Machine Learning Methods for Lung Tumour Classification

Carl Viggo Nilsson Gravenhorst-Lövenstierne

Abstract

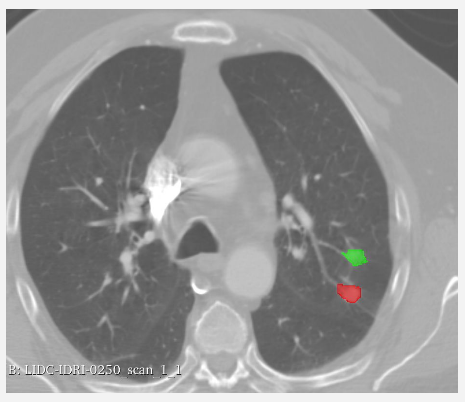
Lung cancer stands as one of the deadliest forms of cancer. In 2020 alone, the disease claimed the lives of more than 1.8 million individuals globally. Advancements in medical imaging technology, such as low-dose computed tomography, have resulted in a global accumulation of high-quality image data. However, the traditional approach of identifying and diagnosing cancer still relies on manual evaluation of such data. This workflow is time consuming and prone to errors.

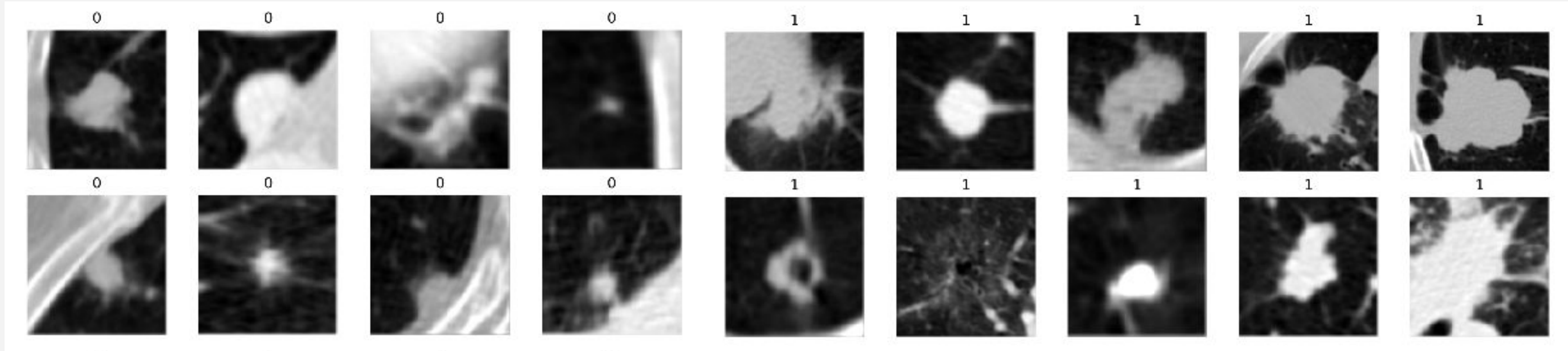
To assist clinicians, computer models specifically designed for tumour classification have recently been developed. This study aims to enhance and evaluate machine learning algorithms, specifically random forests and convolutional neural networks, in diagnosing lung cancer through the analysis of computed tomography scan images.

This study reveals that based on the analysis of the Kaggle Data Science Bowl 2017 data set, conventional machine learning methods outperform fine-tuned deep learning models in distinguishing between malignant and benign tumours, especially in reducing false positives and negatives.

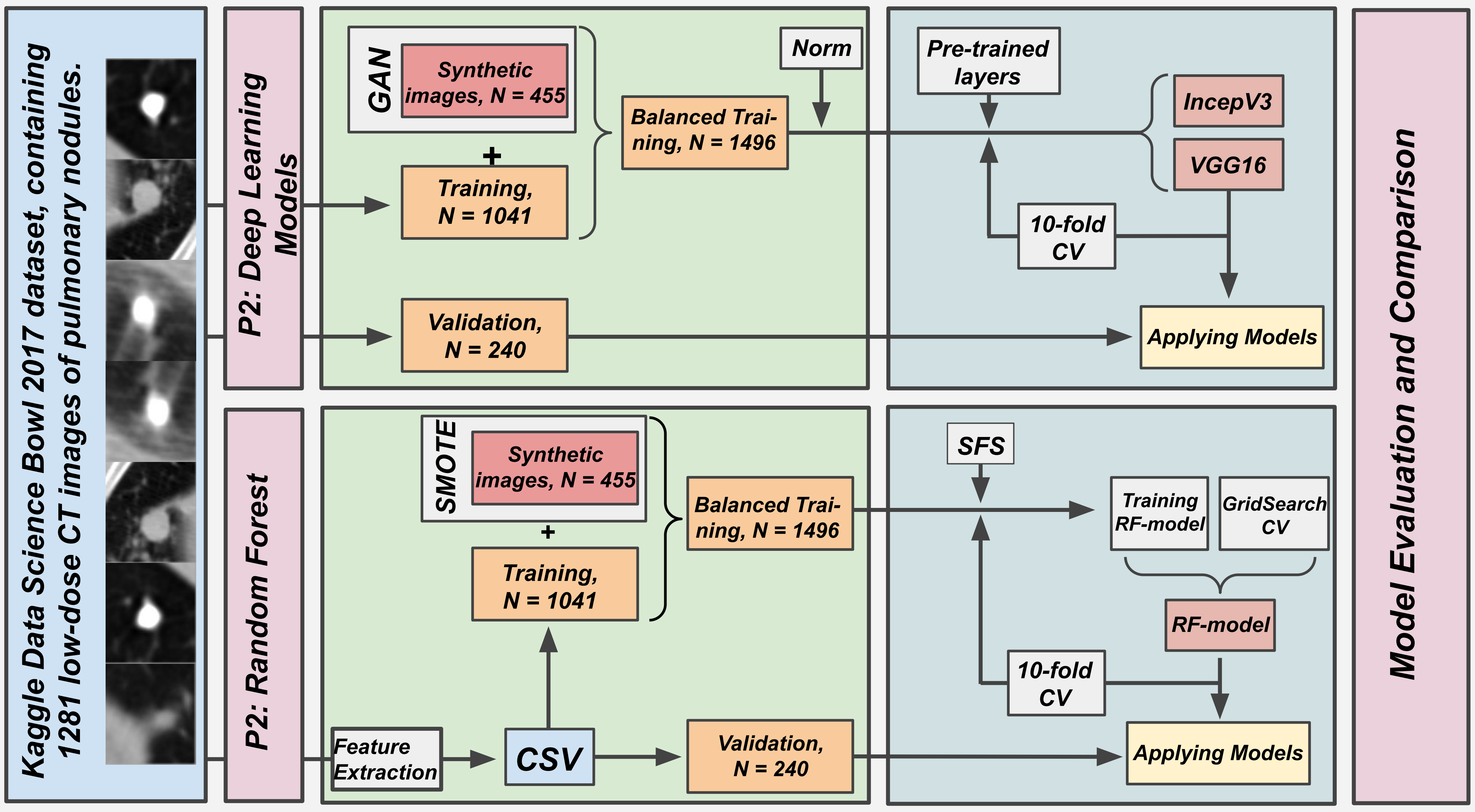
In addition, we can conclude that the random forest model with feature extraction from the PyRadiomics library, combined with the synthetic minority over-sampling technique (SMOTE), proved more efficient on limited datasets. Compared to this, the deep learning model with feature extraction from a pre-trained AlexNet model, fine-tuned on data partly generated by a generative adversarial network (GAN), was less effective. Nonetheless, if the dataset had been larger and denser in terms of information, then the deep learning approach might have yielded improved performance.

CT Images of Pulmonary Nodules





Method

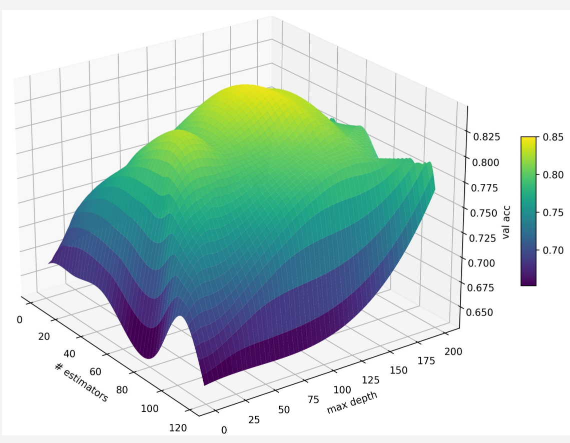


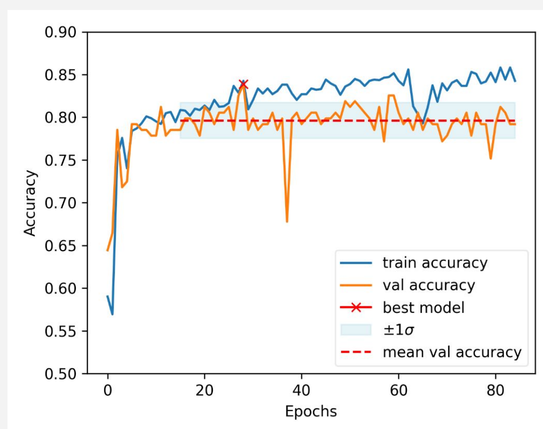
Results

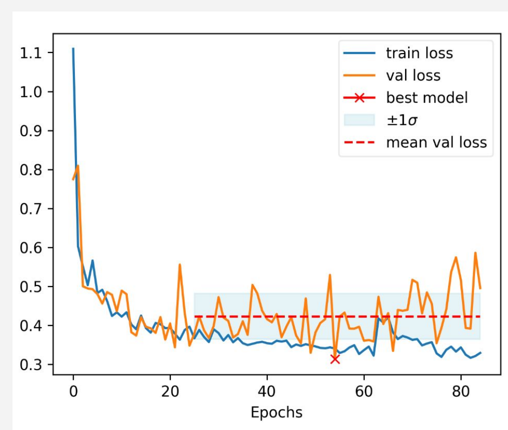
The table below summarises the mean validation accuracy, mean loss, specificity and sensitivity for the AlexNet and Random Forest models. This data is visually represented as two violin plots, where the violin body represents the distribution of model accuracy.

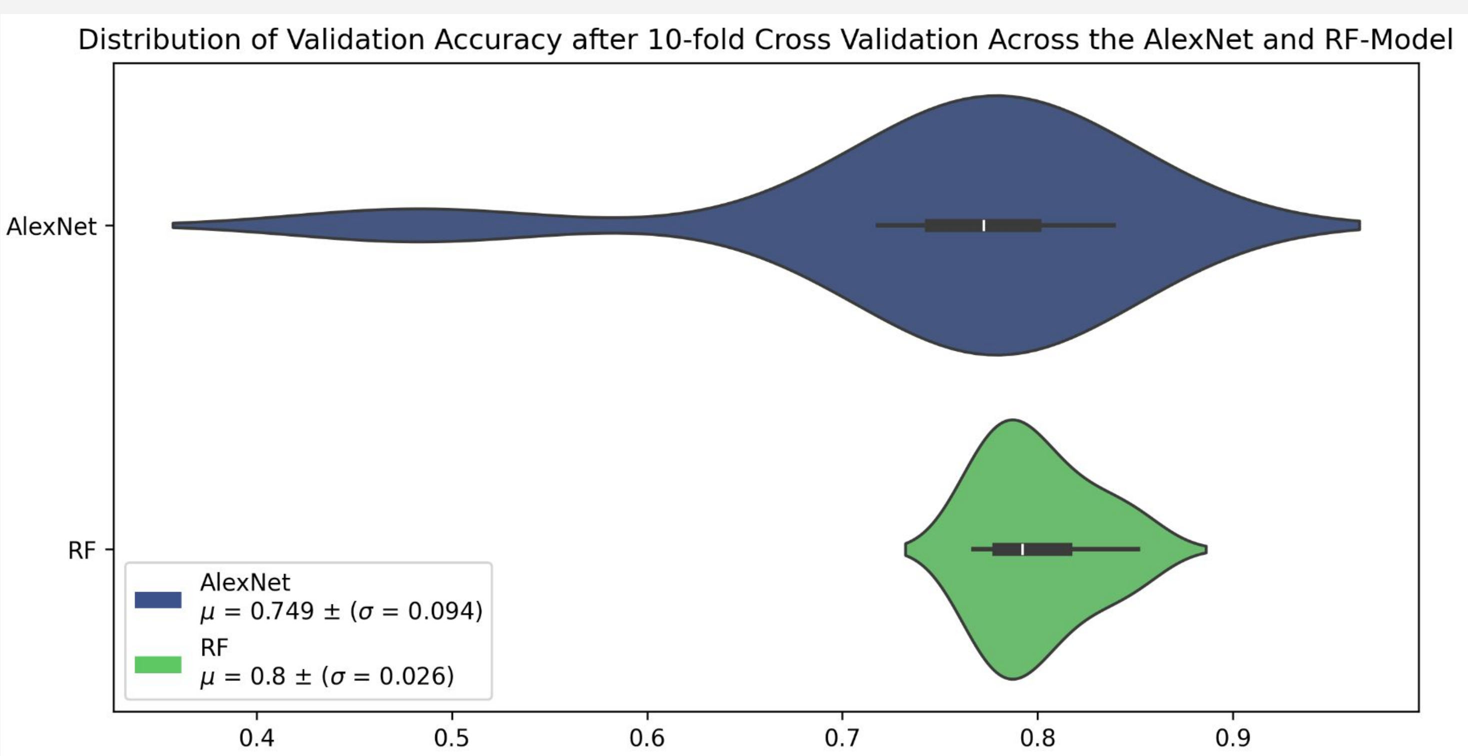
The three figures presented below display the learning curves of the top-performing Random Forest (left) and AlexNet (right) models after 10-fold cross validation. Confusion matrices have then been computed for these models, with a sample size of 240.

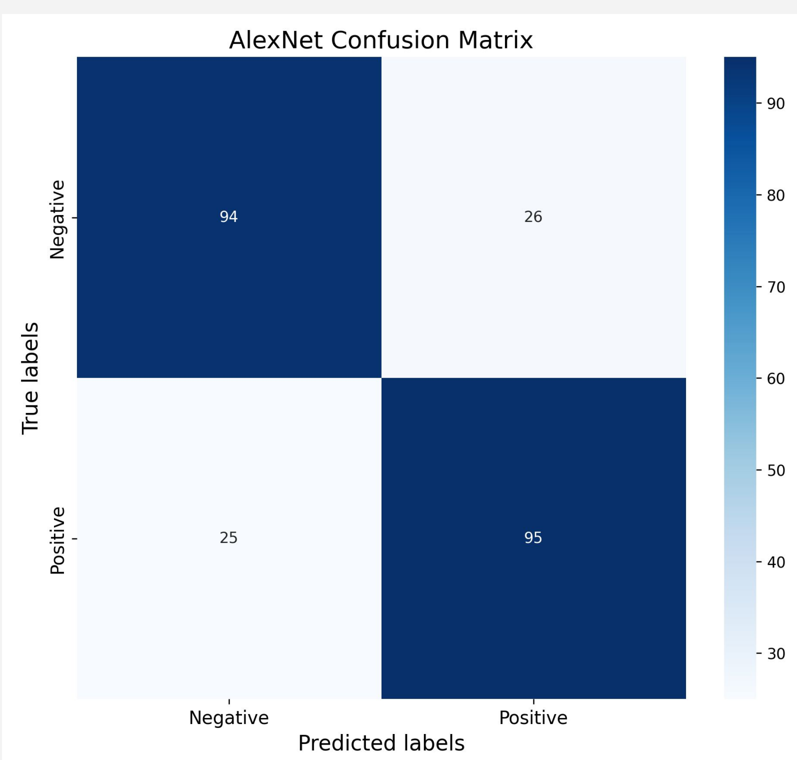
Metric	AlexNet	Random Forest
Validation Accuracy, 10-fold	0.748 ± 0.094	0.800 ± 0.026
Validation Loss, 10-fold	0.59 ± 0.17	–
Specificity, best-model	0.78	0.97
Sensitivity, best-model	0.79	0.93

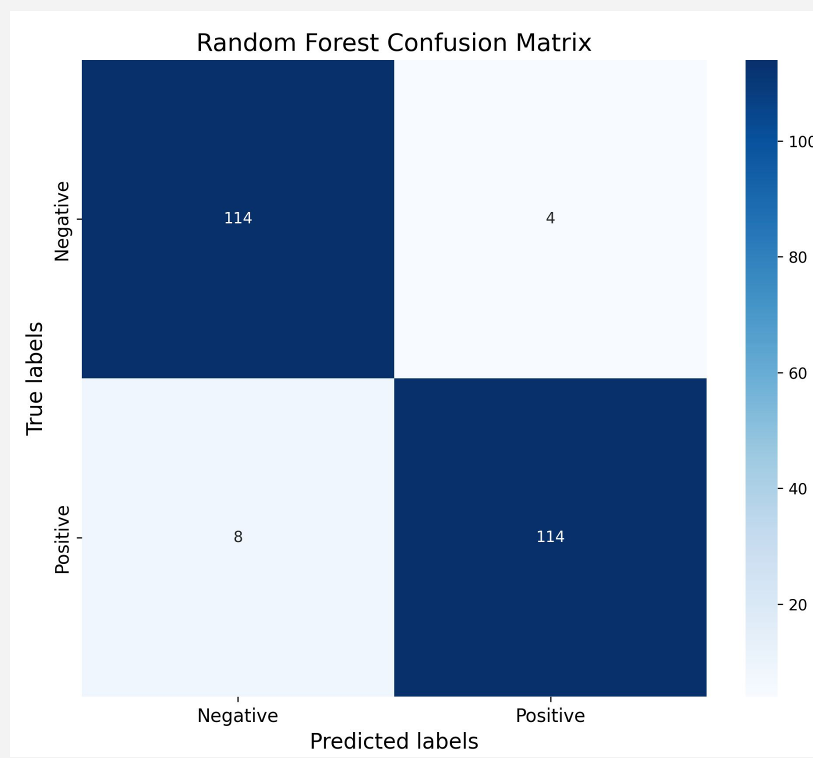












Conclusions

This study reveals that based on the analysis of the Kaggle Data Science Bowl 2017 data set, conventional machine learning methods outperform fine-tuned deep learning models in distinguishing between malignant and benign tumours, especially in reducing false positives and negatives.

In addition, we can conclude that the random forest model with feature extraction from the PyRadiomics library, combined with the synthetic minority over-sampling technique (SMOTE), proved more efficient on limited datasets. Compared to this, the deep learning model with feature extraction from a pre-trained AlexNet model, fine-tuned on data partly generated by a generative adversarial network (GAN), was less effective. Nonetheless, if the dataset had been larger and denser in terms of information, then the deep learning approach might have yielded improved performance.