Using a Multi-Modal Approach to Increase Accuracy on Predicting Progression of Lung Tumours

Benjamin Åstrand
Department of Computer Science
The Institute of Technology
Linköping University
benas022@student.liu.se

Johan Christiansson
Department of Computer Science
The Institute of Technology
Linköping University
johch241@student.liu.se

Sven Grinneby
Department of Computer Science
The Institute of Technology
Linköping University
svegr833@student.liu.se

Oscar Eriksson
Department of Computer Science
The Institute of Technology
Linköping University
oscer447@student.liu.se

Olof Swedberg
Department of Computer Science
The Institute of Technology
Linköping University
olosw720@student.liu.se

Abstract

Deep learning models, particularly in the realm of computer vision, have emerged as powerful tools in medical analysis and diagnosis, offering the potential for faster and more accurate diagnostic tasks while mitigating human biases. However, their widespread adoption in the medical field is impeded by challenges in data collection and availability. This paper addresses these challenges by exploring methods to train models with reduced data requirements, focusing on the integration of pretrained Resnet50 models for feature extraction alongside patient tabular data to create a multi-modal model. By augmenting the architecture with readily available data sources, the aim is to achieve comparable or superior accuracy to existing models trained on extensive datasets. The study builds upon previous research on Resnet50 utilization in medical diagnosis and proposes novel approaches to improve model accuracy by incorporating additional patient information. Results indicate that multi-modal models combining image and tabular data outperform solely image-based models, suggesting that clinical variables complement the information extracted from medical images. Additionally, an analysis of precision and recall metrics is provided, shedding light on the trade-offs between false positives and false negatives in diagnostic accuracy. Future research directions include exploring combinations of image data and clinical variables to optimize model performance while minimizing input requirements. Ethical considerations, including patient privacy and the impact of model predictions on treatment decisions, are also discussed. This research contributes to advancing computer vision-assisted diagnostics in healthcare, paving the way for more efficient and accessible medical practices.

1 Introduction

Deep learning models and other machine learning algorithms are becoming increasingly efficient in performing medical analysis and diagnosis [4]. Specifically, computer vision, the area of computer analysis of video and images, has proven useful in medical cases. The deployment of deep learning models can reduce the time required to perform various advanced diagnostic tasks, increase accuracy, and remove human biases [2]. Recent findings even show that many models perform with accuracy comparable to that of medical professionals [9]. However, an issue limiting the practical applications of deep learning in the medical field is the data collection and availability. The resources required to

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

collect the extensive amount of data needed to train models coupled with legal issues pertaining to saving and using client information have slowed down the adoption and development of models into practice [10].

It is therefore of utmost relevance to evaluate ways to train models with fewer samples and less expensive data collection. The usage of pre-trained Resnet50 models assisting in feature extraction from images coupled with subsequent classification heads is one way to address this issue. This paper develops on an earlier model created for a specific dataset, adding less expensive and more readily available patient tabular data to the architecture and creating a multi-modal model. Achieving comparable, or higher accuracy by adding more easily collected data would imply a significant step toward better, computer vision-assisted diagnostics and predictions.

2 Previous related work

This section compiles relevant previous materials pertinent to this project. It includes a description of Resnet50 and the medical paper from which this project draws its foundations.

2.1 Resnet50

Resnet50 is a deep convolutional neural network made by Microsoft in 2015. The name Resnet50 is derived from it being a Residual Network architecture with 50 layers. It consists of 4 main parts: the convolutional layers, the identity block, the convolutional block, and the fully connected layers. Utilizing skip-connections, the issue of vanishing gradients when adding multiple layers is addressed, which allows the model to become deeper (having a higher amount of convolution layers). When trained on the ImageNet dataset consisting of 14 million images and over 1000 classes, Resnet 50 achieves an error rate comparable to humans (22.85% vs 5.1%). [8]

2.2 Resnet50 in medical diagnosis

There have been multiple research papers specifically exploring the utilization of Resnet50 in medical diagnosis, for non-cancer and cancer diseases. The applications range from predicting disease progression to classification of current states. [11], [3], [5], [1], [7].

Three studies have been investigated pertaining to classification of breast tumors specifically using Resnet50 [11], [5], [7]. The first study aims to accurately classify breast cancer tumors into the categories benign or malign using Resnet50 and a convolutional neural network classification head on histopathology medical image data. In this study, Resnet50 is used exclusively for feature extraction. They achieved a maximum accuracy of 91,7%. [11] In another study, three different attention modules for classification are tested after the Resnet50 feature extraction: firstly a spatial attention module (SAM), secondly a channel attention module (CAM) and finally a Convolutional block attention module (CBAM) [7]. The addition of the CBAM as classification head, which is a combination of the SAM and CAM attention modules, provided significant improvement to the model and resulting in a final classification performance of 87% and an overall accuracy of 78% [7]. The third study investigated on the subject utilizes Resnet50 and a subsequent classification head with fully connected convolution layers, the activation function ReLU and finally a binary classifier through the Sigmoid activation function [5]. This architecture achieved the exceptionally high accuracy of between 98 to 99,8% [5].

Another study investigates different machine learning models ability to classify the progression of Alzheimer disease. This study compares the classification accuracy of Resnet50 trained on MRI-scans versus the classification accuracy of an extreme gradient boosted tree model that is trained on MMSE (Mini Mental State Data). In this study, the Resnet50 model outperforms the extreme gradient boosted tree model with an accuracy of 98,99% compared to 91,3%. [3]

In another study, different convolutional neural networks (GoogleNet, AlexNet, VCG-16, Resnet50, Inceptionv3) were tested to classify brain tumors. Resnet50 outperformed the other networks resulting in a final accuracy of 98,14%. [1].

2.3 Multi-modal models

Multimodal models can be categorized into early-, intermediate- and late fusion models. In early fusion architectures, different datasets are sent to the same model. In intermediate fusion models, the output of one model trained on one dataset is sent into another model as input combined with another dataset. Late fusion models combine the output of multiple different models that are trained on different inputs. [14]

The effectiveness of multimodal models vary. In one study aiming to predict neurocognitive impairment for HIV-patient, the accuracy of the multimodal model was significantly higher at 80% compared to 72% and 65% for the inputs trained separately [13]. In two other studies, one reviewing prediction of left ventricular dysfunction [15] and the other classification of biometric signals [6], the improvement to accuracy after adding multimodality was only minor at 3% and 2% respectively.

3 Problem formulation

The main focus of this paper is a development on a previously conducted study predicting lung cancer progression based on cell-tissue samples analyzed with Resnet50 and a classification head. Figure 1 depicts the architecture used in this study. The 18 images per sample, which show different markers, were used as input for the model. When performing the study, the research team identified that using only 8 out of the 18 markers as input achieved the same accuracy as adding more markers, indicating a minimum threshold of markers needed to accurately ascertain lung cancer progression. The study achieved a progression prediction accuracy of 95,9%. Using the clinical markers (tabular data) as input did in this study alone not achieve an accuracy above the baseline limit of 85%. [12]

In the quickly expanding field of deep learning-assisted medical analysis, an issue is as stated previously the data collection and training limitations. One way to achieve higher accuracy whilst utilizing existing resources is as demonstrated in the multiple studies referenced in this paper to use the pretrained Resnet50 for feature extraction from a smaller data set. This approach addresses the issue of smaller available datasets.

Furthermore, due to high costs and time needed to collect high-quality medical imagery, it is of high relevance to investigate potential improvements to existing models that does not require advanced imagery to the same extent. By including more easily collected and readily available data to existing models, such as tabular data pertaining to more general patient information such as age, sex and BMI through multi-modal approaches, the goal is to improve the accuracy of existing models without significantly increasing input data load or cost. It is of further interest to investigate the potential to substitute expensive medical imagery with cheaper tabular data while keeping comparable acccuracy of the models predictive capabilities.

Through the model design planned in this paper, general tabular data on the patients whose lung tissue corresponds to the respective TIF-images is combined with the Resnet50 architecture. Our goal is therefore to investigate the improvement potential to the existing models by adding more easily collected and readily available data.

4 Method description

The method for this paper was to first replicate the medical paper's method by utilizing the attached code from the author's GitHub repository [12]. The dataset consists of highly multiplexed imaging mass cytometry (IMC) images of the tumor immune microenvironment (TIME) from 416 patients with lung adenocarcinoma (LUAD) (see Figure 6). Therefore, each patient had 18 images from different markers illustrating phenotypes and spatial interactions within the tumor microenvironment. These images had a corresponding tabular dataset of clinical variables containing information about the patient's sex, age, BMI, smoking status, pack-years, cancer stage, histological pattern, and the progression variable which the model aimed to predict. The clinical variables also contained information regarding if the patient survived, this was removed to ensure the model's focus solely on predicting future outcomes.

Since 84.16% of the patients were cases for which progression was true, the data was over-sampled to address the class imbalance. This was acheived by using RandomOverSampler from the imbalanced-learn library.

Each IMC image (markers) was resized to 224x224 pixels and copied to three copies as input to the pre-trained ResNet50 model to extract a 2048 embedded feature vector. These were concatenated to a single vector of size *number of markers* x 2048.

The concatenated feature vector underwent dimensionality reduction using Principal Component Analysis (PCA) with a target dimensionality of 9, a more manageable size for further processing. Finally, the reduced-dimensional feature vector resulting from PCA was fed into a Support Vector Machine (SVM) classifier. The SVM was trained to classify samples into two categories: *progression* or *no progression*.

Three tabular classification models were evaluated to identify the most suitable one for integration into the multi-modal model. These were TabNet with a deep learning architecture, XGBoost with a gradient boosting algorithm, and a Multilayer perceptron (MLP) consisting of 1 hidden layer with 64 units. The MLP architecture is visualized in Figure 3.

Previous related work, specifically regarding multi-modal approaches, was investigated to determine what alternatives to implement when using both image and tabular data. First, tabular data was cleaned by removing NaN values, and therefore lowering the number of samples to 334.

Three multi-modal approaches were employed in the study. The first approach was an early fusion model which involved concatenating the cleaned data (*raw*) with the embedded feature vector extracted from ResNet50 and subsequently passing it through PCA. This architecture is displayed in Figure 2.

The second multi-modal approach was an intermediate fusion model which involved selecting the most appropriate tabular model based on its functionality and accuracy, and then integrating it with the embedded feature vector from ResNet50 through concatenation, following the same concatenation procedure as with the *raw* data. Finally, prediction probabilities from both the tabular and image models were aggregated in a late fusion model using logistic regression, which was trained to refine the combined predictions.

Lastly, the accuracy impact of individual clinical variables was investigated by testing every variable individually on the best concatenation model.

4.1 Limitations

Due to limited computational time, different combinations of markers will not be investigated to determine the optimal combination of markers and clinical variables. Instead, the highest accuracy combination of markers concluded by the medical paper will be used [12].

5 Results

In this section, the results from the study will be presented by first by describing the implementation part of the project, followed by the actual accuracy results.

5.1 Implementation

When replicating the medical paper's method [12], several challenges were faced due to absence of data and discrepancies between the paper's method and the attached code. Therefore, several code adjustments were made to ensure proper functionality to the designated purpose.

The absence of validation data led to the prediction tests being solely performed on test data using K-fold cross-validation. Additionally, certain image data crucial for analysis, such as the significant αSMA marker, was unavailable, thus impeding its utilization. Consequently, full validation of the medical paper's findings couldn't be achieved since the reported accuracy was based on different markers. However, the results hint at similar outcomes (as indicated in Section 5.2). Therefore, our project focused solely on predicting progression using the six available markers that yielded the highest accuracy according to the medical paper: CD68, CD163, HLA-DR, CD11c, CD14, and CD16 [12].

Other code adjustments were made to make the code more readable and flexible to adjustments. The multi-modal approach also resulted in several changes to fuse the different datasets.

The attempt to implement XGBoost and TabNet proved unsuccessful, primarily due to the difficulty encountered in extracting the embedded feature vector. Consequently, these models could not be utilized. Instead, the multi-modal approach was tested solely on the MLP model, since this allowed for comparison between the aggregated and concatenated results. This architecture is visualized in Figure 4 and Figure 5.

5.2 Model performance

The results presented in Table 1 show the accuracy of the different models that were trained and tested in this study. The *Recreated Image Model* is the replicated model from the medical paper [12] using the 6 markers described in Section 4.1. All *Tabular Models* were trained on 7 of the clinical variables as described in Section 4 and the *Multi-modal Models* were trained on the 6 markers and the 7 clinical variables.

Table 1: Comparison of Baseline and 5-fold Cross Validation Results (%)

	Baseline	**5-Fold Cross Validation Accuracy**					**Mean**	**Std. Dev.**	**Precision**	**Recall**
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5				
Image models										
The Paper's Image Model	84.16	-	-	-	-	-	95.9	-	-	-
Recreated Image Model	84.16	93.38	93.38	97.06	94.85	92.65	94.26	1.57	90.61	98.84
Tabular models										
MLP	85.63	44.35	50.43	43.86	35.96	50.00	44.92	5.25	45.28	48.78
XGBoost	85.63	63.04	63.04	62.64	60.44	65.93	63.02	1.96	62.93	68.07
TabNet	85.63	39.78	54.03	46.87	48.75	24.37	42.76	11.48	58.53	69.80
Multi-modal models										
Raw Concatenation	85.63	97.39	96.52	95.61	93.86	97.37	96.15	1.32	94.67	97.95
MLP Concatenation	85.63	95.65	94.78	92.11	91.23	98.25	94.40	2.52	91.63	97.95
MLP Aggregation	85.63	96.52	95.65	99.12	97.37	99.12	97.56	1.39	97.22	97.95

Next, the impact of each clinical variable on performance was assessed individually by training the *Raw Concatenation* model on images from the 6 markers and 1 of the 7 clinical variables at a time. The results are shown in Table 2.

Table 2: Comparison of individual clinical variables using Raw Concatenation (%)

	Baseline	**5-Fold Cross Validation**					**Mean**	**Std. Dev.**	**Precision**	**Recall**
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5				
Sex	84.16	94.85	93.38	94.12	94.85	93.38	94.12	0.66	90.29	98.84
Age	84.16	94.85	93.38	94.12	95.59	93.38	94.26	0.86	90.53	98.84
BMI	84.16	94.85	93.38	94.12	94.85	93.38	94.12	0.66	90.29	98.84
Smoking Status	84.37	91.91	96.32	95.59	97.06	97.06	95.59	1.92	92.54	99.38
Pack Years	85.37	95.65	93.91	94.74	93.86	98.25	95.28	1.62	93.01	97.95
Cancer Stage	84.37	93.38	96.32	94.12	97.79	96.32	95.59	1.61	92.63	99.09
Histological Pattern	84.16	97.79	94.85	94.12	97.06	94.85	95.74	1.43	93.06	98.84

6 Conclusion / Discussion

In this section the conclusion and discussion of the results will be presented.

6.1 Model performance

As showcased in Table 1, neither of the *Tabular models* achieved a mean accuracy above baseline. The previous paper [12] did not achieve an accuracy significantly higher than baseline when predicting progression based on clinical variables, suggesting that the clinical variables do not contain enough information to accurately predict progression.

Furthermore, all *Multi-modal models* achieved equal or higher performance compared with the *Recreated Image Model*. This result suggests that although the clinical variables do not contain

enough information to accurately predict progression on their own, they contain information that complements the information in the IMC images.

The *MLP Aggregation* model achieved the highest mean accuracy over the 5 folds. This finding indicates that the MLP aggregation method is particularly effective for this application, likely due to its ability to capture relationships between the different types of data. The superior performance of the *MLP Aggregation* model highlights the advantages of using sophisticated aggregation techniques to effectively combine multi-modal data.

Additionally, our analysis indicates that multiple clinical variables are necessary to achieve optimal accuracy. As shown in Table 2, the *Raw Concatenation* model using individual clinical variables yielded varied results: the lowest mean accuracy was obtained with the variable 'sex', while the highest accuracy was achieved with 'histological pattern'. This indicates that 'histological pattern' is the best individual predictor among the clinical variables. However, as evidenced by Table 1, combining multiple clinical variables with image data results in the best performance, highlighting the value of integrating diverse data sources for improved accuracy.

Moreover, false negatives pose a greater risk than false positives, as misclassifying progression as non-progression when it is present can have severe consequences for individuals' health and well-being. Therefore, prioritizing high recall becomes imperative. In this context, the *Recreated Image Model* emerges as the model with the least potential for such adverse outcomes (see Table 1). Examining individual variables in Table 2, smoking status stands out with the highest recall. This signifies that smoking status is associated with the highest certainty of correctly identifying patients with progression, albeit at the expense of potential false positives. As evident from Table 2, smoking status resulted in a precision of 92.54%, implying some patients may be falsely diagnosed with progression, which could adversely impact their well-being. Comparing precision and recall, it becomes evident that the models generally excel in recall over precision, a favorable trend. While optimal performance would ideally entail high values in both metrics, the emphasis on recall, given its critical health implications, remains justified.

6.2 Future research

For future research, it would be interesting to examine what combinations of image data and clinical variables could achieve the highest accuracy to minimize the number of required inputs. This paper only examines which individual clinical variables impact accuracy the most, although other combinations could have potential to improve accuracy.

Moreover, exploring combinations of clinical variables that ensure minimal false negatives, i.e., high recall, would be valuable for further research. Employing a weighted confusion matrix can amplify loss for false negatives. Such investigations could yield insights into developing models that reliably identify all instances of progression, thereby enhancing patient outcomes and minimizing diagnostic oversights.

It would also be interesting to investigate the impact of individual clinical variables in the *MLP Aggregation* model since this was the model that resulted in the best outcome.

7 Ethical considerations

In general, the medical field is a sensitive area due to the personal information that is being processed. One important ethical consideration is therefore information security - it is very important to consider both patient confidentiality and privacy.

Furthermore, the importance of the decision that stems from the recommendation that the model produces is monumental. The likelihood of malign cancer progression determines both very expensive and for the patient impactful treatments such as chemotherapy or radiation treatment. When the decision involves life and death, an accuracy of anything below 100 percent indicate that some individuals are exposed to fatal risk. In this case, false negatives are especially harmful - if the model predicts that the cancer will not progress and that treatment is thus abstained, the cancer will progress without appropriate remedies which can be fatal.

Since the model is trained on a set which might exhibit some biases, the prediction when changing underlying factors such as demographics might be skewed and thus produce inaccurate results. In

this case specifically, when the data sample is limited to 416 individuals, the generalization ability of the model might be insufficient when performed on the general population.

There is also the aspect of human interaction - some patients might be uncomfortable having a consideration of this personal magnitude to be made by a computer rather than another human.

8 Contribution statement

The main parts in replicating the original study have been performed by the group as a whole - to gain a comprehensive understanding of the original report and the associated code, we have taken turns reading documentation, research the topic and writing code. It was based on this and our problem statement that we as a group worked on structuring different multimodal architectures, discussing potential ways of implementation.

However, we divided the subsequent implementation work and writing of the report between us. This mainly implied that the individual was responsible for the completion of the segment and not individually responsible for doing the work in its entirety. Therefore, dictated by availability and other circumstances, everyone has essentially worked on every part.

8.1 Implementation responsibility areas

Benjamin Åstrand: Implementing intermediate fusion model and late fusion model. Interpreting code from the medical paper GitHub and recreating code skeleton
Johan Christiansson: Creating multi-layered perceptron and analyzing the tabular data
Olof Swedberg: Replicating old study's model
Oscar Eriksson: Implementing early fusion model
Sven Grinneby: Analyzing the most useful images (markers) from the 18-images dataset

8.2 Report responsibility areas

Benjamin Åstrand: Conclusion/Discussion
Johan Christiansson: Problem formulation
Olof Swedberg: Method description
Oscar Eriksson: Ethical considerations and results
Sven Grinneby: Introduction and previous work

9 Supplementary Material

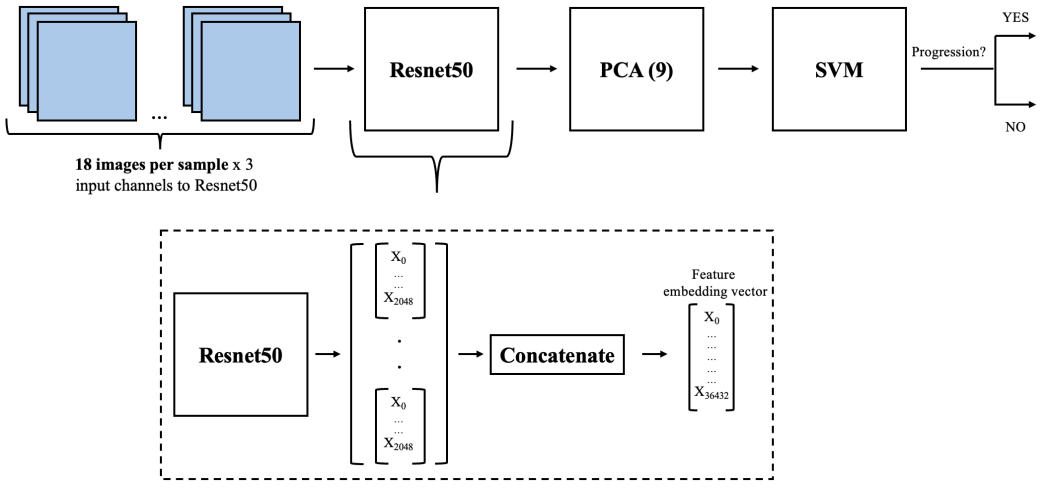

Figure 1: Original architecture from previous study

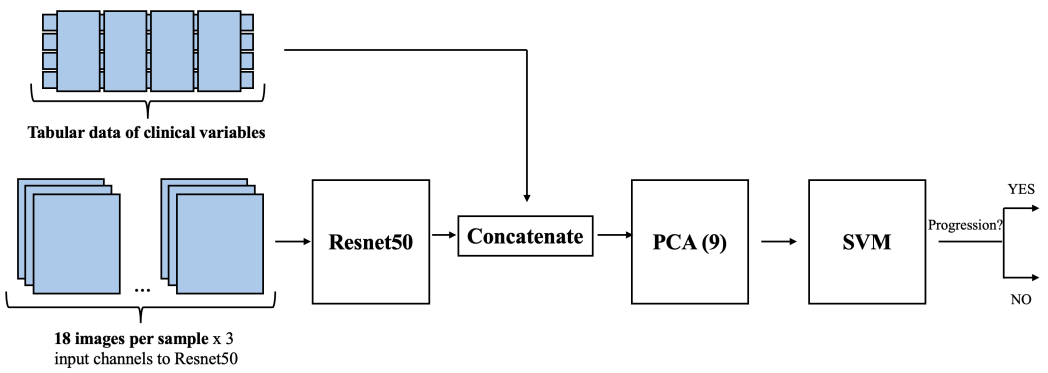

Figure 2: Architecture with early fusion multimodal model

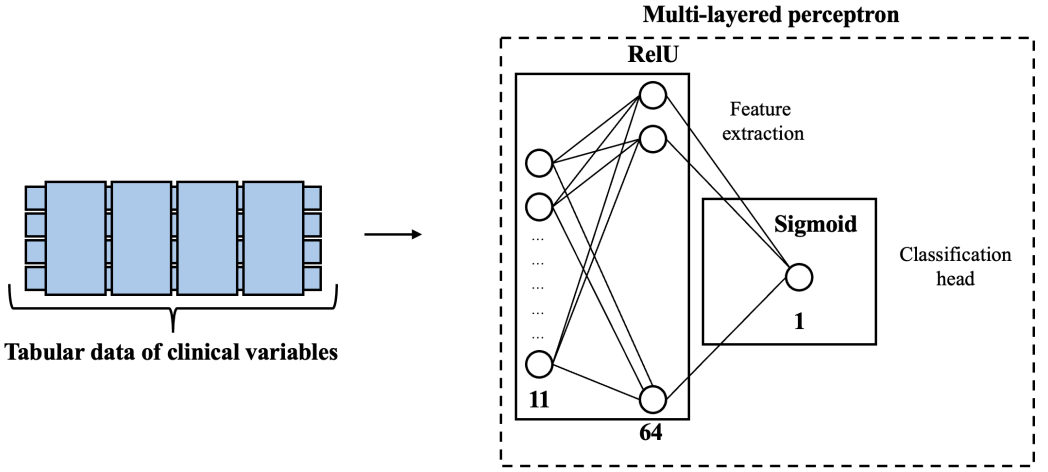

Figure 3: Multi-layered percepetron architecture

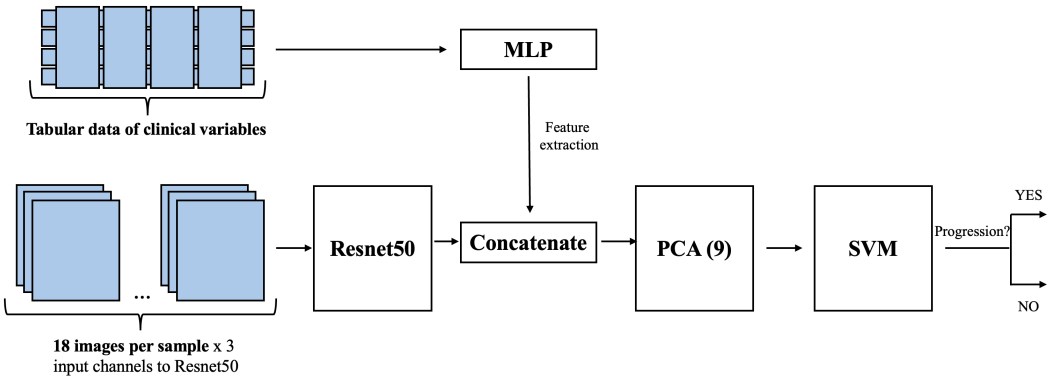

Figure 4: Architecture using intermediate fusion multimodal model

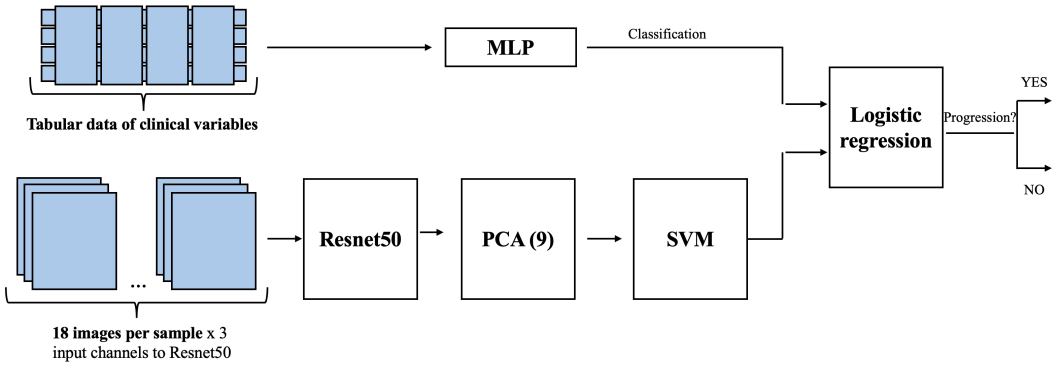

Figure 5: Architecture using late fusion multimodal model

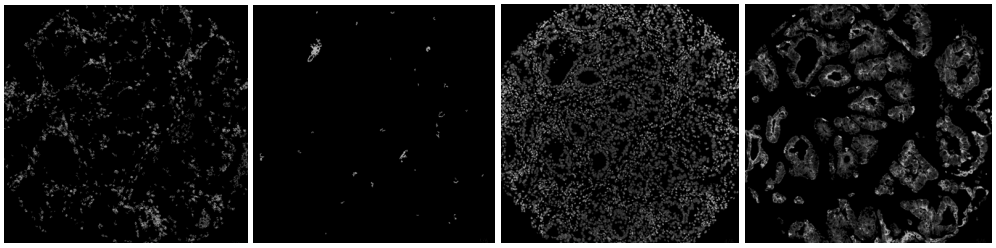

Figure 6: Sample of IMC images

References

[1] Anand Deshpande, Vania V Estrela, and Prashant Patavardhan. "The DCT-CNN-ResNet50 architecture to classify brain tumors with super-resolution, convolutional neural network, and the ResNet50". In: *Neuroscience Informatics* 1.4 (2021), p. 100013.

[2] Andre Esteva et al. "Deep learning-enabled medical computer vision". In: *NPJ digital medicine* 4.1 (2021), p. 5.

[3] Lawrence V Fulton et al. "Classification of Alzheimer's disease with and without imagery using gradient boosted machines and ResNet-50". In: *Brain sciences* 9.9 (2019), p. 212.

[4] Hafsa Habehh and Suril Gohel. "Machine learning in healthcare". In: *Current genomics* 22.4 (2021), p. 291.

[5] Qasem Abu Al-Haija and Adeola Adebanjo. "Breast cancer diagnosis in histopathological images using ResNet-50 convolutional neural network". In: *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE. 2020, pp. 1–7.

[6] Mohamed Hammad, Yashu Liu, and Kuanquan Wang. "Multimodal biometric authentication systems using convolution neural network based on different level fusion of ECG and fingerprint". In: *Ieee Access* 7 (2018), pp. 26527–26542.

[7] Warid Islam et al. "Improving performance of breast lesion classification using a ResNet50 model optimized with a novel attention mechanism". In: *Tomography* 8.5 (2022), pp. 2411–2425.

[8] Nitish Kundu. *Exploring Resnet50: An in-depth look at the model architecture and code implementation*. Jan. 2023. URL: https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at-the-model-architecture-and-code-implementation-d8d8fa67e46f.

[9] Xiaoxuan Liu et al. "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis". In: *The lancet digital health* 1.6 (2019), e271–e297.

[10] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. "Deep learning for medical image processing: Overview, challenges and the future". In: *Classification in BioApps: Automation of decision making* (2018), pp. 323–350.

[11] Shadan Alam Shadab et al. “Detection of cancer from histopathology medical image data using ML with CNN ResNet-50 architecture”. In: *Computational Intelligence in Healthcare Applications*. Elsevier, 2022, pp. 237–254.

[12] Mark Sorin et al. “Single-cell spatial landscapes of the lung tumour immune microenvironment”. In: *Nature* 614.7948 (2023), pp. 548–554.

[13] Yunan Xu et al. “Machine learning prediction of neurocognitive impairment among people with HIV using clinical and multimodal magnetic resonance imaging data”. In: *Journal of neurovirology* 27 (2021), pp. 1–11.

[14] Keyue Yan et al. “A review on multimodal machine learning in medical diagnostics”. In: *Math. Biosci. Eng* 20.5 (2023), pp. 8708–8726.

[15] Yajing Zeng et al. “A multimodal parallel method for left ventricular dysfunction identification based on phonocardiogram and electrocardiogram signals synchronous analysis”. In: *Mathematical Biosciences and Engineering* 19.9 (2022), pp. 9612–9635.