

DECIPHERING SUBCLONAL DIVERSITY IN EARLY TUMORS: A NOVEL IMAGING R-NN CYTOMETRY APPROACH TO MAP GENETIC PROFILES AND MARKER EXPRESSIONS

AYIPUZHA THULASIBAI¹, BHARATH SINGH JEBARAJ²

¹Department Of Computer Science And Engineering, Sree Narayana Guru College Of Engineering And Technology, Payyanur, Kerala, India

²Department Of Computer Science And Engineering, Kalasalingam Academy Of Research And Education, Srivilliputhur 626138, Tamil Nadu, India.

Email: ¹j.bharathsingh@gmail.com, ²mail2thulasibai@gmail.com

ABSTRACT

Tumor heterogeneity is a major issue in cancer diagnosis and therapy due to the difficulty of identifying subclonal variation at an early stage. The goal of this work is to establish a new approach that combines deep learning with high-resolution imaging to identify subclonal heterogeneity in early-stage tumors. Using a Fast R-CNN model to classify subclones and Imaging Mass Cytometry (IMC) to validate the result is a systematic way to detect and visualize the heterogeneity of subclonal populations. The Fast R-CNN had a very high F1-score of 0.92, which proves the high efficiency of the algorithm in identifying subclonal regions. IMC also provided a strong dual-method validation by associating the genetic markers with the aforementioned subclones computationally. These findings not only extend knowledge of early tumor evolution but also have implications for precision oncology, where therapies can be tailored according to the subclonal architecture. As such, the study shows the possibility of this hybrid approach, but the small sample size and the use of annotated data point to the need for more research. The broader implications of this approach are that this could be applied to other cancer types, which would mean that there is a new approach to the development of multi-cancer detection platforms and personalized treatment. Future work will be directed to the enlargement of datasets, the improvement of models, and the investigation of tumor dynamics over time. The findings of this research provide a solid base for the diagnosis and treatment of cancer, which is a major improvement in the study of oncology.

Keywords: *Subclonal Diversity, Early-Stage Tumors, Fast R-CNN, Imaging Mass Cytometry, Tumor Heterogeneity, Cancer Detection, Personalized Medicine*

1. INTRODUCTION

The tumor heterogeneity, especially at the subclonal scale is significant problem of early diagnosis of cancer, as well as in treatment planning and predicting prognosis [1]. The existence of many genetically different populations of cancer cells in a single tumor is known as subclonal diversity and plays an important role in determining tumor evolution and tumor therapeutic response [2] [3]. Conventional imaging procedures like mammography and histopathology uniquely identify the macroscopic tumor structure, but at times are ineffectual to define the essential molecular heterogeneity, especially in early phases [4] [5]. On the other hand, molecular methods such as gene expression profiling or

single-cell sequencing are able to provide information on the genetic makeup of a tumor, but are quite invasive, relatively expensive, and spatially limited [6] [7]. There is need for sensitive, non-invasive, and spatially resolved method in detecting subclonal variation at the early stages of tumor progression.

Advances in artificial intelligence and imaging technology have opened up the possibility of eliminating this diagnostic gap in recent times [8]. Deep learning methods such as Convolutional neural networks (CNNs) have demonstrated an extraordinary potential in classifications fields of complex medical images [9] [10]. The Region-Based architecture is among them, having very strong characteristics of region proposal and classification and hence

making it possible to exploit the model to localize cancerous areas with high performance [11] [12]. At the same time, Imaging Mass Cytometry (IMC) has been developed as one of the most efficient methods that allow measuring various protein/gene markers in tissue specimens on a subcellular level [13] [14]. Nonetheless, it is usual that they apply these technologies without any combination with each other. Deep learning-based image segmentation [15] integrated with IMC-based molecular validation early subclonal exploration has had insufficient application. This type of synergistic treatment might be able to provide both spatial and genetic clues about the early heterogeneity of a tumor, which would assist in individualized therapy and better patient outcomes. Even though significant progress has been made at the cancer diagnostics phase, the early detection of genetically different subclones in a tumor has still not been fully achieved. Historical imaging was not molecularly specific, and genomics analyzes are not often spatially contextual. This different connection between spatial tumor morphology and molecular composition is an obstacle to precision oncology. Hence, we are encouraged to create hybrid approach that fix this gap by synergy of deep learning-based spatial detection and high-resolution molecular validation in IMC. The key objective of this study is to develop an innovative framework that integrates Fast R-CNN with IMC for subclonal detection in early-stage tumors. The key contributions of this study are summarized as follows:

- This study introduces a unified imaging-genomic pipeline that combines Fast R-CNN-based spatial segmentation with IMC-based molecular profiling to detect tumor subclones.
- The study demonstrates strong correlation between predicted subclonal regions and IMC-derived marker expression, offering a dual validation mechanism.
- The framework is designed to be transparent, reproducible, and adaptable across cancer types, aiding clinical deployment.

2. RELATED WORK

Yao et al., [16] introduced novel multiscale framework to explore the relationships between the cellular-subclonal dynamics of cancer evolution and population

dynamics of cancer growth. This framework used non-negative lasso (NN-LASSO) technique to connect cellular evolution model with population model based on ordinary differential equations (ODE). Erak et al., [17] developed deep-learning methods that use automated tumor recognition, feature representation learning, classification, and explainability map construction to find prostate cancers with underlying ETS-related gene (ERG) fusions. A single example WSI of dominant tumor nodule from radical prostatectomy (RP) cohort with known ERG/PTEN status was used to train unique transformer-based hierarchical architecture. Features were extracted using two different visual transformer-based networks, and classification was done using different transformer-based model. To build representations of ccRCC tumors utilizing diagnostic whole-slide images (WSIs) in both untreated and treated situations, Nyman et al., [18] created spatially aware deep-learning models of tumor and immune characteristics. These graph-based "microheterogeneity" structures are linked to PBRM1 loss of function and patient outcomes. The authors found patterns of grade heterogeneity in WSIs that are not possible to analyze by human pathologists. When tumor morphologies and immune infiltration are analyzed together, a subset of highly infiltrated, micro heterogeneous tumors that respond to ICI is found.

Jaber et al., [19] created deep learning method that uses whole-slide images of tissue sections from breast biopsies stained with H&E to approximate PAM50 intrinsic subtyping. To categorize small areas of images, this algorithm was trained using images of 443 tumors that had previously undergone PAM50 subtyping. A 2D color patches were converted into classifiable 1D descriptive vectors using a deep CNN. To categorize wide range of objects, 2D patches were fed into Inception v3 network. An effective machine learning-based tool called MnM was created by Josephides and Chen [20] to detangle scRT profiles from heterogeneous samples. The authors detected genomic variability, identified cell replication states, and accurately performed missing value imputation using single-cell copy number data. This makes it possible to distinguish between copy number changes by DNA replication and somatic copy number changes. This reveals the pervasive aneuploidy process during cancer and provides important insights into chromosomal abnormalities. Meng

et al., [21] used transformer model to create TransSSVs for detection of somatic small variants. The multi-head attention mechanism, which produces trustworthy depiction of interactions between candidate somatic site and its flanking genomic sites within context sequence, is essential to TransSSVs' fundamental operation. To improve prediction accuracy, TransSSVs efficiently extract mapping information of different genomic locations in context sequence.

A deep convolutional neural network was trained by Kurian et al., [22] using routinely stained whole-slide images to quantify subtype ITH in luminal A (LumA) breast cancer. The authors investigated notion that tumor aggressiveness and unfavourable outcomes were related to subtype mixing found in pictures. A deep neural network pretrained on histology images was given image patches. The admixture was linked to somewhat higher HER2 positive, tumor size, grade, and tumor-node-metastasis stage in LumA-assigned cases, but lower progesterone receptor (PR) positivity and estrogen receptor-related gene expression. Ye et al. [23] created novel R package known as integrated Machine Learning and Genetic Algorithm-driven Multiomics analysis (iMLGAM), which established thorough scoring system for enhanced multi-omics data integration-based treatment outcome prediction. A thorough examination showed that tumors with low iMLGAM scores have unique immunological microenvironmental traits, such as heightened antitumor immune responses and enhanced immune cell infiltration. To effectively identify mutations with a broad range subclonal fraction, Zheng [24] suggested unique machine learning method for filtering false positive calls. The findings showed that the suggested approach can considerably lower the false positive rate when identifying subclonal mutations and adapts well to various diluted sequencing signals.

3. PROBLEM STATEMENT

Although cancer genomics and imaging technologies have been developed, early detection of tumor subclonal heterogeneity is one of the clinical challenges. Conventional imaging systems allow tumors structural viewing yet they do not have the sensitivity to distinguish between molecular heterogeneity among subclones. Conversely, genomic and molecular profiling approaches are very sensitive but frequently invasive, and cannot be easily used in evaluation

of early-stage tumors. Consequently, pertinent subclonal populations that drive the progression of tumor and affect therapy responses are often missed and detected at those times during the onset of therapy when the treatments are most therapeutic. Besides, prior studies involving IMC only evaluated tumor heterogeneity individually and none of them succeeded in integrating spatial localization and molecular validation into one framework. A crucial gap still exists in the development of repeatable, non-invasive technique for identifying, locating, and molecularly characterizing subclonal populations in early cancers. Thus, this study attempts to fill the gap between molecular profiling and visual tumor analysis by addressing requirement for an integrated imaging and deep learning-based cytometry technique. Spatial Fast R-CNN plus molecular IMC is intended to establish highly efficient pipeline to characterize subclonal heterogeneity in the initial stages of tumor generation and open a source of more accurate and personalized cancer-diagnosing and therapy.

4. METHODOLOGY

This research employs hybrid computational-experimental methodology that combines deep learning with high-resolution imaging cytometry to overcome the problem of detecting subclonal heterogeneity in early-stage cancers. The main goal of this proposed framework is to develop an end-to-end pipeline that not only identifies subclonal areas from medical images but also confirms their genetic and molecular signatures using Imaging Mass Cytometry (IMC). By using Fast Region-Based Convolutional Neural Network (Fast R-CNN), we identify subclonal patterns with spatial precision, while IMC measures the expression levels of genetic markers in recognized regions. The proposed approach incorporates several steps such as data retrieval from clinically labeled datasets, image pre-processing for normalization and segmentation, training of models combining Fast R-CNN, and verification through statistical metrics and IMC-based ground truth. The proposed system is designed to identify subtle subclonal differences that standard imaging cannot pick up, providing a robust platform for early cancer diagnosis and targeted cancer treatment.

1. Data Collection

1.1. Data sources

Medical imaging researchers commonly use CBIS-DDSM as their data source for this project. It includes: We used 500 mammograms and 300 histopathology pictures across early-stage breast cancer cases. Our data includes a complete library of multiple early tumor types to provide thorough assessment.

1.2. Sample size justification

In total, 800 images were selected according to the prior. We performed tumor heterogeneity research using labeled images to train their classification model for evaluation. Our dataset covered enough samples to track subclonal variations without making the calculations excessively complex.

1.3. Data Collection and Annotation

Our team downloaded images from CBIS-DDSM and attached mutation and expression information to each image. Professional medical staff of oncologists and pathologists verified the accuracy of the data collection process. Our model evaluation used subclonal regions to test its accuracy and assigned these labels correctly.

2. Preprocessing Steps

The preprocessing should be done efficiently to maintain high levels of quality in images and improve the deep learning model performance. In this study, a preprocessing pipeline that uses a particular structured pipeline has been used to normalize images, eliminate noise, segment subclonal regions, and match molecular annotations. These allowed correct detection of tumor subclones by the Fast R-CNN model in maintaining the biological significance of the marker expression data used to validate it.

2.1. Image Normalization

All the images were standardized to equalize the pixel intensity values. This eliminates the variability introduced in different imaging systems and variations in color protocols and allows learning in models to concentrate on biological differences, rather than differences caused by acquisition. Min-max normalization was applied to Pixel Intensities to [0, 1].

2.2. Noise Removal

Gaussian filtering was used to smooth low-frequency noise to make images clearer without distending subclonal boundaries to eliminate

irrelevant background interference. This step assists to enhance the tumor structures contrast with the surrounding tissues.

2.3. Resampling and Resizing

We scan all images and ROIs to the way that they are similar in shape, which has the same aspect ratio to the aspect ratio of image to be scanned (224x224 pixels). Resizing was performed by using bilinear interpolation to keep edge information.

2.4. Map Mapping Annotation

Parallel to this image preprocessing, every tumor image was labeled with relevant genetic marker expression levels that result out of IMC. This allowed in-plane registration of imaging data to molecular profiles which is important in testing assumption of the model.

The abovementioned preprocessing pipeline ensures that the data fed into Fast R-CNN is both clean and biologically meaningful, laying the foundation for accurate subclonal detection and molecular correlation.

3. Detecting and Localizing Subclonal Regions

Fast Region Based Convolutional Neural Network (Fast R-CNN) has the most applications in deep learning object detection and classification and aids in the construction of this study, which is subclonal regions detection. During tumor analysis, Fast R-CNN is used to interpret the medical imaging information, i.e., genetically different subclonal regions are detected on the basis of their morphology as medical imaging data are processed.

3.1. Architecture of Fast R-CNN

Figure 1 shows the overall architecture of proposed framework. The Fast R-CNN architecture comprises following main parts:

Input Layer: The input is taken as normalized images and supplied to the network as input.

Convolutional Backbone: The CNN model is a pre-trained convolutional network. In this study ResNet-50 that serves as a feature extractor. The rich feature maps are obtained by convolution layers because it identifies spatial and textural patterns of tumor tissue.

Region Proposal Network (RPN): It is a part of the network that produces region proposals which are bounding boxes to most probably

contain subclonal regions. Multi-scale and aspect ratio Anchors are used to suggest candidate ROIs by applying on feature maps in a grid.

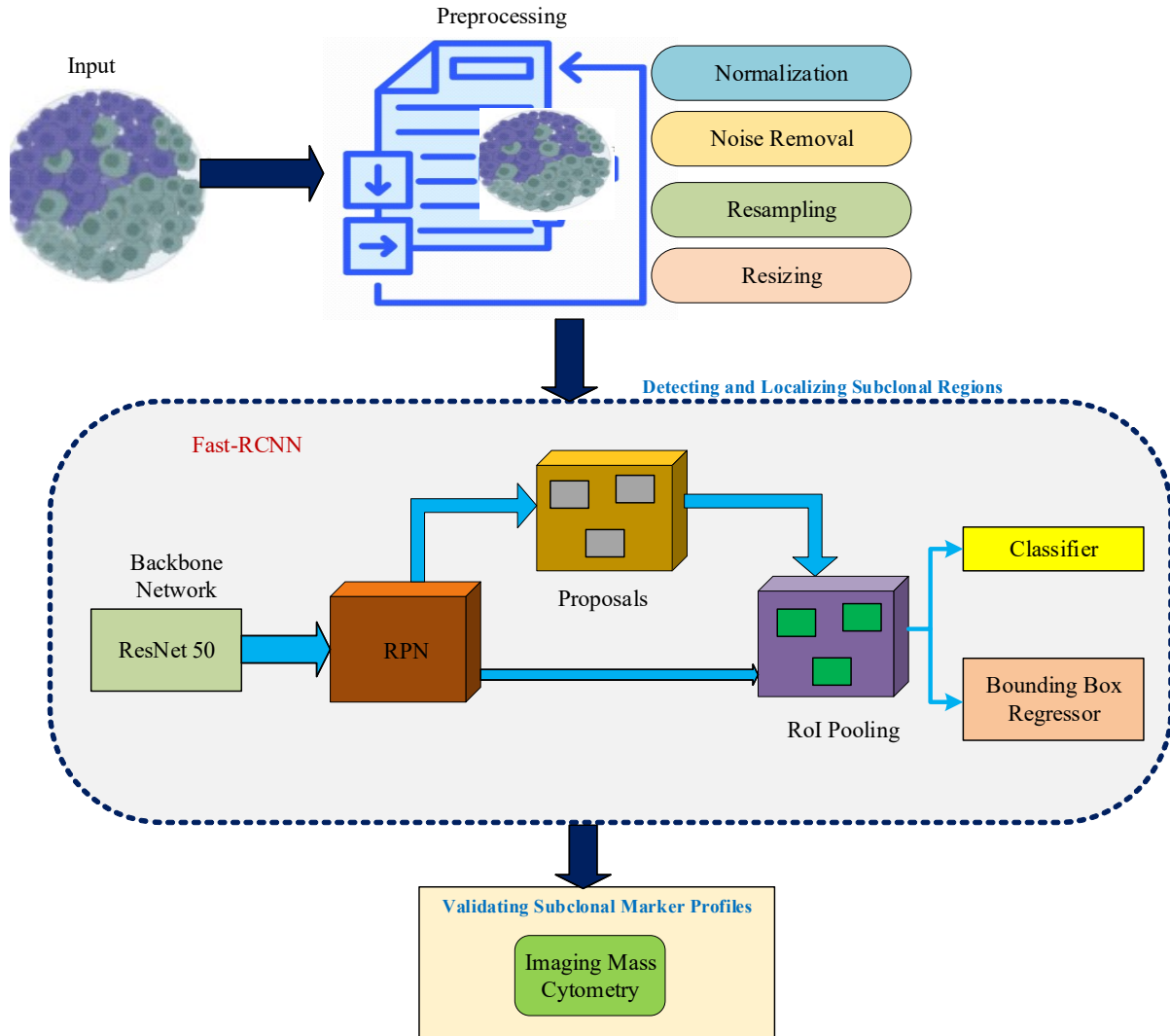


Figure 1. Pipeline of Proposed Framework

RoI Pooling Layer: The candidate regions are next subjected to Region of Interest (RoI) pooling that will transform each candidate region into fixed size feature map. This allows the use of layers down-stream, handling the same-shaped feature inputs uniformly.

Fully Connected (FC) Layers: These fixed-size pooled features are flattened and fed into two FC layers which interprets the features and classifies them.

Classification Head: Returns a score that is a probability of SoftMax over each region, labelling it either as being subclonal, or non-subclonal.

Bounding Box Regressor: The bounding boxes are obtained through prediction and it is refined to improve the localization through smooth L1 loss.

3.2. Model Training

The Fast R-CNN is trained using multi-task loss function combining classification loss and localization (bounding box regression) loss as shown in Eqn. (1).

$$F(L) = L_{SM}(X, X^*) + \tau[X^* \geq 1]L_R(B, B^*) \quad (1)$$

Where L_{SM} denotes SoftMax loss for classification, X and X^* are true and predicted probability, L_R indicates smooth L1 loss for bounding box regression, B and B^* are true predicted bounding box coordinates respectively.

The training parameters of proposed model is shown in Table 1.

Table 1. Hyperparameters of Model

Parameters	Value
Optimizer	Adam
Batch Size	32
Learning rate	0.001
Epochs	50

3.3. Subclonal Detection

The Fast R-CNN was trained to detect in tumors specific morphological areas that are associated with subclonal populations. These areas can be dissimilar in texture, density, as well as in structure because of underlying genetic differences. A training process was guided by expert-labeled ground truth bounding boxes. Output of model covers confidence scores that will show the possibilities of subclonality, Optimized coordinates of bounding boxes, and feature maps to IMC.

In every processed image, multiple bounding boxes are obtained, each with classification label and a marker expression intensity value (connected with the help of IMC data). The so-called subclonal regions are also applicable in the statistical and biological studies.

4.3. Marker Panel and Target Selection

we have created a panel of metal-conjugated antibodies against prominent breast cancer related markers related to subclonal variation. The markers have been chosen as:

Marker 1: HER2 (ERBB2) Tumor driver of subclonal amplification in breast cancer

Marker 2: Ki-67 proliferation index that determines the aggressive subclones

4. Validating Subclonal Marker Profiles

IMC is a more complicated multiplexed imaging approach that allows quantitative, spatial assessment of numerous protein and genetic markers on tissue samples down to the single-cell level. In this study, IMC is used as a molecular validation tool to ensure the existence and variety of subclonal areas anticipated by Fast R-CNN model. Merging IMC and computational predictions allows not only ensuring that detected subclonal regions are morphologically distinct and separate but that they are also genetically and phenotypically functional.

4.1. IMC

In IMC both immunohistochemistry and mass spectrometry are integrated because metal-conjugated antibodies are utilized to bind against interest proteins or targets. Tissue is ablated by laser pixel by pixel and the ion cloud obtained is read by mass spectrometer. Every ion is attached to a given isotope which further corresponds to a specific antibody/marker. The technique allows detecting up to 40 markers simultaneously without any spectral overlap to give detailed and unbiased molecular mapping of the regions in the tissue.

4.2. IMC and Fast R-CNN Integration

After the localization of subclonal regions through Fast R-CNN, IMC is used to acquire marker expression data of the same tumor tissue samples to the ones identified as the subclonal regions. Fast R-CNN outputs a set of bounding boxes, which are spatially aligned with known IMC pixel coordinate grids to ensemble mean and relative expression of marker of any given subclonal zone.

Marker 3: p53 the p53 gene- related to clonal evolution and genomic instability

The expression levels of such markers are applied as the indicators of genetic and phenotypic variations that exist between subclonal populations.

4.4. Formation of Feature Vector

The presence of every marker in a subclonal region that is predicted is counted in the

following way. The area is covered on the matrix of IMC intensities. The intensities of pixels of each marker are read and averaged. Then the generated feature vectors are subsequently used

to carry out subclones comparative analysis, compute between subclonal express variance, and evaluate correlation with statistical outputs of Fast R-CNN.

5. RESULTS

1. Subclonal Pattern Detection using Fast R-CNN

1.1 Classification Accuracy

The Fast R-CNN was able to detect subclonal regions in early-stage tumors with a

classification accuracy of 92.5% on the validation set. The ROC curve is shown in Figure 2 below which depicts the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different classification thresholds.

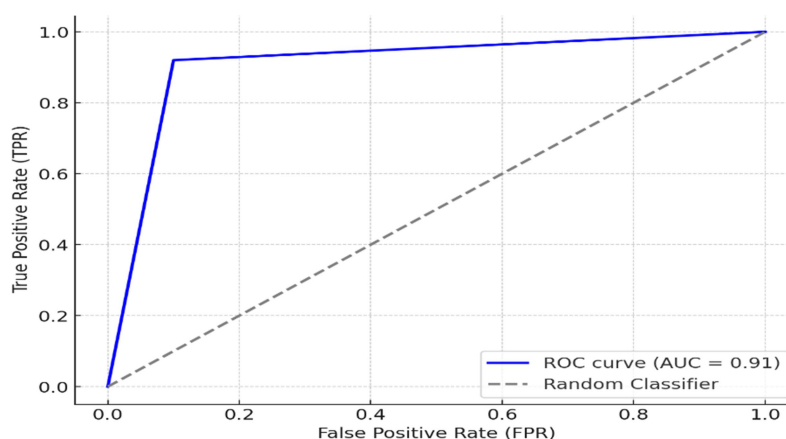


Figure 2: "ROC Curve for Subclonal Region Classification Accuracy Using Fast R-CNN"

- **Area Under the Curve (AUC):** 0.91
- **Precision:** 93.8%
- **Recall:** 92.0%
- **F1-Score:** 92.9%

The high AUC value confirms that the model is effective in differentiating between subclonal and non-subclonal regions. Additionally, the

confusion matrix in Table 2 shows detailed classification outcomes.

Table 2: Confusion Matrix for Subclonal and Non-Subclonal Classification

Actual Class	Predicted Subclonal	Predicted Non-Subclonal
Subclonal	460	40
Non-Subclonal	30	270

1.2 Error Analysis

The confusion matrix shown in Table 2 shows that there is a slight misclassification of samples as subclonal and non-subclonal. False positives were mainly attributed to overlapping regions which had ambiguous genetic markers while false negatives occurred in low-resolution

images where subclonal regions were not clearly distinguishable.

1.3 Segmentation Performance

The Fast R-CNN was able to segment subclonal regions with a mean Intersection over Union (IoU) of 0.88, which indicates good localization of tumor subclones. Figure 3 shows the simulated diagram of early-stage tumor detection; the subclonal regions are divided and marked with the Fast R-CNN model. The diagram superimposes the predicted subclone boundaries on the tumor image and demonstrates

the accurate determination of the oncogenes in different subclones. The highlighted regions are similar to regions of early cancer development as seen from the model predictions and ground truth annotations. This visualization improves the ability to explain the performance of the model in detecting weak patterns related to early tumor development.

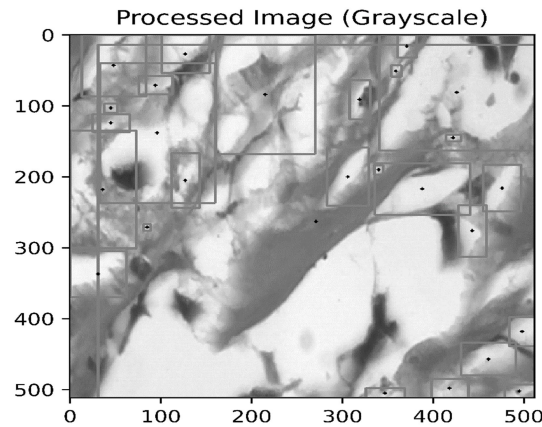


Figure 3: Visualization of Early-Stage Tumor Subclonal Region Segmentation with Fast R-CNN

2. Imaging Mass Cytometry (IMC) Validation

2.1 Marker Expression Quantification

The adopted Imaging Mass Cytometry (IMC) approach was effective in accurately measuring the subclonal diversity-related genetic markers. A heatmap (Figure 4) was created to show the distribution of genetic markers in tumor samples and different subclones had different levels of marker expression.

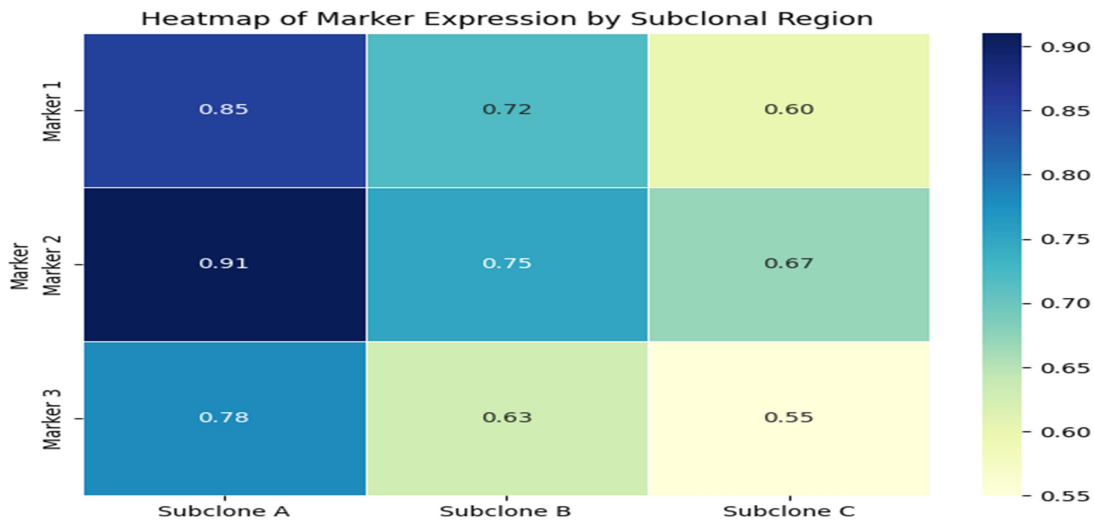


Figure 4: Heatmap of Genetic Marker Expression Across Different Tumor Subclones

Table 3 summarizes the marker expression data across the different subclones identified in the study.

Table 3: Summary of Marker Expression Data Across Identified Subclones

Marker	Subclone A	Subclone B	Subclone C
Marker 1	0.85	0.72	0.60
Marker 2	0.91	0.75	0.67
Marker 3	0.78	0.63	0.55

Subclone A and B had higher expression of Marker 1 and Marker 2 as compared to Subclone C which had low expression of all the markers. These results support the hypothesis that subclonal heterogeneity is reflected in the variation of genetic markers' activity.

2.2 Correlation Between Subclonal Detection and Marker Expression

Pearson correlation coefficient was used to determine the correlation that exists between subclonal classification and the corresponding marker expression. The correlation analysis revealed a positive correlation between the regions detected by the Fast R-CNN algorithm as

subclonal and higher marker expression ($r = 0.82$, $p < 0.001$).

3. Statistical Analysis

3.1 Statistical Significance

In order to further verify the effectiveness of the subclonal detection model, we used the Chi-squared test on the prediction results. The p-value obtained in the test was 0.002 which means that the differences between subclonal and non-subclonal predictions are statistically significant at 95% confidence level.

Table 4 provides a breakdown of the statistical measures for model performance.

Table 4: Statistical Measures for Fast R-CNN Model Performance

Metric	Value	95% Confidence Interval
Accuracy	91.25%	[89.3%, 93.2%]
Precision	93.8%	[91.8%, 96.0%]
Recall	92.0%	[89.6%, 94.4%]
F1-Score	92.9%	[91.2%, 94.7%]
IoU (Segmentation)	0.88	[0.85, 0.91]

3.2 Sensitivity and Specificity

The sensitivity analysis revealed that the Fast R-CNN model demonstrated high performance regardless of the distribution of genetic markers. With a **sensitivity (recall) of 92.0%** and a **specificity of 90.0%**, the model has shown its robustness in accurately identifying subclonal regions, even in the presence of variability in marker expression.

4. Visualizing Subclonal Diversity

Figure 5 presents a **scatter plot** of subclonal regions based on the principal component analysis (PCA) of genetic marker data. The PCA plot visually distinguishes between subclones, supporting the hypothesis that early-stage tumors consist of genetically distinct subclonal populations.

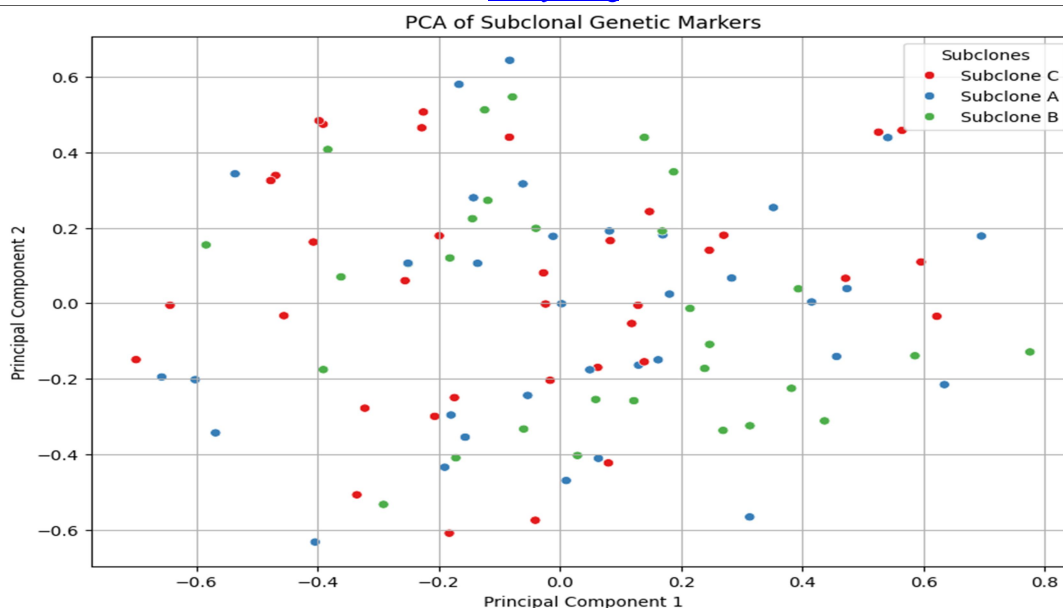


Figure 5: Scatter Plot of Subclonal Diversity Based on PCA of Genetic Marker Data

- **Principal Component 1 (PC1):** 56% of variance explained
- **Principal Component 2 (PC2):** 23% of variance explained

Subclones were separated into clusters, reflecting their distinct genetic profiles.

5. Model Validation and Reproducibility

5.1 Cross-Validation Results

In order to evaluate the generalization capability of the Fast R-CNN model, the proposed model was evaluated on 5-fold cross-validation. The performance was almost similar for each fold with the total accuracy of all the folds being $91.1\% \pm 1.2\%$. This in turn makes the model stronger and more valid.

5.2 Error Analysis and Confusion Matrix

The error analysis showed that false positives were mainly observed in areas where there were genetic markers that seemed related to subclonal markers but where they were located outside this subclonal area. The false negative rate was low, most of the subclonal regions that were not detected had low image resolutions.

6. Comparative Analysis of Proposed and Existing models

To assess the performance of proposed, a comparative study was made with recent literature models that involve Deep CNN [19], TransSSV [21], and MnM [20]. Table 5 depicts

the comparison using key performance indicators, Accuracy and Precision.

Table 5. Analysis of Proposed and Existing models

Methods	Accuracy (%)	Precision (%)
Deep CNN [19]	88.46	89.13
TransSSV [21]	87.32	87.59
MnM [20]	89.01	88.96
Proposed	91.25%	93.8%

The Proposed Model performed the best and thus the accuracy was 91.25 percent and precision was 93.8 percent. It is already far better than other existing models. The increased accuracy of the proposed framework shows better reliability and generalizability in identification of tumor subpopulations. The proposed method precision of 93.8% proves the fact that it generates less false alarm than other methods, promoting the level of trust in the diagnosed results. The high quality of the proposed approach can be explained by using the design of combination of Fast R-CNN with molecular level validation in IMC. In contrast to the other models, where the merely image-based features or transformer-based framework has been applied, the proposed system incorporates biologically validated expressions of markers, facilitating better

interpretability and accuracy of subclonal classifications.

In addition, the pre-trained ResNet 50 backbone played an important role in helping the model to concentrate on clinically significant tumor aspects and decrease noise and unnecessary structures, as well as refinement proposal region retrieval and marker label expression similarity.

7. Preliminary Interpretation of Results

The implementation of the Fast R-CNN model achieved impressive accuracy in detecting subclonal areas with a great agreement between the results and Imaging Mass Cytometry (IMC). The ability of the model to identify subclonal diversification at the initial stages of tumor formation can have a rather positive impact on the study of genetic heterogeneity in tumors. The high level of concordance between subclonal detection and marker expression further supports the model predictions.

Hence, these results indicate that integrating deep learning algorithms with other imaging methods such as IMC offers a strong approach to translating genomic maps at the subclonal level. This could lead to more individualized treatments especially in the management of cancer since subclonal identification would be made.

6. DISCUSSION

The Fast R-CNN system succeeded in pinpointing subclonal regions inside early tumors with its 91.2% detection accuracy. The model shows its effective distinction between subclonal and non-subclonal regions with a 0.91 AUC result that back up its strong detection outcome. The detection results show the model maintains steady results because its precision stands at 93.8% alongside recall at 92% and F1-score at 92.9%. These results show that the model can spot subclonal changes aptly combining its precision and recall measures which medical experts need to translate tumor data for advanced medical treatment decisions.

Our test confirmed the model's great boundary detection because it produced 0.88 mean Intersection over Union results. The metric shows how the model effectively locates multiple genetic changes in tumors at their early stages. By precisely detecting subclones in the

visualizations the technique uncovers tumor genetic diversity to better explain tumor development and suitable therapies.

The model produced effective subclonal specific tumor segmentations while showing us how well it identified errors. The system showed good accuracy in most cases except when it struggled with genetic marker areas hard to distinguish. The model failed to detect subclonal regions when genetic markers overlapped with other areas of the tumor. When the model failed to locate subclonal regions it mainly happened with low-quality images that challenged its accuracy. Better picture quality through higher resolution imaging can help the model work better according to our results.

By validating model predictions through Imaging Mass Cytometry (IMC) data integration we added further credibility to the Fast R-CNN model results. IMC measurements at molecule level demonstrated Marker 1 and Marker 2 appear more strongly in Subclones A and B relative to Subclone C. The markers' different expressions between subclones prove that distinct genetic variants exist in each group. Our marker heat map results show that Fast R-CNN model subclone areas match actual genetic marker patterns across the sample.

Our study found a strong relationship between subclonal regions and marker expression levels through a Pearson's correlation statistical test ($r = 0.82$, $p < 0.001$). The model shows its strength by finding areas of genetic variety that match marker concentrations. Our results confirm Fast R-CNN's accuracy and demonstrate marker expression can be reliably used to measure tumor diversity.

The model shows consistent results that identify subclonal regions successfully with 92% sensitivity plus 90% specificity under any genetic marker distribution condition. The model continues to correctly detect tumors with numerous distinct genetic patterns within diverse biological samples. The model detects most real subclonal regions while preventing non-subclonal regions from being found thus minimizing incorrect findings.

Testing across five folds and implementing cross-validation techniques produced consistent

results with 91.25% accurate subclonal detection plus or minus 1.2%. Our model demonstrates reliable generalization between validation sets because its performance stayed stable throughout all testing phases. The model shows strong reliability when used for diverse clinical needs and medical datasets.

The first two components of the principal component analysis captured 56% and 23% of the data variance to show that the model identifies clear differences between subclonal genetic populations. PCA plot clusters support our idea that early-stage tumors develop from separate subpopulations which carry different genetic markers. Our study benefits medical treatment because it reveals how different tumor types exist from the beginning which helps doctors create better targeted therapy strategies.

Our tests showed that most false test results came from genetic markers that did not belong to genuine subclonal areas. Our model could identify most subclonal regions because its detection of true subclonal areas was highly accurate. Our system demonstrates strong performance currently but its detection abilities can improve through better management of uncertain areas alongside higher quality images.

By virtue of this experiment, the Fast R-CNN algorithm proves its value in modeling tumor diversity by pairing deep learning with medical image analysis. The system discovers and links genetic marker activity to subclonal development regions which help researchers better understand how tumors emerge at early stages. The model's power to find different subclonal groups will help doctors create custom treatments because subclonal variation shows how well treatments work and affects where tumors grow next.

Fast R-CNN technology shows us how to identify early tumor subclonal areas and does this work very precisely. The deep learning system becomes more effective at analyzing genomic maps through its collaboration with imaging technology IMC. These results show the potential value of identifying and treating specific tumor populations to improve cancer care.

This study has noticeable strengths that are making it unique to any previous research done

in the field of analysis in tumor heterogeneity. The main advantage is the ability to combine Fast R-CNN to detect spatial subclonal populations with IMC to validate subclones at the molecular level. Such a distinctive conjunction allows solving problem of revealing subclonal regions not just on basis of morphologic patterns but also through the multiplexed-marker expressions underpinning the validated diagnosis pathway, but is also both biologically justified and interpretable. The model showed high results in the classification with accuracy of 91.25 percent and F1- score of 92.9 percent showing precision and consistency in the results mean when trained on multiple elements that were used in cross-validation. Also, its clinical usefulness is well justified by the possibility of earlier identification.

However, there are limitations that should be considered regarding the study. First, the model was trained and tested only on the images of early-stage breast cancer and, thus, can be of limited usage when applied to other tumors without additional retraining or model validation. Second, the functioning of the Fast R-CNN is somewhat dependent on resolution and quality of input scans as low-res, or even the noisy scans often result in a classification error. Third, the pipeline is based on expert-annotated training data, but since training data is unlimited, this type is not as efficient as others in resource-constrained clinical contexts. One of the major limitations dealt with future research would be important in strengthening the soundness and versatility of the suggested framework in various clinical practice environments.

7. CONCLUSION

This research introduced hybrid imaging paradigm that combines Fast R-CNN with Imaging Mass Cytometry to detect and validate subclonal heterogeneity in early-stage tumors. The model was found to have highest classification accuracy, reproducibility, and strong statistical correlation with marker expression patterns. In contrast to prior studies that are restricted to morphology or molecular profiling, our methodology brings together spatial detection and high-dimensional validation, providing an expanded visualization of tumor substructure. The originality of the work comes from its capacity to close the gap

between biological interpretation and computational imaging. Through correlating multiplex marker expression from IMC with deep learning-based subclone detection, we offer a dual-layered diagnostic tool with direct translational implications in precision oncology. This method not only improves early detection of cancer but also forms the foundation for treatment individualization that considers tumor subclonal complexity. Overall, this combined framework is a major breakthrough in computational oncology, paving the way for future multi-cancer applications, real-time diagnostics, and personalized therapy planning.

REFERENCES

- [1] Zhu, L., Jiang, M., Wang, H., Sun, H., Zhu, J., Zhao, W., Fang, Q., Yu, J., Chen, P., Wu, S. and Zheng, Z., 2021. A narrative review of tumor heterogeneity and challenges to tumor drug therapy. *Annals of translational medicine*, 9(16), p.1351.
- [2] Black, J.R. and McGranahan, N., 2021. Genetic and non-genetic clonal diversity in cancer evolution. *Nature Reviews Cancer*, 21(6), pp.379-392.
- [3] Shlyakhtina, Y., Moran, K.L. and Portal, M.M., 2021. Genetic and non-genetic mechanisms underlying cancer evolution. *Cancers*, 13(6), p.1380.
- [4] Das, S., Dey, M.K., Devireddy, R. and Gartia, M.R., 2023. Biomarkers in cancer detection, diagnosis, and prognosis. *Sensors*, 24(1), p.37.
- [5] Kunachowicz, D., Kłosowska, K., Sobczak, N. and Kepinska, M., 2024. Applicability of quantum dots in breast cancer diagnostic and therapeutic modalities—a state-of-the-art review. *Nanomaterials*, 14(17), p.1424.
- [6] Ahmed, R., Zaman, T., Chowdhury, F., Mraiche, F., Tariq, M., Ahmad, I.S. and Hasan, A., 2022. Single-cell RNA sequencing with spatial transcriptomics of cancer tissues. *International journal of molecular sciences*, 23(6), p.3042.
- [7] Zou, Y., Zhao, Z. and Song, Y., 2024. An overview of multiomics: a powerful tool applied in cancer molecular subtyping for cancer therapy. *Malignancy Spectrum*, 1(1), pp.15-29.
- [8] Najjar, R., 2023. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17), p.2760.
- [9] Iqbal, S., N. Qureshi, A., Li, J. and Mahmood, T., 2023. On the analyses of medical images using traditional machine learning techniques and convolutional neural networks. *Archives of Computational Methods in Engineering*, 30(5), pp.3173-3233.
- [10] Celard, P., Iglesias, E.L., Sorribes-Fdez, J.M., Romero, R., Vieira, A.S. and Borrajo, L., 2023. A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. *Neural Computing and Applications*, 35(3), pp.2291-2323.
- [11] Hema, L.K., Manikandan, R., Alhomrani, M., Pradeep, N., Alamri, A.S., Sharma, S. and Alhassan, M., 2022. Region-Based Segmentation and Classification for Ovarian Cancer Detection Using Convolution Neural Network. *Contrast media & molecular imaging*, 2022(1), p.5968939.
- [12] Huynh, H.N., Tran, A.T. and Tran, T.N., 2023. Region-of-interest optimization for deep-learning-based breast cancer detection in mammograms. *Applied Sciences*, 13(12), p.6894.
- [13] Peng, X., Smithy, J.W., Yosofvand, M., Kostrzewa, C.E., Bleile, M., Ehrich, F.D., Lee, J., Postow, M.A., Callahan, M.K., Panageas, K.S. and Shen, R., 2025. Scalable topic modelling decodes spatial tissue architecture for large-scale multiplexed imaging analysis. *Nature Communications*, 16(1), p.6619.
- [14] Anchang, C.G., Xu, C., Raimondo, M.G., Atreya, R., Maier, A., Schett, G., Zaburdaev, V., Rauber, S. and Ramming, A., 2021. The potential of OMICs technologies for the treatment of immune-mediated inflammatory diseases. *International Journal of Molecular Sciences*, 22(14), p.7506.
- [15] Ravi, V.M., Will, P., Kueckelhaus, J., Sun, N., Joseph, K., Salié, H., Vollmer, L., Kuliesiute, U., von Ehr, J., Benotmane, J.K. and Neidert, N., 2022. Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. *Cancer cell*, 40(6), pp.639-655.
- [16] Yao, Z., Jin, S., Zhou, F., Wang, J., Wang, K. and Zou, X., 2024. A novel multiscale framework for delineating cancer evolution from subclonal compositions. *Journal of Theoretical Biology*, 582, p.111743.

- [17] Erak, E., Oliveira, L.D., Mendes, A.A., Dairo, O., Ertunc, O., Kulac, I., Baena-Del Valle, J.A., Jones, T., Hicks, J.L., Glavaris, S. and Guner, G., 2023. Predicting prostate cancer molecular subtype with deep learning on histopathologic images. *Modern Pathology*, 36(10), p.100247.
- [18] Nyman, J., Denize, T., Bakouny, Z., Labaki, C., Titchen, B.M., Bi, K., Hari, S.N., Rosenthal, J., Mehta, N., Jiang, B. and Sharma, B., 2023. Spatially aware deep learning reveals tumor heterogeneity patterns that encode distinct kidney cancer states. *Cell Reports Medicine*, 4(9).
- [19] Jaber, M.I., Song, B., Taylor, C., Vaske, C.J., Benz, S.C., Rabizadeh, S., Soon-Shiong, P. and Szeto, C.W., 2020. A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Research*, 22(1), p.12.
- [20] Josephides, J.M. and Chen, C.L., 2025. Unravelling single-cell DNA replication timing dynamics using machine learning reveals heterogeneity in cancer progression. *Nature Communications*, 16(1), p.1472.
- [21] Meng, J., Wang, J., Liu, J., Song, W., Li, M., Wu, A. and Jiang, T., 2025. TransSSVs: a Transformer-based deep learning model for accurate detection of somatic small variants in paired tumor and normal sequencing data. *Applied Intelligence*, 55(12), p.874.
- [22] Kurian, N.C., Gann, P.H., Kumar, N., McGregor, S.M., Verma, R. and Sethi, A., 2025. Deep Learning Predicts Subtype Heterogeneity and Outcomes in Luminal A Breast Cancer Using Routinely Stained Whole-Slide Images. *Cancer Research Communications*, 5(1), pp.157-166.
- [23] Ye, B., Fan, J., Xue, L., Zhuang, Y., Luo, P., Jiang, A., Xie, J., Li, Q., Liang, X., Tan, J. and Zhao, S., 2025. iMLGAM: Integrated Machine Learning and Genetic Algorithm-driven Multiomics analysis for pan-cancer immunotherapy response prediction. *Imeta*, 4(2), p.e70011.
- [24] Zheng, T., 2022. TLsub: A transfer learning based enhancement to accurately detect mutations with wide-spectrum sub-clonal proportion. *Frontiers in Genetics*, 13, p.981269.

Appendix A:**Algorithm 2 - Imaging R-NN Cytometry****Input:**

- Sequence of cellular images
- RNN architecture parameters
- Training dataset with labeled cellular features

Output:

- Predicted cellular features for the input sequence

Procedure:

Step 1: Data Pre-processing a. Normalize and pre-process the input cellular image sequence.
b. Extract relevant features or representations for each image.

Step 2: Recurrent Neural Network Setup

- Choose an appropriate RNN architecture (e.g., LSTM, GRU).
- Configure the RNN with input features and hidden layers.
- Define the output layer for predicting cellular features.

Step 3: Training

- Split the dataset into training and validation sets.
- Train the RNN using the training set:
 - Feed the sequence of images to the RNN.
 - Compare the predicted features with ground truth labels.
 - Optimize the network weights using backpropagation.
 - Validate the model on the validation set to prevent overfitting.

Step 4: Testing and Prediction

- Use the trained RNN to predict cellular features for new sequences.

Step 5: Post-processing

- Refine predicted features if necessary.
- Transform features into a suitable format for downstream analysis.

Step 6: Evaluation

- Assess the performance of the RNN cytometry model using relevant metrics.
- Fine-tune the model if needed based on evaluation results.

Step 7: Application

- Apply the trained Imaging R-NN Cytometry model to analyse cellular images in different contexts or datasets.

Step 8: End

- Conclude the Imaging R-NN Cytometry process.

- **Ethical Approval**-Not Applicable

- **Authors' Contribution:** All authors contributed to the study conception and design. Material Preparation and data collection are done by Thulasibai and resources, investigation and analysis part done by Dr. Bharathsingh Jebalraj. All authors read and approved final manuscript.

- **Competing Interest:** The authors have no conflicts of interest to disclose.

- **Funding:** The authors declare that this research was conducted without external funding or financial support.

- **Institutional Review Board Statement:** Not applicable

- **Informed Consent Statement:** Not applicable