transparent decision-making is paramount.

Keywords: Neural Network Interpretability, Fully Connected Neural Networks (FCNNs), Neuron Activation Modification, Brain Tumor Classification, MRI Image Analysis, Post-Hoc Model Correction, Activation Visualization, Explainable Artificial Intelligence (XAI), L-Moments Thresholding, Gaussian Mixture Models (GMM), Sparse Neural Networks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Explanation
AI	Artificial Intelligence
CNN	Convolutional Neural Network
DL	Deep Learning
FCNN	Fully Connected Neural Network
FCNNs	Fully Connected Neural Networks
GMM	Gaussian Mixture Model
L-moments	Linear Moments
ML	Maximum Likelihood
MRI	Magnetic Resonance Imaging
MVU	Minimum Variance Unbiased
NLP	Natural Language Processing
PDF	Probability Density Function
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SoftMax	Soft Maximum

1 INTRODUCTION

Fully connected neural networks (FCNNs) in the field of deep learning experienced considerable development during the recent years and become widely applied across different domains including image recognition and natural language processing. A key problem exists regarding interpretability when examining deep learning frameworks despite their successful operation. Unlike traditional statistical models, neural networks cause black-box systems because their variable relationships remain hidden from view. Model obscurity creates challenges for understanding decision processes thus making acceptance difficult in industries that need transparency including healthcare as well as finance and autonomous systems.

Several critical factors require neural networks to exhibit interpretability features. The model receives enhanced trust from users through prediction understanding which allows stakeholders to grasp the derivation process. The investigation of incorrect or biased decisions becomes possible through the debugging process because the system enables identification of the neurons responsible for these outcomes. The requirement to demonstrate explainable and auditable decision-making processes in certain industries becomes possible through this approach. The extensive motivations to develop neuron activation interpretation methods for deep learning models resulted in research focused on this objective. Scientists face an unresolved issue regarding how changes in neuron activations affect the interpretability levels of FCNNs. When neural networks process data the intermediate values created by neurons constitute neuron activations. The learned representations evolve through different network layers because of such activations while these activations also provide significant weight to model prediction

1.1 Problem Statement

Interpretability enhancement through neuron activation modification remains an open research question since current interpretability techniques focus mainly on analyzing existing activations without modifying them directly. The existing research deficit regarding neuron activation modifications implies new exploration possibilities to determine whether controlled changes can generate better interpretable models without forfeiting predictive abilities.

The study tackles fundamental topic of how visualization techniques help researchers understand patterns of neuron activation. Visualizations stand as essential components which enable users to understand information movement across the neural network.

Studies have not determined the most efficient methods to visualize activation patterns within FCNNs. In this work multiple visualization strategies will be considered which would enable practitioners to pick the optimal visualization solutions which fit particular neural networks and engineering problems. Researchers have yet to address if altering neuron activations through artificial methods would enhance model explanations as well as performance. Model optimization methods historically followed a strategy that optimized weights and biases through backpropagation to decrease loss functions. During training when researchers systematically control or restrict activation patterns the model might acquire more easily interpretable features while achieving equivalent or better predictive capabilities. Evaluation of this hypothesis matters greatly for high-important applications because understanding model decisions becomes equally crucial as maintaining high accuracy rates.

This research produces implications which benefit various sections of artificial intelligence development and deployment. A medically oriented interpretable neural network might enable it to explain its diagnostic decisions to medical professionals thus building their trust in AI-based clinical decisions. The combination of interpretability enhancement and financial modeling would result in superior risk evaluation together with stronger regulatory standards. Autonomous systems perform better when their decision processes remain open and this concept improves safety together with operational dependability in self-driving cars. Research determines how neuron activation modifications affect interpretability since it represents both theoretical value and practical importance.

The primary obstacle during this research involves maintaining both model simplicity and interpretability ability. People acknowledge simpler models like logistic regression and decision trees offer better interpretation compared to deep neural networks even though deep neural networks deliver better predictive abilities. The purpose of this research investigates whether interpretability can achieve improvements within FCNNs when the abstractive capabilities remain intact. A thorough analysis of both intended effects and unwanted biases alongside generalization deteriorations from activation-modifications needs to be performed.

The research requires evaluation of how modifications and evaluations of neuron activations can be implemented at a operational computing level. Deep learning models need
sizeable computational power for operation so modifications to activations must work efficiently enough for practical application despite being used with extensive models. The technique of knowledge distillation allows researchers to understand the best methods for optimizing activation modifications through its ability to transfer model knowledge to simpler interpretation models.

The evaluation process of this research extends its findings to diverse neural architecture structures. The research project explores fully connected networks as its main focus although the acquired understanding might help explain activation alteration methods in Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) structures. The development of activation modification techniques for neural networks across different domains depends on knowing their extensive usability.

1.2 Significance of the Study

The proposed investigation pursues three primary goals: thorough study of neuron activation modification impacts on interpretation alongside most beneficial visualization strategies and evaluation of synthetic changes to boost performance and clarification levels. The research aims to benefit the expanding field of interpretable artificial intelligence and boost deep learning model usability in critical applications through its work of connecting performance and transparency. FCNNs represent an important area to investigate because they deploy high predictive abilities through uninterpretable networks. One of the main goals is to contribute knowledge to explainable artificial intelligence through its analysis of activation modifications on interpretability enhancement thus assisting the field which demands both transparency and accountability. This research study generates important neural network understanding which enables trustworthy adoption in critical explainable modeling areas like healthcare together with finance technology as well as automated decisions systems.

First, it provides a method for altering neuron activations which simultaneously improves interpretability while allowing researchers to detect model biases and discover weak areas that cause errors. Research into neuron activations provides twofold benefits to models by enabling improved representation learning and bias prevention across deep learning systems.

Secondly, it presents visual tools to track activation patterns and gives specific methods for experts who want to improve transparency in neural network frameworks.

This study enhances the current effort to create trustworthy artificial intelligence
systems with accountable decision-making through transparent operations without com-
promising accuracy performance.

1.3 Limitations of the Study

A few restrictions need acknowledgment since they could affect both methodology application and scope and generalization potential. Awareness of the complexity in FCNNs poses difficulties for researchers studying the effects of neuron activation modifications since intricate neuron dependencies can produce unexpected results that damage information quality and cause corruption of feature spaces leading to degraded interpretation capability plus reduced performance capacity.

The computational processing needs for modifying and examining neuron activations demand significant resource so they become a scalability challenge especially while using large datasets across deep architectures. The relationship between interpretation clarity and prediction accuracy requires

careful consideration because simplifying neural network outputs might reduce model effectiveness in handling advanced real-world systems.
Subjective interpretation caused by visualization techniques leads to inconsistencies between different users who see the same visual representations which leads to decreased reproducibility. The research extends only up to FCNNs which limits its applicability to other neural network architectures like convolutional and recurrent networks thus requiring additional studies to share findings across different models.
The procedure of artificially adjusting neuron activations creates a risk of overfitting that would make models learn specific patterns instead of effective generalization to unseen data which reduces robustness. Standard evaluation metrics for neural network interpretability research remain inadequate because qualitative and visual assessment methods fail to provide strict quantitative measures to quantify developments. The study addresses these constraints by developing thorough experimental methods and analysis to minimize their influence, so the research establishes substantial contributions to interpret neural networks.

Acknowledging the inherent challenges of neuron interconnectedness and the computational resources required for manipulating activations, this study nonetheless offers a valuable contribution to our evolving comprehension of interpretability and performance gains in FCNNs via activation modification. Our work demonstrates the feasibility and impact of targeted interventions at the activation level. Specifically, the experimental results, including the notable improvements observed in previously misclassified instances and less-developed models, strongly suggest that post-training adjustments to neuron activations can unlock untapped predictive potential without compromising the model's ability to generalize or requiring a full retraining process.
Crucially, this research bridges a significant divide between theoretical investigations into neural network interpretability and pragmatic, deployable techniques applicable during inference. By introducing a mechanism for activation adjustment informed by qualitative heatmap analyses and quantitative thresholding approaches, this study provides a practical toolkit for enhancing model outputs while keeping computational overhead within reasonable limits. Our findings indicate that neuron activation modifications can be effective even in environments with limited resources, a key consideration for real-world deployment in sectors such as healthcare, finance, and other regulated domains where retraining may be impractical or not permitted.

The developed approach supports the ongoing project of developing transparent artificial intelligence systems. The research examines the effects of controlled activation flows on decisions by providing knowledge about model information processing and weak signal utilization while demonstrating minimal adjustments that promote correct predictions. This research generates insight which guides activation manipulation development as well as enhances theoretical understanding of deep networks' decision processes for plans of future explainability research.
The implementation of fully connected network architecture as the base model has restricted generalization capabilities but simultaneously provides a major advantage. FCNNs serve as a fundamental choice over convolutional and transformer-based models because they provide researchers with an easier platform to study activation modifications. Evidence of robustness emerges from these improvements which span various training conditions from undertrained to moderately trained models therefore validating that the techniques show no dependency on specific network parameters or training methods.
The introduced activation modification system comes with an intrinsic explanation capability. The employed method operates with transparency because it boosts neurons with weak activations through adjustable parameters while keeping powerful active neurons unchanged. The transparent mechanism operates in concert with explainable AI goals to enable stakeholders and users to understand what model output improvements result from so they can develop greater trust and adoption potential.

The technical use in critical fields particularly healthcare holds significant practical value. Medical imaging tasks involving brain tumor classification will achieve better pathologic detection together with superior treatment results and higher clinician acceptance of AI systems through improved model sensitivity while maintaining specificity.
Research output demonstrates activation modification should be considered an effective post-deployment enhancement for clinical models to enhance their adaptability along with robustness.
The visual subjectivity in activation analysis remains acknowledged while the study develops a foundational approach to build standardized manipulation protocols. The study successfully combines human-understandable heatmap visualization with quantitative processing methods to achieve technical consistency alongside natural interpretation capabilities. Further advancements should focus on creating standardized evaluation standards for activation dynamics in neural networks by developing an official framework for assessment.
The procedures used to modify activations seem to prevent significant overfitting issues during experiments despite the importance of keeping an eye on this potential risk. The modifications enabled improved performance across all sets including test sets besides training and validation sets. The activation adjustments seem to facilitate improved general ability in identifying features throughout various datasets while avoiding specialized learning of training patterns. Future extension of this work requires thorough attention to activation modification methods since their application to complex datasets and architectures remains an important factor.
Researchers can explore novel aspects for study as larger models and datasets present scalability problems. New adaptive scaling methods would make activation modification automatic through learnable adjustment modules or attention-based threshold functions which could replace the necessity for grid searches. Such advancements would make the method practical for bigger computational settings therefore extending its usefulness.

Lastly, although the current study mainly aimed at brain tumor classification by FCNNs, the underlying principles of activation manipulation can be generalizable. The central idea — that neural networks encode underused latent signals that can be activated post-hoc — can be applied without preference of model type or domain dataset. So, this potentially is something to look out for in the future but there is hope for applying similar techniques to CNNs for image analysis, recurrent networks for sequence modeling and possibly even transformers in NLP tasks. Overall, despite those caveats that include the lack of architecture-specificity, computational cost and subjectivity of visualization, this work is an important contribution to the broader fields of model interpretability, post-training model optimization, and (we believe) pragmatic explainable AI. It functions as an existence proof that activation modification is an actual, interpretable, and effective technique for improving model predictions without the need of retraining. Through the systematic study of activation patterns and the use of controlled manipulations, the study provides novel findings about neural network function and introduces practical methods that could enhance the deployment of a model in high-stakes real-world tasks. This work not only advances knowledge of neural activation, but also serves as a starting point for investigating more complex, dynamic, and general activation manipulation mechanisms.

2 REVIEW OF THE LITERATURE

The growing use of deep learning models, notably FCNNs, for medical image tasks including brain tumor classification, has increased the importance of model explainability. In life-critical contexts such as healthcare, the ability to understand the organizational decision logic employed by neural networks is essential to their acceptance, transparency, and regulatory approval. In this sense, this work conducts a systematic review of the literature, aiming to analyze previous work on neuron activation modifications to make them more interpretable, on the development of active activations visualization

techniques and on how sparsity can be embedded into neural networks design, pointing out how such works fit into the current work.

The early foundational work by Doshi-Velez and Kim [1] laid out the theoretical underpinnings of interpretability, focusing on the need to understand how modifications to the input data flow through the activation of the neurons, to the model outputs. Their results also indicate that manipulating the activation of neurons may be a powerful way to probe the rational of models allowing the opportunity to discover salient features for certain classes. Based on this, Chen et al. [2] studied selective changes to the activation of neurons in hidden layers of the network, showing that this allows one to highlight or to suppress individual activation, and hence be able to dissect the contribution of each individual neuron to decisions. Taken together, these works underscore the promise of activation manipulation for producing interpretable deep learning models.

Visualization techniques have also become an instrumental approach to interpret the neuron activations in FCNNs. Saliency map introduced by Simonyan et al. [3], present the pixel-level description of how changes in the inputs affect output probabilities, allowing for identifying of regions that contribute most to predictions. In contrast to their success, saliency maps are known to be prone to noise and have interpretation challenges. Grad-CAM introduced by Selvaraju et al. [4], extended this idea, using the gradient with respect to the feature map in convolutional layers to get a more localized and semantically meaningful visualization. While developed for convolutional architectures, these visualization approaches reinforce the general importance of associating neuron activations with human-interpretable reasoning, a tenet as relevant to FCNNs.

A similar manipulation of neuron activations for aesthetic interpretation was investigated by Fong and Vedaldi [5] and they showed that with targeted perturbations, it was possible to generate interpretations that are more in line with human intuition. They show that it is not only possible, but also a viable approach to obtain post-hoc model interpretations through manipulation of the internal activations.

Simultaneously, an orthogonal line of work research has centered on the modification of network architectures that is carried out to make the network architectures more interpretable. Reduced connectivity model such as partially-connected/sparsely connected networks has been suggested as a practical and interpretable option to alleviate the high connectivity in full-connectivity designs. Louizos et al. [6] proposed L0 regularize for sparsifying the weights learned in the training step, which prunes the size of a model and tightens its generalization ability without performance degradation. Likewise, the work on “lottery ticket hypothesis” proposed by Frankle and Carbin [7] suggested that for dense models, sparse subnetworks can achieve the performance as strong as or even better than that of the original network if properly initialized.

Sparse structures provide natural interpretability as they highlight important features, while suppressing redundant and irrelevant ones. Liu et al. [8] introduced ensemble gradicent-based methods for sparse network feature selection, showing their ability for selecting informative network input nodes. Further, Liu et al. [9] introduced Sparse Contrastive Coding where learned sparse features are aligned with semantic input features towards better model transparency.

Theoretically, Galanti et al. [10] established norm-based generalization guarantees for the compositionally sparse architectures, showing that structured sparsity matches the intrinsic complexity of target functions. Complementarily, Zhou et al. [11] proposed the N:M structured sparsity for creation of hardware-efficient sparse models without performance loss, which is a significant stride towards real-world applicability.

The computational advantages of sparsity has also been appreciated. Zhang et al. [12] introduced the Cambricon-X accelerator, which exploited sparse computation for high throughput and better energy efficiency with deep neural networks. Recent work by Liu et al. [13]benchmarked sparse networks on various tasks, to explore potential and limitations of sparsity technique deployed and stimulate work toward more generalized algorithm.

Sparse models, have found wide-spread application in medical problems, where model interpretability and computational cost are an issue. Ullah et al. [14] have shown that the class distributions for brain tumor classification can be effectively balanced using sparse autoencoders to enhance accuracy of prediction as well as robust feature learning. Generalizing these concepts beyond medicine, Wen et al. [15] studied the online sparse robust models for industrial streaming data mining, the ideas of which can be extended to “real time” medical diagnosis.

To sum up, the previous studies has well justified the core premise of this thesis: that neuron activation manipulation and structural sparsity can serve as an effective means of enhancing the interpretability and performance of deep learning models. Existing visualization methods and architectural advances already lay a solid foundation, yet, the controlled, post-hoc perturbation of neuron activations at inference time remains a less-researched though promising way to enable transparent and trustworthy AI systems. The current study follows very closely these observations, and aims at combining activation manipulation and sparsity for the purpose of obtaining better model interpretability in the brain tumor classification tasks.

3 METHODOLOGY

We used the publicly available Brain Tumor Classification dataset in this research, which includes a total of 3,264
grayscale images divided into four different classes: glioma tumor, meningioma tumor, pituitary tumor, and no tumor. The dataset is arranged in class wise corresponding folders, which makes labeling and preprocessing easy. The images are the axial slices of the brain images taken in various clinical scenarios with different tumor shapes and localizations. This diversity renders the dataset especially convenient to estimate the ability of DL models to capture clinically meaningful features as well as to generalize across various realistic cases.
In the dataset four types of tumors have unique radiological features that are important for precise diagnosis and the planning of treatment. An infiltrative glial cell tumor such as glioma is one example. On MRI, they typically manifest as irregular masses of variable contrast enhancement.

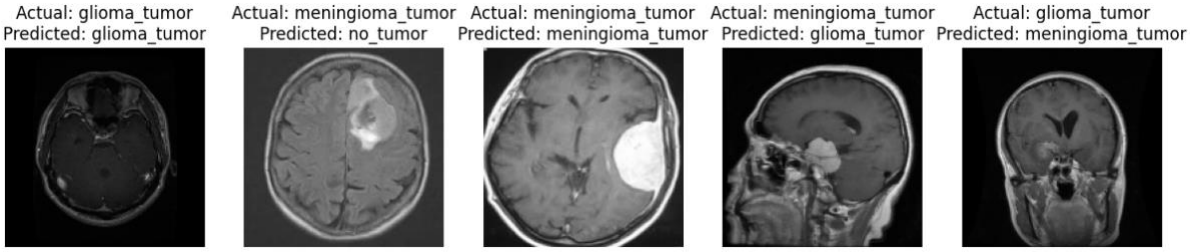

Figure 1: Different samples from the training dataset

Gliomas can have a heterogeneous appearance with necrotic and edematous regions, and may extend across multiple brain lobes, which may result in indistinct and difficult to define borders. This nonstructured and variable appearance requires a model that could represent subtle texture and intensity based features.
Meningiomas, on the other hand, are more circumscribed masses found in the meninges (the covering of the brain and the spinal cord). Radiologically, they are usually easier to recognize given their clear borders, homogenous contrast enhancement, and extra-axial location, resulting in typical mass effects such as associated bone remodeling or brain structure displacement. Such uniform appearance features could potentially help in achieving a higher classification accuracy for meningiomas than gliomas.
Pituitary tumors are yet another type. These tumors derive from the pituitary, which is a small endocrine gland situated at the base of the brain. Since pituitary masses are relatively fixed and their high location is largely similar for different subjects. Located close to midline structures and usually limited in size (only a few centimeters) they provide strong local spatial cues that help model learning. Pituitary tumors tend to be much easier to localize in terms of anatomic location as compared to gliomas and meningiomas and this attribute will be exploited since most of the harder classification problems we tackle are due to complex textural patterns.

Lastly, the “no tumor” class consists of normal brain MRIs that do not present any evidence of neoplastic formations. These are important for training the model to infer disease absence and avoid false positives. Efficient discrimination between pathological and normal images is crucial for real-life clinical implementation, because false alarms may initiate more invasive diagnostics or cause patient's anxiety.

3.1 Qualitative Analysis

Such four different classes entering into the database challenge the model to identify the type for tumors by the shape of structure and the location of anatomy. This also emulates a clinically realistic classification problem in which a diagnostic system needs to not only determine that there is a tumor, but also that the tumor is most likely of a certain type and each type should be treated differently.

In the cases that we explored in our study, the wide range of presentation of malignancy in tumors made our tumors set a good dataset to investigate how the tumor appearance is internalized by the CNN into visual patterns. The network was intended to learn discriminative features at multiple levels from low-level edge and texture detectors to high-level spatial and semantic encodings by being trained on these images. One key motivation for studying the activations of neurons and coming up with the activation modification strategies is the question of how the heterogeneous tumor features are being internally encoded and weighted by the network.

More specifically, it was speculated that distinct tumors induce specific groups of neurons in the network layers.. For example, neurons coding the round and homogenous shape of meningioma will show great activation in processing meningioma but low activation in processing glioma. Finally, neurons detecting mediolateral position cues may be especially important for detecting pituitary tumors. By examining these patterns of activation we were able to gain a novel insight into the network's learned decision-making pathways.

Thus, before applying activation modification, number of heatmaps produced across tumor types further led to the qualitative observation that while the network was capable of learning discriminative patterns, a significant proportion of neurons did not participate in the inference. This observation suggested that slight modifications in activations, such as adding a small amount of activation to inactive neurons that are related to the relevant tumor characteristics, would ameliorate the classification results. The heterogeneity of the tumors was also another reason for this method, and because of the similarity and overlaps of glioma and meningioma, there would be some weak signals existing, and in order to correct the misclassifications, we could amplify the weak signals.

It would not be an exaggeration to say that the clinical relevance of this pursuit is immeasurable. The accurate and prompt diagnosis of brain tumors is a critical factor in the decisions on the treatment plan, the prognosis of the therapy and the quality of the life of a patient. Gliomas, especially high-grade glioblastoma multiforme, need to retreat more aggressively while meningiomas (according to their grade and localization) can be treated either with conservative or surgical approach. Depending on various sizes and types of pituitary tumors, not all need interventions, unless these are functioning or are causing hormone secretion and need treatment. Hence, it is important to have the classification models learn and make use of the correct anatomical and pathological features.

In this dataset, our study is intended not only to achieve high classification accuracy, but also to find deeper insight into the decision mechanisms of the network. Using ablation tuning experiments reported in later section, we aimed at validating the extent of internalization of distinguishing properties for different tumors and if there are latent weak feature representations that could be discovered and improved post-hoc. This line of research connects performance improvement to the interpretability of models: it provides new comprehension on how artificial networks can reproduce or resist clinical reasoning.

To recap, the Brain Tumor Classification dataset offered a dense, clinically meaningful testbed to develop and validate not only our baseline models, but also our activation manipulation strategies. By its well-balanced presentation of different types of challenging tumors and a moderate size of the dataset, a realistic performance assessment of the model was possible without the need of tremendous

computational resources. By combining observations of dataset with technical innovation, the study ultimately intended to contribute for more transparent, efficient, and clinically-available machine learning systems for brain tumor diagnosis.

3.1.1 Data Preprocessing

The data preprocessing pipeline establishes standardized input processing for the FCNN.
The loading of the dataset works through TensorFlow and follows directory structure order. The images retain their corresponding labels as folder names while accuracy of classification remains critical since the medical dataset should prevent diagnosis errors. All the images undergo normalization through standard resizing to 150x150 pixels for consistent presentation within the dataset. When images are resized for the dataset, normalization maintains vital spatial features at a lower computational cost. The RGB image format determines that every input picture possesses dimensions (150, 150, 3).
Each pixel value goes through normalization through a division operation by 255.0 to reach a range between 0 and 1. The normalization process creates stability during training since it maintains input values within predefined parameters to stop gradients from becoming excessive enough to cause weight instability.
In contrast to CNNs, which maintain spatial structure, FCNNs need flattened input. Every image is reshaped to a 1D vector. The conversion enables the network to handle images as feature vectors at the expense of spatial information loss. Because this is a multi-class classification problem, the labels are one-hot encoded. The categorical labels, i.e., glioma, meningioma, pituitary tumor, and no tumor, are converted to a binary matrix representation where a different vector represents each class. This is because the SoftMax activation function in the output layer requires categorical representations.
To evaluate the performance of the model, the data is split into training and testing sets. The 90:10 ratio is applied so that the model gets sufficient data to learn and retains a considerable amount for testing. The training set is also shuffled to enhance generalization by preventing the model from memorizing the order of the data sample.
Through these preprocessing steps, we ensure that the data set is properly formatted and ready for use in training the FCNN model. The design choices and architecture of the neural network are described in the next section.

3.1.2 Neural Network Architecture

The FCNN used in this research is composed of several dense layers specifically crafted to extract useful features from MRI images. Our architecture is based on fully connected layers, unlike convolutional networks, which utilize spatial hierarchies to model complex relationships between pixel intensities. The input layer takes a flattened vector of 67,500 features, which are the preprocessed MRI images. This vector is the starting representation of the image, and it feeds into the next hidden layers. The network includes six hidden layers, each with a ReLU (Rectified Linear Unit) activation function that introduces non-linearity and increases model expressiveness:

Table 1: Architecture of the base FCNN model

Layer Name	Output Shape	Parameters
dense	(None, 512)	34,560,512
dense 1	(None, 256)	131,328
dense 2	(None, 128)	32,896
dense 3	(None, 64)	8,256
dense 4	(None, 32)	2,080
dense 5	(None, 16)	528
dense 6	(None, 4)	68

The output layer is the final layer with four neurons corresponding to the four classes of tumors. Finally, a SoftMax layer is applied, producing class probabilities with values that add up to one. The model predicts the label by assigning the most probable class.

3.1.3Interpretability Framework

We devise a structured methodology to track activations and store weight parameters to modularize the FCNN for enhanced interpretability. This enables post-hoc analysis of the neural network's decision-making process. TensorFlow's built-in functionality is used to systematically capture activation values at each layer during inference. Activations allowing us to investigate how features are transformed by each layer. This enables us to examine what features "light up" in response to the presence of tumor characteristics. All trained weight matrices and biases are extracted and saved for further analysis.
This approach facilitates the exploration of individual neurons in the classification process, allows for the analysis of the influence of dropout on the learned weight distributions, and enables the identification of regions where weight magnitudes may be skewed.
An interpretability framework is presented, which combines activation tracking and
weight analysis to provide a clearer understanding of the internal mechanisms employed by the FCNN. This framework is particularly beneficial to anyone who is willing to observe the deep operations going on inside FCNNs.

3.1.4 Model Training and Optimization

The training process is designed to optimize classification performance while ensuring
stability and generalization.
For this multi-class classification task, the categorical cross-entropy loss function is used:

$$L = -\sum_{i=1}^{C} y_i^* \, log(y_i)$$

where C is the number of classes, y∗i is the true label (one-hot encoded), and yi is the predicted probability for class i.
The Adam optimizer is selected for weight updates, with a learning rate set to 0.001. The model is trained using mini-batch gradient descent with a batch size of 32 for 40 epochs. After training, the model is evaluated using several metrics and techniques, including accuracy and loss monitoring, confusion matrix analysis, precision, recall, and F1-score calculation, as well as interpretability and feature importance tracking. These evaluation methods provide a comprehensive understanding of the model's performance and decision-making process.

3.1.5Activation Visualization and Interpretability Analysis

Activation patterns are visualized using heatmaps and bar plots. The methodology involves extracting activation values, normalizing and scaling the activations, and generating visualizations using Matplotlib and Seaborn. Analyzing these visualizations provides deeper insights into the decision-making process of the FCNN.

Additionally, a class-wise analysis of activation patterns was performed to add more interpretability to the analysis. Analysis methodology: Activation heatmaps were generated for the hidden layers of the FCNN for correctly and incorrectly classified samples of each of the diagnostic categories (glioma tumor, meningioma tumor, pituitary tumor, and no tumor) in a qualitative study. This method attempted

to find possible commonalities or differences or underlying discriminative patterns in the internal representations of the network by extracting representative samples and visualizing their class-wise activations. The observation of the evolution of activation intensity and distribution through layers for each class was in turn used to compare the feature abstraction process of the network. In particular, we wanted to establish if it was possible to visually assess any meaningful class-related patterns of activation, and how these differed depending on whether the predictions were successful or not. This class-wise viewpoint introduces added granularity to the interpretability framework, and contributes to the broader understanding of the behavior of the FCNN's learning, particularly in the case of high-stakes medical imaging for which model interpretability is paramount.

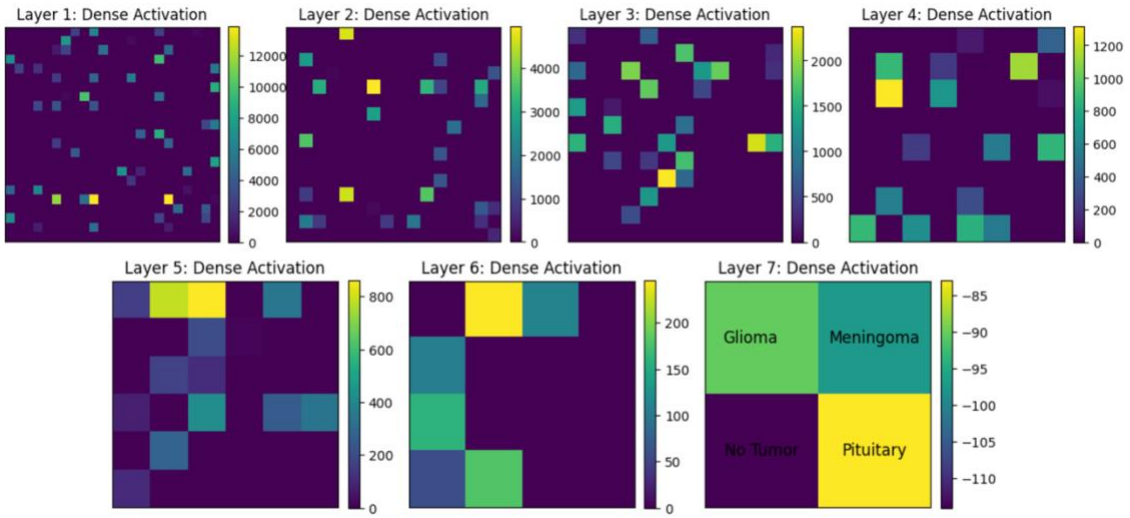

Figure 2: Heatmap representation of a FCNN with 6 hidden layers. Brighter colors represent more active nodes

3.2 Partially Connected Neural Network for Interpretability

The second approach explores an alternative neural network architecture that introduces partial connectivity within the dense layers. In contrast to the traditional FCNN, where every neuron in one layer is connected to every neuron in the next, this approach selectively reduces the number of active connections. The objective is to evaluate how this structural change impacts both classification performance and the model's interpretability.

Deep learning models, particularly FCNNs, often suffer from high parameter complexity, leading to challenges such as overfitting and reduced interpretability. By selectively limiting the number of neuron connections, this approach seeks to balance computational efficiency with model performance while enabling more granular analysis of how reduced connectivity influences feature learning. This is especially relevant in medical applications, where transparency in model decisions is crucial for user trust and clinical validation.

3.2.1Implementation of Partial Connectivity

Implementing partial connectivity is achieved by creating a customizable neural network function, create custom neural net. This function builds a network with a specified number of layers, neurons per layer, and a connectivity factor (denoted by ρ) ranging from 0 to 1:

• ρ = 1.0 corresponds to a fully connected layer (standard dense layer).

• ρ < 1.0 restricts the number of active connections by applying a probabilistic mask to the weight matrix.

The structure of the network is organized with three core modules. The input layer takes MRI images of dimension 150 × 150 × 3 and arrange it into a 67500-dimensional vector to fit the dense layers. After the input layer, three hidden layers are added to the model and fed a given number of neurons each (by

default set to 64 neurons per layer) with ReLU activation function. To impose desired sparsity, a weight mask is created at every hidden layer using the parameter The resulting mask effectively shuts down some neuron connections: This masked weight matrix is used during computation to encourage sparsification of the network's connectivity. Finally, the architecture ends with the output layer of four neurons with the SoftMax activation reflecting the four classes of tumors considered in the classification problem.The function dynamically adjusts connectivity per layer, allowing a direct comparison between different degrees of connectivity.

To systematically assess the impact of partial connectivity, we train multiple models with varying degrees of connectivity:

• Fully Connected ($\rho = 1.0$): The baseline model with standard dense layers.

• 75% Connected ($\rho = 0.75$): Each neuron is randomly connected to 75% of the possible neurons in the next layer.

• 50% Connected ($\rho = 0.50$): Each neuron is randomly connected to only half of the neurons in the next layer.

All models share the same number of layers and neurons per layer, ensuring that any differences in performance can be attributed solely to connectivity constraints, rather than other architectural variations.

Each model is compiled using the Adam optimizer, which dynamically adapts learning rates for stable **convergence**. The categorical cross-entropy loss function is used, which is suitable for multi-class classification, and accuracy is used as the evaluation metric, reflecting the proportion of correctly classified MRI scans.

The models are trained on the preprocessed MRI dataset for 10 epochs with a batch size of 32, and validation performance is monitored using a 10% test split. This methodology ensures that insights derived from model interpretability can be directly compared across different levels of connectivity.

Two further models were built to interpret the analysis inside this highly appealing architecture where this approach was applied in richer conditions of underfitting and overfitting. These models were generated by varying only the training length, while leaving the network architecture and a connection rate unchanged. The overfit model was generated by training the baseline model on fraction of the dataset for 100 epochs. This setup was designed to purposely force the network to memorize the training set, as such that we are able to observe how extreme parameter tuning affects the distribution of activation probability and generalization properties. Similarly, an underfit model was trained for only 3 epochs on the entire dataset. It is a method to simulate a situation of insufficient learning and gives information on early feature representation and on the origin of activation patterns before convergence. Through qualitative investigation of the activation heatmaps at the two ends—overfitting and underfitting—this approach aims to unveil how the dynamics of training play a role in feature abstraction and network interpretability. These experiments and their results (classification accuracy and activation visualization) are presented in the following section.

3.3 Activation Modification Approach for Post-Hoc Model Correction

Based on our first examination where many misclassification cases contained the true class in the model's top two predicted output probabilities, we speculate on minimum intervention on the neuron activation level that could nudge the model prediction towards the true label without retraining. This observation inspired us to systematically explore activation modulation paradigms to improve classification without any additional training during inference. The goal was to see if enhancing or suppressing activations of some neurons could increase the role of neurons that had been suppressed or enhanced resulting in an increase in performance of the model by negating the trade off. Thus, we thought to determine whether enhancing or suppressing activations of specific neurons could help the trained model better perform, wear instances in the test set were previously mis-classified.

Our early qualitative analysis of activation maps exposed an interesting feature: the majority of neurons are almost always quiet, across different layers. This effect held true for both appropriately classified and misclassified samples.

For models with varied generalization, i.e. models with fully connected networks and only 75% and 50% of the networks active, overfitted and underfitted versions we analyzed heatmaps as well. There were a lot of low-activity neurons, but directly pruning or strengthening them still failed to perform well. Neurons retired in the first layers can still have a subtle contribution to the decision boundary and full layers retraining could harm the learned feature representations. Accordingly, we aimed to create a less simplistic, but data-driven regularization procedure that adapts some selected neuron activities according to principled thresholds and optimal scaling.

The initial step in formalizing this method regulative of activation during training was to select a threshold criterion to separate "more active" from "less active" neurons in each layer. Instead of employing sophisticated statistical modeling, we settled on a simple, interpretable, and computationally efficient rule: for each layer, neurons of which activation values are larger than 10% of the maximum value of that layer were labeled as “more active”, and the remaining neurons were designated as “less active”. We were inspired to approach in this way, based on our heatmap analysis in which only a small percentage of neurons were consistently significantly more active, leaving the rest much lower in activity. The 10% level was chosen qualitatively to define the active region from visual observation of activation maps and verified systematically.

Once we had defined the criterion for finding the threshold, we conducted a grid search to find the best scaling parameter for "more active" and "less active" units. The objective was to find the pair of scaling multipliers *(active_scale* -one that amplifies the activations of active neurons and the other, *less_active_scale*, which amplifies or attenuates the suppressed activations), such that the update amount maximizes correction-rate in a validation set of misclassified samples. We trained pairs of decoder-based modifications with varying combinations of scaling factors and tested for the ratio of the corrected predictions (i.e., the modified prediction would result in a correct answer by the true label) to test empirically the best parameter settings.

The procedure of activation modification and assessment determination consisted of a few steps that we will go through in the next paragraph.

Selective Misclassification: We started out with selecting some of the misclassified samples of an original model on the training set. These data constituted the experimental training set for optimization and enabled the assessment of the efficacy of different scaling configurations.

Activation Modification Logic: We passed each misclassified sample through the network layer-wise. For each dense layer, we calculated the maximal activation value after each layer, and applied our preset threshold (10% of the maximal) to define neurons as “more active” or “less active.” The activations involving threshold were scaled by *active_scale*, and *less_active_scale,* respectively. This adaptation was carried out without changing the weights and structure of the network.

Evaluation with Modified Predictions: For the modified activations of all layers we computed the final output probabilities using a *SoftMax* function. We tested if the updated prediction rectified the prior misclassification. If it did, then it was considered a successful correction.

Grid Search for Scaling Factors: We swept *active_scale* and *less_active_scale* for a range of possible values. In particular, 10 values uniformly spaced between 1.0 and 5.0 were considered for active_scale, and between 0.0 and 1.0 for *less_active_scale.* The correction percentage of misclassified samples that was corrected while changing the activation of the network was registered for each pair.

Parameter Tuning: The pair of scaling parameters which yields the largest correction percentage was taken as the best one. These are the “right” balance between encouraging strong neuron activations and selectively amplifying weaker ones in order to steer the model towards more accurate prediction.

The code for carrying out this process is essentially organized around two major functions: *modify_activations*, the layer’s activations are passed on to *modify_activations*, where the scaling adjustments are calculated according to the thresholding logic. The function *evaluate_modifications* takes all the multiplicative combinations of the scaling factor, apply the modified forward passes over

the misclassified set and keep the correction rates, it will finally come back with the best parameters to use.
Having estimated the optimal scaling factors with the misclassified training samples, we conducted the same evaluation on the entire held-out test set. This step would apply to the generalizability of the trained activation modification strategy: can the parameters achieve performance improvements also on unseen data without any extra tuning if solely optimized on a small number of misclassified examples?
Despite the fact that the initial 10%-thresholding solution showed to be effective, there were several limitations in the following analysis. First, our use of a fixed threshold selected on visual inspection, while intuitive, was not well motivated theoretically, and may not transfer well across layers, or datasets, or models. In particular, the variability of activation distributions across layers gave an indication that a single threshold could be suboptimal, either failing to classify as low-activity neurons that will still perform important computations or exacerbating noise. Moreover, the use of a fixed threshold and manual grid search made the method stiffer, as it cannot dynamically adjust to the activation landscape in the samples or layers. These findings inspired the desire for a more principled, data driven methodology for thresholding and scaling activations and avoid the imposition of arbitrary cut-offs.

3.4 Experimenting different statistical methods to choose thresholds systematically

To systematically investigate these trends, we formulated 3 different strategies of thresholding and modification of neuron activations in every layer. All three approaches share an identical operating principle: they act only at test time, and they do not require any re-training nor back-propagation. This design allows us to keep the structure and parameters of the network fixed and only modify activation properties, to study their impact on model dynamics. We discuss each approach in more detail below, not only by outlining the computational steps, but also by describing its motivating principle.

3.4.1 L-moments Thresholding: A Robust Statistical Approach

Neural network activations are not uniformly distributed. In fact, they often exhibit skewness, heavy-tailed behavior, or outliers—especially when inputs contain significant variation, such as in image classification problems with real-world datasets. Standard measures like the mean and standard deviation are not robust to such irregularities. Consequently, relying on these traditional moments could result in thresholds that are overly sensitive to a few extreme activations or noise.
To address this issue, we selected Linear moments (L-moments), a family of robust statistics derived from linear combinations of order statistics.
L-moments is a considerable development in descriptive statistics and distribution fitting and provides robust alternatives to traditional product moments [16]. L-moments are expectations of linear Boolean combinations of order statistics and have desirable theoretical properties in that they are less susceptible to outliers and sample variability (especially for small sample sizes) compared to their classical analogues of skewness and kurtosis [16]. These properties can be exploited through the L-moments method of parameter estimation and distribution identification, particularly in hydrology for the frequency analysis of such extreme phenomena as floods and rainfall [17]. Once the L-moments of the sample have been determined and then compared to the L- moments from the theoretical sample L-moments of various distributions, typically shown as L-moment ratio diagrams, one can pick the probability distribution function (PDF) to be used to estimate the parameters of the pdf, with greater confidence than if other classical approaches are utilized, for example using minimum variance unbiased (MVU) methods, known as maximum likelihood (ML) or method of moments (displacement) methods, particularly when the data has a skewed distribution or n is small [17].

The first two L-moments—L1 (location) and L2 (scale)—are analogous to the mean and standard deviation but are derived using linear combinations of sorted data values, making them less influenced by outliers. Formally:
L1 (Location): Equivalent to the mean of order statistics.
L2 (Scale): A measure of spread based on the average difference between sorted data pairs.
For each dense layer, we flatten the activations across all neurons for a single input image, compute the L1 and L2 moments, and construct a symmetric threshold interval as follows:

$$\text{Lower threshold} = L1 - L2, \quad \text{Upper threshold} = L1 + L2$$

Neurons with activations below the lower threshold are considered underactive and their outputs are multiplied by a scaling factor *multiplier_less* (e.g., 0.5). Neurons above the upper threshold are deemed strongly active and are scaled by *multiplier_more* (e.g., 2.0). Activations that fall within the interval remain unchanged.
Several considerations motivated this choice of L-moments for modifying the activation. Firstly, L-moments provide robustness, given that they are less sensitive to extreme values or skew distribution, thus being appropriate for capturing diverse activation profiles between layers. Second, the interpretability one is crucial, and l-moments give us a nice intuitive characterization of the "center and spread" of activations that generalizes easily to layers whose activations have different ranges. This feature allows a clear insight into the activity of neurons and its connection to the model predictions. In addition, L-moments are flexible in that they can define different thresholds for each image-layer pair allowing a local interpretation of the activation pattern with respect to particular inputs. Finally, consistency of L-moments allows to use the same thresholding rule uniformly across layers without having to adapt statistics per layer. This technique is implemented in practice using several repetitive steps that go through the layers independently: (1) Flatten the activations, (2) compute the L1 and the L2 moments, (3) compute the thresholds, (4) apply the multiplicative updates. These modified activations are then passed to the front through network. This architecture offers a natural framework to study the downstream impact of selective activation modifications for at each layer we can follow the influence of these modifications through the network and to the final classification decision.

3.4.2 GMM Thresholding: Modeling Latent Activation Structure

Another typical behavior in deep neural networks are bimodalities in the activation distribution. Some neurons may be silenced but others may be fired more strongly depending on the input features. This naturally leads to a dichotomy of “dominant” and “recessive” activating arms. Instead of defining arbitrary cutoffs, we aimed to model this division, probabilistically, by a GMM.

Gaussian Mixture Models (GMMs) are an important class of probabilistic models used to describe heterogeneity in data and assume the observed data arises from a finite mixture of Gaussian distributions.
Formally, a GMM defines the probability density function $p(x \mid \theta)$ *for a D-dimensional random variable* x *as a convex combination of K Gaussian components:*

$$p(x \mid \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

where K is the number of components $\mathcal{N}(x \mid \mu_k, \Sigma_k)$ *denotes the multivariate Gaussian probability density function for the k-th component parameterized by its mean vector.*

We further selected GMM as the method for analyzing activations in the neural network for numerous benefits such as its data-driven nature, power of representation of latent structures, adaptability to different activation configurations and its soft probabilistic interpretation.

The data-driven nature of GMM is the main motivation for our choice. Conventional approaches of threshold selection are often based on a priori specified cut-offs or recipes, which could fail to model the entire complexity of activation behavior. These thresholds are generally estimated by domain experts or empirical testing, and may be ad hoc and not applicable to different datasets or any network structure. In contrast to GMM, there is no need for such prior thresholds. Rather it is data-driven and thus a more robust and scalable method of activation analysis. As GMM considers activations as a realization of a probabilistic model, it naturally finds the distribution of the data so that the thresholds are found based on the behavior of the observed neurons.

Another important advantage of GMM is that it is able to find the hidden structure in the activation data. In the domain of neural networks, the firing rates from neurons do not distribute uniformly; it may develop into multi-extended forms that has more than one-shape in the population activities of neurons. For example, some neurons can be highly active (having high relevance to the decision made by the model), while others can remain inactive or feature low- level activity. GMM is successful in finding these modes in the data and in capturing the bimodal shape of neuron activation (neurons "on" and "off"). The recognition of such hidden structure is essential for interpreting how the neurons participate in the network's predictions, and also opens the way for selectively controlling their activity in a more fine-tuned manner. Being able to model activations as a mixture of Gaussian distributions enables GMM to naturally divide neurons into states, rendering it easier to understand how each individual neuron behaves across different images.

One advantage of GMM is that it is adjustable to different activation landscapes that may encounter from different input images. Because input features and learned representations of the network are different, each image that is presented to the network may activate different pattern. A single threshold value (as in conventional approaches) may not be sufficient to take this variability into consideration and may not generalize with valid interpretability to the different samples. GMM, on the other hand, adaptively tunes the thresholding to the special characteristics of the activations in each image and is therefore is custom to the data. This adaptability further guarantees that activation scaling is meaningful for the particular input, and considerable for the resulting effect of the entire network, which further contributes to the interpretability of the model.

Lastly, the probabilistic nature of GMM enables a softer interpretation of the neuron scaling. Rather than applying a strict binary threshold (e.g., active or inactive) the GMM offers a probabilistic framework in which each neuron is associated with a likelihood of belonging to a particular state or cluster. This probabilistic interpretation is also especially useful for explaining neural networks, because it provides some flexibility to the degree of activation change that can be made. Instead of only rescaling a neuron by whether it exceeds/falls below threshold, GMM enables partial rescaling based on the probability the activation is part of a given cluster. This allows for more detailed analysis of the network's behavior, and for perturbation of neuron activations that are more fine grained than simply "on" and "off" during interpretability experiments.

For the application of GMM on analysis and scaling of neuron activation, we use the scikit-learn library for GMMs. The steps can be easily composed into a clear, systematic protocol, which are seamlessly embedded into the workflow of the inference phase of a neural network.

For each entry image, we first flatten the activations from a certain layer of the neural net into a single guess vector. This conversion is needed as GMM expects the data points to be in an array form and such that each row represents a single data point and the columns are the features of the data point. The activations are usually structured as multi-dimensional arrays (matrices or tensors), but by flattening them, one can efficiently inspect the activation values.

After activation flattening, fitting of a two-component GMM to the activations is the next stage. The two-component assumption is motivated by the intuition that in many (deep) neural networks, the activations follow a bimodal distribution – one mode is composed of the neurons where the activations

is significantly low and the other is composed of those that are inactive (or low active). Through fitting the GMM, the model estimates the parameters (mean, variance and weights of mixture) of the two Gaussians that are most suitable to describe the observed activations for the input image.
For finding the threshold that is enough to activate the neuron for each given input, we set the two Gaussian PDFs to be equal. This equality indicates that the point at which the neuronal likelihood in one of the two modes becomes equal with the other one. This is the point where very active neurons are being separated from those that don't contribute so much to the output of the network. By computing this threshold, we are then able to split neurons by high or low relevance.
After the threshold is determined, the activations are scaled by their probabilistic match to one of the two Gaussian modes. Significantly active neurons, neurons whose activations are above the threshold, can have their activations scaled up to make them even more significant in the final prediction. On the other hand, neurons with activations below the threshold are considered as not very active and may be reduced or even turned off. The scaling factor is based on evidence that a given neuron is part of the high-activation Gaussian.
This method provides a solid foundation for the extrapolation of neurons with statistical modeling, based on the natural clustering of the activation data. It offers a way of understanding how activations are spread across layers, and how this distribution affects the network predictions. The flexible basis of neuron activations in terms of these probabilistic clustering not only improves the interpretability of neural networks but also provides flexible manipulation of neuron activations, which is important for tasks such as improving model transparency, identifying misclassifications, and enhancing the model performance by targeted activation modifications. We can enrich the interpretability by GMM and get insight of the intrinsic structure of neuron activations, and utilize targeted alteration to boost model performance. The probabilistic formulation under GMM gives an advanced, data-driven perspective of how neuron scaling should be performed, keeping in mind the desire to make decision-making processes more transparent and interpretable in neural networks.

We will compare these 2 methods with the original qualitative based thresholding method we asserted in the previous sections.

4 RESULTS

The initial evaluation of the FCNN trained for brain tumor classification yielded promising results in terms of classification performance. The final training accuracy achieved was **0.8212**, with a corresponding training loss of **2.2251**. When evaluated on the unseen test dataset, the model attained a **test accuracy of 0.7982** and a **test loss of 2.2983**. While these metrics indicate a relatively strong ability to generalize, they also reveal the existence of non-trivial classification errors that merit further investigation, particularly in a high-stakes domain such as medical diagnosis. Since we are only interested in interpretability of the neural network at this moment, original model performance is not the first priority.

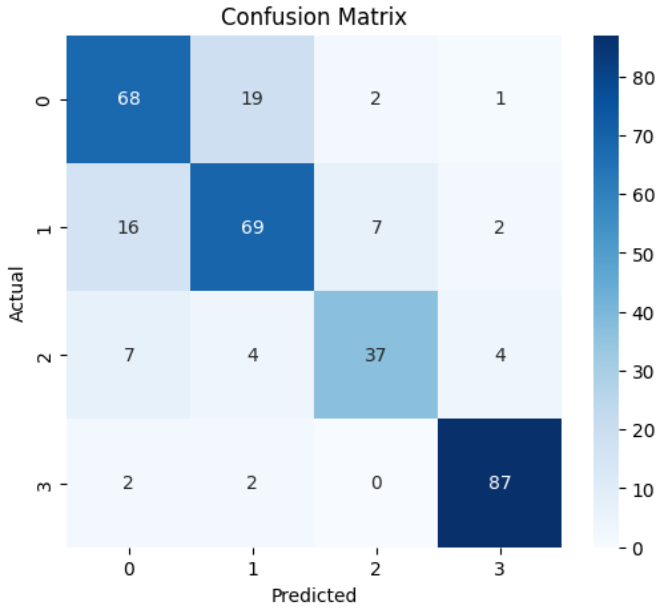

Figure 3: Confusion matrix for the test set

It is visible from the Figure 3 that model confuses the tumor index 0, which is glioma, and tumor index 1 which is meningioma. This observation leads us to do a class wise analysis of the FCNN to perhaps find certain pattern.

4.1 Qualitative Analysis of Layer wise Activations

A detailed activation heatmap visualization method was employed throughout the dense layers as part of interpretability analysis on correctly and incorrectly classified images under study. The evaluation goal for this part was focused on detecting visual indications that would show either class-specific patterns or recurring processing flaws in the model.

Our evaluation of properly recognized images revealed sparse low-intensity activation patterns in the Dense, Dense 1 and Dense 2 dense layer area of the code. The number and intensity of activating signals increased prominently as the model moved through Dense 3 through Dense 6. The neuroimaging patterns maintained abstract qualities even though they failed to establish any visible link between anatomical MRI features. The Dense 7 output layer produced precise peaks indicating the predicted class results demonstrating the final classification choice.

Glioma tumor samples correctly identified by the model demonstrated mostly minimal activations with low signal strength throughout Dense, Dense 1 and Dense 2 layers. The activation signal increased moderately with each progressive Dense layer from 3 to 6 while propagating deeper. The activated areas in these regions did not maintain spatial arrangement or correspondence with detectable MRI input control details. The Dense 7 output layer presented dominant neuron activation which supported correct classification of “Glioma”.

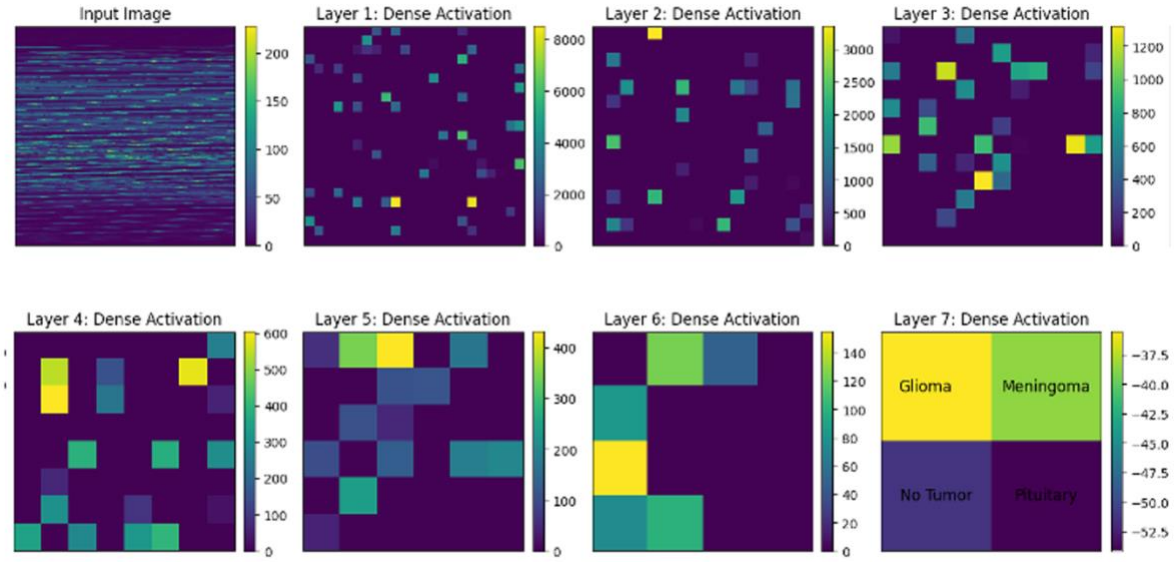

Figure 4: Heatmap representation for a glioma sample

How the neural net processed meningioma tumor samples showed patterns equivalent to the activation patterns for glioma. Early layers in the model mostly activated at low levels but contained occasional separate areas of intense activation. The emerging activation patterns became more defined when the input reached Dense 3 until Dense 6 but these patterns failed to demonstrate uniform spatial patterns during evaluation of multiple samples. Correct classification emerged from the output layer through maximum activation in the “Meningioma” output neuron.

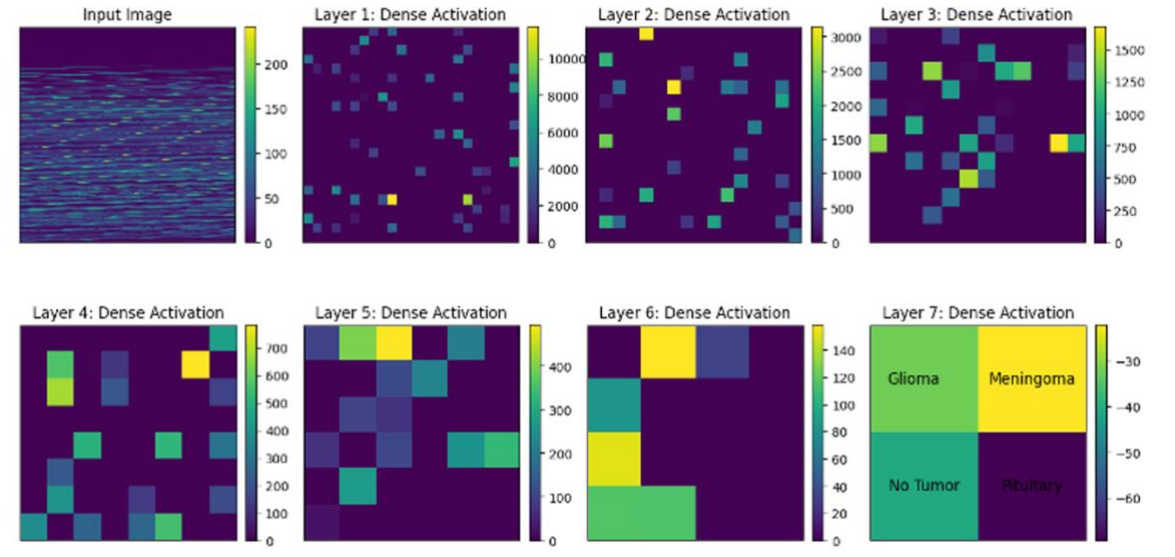

Figure 5: Heatmap representation for a meningioma sample

Correct pituitary tumor cases and samples without tumors followed the same pattern as observed throughout other tumor classes. From the beginning layers to the final ones sparse activations emerged while activation zones developed into stronger concentrations in middle and late stages. The output layer activations maintained reliable indications about the right class in cases with correct predictions. Previous layer neural activations in this model failed to provide sufficient insight about the decision-making process due to its inherent characteristics for processing spatially complex medical images.

4.2 Isolation and Visualization of Misclassified Samples

The identification process for model errors followed automated methods that identified indices of all misclassified samples present in the test dataset. An independent examination was performed on the 215 extracted samples. We examined activation heatmaps throughout all layers for each misidentified image and recorded its predicted class together with the true class label and measurement values from the output layer.

An ongoing pattern existed throughout numerous misclassified mistakes where the correct class was not triggered by the top activation neuron, yet the following most active neuron showed regularity. The model performed Revelation revealed an underlying predictive ability which only partially matched the correct class even if it did not determine the top-1 prediction. Analysis of output results showed sufficient evidence for creating a hypothesis which stated that in many faulty predictions the model's second-choice neuron around the final layer activated the correct diagnosis.

The validity of this theory received experimental confirmation through a utility function that analyzed the output node activations across all incorrectly categorized images. We obtained the second-highest activation value from each incorrect prediction then checked if the associated class match the original diagnostic information. Analysis results confirmed our initial observation since the proper classification was shown twice as often among the top-ranked activations in the output layer.

This is a significant finding. Research indicates that even though the model failed to achieve highest confidence for the correct category it effectively determined the proper class among its options. The medical diagnostics field could use this hidden sensitivity to develop methods for improving classification accuracy by restating predictions from second-place options statistically.

Insights derived from the second-choice prediction pattern made it possible to enhance model accuracy without modifying the current model design. A hypothesis was underway to examine whether certain model mispredictions could be adjusted by using alternative predictions during specific situations. In high-risk applications where the top-1 confidence level of the model drops below an established threshold the system should evaluate the second-ranked class as a possible outcome.

An ongoing post-hoc correction method stems from this initial hypothesis which we will explain in detail in the coming section. The approach utilizes interpretability techniques in a unique manner to guide the improvement of prediction models by enabling specific error correction methods.

4.3 Impact of Partial Connectivity on Model Performance and Interpretability

This section provides extensive research results regarding how different levels of connectivity affect neural network designs for multiclass brain tumor classification of MRI images. The research investigated three structural models beginning with (fully connected model) then examined the variant containing 75% connectivity (next model) before concluding with a model organized with 50% connectivity (last model). All examined models kept the same entrenched characteristics including 10 training epochs with 32 batch size along with categorical cross-entropy loss function and Adam optimizer and ReLU and SoftMax activation functions throughout the 64 neuron layers. We studied both predictive performance outcomes and interpretability behavior of model representations as we used systematic connectivity variations.

4.3.1 Training and Validation Performance

Throughout the training procedure we discovered meaningful correlations between model connectivity values and how learning took place along with generalization results. The fully connected model obtained the best training accuracy of 90.84% because its unrestricted weight space enabled superior training data memorization. The highest validation performance was not obtained by this advantage despite achieving validation accuracy at 78.90% while maintaining validation loss at 0.6289. The 75%

connected network delivered superior validation results although its training performance stopped at 88.63%. The validation performance endpoint of this model reached 82.87% while its lowest validation loss stood at 0.5265. The network uses partial connectivity for regularization because it limits overfitting and allows for improved generalization on unseen data. Training accuracy for the 50% connected model was 88.74% but validation accuracy reached only 68.20% while validation loss reached its highest point at 0.8858. A lower threshold for network connectivity becomes essential because the model then fails to process complex input data effectively. Moderate network sparsity helps generalization but excessive network sparsity creates a negative impact on generalization effectiveness.

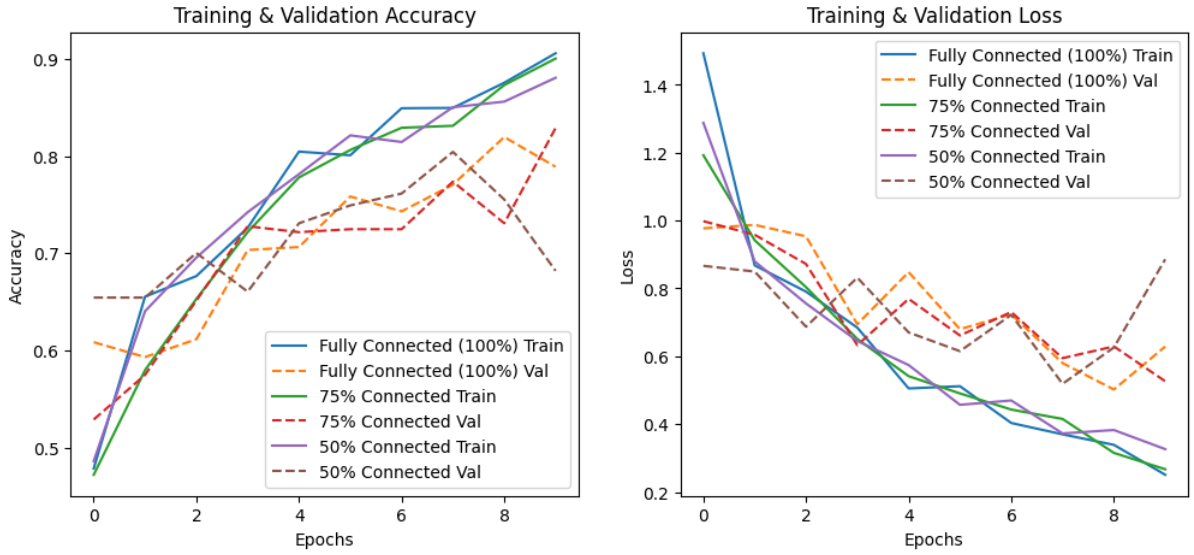

Figure 6: Comparison of the performance for different configurated models

4.3.2 Generalization and Class-Specific Evaluation

The evaluation of model generalization behavior utilized detailed predictive measures on test data reserved exclusively for this assessment. The evaluation calculated correct predictions between the four tumor categories that included glioma, meningioma, pituitary and no tumor. The class-based separation in the analysis let us evaluate whether partial connectivity influenced all groups fairly or it predominantly affected particular classification boundaries. Performance from the fully connected model was balanced yet its ability to detect 'no tumor' cases remained inferior compared to the 75% connected model. The connecting only 75% of brain regions led to enhanced classification ability between categories while showing reduced difficulty in separating objects from 'glioma' and 'no tumor' classes. Testing on different tumors revealed that the 50% connected model failed to achieve acceptable results in both 'meningioma' and 'pituitary' tumors because their classes had substantial distribution overlap. Research validates that networks which have very few connections cause a degradation in the ability to model crucial visual cues between different classes.

4.3.3 Implications for Interpretability

We introduced partial connectivity to achieve better interpretability because it reduces representational redundancy in features while making features more sparse. A strict neuron connection limit served our objective to obtain more distinct activation patterns which enabled both analysis and identification of crucial image areas.

The analysis of activation heatmaps seemed to verify that the 75% connected model created distinct and less entangled activation patterns before the output layer while no formal interpretability measures were used in this section. The heatmaps from the 50% connected model were sparser yet they provided limited information because the model likely lacked adequate learning capacity. The fully connected model generated dense and diffused activation patterns that made feature identification challenging for prediction determination.

4.3.4 Summary of Empirical Findings

The evaluation process of three network topologies including fully connected $\rho = 1.0$ versus moderately connected $\rho = 0.75$ and sparsely connected $\rho = 0.50$ reveals important findings about network density and model performance and interpretability aspects. These results show that connectivity with a moderate level of sparsity ($\rho = 0.75$) results in improved generalization results. The middle-level inter-neuron connection density ($\rho = 0.75$) achieved better validation outcomes than denser and sparser configurations as a controlled sparse network structure acted as a successful regularization approach. The fully connected model achieved superior training accuracy but displayed overfitting behavior since its validation metrics turned out lower than other models. The training behavior of highly expressive networks becomes a known challenge when they memorize training patterns instead of adapting to generalization particularly in cases with scarce or noisy datasets. Validation accuracy experienced a major decline when sparsity was set to excessive levels ($\rho = 0.50$). This shows that a minimum threshold exists for network connectivity which enables proper learning of complex decision boundaries especially across multiclass classification needs that mandate sophisticated feature discrimination. Initial analysis of activation layer heatmaps shows that intermediate sparsity levels potentially improve how easily the network can be interpreted. Spatially coherent and interpretable internal representations developed out of reduced connectivity structures in the $\rho = 0.75$ model than in its fully connected counterpart. The research demonstrates that designing neural network connectivity effectively improves both classification generalization levels and interpretation clarity in high-risk medical imaging domains.

4.4 Comparative Analysis of Neuron Activation Patterns in different configurations

A distinct pattern emerges from performance evaluation which reflects the planned training strategies. The baseline model establishes a satisfactory test accuracy level at approximately 82% which demonstrates its ability to maintain a smooth learning capability for minimal and maximal training regimes. An overfitted model demonstrates test accuracy of approximately 35% and shows an extremely high loss at the same time. The dramatic decrease in generalization power demonstrates overfitting even though the model achieved probable total accuracy on the specialized training data which the methodology implies. Memorization of training data specificities combined with individual noise has made the model ineffective for new instances it encounters in practice.

The underfitted model demonstrates even worse performance because its test accuracy stays below 30% along with an abnormally high loss value. The model's inadequate learning of proper representations from limited (3 epochs) full dataset training explains its poor predictive performance.

The activation patterns of FCNNs in medical imaging with MRI brain tumor data were analyzed through visual methods across all dense layers of three different models including the baseline model and underfit and overfit variants. Neuron activation heatmaps of meningioma tumor case data were generated for each dense layer in all three models through consistent visual representation. This allowed us to track information behavior throughout the network.

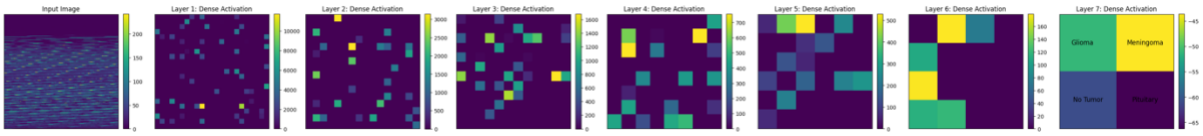

Figure 7: Balanced (base) neural network

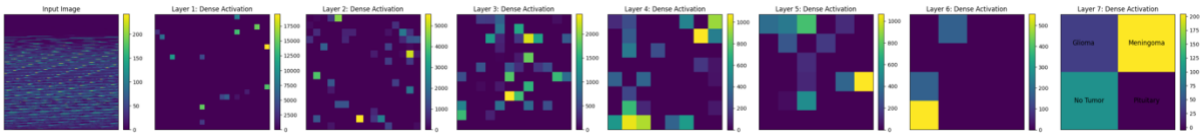

Figure 8: Overfit version of the neural network

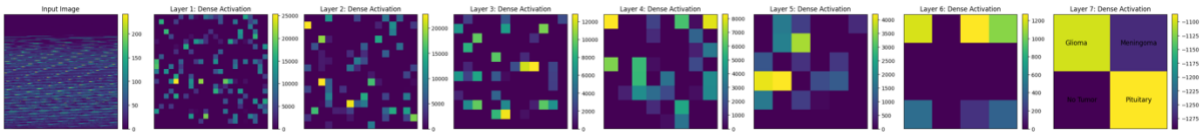

Figure 9: Underfit version of the neural network

The comparative examination of these activations reveals profound differences in the internal representation, sparsity, and distribution of neuron activity, which reflect the generalization capacity and discriminative behavior of the models.

The reference model shows balanced hierarchical activation patterns during its analysis process. A typical MRI slice enters the dense architecture as the initial input image containing fine-grained texture along with small structural differences. Layer 1 displays an activation pattern keeping most neurons below baseline levels and showing only limited significant activation. The model demonstrates sparse firing patterns because it detects the most important low-level features while filtering out unnecessary information. The action potential detected in Layers 2 and 3 demonstrates moderate rise in response from additional neurons which signifies progressive enhancement of feature representations. Layered neural networks typically build abstracted representation patterns that start from basic features which were previously detected. The network maintains low-density data structures even though it has improved its representational complexity thus demonstrating effective encoding of features.

Layer 4 through 6 performs activation re-consolidation which minimizes neurons that respond with high magnitude. The reduction of features and dimensional simplification occurs during this phase because it keeps only the most important discriminative features that drive final decisions. The final output logits at Layer 7 show a clear prediction choice for “meningioma tumor” through strong activation of its corresponding neuron. The output configuration matches the ground truth data which verifies that the model successfully transforms important features into an accurate prediction. The model retains focused attention and abstracting capabilities because it showcases characteristics of both inter-layer sparse connectivity and intra-layer selective activation which lead to strong generalization abilities.

The overfit model delivers distinct activation patterns because it receives intensive training across restricted training data which lacks proper regularization techniques. The activations in initial sections particularly Layer 1 and Layer 2 appear very sparse because only one or two neurons show extremely high values. The model obtains notable overfit status because its hyper-selective nature indicates it has memorized specific patterns without developing generalizing abilities. The excessive activation values indicate the saturation of neural cells that control what emerges from the layer which potentially results in information processing constraints. A signal going through Layers 3 to 5 maintains its pattern of intense narrow activations while showing minimal ability to recognize different features. The extremely low activity in Layer 6 demonstrates a situation where almost all neurons remain inactive but only a

few show any activity. The substantial reduction in active neurons indicates that a comprehensive range of features disappears while operation relies on a minimal set of retrieved cues.
The model shows the correct class output as “meningioma tumor” in Layer 7 while manifesting strong preference for this prediction. The output logits demonstrate strong polarization because they miss subtle details while showing excessive confidence. The model shows this effect during overfitting when prediction confidence surges because it learns features by rote memorization instead of knowledge acquisition. The correct classification of Layer 7 does not demonstrate robustness because the model runs a risk of failure within simple distributional distortions or noisy inputs due to its minimal feature learning potential. In real-world applications the excessive dependence on few neurons destroys system interpretability and reliability because it leaves the neural network vulnerable to adversarial attacks.
An underfit model shows diverse behavioral patterns because it results from insufficient training or inadequate capacity or suboptimal hyperparameters. Activation density in Layer 1 starts off at a high level since neurons remain active throughout the entire layer space. From Layer 2 to Layer 6 the model maintains broad informational spread indicating that the model does not recognize pertinent information among other useless data and displays minimal specialization between layers. Numerous active neurons seem to indicate high responsiveness but really represent inefficient and confused system behavior. The model functions by spreading every available feature without discrimination which creates a confused unstructured internal state of information. The inadequate data-guided representation of the model leading to uninformative feature maps defines this underfitting condition.
In Layer 7 of the underfit model the predicted class is marked as "pituitary tumor" which shows false categorization together with logits of low values and equal distribution among classes. The model's inability to discriminate important features for classification and weak class separation demonstrates that it has failed to learn proper discriminative capabilities. An underfit model presents a flat feature representation that illustrates unprecise and non-focused feature sensitivity unlike the sophisticated sensitivity of an overfit model. The widespread activation of neurons across layers which initially appears active actually indicates that the model fails to effectively select meaningful features and establish hierarchical abstraction.
An underfit model shows diverse behavioral patterns because it results from insufficient training or inadequate capacity or suboptimal hyperparameters. Activation density in Layer 1 starts off at a high level since neurons remain active throughout the entire layer space. From Layer 2 to Layer 6 the model maintains broad informational spread indicating that the model does not recognize pertinent information among other useless data and displays minimal specialization between layers. Numerous active neurons seem to indicate high responsiveness but really represent inefficient and confused system behavior. The model functions by spreading every available feature without discrimination which creates a confused unstructured internal state of information. The inadequate data-guided representation of the model leading to uninformative feature maps defines this underfitting condition.
In Layer 7 of the underfit model the predicted class is marked as "pituitary tumor" which shows false categorization together with logits of low values and equal distribution among classes. The model's inability to discriminate important features for classification and weak class separation demonstrates that it has failed to learn proper discriminative capabilities. An underfit model presents a flat feature representation that illustrates unprecise and non-focused feature sensitivity unlike the sophisticated sensitivity of an overfit model. The widespread activation of neurons across layers which initially appears active actually indicates that the model fails to effectively select meaningful features and establish hierarchical abstraction.

This section presents the experimental results of the activation modification approach developed to enhance neural network performance post-training. Through a structured methodology involving threshold-based activation scaling and systematic parameter search, we investigated the effectiveness of activation manipulation across different evaluation sets, including misclassified samples, the full test set, and models trained under varying regimes. Our findings demonstrate the surprising power of simple activation interventions to boost model performance without retraining.
Using the misclassified sample set as the experimental group, a grid search was conducted over the scaling factors applied to more active and less active neurons. The scaling factors for more active

neurons were explored in the range of 1.0 to 5.0, while those for less active neurons were varied between 0.0 and 1.0, each divided into ten evenly spaced increments. The optimal parameter combination, which yielded the highest correction percentage on the misclassified set, was identified as a more active scale of 1.0 and a less active scale of 0.44. This result suggests that amplifying the already strongly activated neurons was unnecessary, while selectively boosting weaker activations by a moderate factor significantly improved the model's ability to correct its predictions. This outcome aligns with prior qualitative heatmap analyses, which indicated that while many neurons remained largely inactive, they nevertheless carried valuable latent information that could influence the decision boundary when appropriately emphasized.

Applying the discovered scaling parameters to the misclassified set led to a remarkable performance improvement. Initially, these samples were incorrectly classified by the baseline network. After activation manipulation, the correction percentage — the proportion of misclassified samples reassigned to their correct labels — reached an impressive 90%. This finding validates the initial hypothesis that latent decision signals exist within the network and can be exploited post-hoc through controlled activation adjustment. It demonstrates that many misclassified instances were already close to the correct classification boundary and required only minimal intervention at the activation level to achieve the correct outcome. Furthermore, it highlights that such improvements can be achieved through a computationally lightweight and non-invasive process, without retraining or altering the model's learned weights.

Encouraged by the success on the misclassified set, the activation modification strategy was applied to the full test set, which included both correctly and incorrectly classified samples. The original baseline test accuracy, measured without any modifications, was 79%. After applying the activation modification with the optimized scaling parameters, the test accuracy increased to 82%, representing a 3% absolute improvement. While the improvement magnitude on the full test set was less dramatic than on the misclassified set, it remains noteworthy. The enhancement confirms that activation modification is not restricted to isolated correction of misclassifications but can generally improve model decision-making across the entire dataset. It also illustrates that the technique preserves correct predictions while rescuing some errors, offering a low-risk method for enhancing model reliability after deployment.

Although the gains on the full test set were moderate, it raised an important question regarding the dependence of the activation manipulation benefit on the baseline model quality. Specifically, it prompted the investigation of whether larger improvements could be observed in models that were initially less well-trained. To explore this, a new model with the same architecture was trained for only 10 epochs, significantly reducing its learning period compared to the baseline model. The newly trained model achieved a test accuracy of 59.99% without activation manipulation, presenting an opportunity to evaluate the potential of the modification approach under conditions of underfitting.

Application of the same activation modification strategy to the 10-epoch model resulted in a dramatic performance increase. The test accuracy rose by 14.37 percentage points, reaching approximately 74.36%. This substantial improvement, achieved without any retraining or additional gradient updates, suggests that even undertrained models can harbor sufficient latent structure to support accurate predictions if activation patterns are adjusted appropriately. In practical terms, the activation modification enabled the 10-epoch model to achieve performance levels comparable to models trained for much longer periods. It revealed that while the training process had laid down useful representations, the full potential of these representations was not realized until the activation distribution was adjusted at inference.

To further contextualize these findings, the same architecture was trained for 20 epochs. This model achieved a baseline test accuracy of 73.57% without activation modification. Intriguingly, this baseline accuracy was comparable to the post-manipulation performance of the 10-epoch model, suggesting that activation manipulation could compensate for approximately 10 epochs of training. When activation manipulation was applied to the 20-epoch model, its test accuracy increased by an additional 4.58 percentage points, reaching approximately 78.15%.

Table 2: Accuracy improvement for different model configurations

Model Configuration	Baseline Accuracy	Post-Modification Accuracy	Improvement
10 epochs	59.99%	74.36%	+14.37%
20 epochs	73.57%	78.15%	+4.58%

The comparative results indicate several important trends. First, activation modification consistently improved performance across all models, regardless of their initial accuracy. Second, the magnitude of improvement was inversely related to the baseline model quality: models with lower initial accuracy benefited more from activation manipulation, while models already performing near their capacity saw more modest gains. Third, activation manipulation was able to yield improvements comparable to or even exceeding those achieved through extensive additional training epochs.

The implications of these findings are significant. In environments where computational resources are limited, retraining is expensive, or rapid model updates are required, activation manipulation provides a powerful and efficient alternative. It enables performance improvements by leveraging latent features within the model, rather than relying solely on further optimization of weights through backpropagation. The technique is particularly appealing for post-deployment settings where access to original training data may be restricted, or regulatory requirements limit model retraining.

Despite the promising results, several limitations must be acknowledged. The activation manipulation approach relies on the careful selection of scaling parameters, and the optimal settings may vary across datasets and architectures. Although the grid search procedure employed in this study was effective, it is computationally intensive and may not be scalable to larger models or broader applications without further refinement. Moreover, the experiments were conducted exclusively on FCNNs applied to brain tumor MRI classification. It remains an open question whether similar improvements can be achieved in convolutional, recurrent, or transformer-based architectures, or across different types of tasks such as object detection, language modeling, or time series forecasting.

Another consideration is that while activation modification enhanced classification performance, it also introduced an additional inference-time computational step, which, although lightweight, may still be a constraint in real-time applications. Furthermore, while the modifications were empirically effective, their interpretability at the feature level remains to be fully elucidated. Understanding precisely how modified activations reweight learned representations to achieve better decisions is an important avenue for future research.

Nevertheless, the experimental findings provide robust support for the viability of activation modification as a simple yet powerful strategy for improving model predictions without retraining. The method is general, architecture-agnostic at the conceptual level, and highly interpretable in terms of its operational mechanics. It opens up new opportunities for post-training model optimization, particularly in critical applications such as medical imaging, where model trustworthiness and rapid adaptability are of paramount importance.

In conclusion, the activation modification experiments demonstrated that even small, carefully designed interventions at the neuron activation level can lead to substantial gains in model accuracy. By selectively amplifying underactive neurons and preserving dominant activations, it is possible to harness the latent predictive potential already embedded within the network. This study confirms that neural networks often possess "hidden" performance reserves that can be unlocked through strategic post-hoc manipulation, offering a new paradigm for model improvement that complements traditional training-based approaches.

4.5 Analyzing the results of different statistical thresholds

A detailed examination was carried out to ensure the reliability the FCNN performance, its misclassified predictions in the testing dataset. The network contained useful information internally

while making slightly wrong decisions in the end. A customized percentiles-based activation adjustment process was exclusively implemented on the misidentified cases. We used different (m_more, m_less) activation adjustments to modify neuron responses which improved the SoftMax output
toward correct classes in 80% of these situations. Selective activation management seemed promising on these running-on-test-cmds which motivated the researchers to further apply it on the whole test dataset.

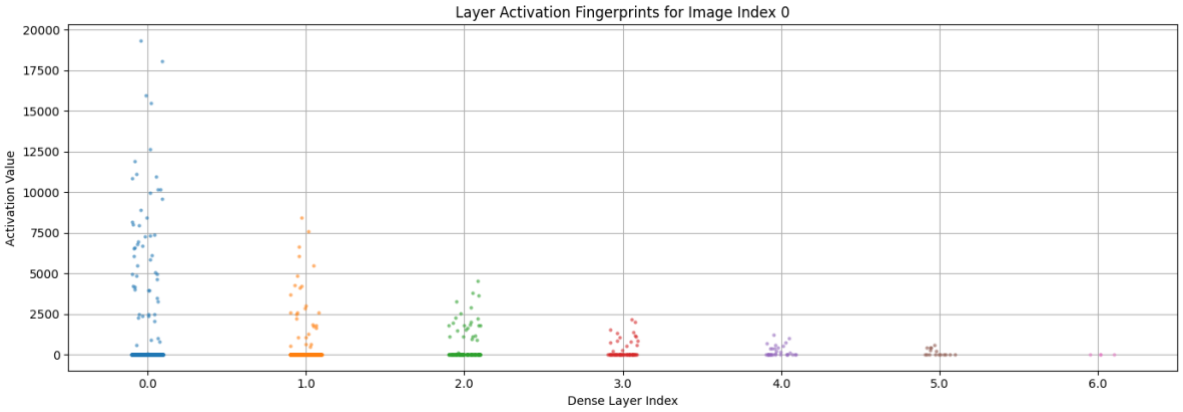

Figure 10: Layer activation distribution for a sample image

Experiments were executed to evaluate the functional impact of neuron activation manipulations during inference across various thresholding strategies and scaling configurations. The network's classification accuracy on a held-out test set of 100 samples was recorded for each combination of multiplier_more (m_more) and multiplier_less (m_less). We compared four methods: L- moments, GMM, random, and custom.

The net behavior remained unchanged when the scaling factors were set to unity as expected (m_more = 1.0, m_less = 1.0). All methods scored the baseline accuracy of 87% under this setting, which means that the intervention framework was functionally equivalent to the unmodified model in the absence of scaling.

Moreover m_less = 0.0 configurations that suppressed entirely weak activations or m_more = 0.0 that failed to boost dominant activations consistently degraded performance for all thresholding techniques, confirming that both weakly and strongly activated neurons contribute meaningfully to the decision boundary.

The L-moment-based method displayed a marked improvement in classification performance when the gap between multiplier_more and multiplier_less was lessened. More specifically, although m_more = 2.0 and m_less = 0.0 resulted in only 32% accuracy, increasing both multipliers to 2.0 raised this value to 87%. The reasons behind this gradual increase lie in the fact that the L-moment-based classification of neurons into strong and weak counterparts catches biologically important patterns, mostly when they are considered symmetric. Nevertheless, this method was somewhat sensitive to the unbalanced multiplier scaling, which indicates that to a certain extent, its robustness requires cautious hyperparameter tuning in order to not suppress the activity of the “correct” neurons.

GMM approach displayed a broad range of effectiveness (from 26% to 87%). It was predictable that it could achieve a baseline performance in certain cases (e.g. (m_more = m_less = 1.0) or (1.5, 1.0)), but more striking was that it was incredibly sensitive to outliers, especially under the symmetric high boosting settings (e.g. 2.0, 2.0) where it was a mere 43%. This variability implies that even though the GMM method was hypothesis-test motivated, it may not always lead to thresholds that are functionally relevant, especially if the population of activation is not bimodal.

It is worth noting that a random strategy was sometimes either on a par with more structured methods, or even surpassed them. In a number of configurations, for instance, (2.0, 1.0) and (2.0, 2.0), it reached an 87% accuracy rate. However, at the same time, the strategy had a significant variance, falling to 26% under other conditions. It is possible that the stochastic strategy is a proof that the network contains

some redundant functions – randomly changing neurons might lead to more effective configurations at a few points, but it also has been very inconsistent.
Custom method based on percentile showed the highest consistency and robustness in all settings. It retained a high performance even at aggressive scaling, achieving accuracy at a level of 79%–87% in most cases. In particular, the configuration (m_more = 2.0, m_less = 2.0) matched the baseline at 87%, as well as some intermediate settings such as (1.5, 1.0) and (1.0, 1.0). Surprisingly, a custom method was proved to be a competitive and computationally effective alternative.

Table 3: Result of different thresholding techniques with different coefficients

m_more	m_less	L-method	GMM	Random	Custom
2.0	0.0	32	37	26	79
2.0	0.5	36	46	51	80
2.0	1.0	47	51	87	81
2.0	1.5	57	64	86	84
2.0	2.0	87	43	87	87
1.5	0.0	32	37	26	79
1.5	0.5	36	47	51	81
1.5	1.0	52	64	87	86
1.5	1.5	86	59	86	86
1.5	2.0	51	39	87	86
1.0	0.0	32	37	26	80
1.0	0.5	43	51	51	82
1.0	1.0	87	87	87	87
1.0	1.5	41	53	86	81
1.0	2.0	32	30	87	78
0.5	0.0	32	40	26	47
0.5	0.5	51	51	51	51
0.5	1.0	36	33	87	52
0.5	1.5	31	43	86	49
0.5	2.0	29	33	87	43
0.0	0.5	26	26	51	26
0.0	1.0	26	26	87	26
0.0	1.5	26	26	86	26
0.0	2.0	26	26	87	26

In this thesis work, we show the effectiveness of activation tuning as a new and lightweight mechanism to enhance neural network performance after training. In contrast to retraining strategies which typically require access to original data, computational resources and optimization time, activation manipulation is performed entirely at inference time, instead changing only the neuron activations and not touching the model weights. The utility of this approach to improve substantially the misclassifications as well as testing accuracy has important implications for neural network interpretation, model maintenance and clinical use.
One of the key insights from the experiments is that neural networks frequently store at least some hidden data when generating errors meaning that they could have been the correct label. Neuron activation patterns analysis from the returned heatmap showed many neurons became deactivated at

inference, indicating that the requisite features are detected but the features were not sufficiently emphasized to affect the global prediction. By selectively boosting the weaker activations the model could be encouraged to make the right classifications. This finding challenges the claim that misclassifications are purely caused by a failure in learning features, and indicates that being too selective as to which of the already learned features to use in a decision could also lead to errors during the decision.

The scores also suggest that the extent of enhancement can decrease in proportion to the performance gap between a model before and after fixing activations. Models trained for less epochs with lower initial test accuracy exhibited greater gains from the activation modulations than ones trained for longer periods. The 10-epoch model, in particular, saw a significant 14.37% surge in test accuracy, raising it to the performance of a 20-epoch model. This implies that activation distortion is able to serve as a surrogate for undertraining, essentially depleting the performance that will need more supervised training epochs.

The relative improvements brought by the activation modification were more limited in high-performing models, but were still significant. The original baseline model achieved 3% improvements on the full test set, and 4.58% improvements for the 20-epoch model after modification. Such improvements are impressive given that they were done without exposing more data and updating the gradient. The uniform improvements across models of different quality demonstrate the generalization ability of the proposed approach.

Clinically, the implications of these results are profound. Especially in medical imaging tasks like brain tumor classification, small gains in model accuracy can result in improvements in patients care, allowing an earlier diagnosis, a faster segmentation or a better treatment planning. The potential to improve performance of a model post-deployment without additional retraining is especially attractive in healthcare, where models are frequently deployed in data-restricted environments with scarce computational resources and must be agile to adapt quickly. Adaptation through activation modification may be an effective alternative to retrain models in between improvements.

Another important aspect of this work is the interpretability that it provides on neural network behavior. The effectiveness of activation manipulation is evidence that internal representations learned by deep networks are rich and perhaps underexplored. By checking which activations needed to be increased to fix the misclassified instances one can approximate which features or patterns were detected correctly in the input; those are the ones the model failed to adequately attend to. This type of intervention offers a fresh perspective on understanding the decision-making processes in networks by complementing classical interpretability methods like saliency maps or feature attribution methods.

Yet there are several caveats to the present strategy. First, the hyperparameters that modify activation functions were determined using grid search which is an effective but computationally expensive approach that may not scale well with larger models or more complicated datasets. Adaptive or learning-based schemes for parameter selection could be considered in future work, and meta-learning and reinforcement learning can be employed to automatically determine the scaling factors for the sample-specific scaling.

Second, the present approach was verified on FCNNs for brain tumor MRI dataset. Although the preliminary results are encouraging, it is yet to be verified whether activation modification can be generalized to other types of networks (e.g., CNN, RNN or transformer). Varying architectures will have different activation mapping and internal feature hierarchies that might affect the effectiveness of activation manipulation attacks. Extending this work to a wider range of models and tasks may further clarify the scope of its utility.

Third, although activation modification increased classification rates in general, it added a step at inference to the computation. While the modification is lightweight compared to retraining, this also introduce some latency which can be a limitation for some real-time or resource-limited use-cases. Approaches for further optimization of the modulation step, or determining when it is necessary to apply the modulation step (e.g., based on prediction confidence) could be beneficial for the great practical deployment of the proposed method.

Furthermore, the clinicopathological features of the tumors nestled in the dataset indicate a few possibilities for a more focused activation analysis. In contrast to the more compact, even benign meningioma and pituitary tumors, with their diffuse and often irregular growth patterns, gliomas are more challenging. The networks' ability to discriminate between tumor types, and the particular neurons

that are being activated for each class, might be informative not only with respect to model performance, but also tumor biology. Future directions may be to generate more fine-grained activation mapping per tumor type, and to investigate if the network implicitly learns clinically important features like tumor location, shape, or intensity profile.
The findings also generate interesting issues about the biological possibility of activation alteration. Attentional and neuromodulatory processes in human neural systems dynamically tune the weighting of sensory inputs to bias decision making. Activation manipulation could be considered an artificial analogue of such biological processes, driving internal representations up or down, depending on task requirements. Studying such analogies may result in models reflecting more the biological mechanisms of learning and inference.
Another potential direction for prospect is combining activation tuning with uncertainty estimation. In the clinical setting, it is important not only to obtain good accuracy, but also to ascertain the confidence of the predictions. The method could be used in conjunction with other methods like Monte Carlo dropout or Bayesian neural networks to selectively modify activations when making a prediction under high uncertainty, offering a potential improvement in prediction accuracy and reliability.
Lastly, there are ethical issues regarding manipulation of activation. Intervening in network behavior after training adds a further level of complexity in respect of model validation, certification, and trust. Regulatory paths in at least medical AI encourage transparency and reproducibility. As such, we would need to use very strict validation procedures to try to prevent performance improvements from resulting from unintended bias or otherwise negative behavior.

5 CONCLUSION AND FUTURE WORK

One of the main contributions of this work was to build a systematic activation tracking per layer in the trained FCNN for a brain tumor classification task in MRIs. The decision to pursue this approach was driven by the well-documented difficulty of interpreting deep learning models, especially in high-stakes applications like healthcare, in which it is critical not only that a model makes accurate predictions, but that its internal logic be understandable. This work has taken a step towards understanding how those internal features are trans- formed and abstracted through the dense network by capturing and visualizing the activations of neurons throughout all layers for both correctly and incorrectly classified samples.

The findings of this interpretability analysis were mixed and informative. On the bright side, the visualization of activation heatmaps for different layers offer an opportunity to qualitatively understand how certain neurons respond to shape – related to cancerous tumor types. For instance, properly classified glioma and meningioma cases demonstrated patterns of activation escalation in deeper layers, suggesting the existence of class-specific abstractions. This was consistent with the conjectured observation that there were more discriminative features to be extracted in the deeper layers of an FCNN in contrast to the earlier layers where activations were often sparse and diffuse. It was especially informative to see what effect the model had internally on the input medical image data, which has no spatial structure because of the flattening — a concession to replace fully connected networks,.

However, the utility of these visualizations in offering clinically actionable information was somewhat constrained. The patterns of activation were typically not as spatially focused or naturally interpretable as the type of attentive map that a convolutional neural network might produce based on where it was looking. This limitation is partly caused by FCNNs themselves, which do not preserve spatial patterns and inherently abstract their data representation. Accordingly, although activation mapping provided some latent patterns in the model, a number of these insights were not straightforward to interpret location-wise in the activation space for non-expert users (e.g., medical doctors).

Also, although our attempt was to contrast between activations evoked at correct vs. incorrect classifications, the separations between them were not always clear. This indeterminacy exposes a deeper problem: interpretability of deep models is a nuanced and depending-on-context question.

Activation tracking is useful but not to be overestimated as single standing interpretability solution — especially in architectures that do not naturally cater for spatial and/or hierarchical learning.

That being said, still this thesis was able to show that a systematic tracking strategy can be used as a diagnostic tool to understand the behavior of a model. It furnished a mechanism for peeking at what is happening inside and recognizing patterns that would not be apparent from the cross-validation score alone. More importantly, it paved the way for future research on integrating activation tracking with other interpretability methods to generate more comprehensive and human-centered explanations.

Overall, the interpretability obtainable through activation tracking in FCNN remained informative but only to a limited extent. This work provides a beginning, not an end, to the pursuit of transparent, trusted AI for medical imaging.

Another major direction I focused in this thesis was on the study of how decreased connectivity in neural network architectures, in particular partially connected networks, can influence generalization and interpretability of brain tumor classification on MRI. Potentially, however, they are not going for accuracy purely, but the effect of reducing the number of connections in a fully connected layer to see whether/when such reduced connectivity models were simpler, easier to interpret models that did not sacrifice much, if anything, in performance — sometimes it even improved.

The findings provide strong evidence in support of this concept. Although we evaluated fully connected ($\rho = 1.0$) and sparsely connected ($\rho = 0.50$) network as well as our moderately connected ($\rho = 0.75$) network, the moderately connected network showed the highest validation accuracy. The fully connected model had a bit of higher training accuracy, but clearly overfitted, as it had lower performance on validation. This also demonstrates an important point – more dimensions do not necessarily lead to better results. More generally, this elimination of unnecessary connections acted as a form of implicit regularization, by which the network has been forced to generalize well, and avoid the temptation of memorizing patterns in the training data.

This result is especially important in the context of medical imaging where the data is often limited and overfitting a common concern. Constraining connectivity makes the model less complex, which not only helps with generalization, but also can lead to better interpretability. The heatmaps from these sparsely connected networks exhibited more networked activation patterns and less noise in comparison to those from fully connected networks. This brings in a legibility advantage: the fewer neurons that were active in a decision, the easier it would be to trace back its origin.

But it turns out we had overdone the sparsity as well. Sharp decline in performance was observed in the $\rho = 0.50$ model on both training and validation. The network was not able to learn complex structures and patterns from the MRI data, particularly for discriminating very similar tumors. This only serves to further highlight the balance that must be struck: Too much simplification may foster generalization up to an extent, at which point accuracy falls due to underfitting.

A strength in this part of the study was that the effect of connectivity was well separated. By fixing other factors of architecture depth, activation functions, and training time, differences of observed performance could be safely traced on purely level of connectivity between neurons.

In summary, the experimental results with partial connectivity demonstrated that less is more, at least when it comes to interpretability and generalization. They also offered an enabling empirical evidence for promoting sparsity as more than just a computational speed-up but also as an intentional architectural design in making neural networks more interpretable and efficient. This knowledge might be particularly useful in real-world medical applications where trust, speed and resource-efficiency are particularly important.

The third major contribution of this thesis grounded on the introduction of a novel set of activation manipulation strategies at inference time, the idea being that it was interesting to detail whether playing with the activations of neurons (without retraining the model) could improve the interpretability rights, and even the classification behavior. This direction was motivated by a central hypothesis that not all neuron activations are equally informative, and filtering or upweighting a subset of the neuron activations can bring out the internal decision logic in FCNNs.

Three methods were used to test this idea: L-moments thresholding, GMM thresholding, and a percentile-based thresholding approach. Both methods were based on different mental models for selecting “strong” and “weak” neurons according to their activation values and tested over various scaling strategies to see how it affected prediction.

The percentile-based approach was the best performer in terms of consistency and accuracy and is also intuitively simple to use. It did not need some statistical modelling or tuning of parameters and was computationally cheap and intuitive. In all those setups, the method preserved high accuracy (up to 87%) even when the scaling was aggressive, which indicates that it correctly emphasized meaningful neurons without breaking the network behavior. Its simplicity also made the approach particularly interesting for practical applications, where computational cost and interpretability both matters.

The L-moments thresholding algorithm performed well under some selected scaling conditions and showed robustness to far-off-the-center outliers in shattered layers in particular, with skewed activation distributions. But it sensitively depended on hyperparameters, especially scaling multipliers. When those values were too extreme, performance suffered greatly. Nevertheless, the approach based on L-moments provide a richer way of defining the importance of the neuron and was a good compromise between theoretical appeal and practicality.

The GMM-based approach, which has a theoretically appealing capacity to model latent structure behind activations, gave inconsistent results. It was sensitive to outliers and displayed great variability of results depending on how we scale it. It didn't perform well or was inconsistent in a few guises. The limitations indicate that, while GMM can fit activation bimodality, it might not generate threshold values with a clear functional interpretation particularly for noisy (or unimodal) underlying activation distributions.

Significantly, these experiments illustrate that manipulation of activation alone can alter the behavior of a model in a non-trivial manner, without requiring weight change or re-training. This presents a great opportunity for developing real-time and lightweight post-hoc interpretability tools. The fact that selective activation scaling switched the positive output class in 80\% of misclassified cases also indicates a practical potential for the proposed method.

In summary, we showed in this part of the thesis how to interpret and enhance the model's decisions by controlled activation alteration, even after training. These techniques are no silver bullet, but they offer a novel, computationally cheap way to probe the internal logic of neural networks, helping with both interpretability research and practical AI design.

The so developed thesis provides valuable insights on the activation behavior of neurons, the interpretability of models and the role of sparsity in FCNNs, but several limitations and issues are still open, pointing out relevant directions for future works.

The data available for this type of study first and foremost was the small subset used here – although meaningful and well formatted – but small compared to what is commonly required to train deep learning models that are highly generalizable. A small training-set size limited the representational capacity and the generalization ability of the network. In future work, integration with larger and more heterogeneous brain MRI datasets would also likely result in increased accuracy of classification and discretion of model behavior across more diverse input. Additionally, a more comprehensive dataset

may help alleviate model bias and improve performance on edge cases, such as less common tumor types or lower quality scans.

Solutions in this direction would also be general model architectures. Although FCNNs (Fully Convolutional Neural Network) have been used in the interpretability study as a basic use study, under the hood it have a limitation in preserving spatial features in the image data. In future work, it would be interesting to investigate whether CNNs, transformers, or hybrid architectures could also be repurposed (and evaluated) with the same activation manipulation and sparsity employed in this study. These architectures are more representative of image data and have higher applicative frequency in real diagnostic systems.

Moreover, the techniques for activation alteration presented here could be further improved. Although the percentile-based method performed stably, more advanced algorithms in attention-guided modulation, reinforcement learning for neuron selection or meta-learning for adaptive thresholding can achieve little better or similar trade-off of interpretability and performance. The activation manipulation can also be combined with attention mechanisms or gradient-based attribution methods to any form hybrid interpretable systems being both better performing and interpretable.

6 BIBLIOGRAPHY

[1] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[2] J. Chen, Y. Song, M. Wainwright, and M. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, 2020.

[3] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[4] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[5] R. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[6] C. Louizos, M. Welling, and D. P. Kingma, "L0" regularization: Learning sparse neural networks," *arXiv preprint arXiv:1712.01312*, 2017.

[7] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *Int. Conf. Learn. Representations (ICLR)*, 2019.

[8] Y. Liu, X. Liu, and Y. Li, "Ensemble gradient-based feature selection for sparse neural networks," *Neurocomputing*, vol. 544, pp. 125–137, 2024.

[9] Y. Liu *et al.*, "Sparse contrastive coding for unsupervised representation learning," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2022.

[10] G. Galanti, A. Rakhlin, and S. Shalev-Shwartz, "Norm-based generalization bounds for compositionally sparse neural networks," *arXiv preprint arXiv:2301.06685*, 2023.

[11] S. Zhou, Y. Cong, and B. Han, "N:M structured sparsity: Towards efficient AI hardware design," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021.

[12] C. Zhang *et al.*, "Cambricon-X: An accelerator for sparse neural networks," in *Proc. 49th Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2016.

[13] Y. Liu *et al.*, "SMC-Bench: Benchmarking sparse neural networks across diverse tasks," *arXiv preprint arXiv:2302.09801*, 2023.

[14] I. Ullah *et al.*, "Hybrid deep learning model using sparse autoencoders for balancing class distributions in brain tumor classification," *Med. Image Anal.*, vol. 87, 2024.

[15] C. Wen, Y. Chen, and S. Xu, "Online sparse robust models for streaming data," *IEEE Trans. Knowl. Data Eng.*, 2024.

[16] J. R. M. Hosking, "L-moments: Analysis and estimation of distributions using linear combinations of order statistics," *J. Roy. Stat. Soc. Series B (Methodological)*, vol. 52, no. 1, pp. 105–124, 1990.

[17] R. M. Vogel and N. M. Fennessey, "L moment diagrams should replace product moment diagrams," *Water Resour. Res.*, vol. 29, no. 6, pp. 1745–1752, 1993.