N	I have used GenAI tools for developing ideas.
Y	*I have used GenAI tools to assist with research or gathering information.*
N	I have used GenAI tools to help me understand key theories and concepts.
N	I have used GenAI tools to identify trends and themes as part of my data analysis.
N	I have used GenAI tools to suggest a plan or structure for my assessment.
N	I have used GenAI tools to give me feedback on a draft.
N	I have used GenAI tool to generate image, figures or diagrams.
Y	I have used GenAI tools to proofread and correct grammar or spelling errors.
N	I have used GenAI tools to generate citations or references.
N	*Other [please specify]*
Y	I have not used any GenAI tools in preparing this assessment.

Abstract

Brain tumours are one of the most lethal diseases in the world, often leaving patients with as low as a 5% [2] survival rate. Studies have suggested that early detection can improve the survival rate of patients significantly [25]. This report proposes a rapid deep learning approach to improve the detection of brain tumours in MRI images.

The model scored an accuracy of 95%, very similar to that of reported radiographers [23]. The model has a high understanding of the dataset, demonstrated by its flawless AUC and AUPRC of 1. This demonstrates the ability of such a model to revolutionise how we treat and diagnose this disease.

Introduction

Brain tumours are highly dangerous due to the complexity of the brain. With survival rates as low as 5%[2] it is a critical area of research to aid the prognosis of the patients. Studies have suggested that early diagnosis of brain tumours is associated with increased survival rates [25], underscoring the need for efficient and effective diagnosis.

Historically, methods of classical machine learning (ML) including random forests [3] and support vector machines (SVM) [4] are used in the diagnosis of tumours. However, contemporary research suggests transfer learning outperforms these in both speed and sample size requirement [5][6] [7][8]. Hence, this report proposes an approach based in transfer learning that automates the brain tumour detection in MRI scans; using a dataset of labelled samples.

This report identified the best hyper-parameters for training a deep convolutional neural network (DCNN) with transfer learning to classify samples from a dataset of brain tumours (n=222) with an accuracy of 95.2% and an ROC AUC = 1. This result demonstrates that on small datasets, deep CNN

trained via transfer learning could aid in the diagnosis of the 35 new brain tumour patients a day in the UK [10].

This report is split into 3 sections: methodology (pre-processing, hyperparameter selection, model training), results and summary.

Methods

To begin, the images are pre-processed. In the pre-processing phase, the images underwent multiple transformations to aid the learning process [11].

The process consists of the following steps:

- Standard step of normalising the pixel values between 0 and 1.
- Equalise the brightness of the pixels. This helps to reduce the variation between different MRI machines, resulting in the images being more standard across the dataset.
- The scans are cropped to focus on the brain and eliminate the black border around it. This reduces the amount of unnecessary data in the image, as well as making the images more standard, resulting in more efficient training.
- The images are resized to 224x224 pixels as this corresponds to the dimensions used in the imageNet database and is the expected size for the pretrained model.
- We remove 10% of the data as test data.

Proposed Network Architecture

This report uses the widely adopted transfer learning architecture (Fig. 1) for classification involving a pre-trained CNN (like alexNet) [31] followed by global average pooling [32], a dense layer, and a dropout layer to reduce overfitting [33].

Fig. 1. A table showing the structure of the final network

Layer (type)	Output Shape
resnet50v2 (Functional)	(None, 7, 7, 2048)
global_average_pooling2d_2 (GlobalAveragePooling2D)	(None, 2048)
dense_4 (Dense)	(None, 1024)
dropout_2 (Dropout)	(None, 1024)
dense_5 (Dense)	(None, 1)

Hyperparameter Search

Hyperparameters are fixed before training and greatly influence model performance. This report seeks to identify the optimal hyperparameters for this classification task.

Fig. 2. A table of hyperparameters, explaining their options and range.

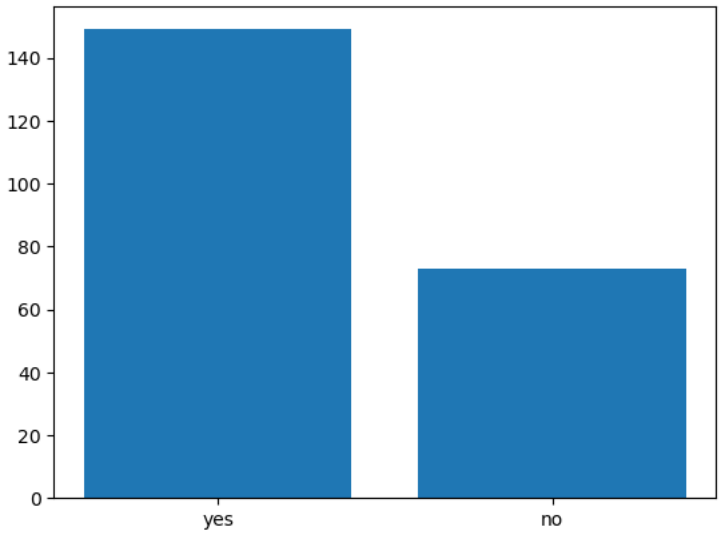
hyperparameter	Options/range	Explanation
Type of network	[VGG16, VGG19, MobileNet, DenseNet121, EfficientNet, xception, InceptionV3 and ResNet-50V2]	These models have seen success [11][12][13] in other medical image classification
Number of neurons in the dense layer	[32,64,128,256,1024]	This set of options was chosen as they need to be multiples of 8 in order to benefit from the accelerated matrix multiplication [20]. 1024 was the largest value here due to limited compute resources
Learning rate	Range [1e-5,1e-2]	This is because the learning rate neds to be a very small number in order for the model to identify the best function to fit the data
Batch size	[16,32,64]	Batch size has a large influence on model performance [18]

		the selection of sizes are powers of 2 to take full advantage of the GPU [17]
Optimiser	[ADAM,SGD]	The optimiser was chosen from ADAM or SGD. These represent adaptive optimisation methods and stochastic gradient decent respectively. Generally, ADAM performs better in initial training, whereas SGD excels in longer training processes and generalises better [19]. However, due to the small dataset, it is possible that SGD did not have the opportunity to outperform ADAM because of the low number of epochs, as over fitting occurred with larger ones.

This report utilises Bayesian search for its efficiency over grid search. It achieves this by cutting down the search space by beginning with random settings and evaluating their performance. As the trials go by, it selects settings that are closer to the highest performing settings [14]. This is critical because the number of HPs are large.

The use of stratified k-fold here was due to the small sample size. Using k-fold, it is possible that more of the dataset is used compared to conventional training-test split. The stratification of the folds is used to reduce the impact of the class imbalance in the data (positive samples =149, negative samples = 73).

Fig.3. A bar chart comparing the number of positive and negative samples, highlighting the class imbalance.



Pre-feature extraction training

When training on a set of samples and labels, this report proposes that the data is augmented using the augment2a function. This function ensures that the data is balanced by augmenting the minority class creating more samples than the majority. The augmentations include left and right shifts, brightness changes, zooms and rotations. These introduce variation back into the samples, improving the generalisability of the network on a small training dataset. Further, it reduces overfitting [15] and generally improves the accuracy of the model[11].

The features of the augmented image set are extracted using the pre-trained model. This creates the training data for the task-specific layers to learn from. This is possible because the pre-trained models are frozen during the HP search process as the choice was made to not fine-tune the model here, to reduce the computational resources required.

Computing the features once and training the much smaller task-specific networks saves heavily on computations. The ability to compute the features only once also influenced the decision for the input image size, as it meant that the images could be inputted directly to the pre-trained model, eliminating the need for a task-specific layer above the pre-trained model. Early stopping was used to prevent the

models from overfitting. This was important due to the small dataset. Early stopping ceases the training if the validation loss doesn't improve for a certain number of epochs.

The objective function of the Bayesian search provides the score of the network in the form of the average accuracy across folds. This function utilises the standard pre-extraction approach described above. Each set of hyperparameters is trained for 10 epochs with early stopping. After which, the validation is computed for the fold.

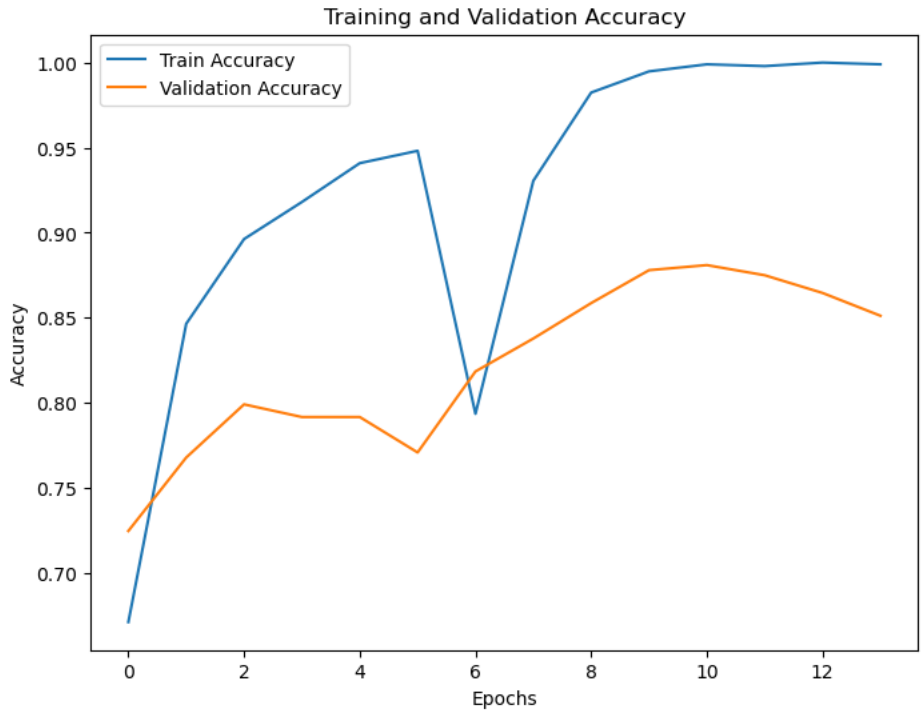
We utilise the pre-feature extraction training model to effectively train and evaluate the performance of the task-specific head on the data in each of the k-folds. The Bayesian search algorithm tries to maximise this average score over all folds.

The training process

The Bayesian HP search returned the final network structure (Fig. 1)

The model was then trained on all of the data except the training data. The layers were frozen while the model trained for 10 epochs (early stopping after 3 epochs with no improvement). The model is then finetuned. This is the process where the pre-trained model is unfrozen with a learning rate ten times less than before to prevent the target network from overfitting. This process is found to increase model performance by as much as 50% [34]

Fig. 4. A line graph to show the relationship between training accuracy and validation accuracy as the epochs increased.



Results

Fig. 5. A table showing key result metrics.

Metric	Class 0.0	Class 1.0	Macro Average	Weighted Average
Precision	0.875	1	0.9375	0.958333
Recall	1	0.928571	0.964286	0.952381
F1-Score	0.933333	0.962963	0.948148	0.953086
Support	7	14	21	21
Accuracy				0.952381
ROC AUC Score				1
AUPRC				1

The Bayesian search ran for 150 iterations, and the best hyper parameters it found was learning rate = 9.558e-5, batch size = 32, ADAM optimiser, neurons in the dense layer of the task specific head = 1024, dropout rate of 0.6775 and pretrained model was Resnet.

The evaluation process

The evaluation process involved selecting 10% of the data after the preprocessing phase (pixel value normalisation, resizing and cropping of the image) that the network has never seen before, nor was it used as validation in HP search. This data was also excluded from the hyperparameter search. The network then generates its predictions. These are probabilities that each validation sample is positive or negative. We then create a list of predicted labels. These are 1 if the probability is greater than 0.5 and 0 if it is less than 0.5. This allows us to create our confusion matrix, precision and recall using scikit learn using the usual formulae. Area under the curve (AUC) and area under the precision recall curve (AUPRC) were also calculated. AUPRC is a metric that works similarly to AUC but is more accurate for heavily imbalanced data [22].

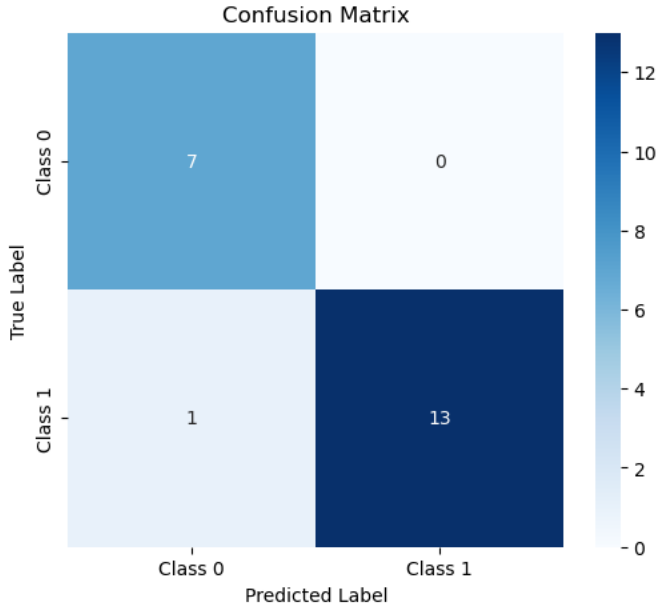


Fig.6. A confusion matrix of the final model on the test dataset

Analysing the result, we see the accuracy is 95.23% leaving only one sample from the testing set classified incorrectly. This is reflected in the extremely high precision and recall metrics at 1 and 0.929 respectively. All class 0 samples were classified correctly with one false positive. This is reflected in the confusion matrix (Fig.6) that was generated from this data. ROC AUC was calculated to determine how effective the model was at distinguishing between classes. With an AUC and

AUPRC score of 1, the model is able to distinguish perfectly the two classes. The model outperformed Random Forests (86% accuracy) [21].

Limitations and future research

These results may not generalise well to larger more diverse datasets. With a small database (n=222) only 21 images remained for testing. This may not cover enough cases for the model to test how well the model generalises.

Alongside this, the evaluation of the ensemble method would have given valuable insight into its performance increase over standard methods as it has been suggested to be much better [13]. Additionally [15] proposes many better ways of data augmentation, like GANs, which could be used in conjunction with the simple augmentations used in this experiment to further increase the size of the dataset.

As previously mentioned, future research should be conducted into the accuracies of radiographers on brain scans to compare to the accuracy of a model. The study may also look into the effect of AI assistance on the accuracy of radiographers.

Furthermore, future studies should leverage larger, more varied datasets, utilising more advanced data augmentation strategies to improve the models generalisability.

Summary

This report details the process followed to train a DCNN to classify brain tumour MRI scans on a small dataset (n=222), achieving an accuracy of 95.23% and perfect ROC AUC and AUPRC scores.

This high level of accuracy suggests that the network is comparable to the diagnosis accuracy of some radiologists. [ref] reports that radiologists diagnose various diseases (including brain tumours) with accuracies at 87% [23]. However, further data must be collected regarding radiologist's accuracies because this trial had a limited dataset (60 images, 27 radiographers)

The potential of integrating this such a model into the diagnosis loop may assist professionals to reduce the workload for staff and act as a failsafe for human error in diagnosis, as well as making diagnoses faster, as one study suggests that a model comes to a decision in three seconds compared to a humans five minutes [24] however this avenue should be pursued with caution.

As previously mentioned, future research should be conducted into the accuracies of radiographers on brain scans for comparison. Furthermore, future studies should leverage larger, more varied datasets, utilising more advanced data augmentation strategies to improve the models generalisability.

Bibliography

[1]
Madhura Kalbhor, S. Shinde, Daniela Elena Popescu, and D. Jude Hemanth, “Hybridization of Deep Learning Pre-Trained Models with Machine Learning Classifiers and Fuzzy Min–Max Neural Network for Cervical Cancer Diagnosis,” vol. 13, no. 7, pp. 1363–1363, Apr. 2023, doi: https://doi.org/10.3390/diagnostics13071363.
[2]
“Survival | Brain and spinal cord tumours | Cancer Research UK,” *www.cancerresearchuk.org*. https://www.cancerresearchuk.org/about-cancer/brain-tumours/survival
[3]
D. Jareena Begum and S. P. Chokkalingam, “MRI-Based Brain Tumour Detection and Classification Using Random Forest Algorithm,” *Smart Innovation, Systems and Technologies*, pp. 77–91, 2025, doi: https://doi.org/10.1007/978-981-97-8355-7_7.
[4]
A. K. Sharma, A. Nandal, A. Dhaka, and A. Sinhal, “A Novel Brain Tumor Classification Algorithm using SVM Classifier,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 12, no. 11, pp. 175–182, Nov. 2022, doi: https://doi.org/10.46338/ijetae1122_19.
[5]
A. Payan and G. Montana, “Predicting Alzheimer’s disease: a neuroimaging study with 3D convolutional neural networks,” *arXiv.org*, 2015. https://arxiv.org/abs/1502.02506? (accessed Mar. 20, 2025).
[6]
R. R. Ali *et al.*, “Learning Architecture for Brain Tumor Classification Based on Deep Convolutional Neural Network: Classic and ResNet50,” *Diagnostics*, vol. 15, no. 5, p. 624, Mar. 2025, doi: https://doi.org/10.3390/diagnostics15050624.
[7]
Ishfaq Hussain Rather, S. Kumar, and A. H. Gandomi, “Breaking the data barrier: a review of deep learning techniques for democratizing AI with small datasets,” *Artificial Intelligence Review*, vol. 57, no. 9, Aug. 2024, doi: https://doi.org/10.1007/s10462-024-10859-3.
[8]
M. Romero, Y. Interian, T. Solberg, and G. Valdes, “Targeted transfer learning to improve performance in small medical physics datasets,” *Medical Physics*, vol. 47, no. 12, pp. 6246–6256, Oct. 2020, doi: https://doi.org/10.1002/mp.14507.
[9]
A. Khan, A. Husen, S. Nisar, H. Ahmed, Syed Shah Muhammad, and S. Aftab, “Offloading the computational complexity of transfer learning with generic features,” *PeerJ. Computer science*, vol. 10, pp. e1938–e1938, Mar. 2024, doi: https://doi.org/10.7717/peerj-cs.1938.
[10]
Cancer Research UK, “Brain, other CNS and intracranial tumours statistics,” *Cancer Research UK*, May 14, 2015. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/brain-other-cns-and-intracranial-tumours
[11]
J. Zubair *et al.*, “Advanced AI-driven approach for enhanced brain tumor detection from MRI images utilizing EfficientNetB2 with equalization and homomorphic filtering,” *BMC medical informatics and decision making*, vol. 24, no. 1, Apr. 2024, doi: https://doi.org/10.1186/s12911-024-02519-x.

[12]
Md. Ariful Islam, M. F. Mridha, M. Safran, S. Alfarhood, and Md. Mohsin Kabir, "Revolutionizing Brain Tumor Detection Using Explainable AI in MRI Images," *NMR in Biomedicine*, vol. 38, no. 3, Feb. 2025, doi: https://doi.org/10.1002/nbm.70001.
[13]
K. M. Hosny, M. A. Mohammed, R. A. Salama, and A. M. Elshewey, "Explainable ensemble deep learning-based model for brain tumor detection and classification," *Neural Computing and Applications*, Nov. 2024, doi: https://doi.org/10.1007/s00521-024-10401-0.
[14]
J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, Mar. 2019, doi: https://doi.org/10.11989/JEST.1674-862X.80904120.
[15]
C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, Jul. 2019, doi: https://doi.org/10.1186/s40537-019-0197-0.
[16]
H. Xu, M. Ainsworth, Y.-C. Peng, M. Kusmanov, S. Panda, and J. T. Vogelstein, "When are Deep Networks really better than Random Forests at small sample sizes?," *arXiv.org*, 2021. https://arxiv.org/abs/2108.13637v1 (accessed Mar. 20, 2025).
[17]
D. Masters and C. Luschi, "Revisiting Small Batch Training for Deep Neural Networks," *arXiv:1804.07612 [cs, stat]*, Apr. 2018, Available: https://arxiv.org/abs/1804.07612
[18]
I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," *ICT Express*, vol. 6, no. 4, May 2020, doi: https://doi.org/10.1016/j.icte.2020.04.010.
[19]
N. S. Keskar and R. Socher, "Improving Generalization Performance by Switching from Adam to SGD," *arXiv:1712.07628 [cs, math]*, Dec. 2017, Available: https://arxiv.org/abs/1712.07628
[20]
"Video Series: Mixed-Precision Training Techniques Using Tensor Cores for Deep Learning | NVIDIA Technical Blog," *NVIDIA Technical Blog*, Jan. 30, 2019. https://developer.nvidia.com/blog/video-mixed-precision-techniques-tensor-cores-deep-learning/ (accessed Mar. 20, 2025).
[21]
A. Naseer, T. Yasir, A. Azhar, T. Shakeel, and K. Zafar, "Computer-Aided Brain Tumor Diagnosis: Performance Evaluation of Deep Learner CNN Using Augmented Brain MRI," *International Journal of Biomedical Imaging*, vol. 2021, pp. 1–11, Jun. 2021, doi: https://doi.org/10.1155/2021/5513500.
[22]
L. Peng, Y. Travadi, R. Zhang, Y. Cui, and J. Sun, "Imbalanced Classification in Medical Imaging," *arXiv.org*, 2022. https://arxiv.org/abs/2210.12234v1 (accessed Mar. 20, 2025).
[23]
P. Lockwood and G. Dolbear, "Image interpretation by radiographers in brain, spine and knee MRI examinations: Findings from an accredited postgraduate module," *Radiography*, vol. 24, no. 4, pp. 370–375, Nov. 2018, doi: https://doi.org/10.1016/j.radi.2018.05.009.

[24]
J. K. Ruffle *et al.*, “VASARI-auto: Equitable, efficient, and economical featurisation of glioma MRI,” *NeuroImage: Clinical*, vol. 44, p. 103668, Sep. 2024, doi: https://doi.org/10.1016/j.nicl.2024.103668.
[25]
D. Kawauchi *et al.*, “Early Diagnosis and Surgical Intervention Within 3 Weeks From Symptom Onset Are Associated With Prolonged Survival of Patients With Glioblastoma,” *Neurosurgery*, vol. 91, no. 5, pp. 741–748, Nov. 2022, doi: https://doi.org/10.1227/neu.0000000000002096.
[26]
L. Gao, L. Zhang, C. Liu, and S. Wu, “Handling imbalanced medical image data: A deep-learning-based one-class classification approach,” *Artificial Intelligence in Medicine*, vol. 108, p. 101935, Aug. 2020, doi: https://doi.org/10.1016/j.artmed.2020.101935.
[27]
W. Qu, I. Balki, M. Mendez, J. Valen, J. Levman, and P. N. Tyrrell, “Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 12, pp. 2041–2048, Sep. 2020, doi: https://doi.org/10.1007/s11548-020-02260-6.
[28]
J. I. Peltonen, T. Mäkelä, L. Lehmonen, A. Sofiev, and E. Salli, “Inter- and intra-scanner variations in four magnetic resonance imaging image quality parameters,” *Journal of medical imaging (Bellingham, Wash.)*, vol. 7, no. 6, p. 065501, Nov. 2020, doi: https://doi.org/10.1117/1.JMI.7.6.065501.
[29]
T. B. Smith, S. Zhang, A. Erkanli, D. Frush, and E. Samei, “Variability in image quality and radiation dose within and across 97 medical facilities,” *Journal of Medical Imaging*, vol. 8, no. 05, May 2021, doi: https://doi.org/10.1117/1.jmi.8.5.052105.
[30]
E. Gibson *et al.*, “Inter-site Variability in Prostate Segmentation Accuracy Using Deep Learning,” *Lecture notes in computer science*, pp. 506–514, Jan. 2018, doi: https://doi.org/10.1007/978-3-030-00937-3_58.
[31]
A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2012, Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
[32]
M. Lin, Q. Chen, and S. Yan, “Network In Network,” *arXiv.org*, 2013. https://arxiv.org/abs/1312.4400
[33]
N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[34]
A. Davila, J. Colan, and Y. Hasegawa, “Comparison of fine-tuning strategies for transfer learning in medical image classification,” *Image and vision computing*, pp. 105012–105012, Apr. 2024, doi: https://doi.org/10.1016/j.imavis.2024.105012.