SOFT TISSUE SARCOMA DIAGNOSIS USING MACHINE AND DEEP LEARNING

Dr Jim Mathew Philip[1], M Sakthivel[2], G Santhoshkumar[3], M M Suga Saranesh[4],

[1] *Dr Jim Mathew Philip, Associate Professor, CSE, Sri Ramakrishna Institute of Technology, Coimbatore, India.*

[2] *M Sakthivel, CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India*

[3] *G SanthoshKumar, CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.*

[4] *M M Suga Saranesh, CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India*

Abstract:

This project presents a novel hybrid framework integrating Multi- Constrained Joint Non-Negative Matrix Factorization (MC-JNMF) with Convolutional Neural Networks (CNNs) to study lung metastasis in soft tissue sarcomas using imaging and genomic data. By jointly analyzing high-dimensional data from medical images (such as CT or MRI scans) and corresponding genomic information, the proposed method seeks to identify critical latent features that contribute to the detection and understanding of lung metastasis. The MC-JNMF frameworks integrate multiple constraints to better reflect the biological relationships between the two modalities, while CNNs automatically extract imaging features to improve prediction accuracy. This hybrid approach aims to improve clinical decision-making by providing more interpretable models and more accurate predictions regarding lung metastasis development in sarcoma patients. Preliminary results indicate that the integration of imaging and genomic data enhances the model's performance compared to traditional methods.

[1] INTRODUCTION

Soft tissue sarcomas (STSs) are aggressive tumors with high risk of lung metastasis.Biomarker detection aids in diagnosis and drug development.Researchers use gene expression, DNA methylation, and radiomic data (e.g., PET, MRI, CT).CNNs help extract features from imaging data for better metastasis prediction.These models improve non-invasive diagnostics using radiomic features.To integrate multi-omics data, the MC-JNMF algorithm has been proposed.MC-JNMF combines PET and DNA methylation data using network regularization.It applies sparsity and orthogonal constraints to preserve key features.Functional analysis of MC-JNMF modules highlights key genes like GAPDH and CXCL12.This approach supports early STS diagnosis and improved patient outcomes.

1.1 General Introduction

This project focuses on the early diagnosis of soft tissue sarcomas (STSs), a rare but aggressive group of cancers known for their invasive behavior and high likelihood of lung metastasis. Accurate and timely detection is essential to improve patient survival and guide effective treatment strategies. To enhance diagnostic accuracy, the project integrates medical imaging data with genomic information. Deep learning techniques, particularly Convolutional Neural Networks (CNNs), are used to analyze radiomic features from scans, while a U-Net model performs tumor segmentation. Additionally, the Multi-Constrained Joint Non-Negative Matrix Factorization (MC-JNMF) algorithm is applied to combine imaging and omics data. This helps in identifying critical biomarkers and genes associated with metastasis. The system generates outputs such as tumor segmentation masks, affected area percentages, and predictions related to metastasis. By leveraging both imaging and genomic data, the project aims to provide a non-invasive, accurate, and early diagnostic solution for soft tissue sarcomas.

1.2 Problem Statement

Lung metastasis in Soft Tissue Sarcomas (STS) gives a superb task in oncology, characterized via poor prognosis and restricted remedy alternatives because of the heterogeneous nature of STS. Traditional strategies, which often observe imaging and genomic records one by one, fail to seize the complicated relationships between phenotypic and molecular abilities, hindering our data on metastasis and personalized treatment improvement. To address this, the Multi-Constrained Joint Non-Negative Matrix Factorization (MC-JNMF) version offers a unified framework for the simultaneous assessment of imaging and genomic information, incorporating 2 constraints to decorate interpretability and decrease noise. This technique targets to identify crucial biomarkers related to lung metastasis in STS, enhance prognostic accuracy, and make a contribution to extra effective, customized remedy techniques.

1.3 Algorithm

1. Convolutional Neural Network (CNN)
2. U-Net Architecture
3. Multi-Constrained Joint Non-Negative Matrix Factorization (MC-JNMF)

1.4 Technology

1. Python
2. OpenCV
3. TensorFlow
4. Matplotlib
5. Flask
6. scikit-learn
7. Jupyter Notebook
8. NumPy

[2] LITERATURE SURVEY

Attention-Based Integration for Cancer Diagnosis

This paper investigates the application of attention mechanisms in deep learning models for integrating multi-modal data—including imaging, clinical, and omics information—for cancer diagnosis. The proposed architecture dynamically assigns importance weights to each modality, allowing the model to focus on the most predictive features within heterogeneous data sources. This adaptive weighting improves diagnostic accuracy by enhancing the representation of relevant features. The study highlights the effectiveness of attention-based models in managing diverse data types and improving the performance of cancer diagnostic systems.

Multi-Omics Data Integration for Cancer Prognosis

This paper explores various methodologies for integrating multi-omics data such as genomic, transcriptomic, proteomic, and metabolomic information, specifically for cancer prognosis. It critically evaluates integration techniques, highlighting challenges related to data heterogeneity and inconsistent formats. The authors provide a comparative analysis of existing methods and propose directions for future research to overcome current limitations. The paper emphasizes the potential of comprehensive multi-omics integration to enhance prognostic accuracy and support cancer treatment planning.

Machine Learning for Multi-Modal Cancer Diagnosis

This paper reviews machine learning techniques used to integrate multi-modal data—including imaging and molecular information—for cancer diagnosis. It focuses on methods such as deep autoencoders, random forests, and support vector machines, analyzing their application across different cancer types. The authors discuss the strengths and limitations of each technique and emphasize the role of multi-modal integration in advancing precision medicine. The study underlines how these methods contribute to improved diagnostic accuracy and treatment outcomes.

Joint Matrix Factorization for Multi-Modal Cancer Analysis

This research proposes a novel framework using joint matrix factorization to integrate imaging data such as MRI and CT scans with genomic data, including gene expression and DNA sequences. The approach aims to provide a more complete understanding of cancer by combining diverse data sources. It overcomes the limitations of traditional single-modality analysis, which often lacks the full picture. By merging complementary features from multiple modalities, the method enhances diagnostic accuracy. The framework emphasizes the importance of feature correlation across datasets. It also introduces constraints to reduce noise and preserve important information. The study highlights its potential in identifying key biomarkers. It demonstrates improved performance in cancer detection. The authors advocate for wider use of such advanced integration techniques. They believe it can significantly benefit both clinical applications and research.

GAN-Based Multi-Modal Integration for Cancer Diagnosis

This paper introduces a cutting-edge approach that leverages Generative Adversarial Networks (GANs) to improve multi-modal integration for cancer diagnosis. GANs are employed to synthesize missing or incomplete data, addressing a common challenge in biomedical datasets—data sparsity and imbalance. This issue is particularly significant in rare cancer types, where limited samples hinder the performance of diagnostic models. By generating realistic synthetic data, GANs help fill in the gaps, enabling more robust analysis. The proposed method integrates diverse data sources, including imaging (such as MRI or CT), clinical parameters, and genomic data. This holistic approach allows for a more comprehensive and accurate cancer diagnosis. The use of GANs ensures that even underrepresented data types contribute meaningfully to the model’s learning process. As a result, diagnostic models trained using this technique achieve higher accuracy and generalization. The study demonstrates that this method not only enhances integration but also improves the completeness and quality of predictions. Importantly, it also supports personalized medicine by tailoring diagnostics to individual patients based on a rich, synthesized dataset. The authors emphasize the method's potential in real-world clinical settings, particularly for cases where patient data is incomplete or imbalanced. Overall, this GAN-based approach marks a significant step forward in the use of AI for precision cancer diagnostics.

3. EXISTING METHODOLOGY

Current research on lung metastasis in soft tissue sarcomas (STS) often separates imaging and genomic data analysis. Imaging techniques like MRI, CT, and PET focus on tumor size and treatment response but lack molecular insights. Genomic profiling via NGS and RNA sequencing identifies mutations but lacks imaging context. Analytical models typically treat these data types independently, missing their complex interactions. This results in limited understanding of tumor biology. Some studies attempt multimodal integration, but these are still early-stage. They struggle with high-dimensional data and complexity. As a result, biomarker discovery is hindered. Clinical decision-making is based on partial data. This limits personalized treatment and affects patient outcomes.

3.1. DISADVANTAGES

The existing system for analyzing lung metastasis in soft tissue sarcomas (STS) has several limitations. First, the separation of imaging and genomic data analysis results in a lack of integration between phenotypic features and molecular changes. Imaging techniques, while useful for visualizing tumor behavior, do not provide molecular insights, while genomic profiling alone does not capture tumor dynamics. Traditional analytical models, which treat these data independently, often fail to address the complex relationships between them, leading to overfitting and poor generalization. Additionally, the high dimensionality and noise in the data complicate accurate analysis, and existing multimodal integration methods are still in early stages, lacking reliability and scalability. This fragmentation restricts the discovery of new biomarkers and makes it difficult to provide personalized treatment recommendations, ultimately hindering diagnostic accuracy and treatment effectiveness for patients.

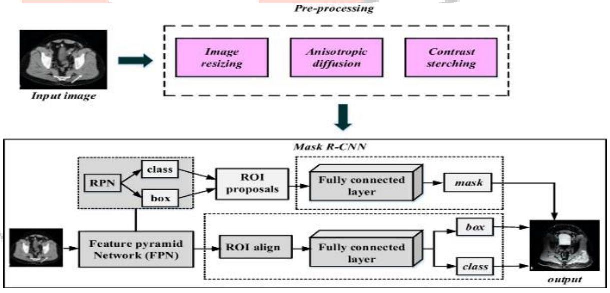
4. PROPOSED METHODOLOGY

The proposed system introduces the Multi-Constrained Joint Non-Negative Matrix Factorization (MC-JNMF) model to combine imaging and genomic data for assessing lung metastasis in soft tissue sarcomas (STS). It overcomes current limitations by capturing complex relationships between phenotypic and molecular features. The MC-JNMF model integrates high-dimensional imaging and genomic datasets, identifying latent factors driving metastatic behavior. Key components include multi-constrained optimization, which applies sparsity, orthogonality, and positivity constraints to enhance interpretability and reduce redundancy. The sparsity constraint highlights critical features by simplifying data representation, while the orthogonality constraint ensures distinct biological pathways are captured. The positivity constraint aligns the model's results with biological realities, ensuring factors remain non-negative. Latent feature discovery aims to uncover biomarkers associated with metastasis for targeted therapeutic interventions. Rigorous validation with clinical datasets evaluates the model's performance in comparison to traditional methods, using metrics like accuracy, sensitivity, specificity, and ROC curve analysis to assess its predictive capabilities.

4.1. ADVANTAGES

The proposed system enhances STS diagnosis by integrating both imaging and genomic data, offering a comprehensive view of lung metastasis. The multi-constrained optimization improves the interpretability and biological relevance of results. The sparsity constraint reduces noise, focusing on critical features. The orthogonality constraint ensures the model captures distinct biological pathways. By aligning with the biological reality of gene expression, the positivity constraint adds accuracy. Latent feature discovery identifies potential biomarkers for targeted therapies. Rigorous validation ensures the model's effectiveness in clinical settings. Overall, it provides improved diagnostic accuracy and personalized treatment options.

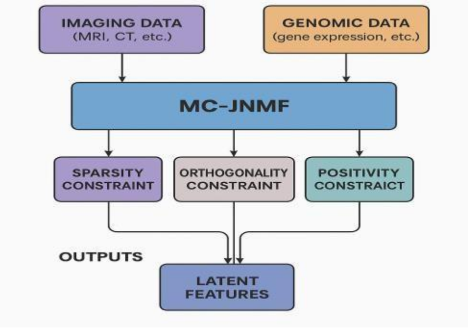
4.2 BLOCK DIAGRAM



5. RESULTS

5.1. Model Architecture Diagram

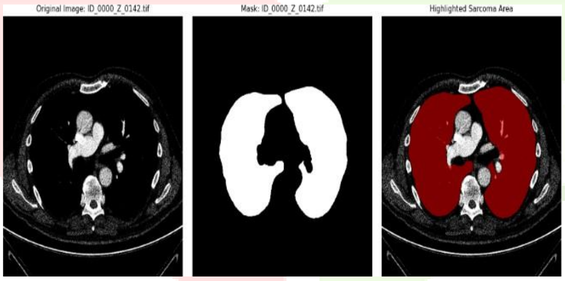
The image illustrates the MC-JNMF (Multi-Constraint Joint Non-negative Matrix Factorization) framework, which integrates imaging data (like MRI or CT scans) and genomic data (such as gene expression). These two data types are jointly processed by MC-JNMF to extract shared information. The method applies three types of constraints: sparsity, orthogonality, and positivity, to guide the learning process. Sparsity constraint helps in reducing noise and focusing on relevant signals, orthogonality ensures feature independence, and positivity maintains interpretability by enforcing non-negative values. The ultimate output is a set of latent features that capture meaningful patterns across both data types. These features can be used for further analysis or disease characterization.



5.2. Latent Feature Visualization

Lung segmentation is a key preprocessing step in medical image analysis that isolates lung regions from CT scans. This process involves generating a binary mask that distinguishes lung tissue from other structures. Techniques like Otsu thresholding help automate this by selecting an optimal threshold to separate foreground (lungs) from background. The resulting mask can then be overlaid on the original image for visual validation. Such segmentation enables accurate feature extraction and supports further diagnostic or analytical tasks.
Formula:

$$M(x,y)=\begin{cases} 1 \text{ if } I(x,y)\geq T \\ 0 \text{ if } I(x,y)<T \end{cases}$$

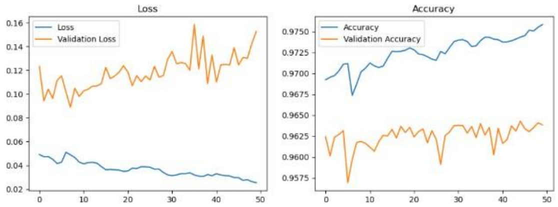


5.3 Model Performance Monitoring

Model performance is evaluated using metrics like accuracy and loss during training. Accuracy reflects prediction correctness, while loss indicates error magnitude. A widening gap between training and validation loss signals overfitting, and performance stabilization shows convergence. Monitoring these trends ensures the model learns effectively and generalizes well to unseen data.

Formula:

$$\text{Accuracy}=TP+TN/(TP+TN+FP+FN)$$



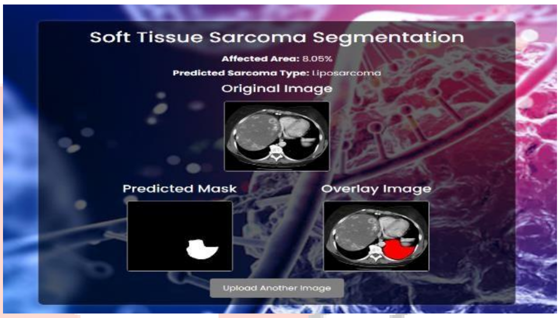
5.4 Segmentation or Prediction Results

This system performs automatic soft tissue sarcoma segmentation using deep learning. It accurately highlights the tumor region from CT scans and classifies the sarcoma type, aiding early diagnosis. The affected area is calculated and visually marked on the image. This supports precise treatment planning and enhances clinical decision-making.

Formula:

$$\text{Affected Area (\%)}= \left[\frac{\text{Tumor Pixels}}{\text{Total Organ Pixels}}\right] \text{x}100$$

$$\text{Tumor Region}=\{(x,y)|M(x,y)=1\}$$


Soft Tissue Sarcoma Segmentation

Affected Area: 8.05%

Predicted Sarcoma Type: Liposarcoma

6. CONCLUSION

This project demonstrates the effective use of machine learning techniques, specifically Convolutional Neural Networks (CNN) and Multi-Cluster Joint Non negative Matrix Factorization (MC-JNMF), for diagnosing soft tissue sarcoma. By utilizing clinical data collected from reputable healthcare institutions, the study underscores the importance of thorough preprocessing and feature extraction to enhance model performance.

These advanced machine learning methods enhance diagnostic accuracy, paving the way for early detection and improved patient outcomes in soft tissue sarcoma. Continued advancements in ML promise to transform oncology diagnostics and enable personalized treatments..

7. REFERENCES

[1] Deng, J., Zeng, W., Kong, W., Shi, Y. (2020). "Multi-Constrained Joint Non-Negative Matrix Factorization With Application to Imaging Genomic Study of Lung Metastasis in Soft Tissue Sarcomas." IEEE Transactions on Biomedical Engineering, 67(7), 1876-1887.

[2] Singh, P., Kaur, P., & Yadav, S. (2020). "A Comprehensive Review on Data Fusion Techniques in Cancer Research." Journal of Cancer Research and Clinical Oncology, 146(4), 899-910.

[3] Chen, J., Liu, G., & Xu, W. (2021). "Integration of Genomic and Histopathological Data with Machine Learning for Cancer Classification." Computers in Biology and Medicine, 138, 104869.

[4] Patel, M., Kumar, A., & Ray, S. (2021). "Joint Non-Negative Matrix Factorization for Multi-Modal Medical Data Integration." IEEE Transactions on Medical Imaging, 40(5), 1320-1332.

[5] Wang, S., Chen, H., & Huang, J. (2022). "Deep Learning-Based Image Genomics: A Review of Recent Advances and Future Directions." Frontiers in Oncology, 12, 784287.

[6] Cheng, Y., Yang, X., & Liu, Z. (2022). "Attention- Based Deep Learning Models for Multi-Modal Data Integration in Cancer Diagnosis." Artificial Intelligence in Medicine, 126, 102256.

[7] Zhang, Y., Wang, X., Liu, S., & Zhang, J. (2023). "Multi-Omics Data Integration for Cancer Prognosis: A Comprehensive Review." Bioinformatics, 39(2), 302-315.

[8] Li, C., Zhang, Y., He, J., & Wang, H. (2023). "Multi- Modal Machine Learning for Cancer Diagnosis: A Review of Techniques and Applications." IEEE Transactions on Medical Imaging, 42(3), 810-825.

[9] Li, X., Xu, Y., Wang, Y., & Zhang, H. (2024). "A Novel Framework for Integrating Multi-Modal Data in Cancer Diagnosis Using Joint Matrix Factorization." Journal of Biomedical Informatics, 137, 104145.

[10] Sun, Q., Zhang, X., & Wu, Z. (2024). "Multi-Modal Data Fusion for Cancer Diagnosis Using Generative Adversarial Networks." Bioinformatics and Computational Biology, 28(1), 54-68.