Identification of DLBCL using multimodal learning

Alejandro Méndez Calvario[1],
Supervisors: Dr. Julio Cesar Pérez Sansalvador[1], Dr.Noureddine Lakouari[1]
[1]Instituto Nacional de Astrofísica Óptica y Electrónica, INAOE, Puebla, México

Abstract

Diffuse Large B-Cell Lymphoma (DLBCL) is one of the most prevalent lymphoma subtypes in both young and older adult populations. Although it has high remission rates, an erroneous or delayed diagnosis can lead to a fatal outcome.

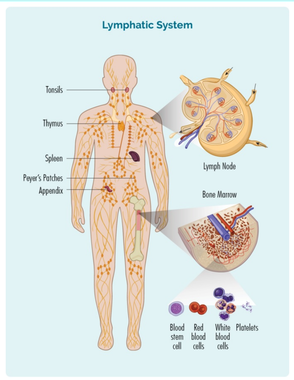
This work proposes an automatic DLBCL identification method using multimodal learning. Unlike conventional machine learning approaches, which typically rely on a single data modality, our proposal integrates information from different sources to emulate the comprehensive clinical diagnostic process.

Initial results, obtained with a convolutional neural network (CNN) architecture, show an accuracy of 0.95 in DLBCL subtype classification. Furthermore, the multimodal combination of features has enabled predictions of patient survival time, achieving an R^2 value of 0.86.

Subsequently, using a genomic database from an independent cohort, the same classification task was performed, yielding an R^2 of 0.71.

As a next step, all available data modalities will be integrated to improve both the accuracy of DLBCL subtype classification and the prediction of survival

Diffuse Large B-Cell Lymphoma (DLBCL)



Lymphoma

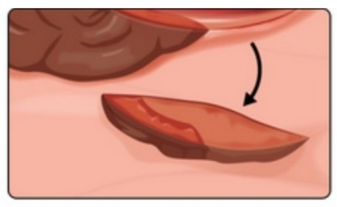
Hodgkin

No-Hodgkin

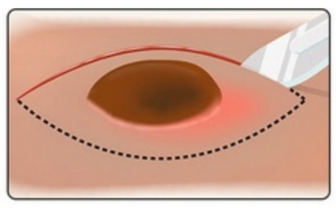
Diffuse Large B-Cell Lymphoma (DLBCL)

Diagnosis

Types of biopsy



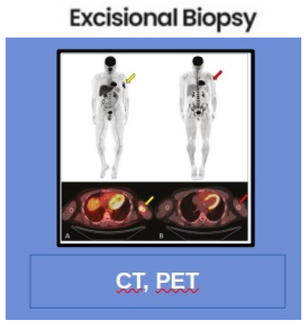
Incisional Biopsy



Excisional Biopsy

Specialists Involved

- Medical hematologist-oncologist
- Pathologist
- Radiologist or nuclear medicine physician
- Radiation oncologist
- Surgeon



Main objective

Develop a method for the identification of DLBCL using multimodal machine learning that achieves results comparable to or better than those reported by monomodal methods.

Specific Objectives

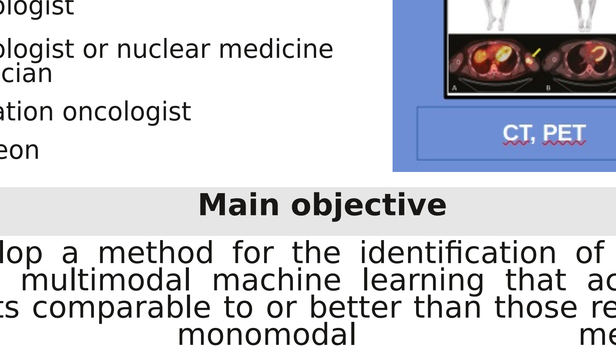
Design an early data fusion strategy for the automatic identification of DLBCL.

Design a late fusion strategy for the automatic identification of DLBCL.

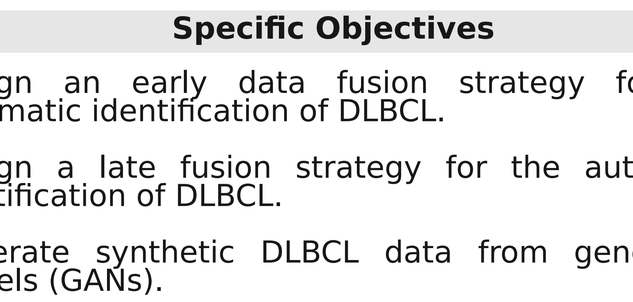
Generate synthetic DLBCL data from generative models (GANs).

Enhance the early and late data fusion strategies designed for the identification of DLBCL.

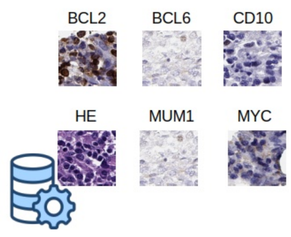
General methodology



Multimodal learning



Previous results

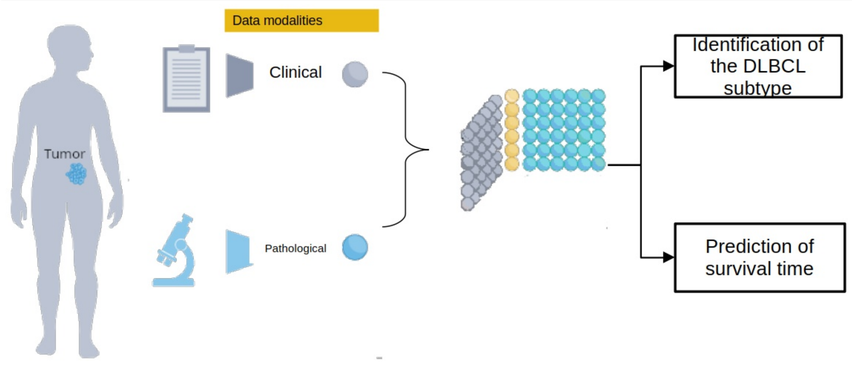


DLBCL-Morph

- 42 tissue microarrays (TMAs), digitized, each with 6 stains (HE, CD10, BCL6, MUM1, BCL2, MYC)
- 150,000 patches of 224x224 pixels

- 209 patients with DLBCL cases
- Age, genetic translocations (MYC, BCL2, BCL6), overall survival, follow-up status, clinical and molecular variables

Identification of DLBCL subtype and survival time



Results

Prediction of overall survival with clinical data

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken (s)
OrthogonalMatchingPursuit	0.45	0.73	1.56	0.11

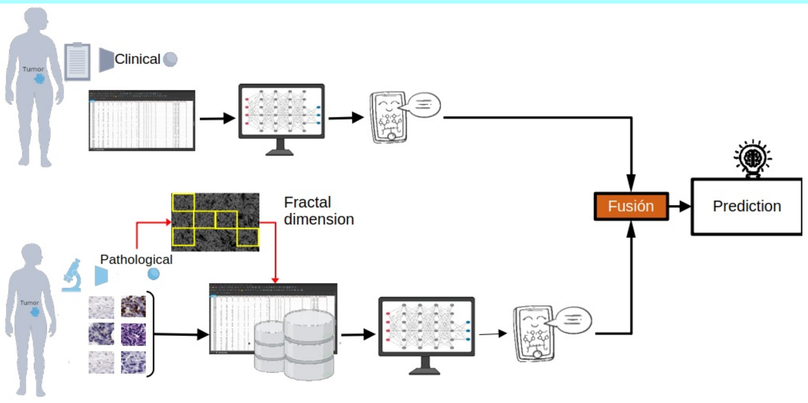
Prediction of overall survival with geometric data

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken (s)
OrthogonalMatchingPursuitCV	3.75	0.08	2.96	0.01

Fusion

Metric	Value
R^2	0.44

Advances of the period



Results

Prediction of overall survival with clinical data

Modelo	Adjusted R^2	R^2	RMSE	Time Taken (s)
ExtraTreesRegressor	0.40	0.66	0.66	0.10
RandomForestRegressor	0.35	0.46	1.98	0.12
HistGradientBoostingRegressor	0.35	0.63	1.98	0.12
XGBRegressor	0.34	0.63	1.98	0.12
LGBMRegressor	0.34	0.63	2.00	0.3

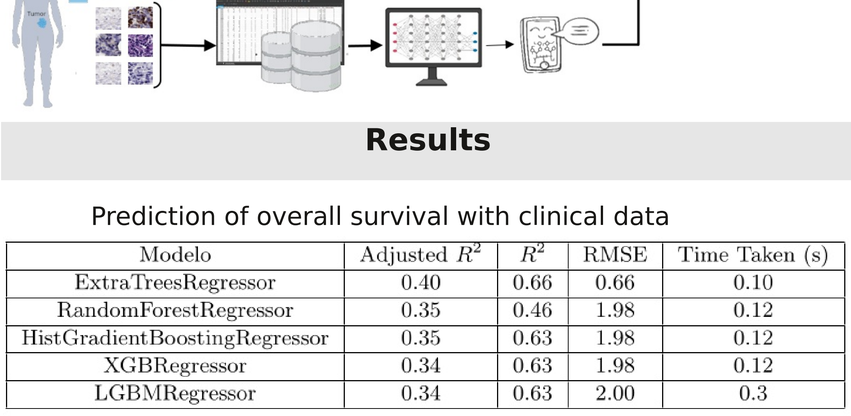
Prediction of overall survival with geometric data

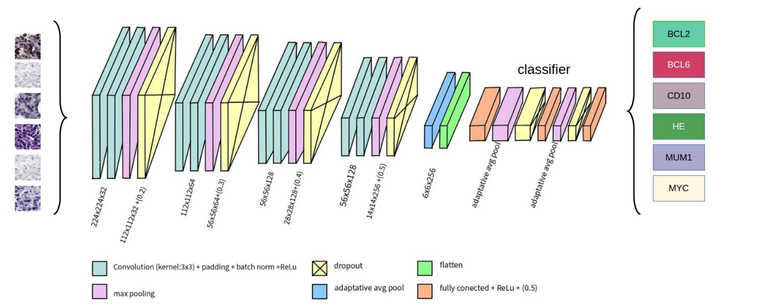
Modelo	Adjusted R^2	R^2	RMSE	Time Taken (s)
GradientBoostingRegressor	0.96	0.96	0.71	23.02
AdaBoostRegressor	0.63	0.63	2.04	14.90
LassoLarsCV	0.30	0.30	2.84	14.90
TransformedTargetRegressor	0.30	0.30	2.80	0.17
LinearRegression	0.30	0.30	2.80	0.18

Fusion

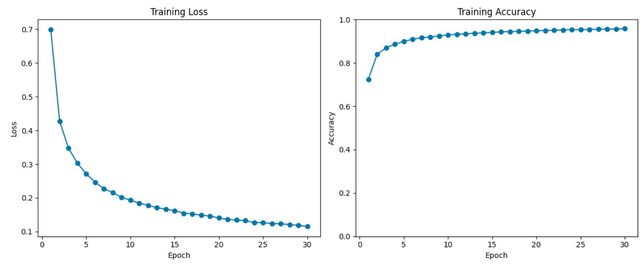
Metric	Value
RMSE	1.2
R^2	0.86

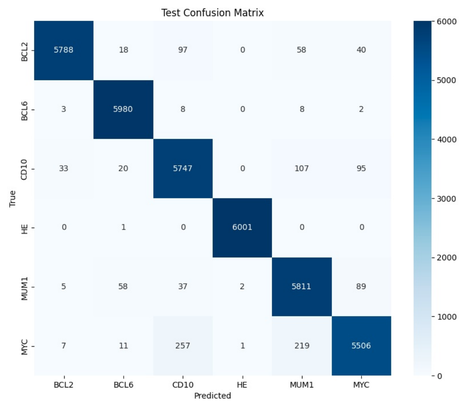
Identification of the DLBCL subtype on digital images





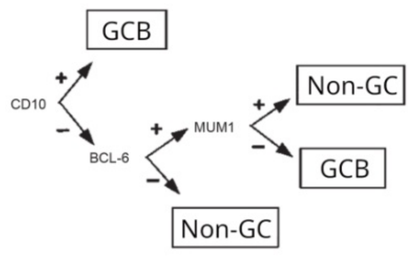
Performance metrics





Metrics	
Loss	0.11
Accuracy	0.95

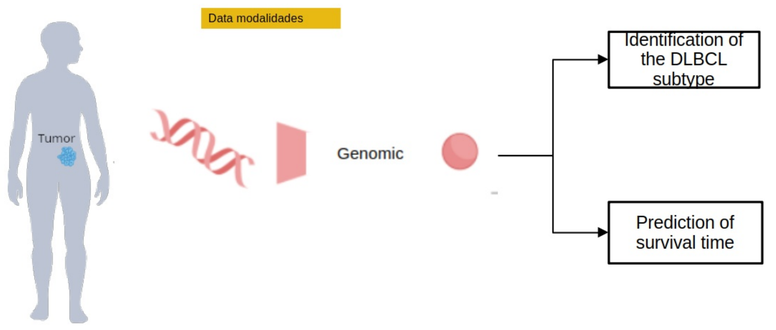
Hans' algorithm



Genomic data

LymphGen

-Clinical data(PAtient age, sex, tumor location, survival,therapeutic response)
- Mutations
- Copy number alterations
- Gene fusions
- NCI Cohort: 574 cases
- Harvard Cohort: 304 cases
- BCC Cohort: 332 cases



Identification of the DLBCL subtype using

Model	Accuracy	Balanced Accuracy	F1 Score	Time Taken
AdaBoostClassifier	0.71	0.57	0.57	0.07
LGBMClassifier	0.76	0.58	0.65	0.09
XGBClassifier	0.64	0.54	0.62	0.07
BaggingClassifier	0.63	0.55	0.62	0.03
RandomForestClassifier	0.62	0.52	0.6	0.14

Schedule of activities

Actividad	2023 Aug.-Dec.	2024 Jan.-May.	2024 May.-Jul.	2024 Aug.-Dic.	2025 Jan.-May.	2025 May.-Jul.	2025 Aug.-Dec.	2026 Jan.-May.	2026 May.-Jul.	2026 Aug.-Dec.	2027 Jan.-May.	2027 May.-Jul.	2027 Aug.-Dec.
Review of the state of the art													
Review and selection of machine learning models													
Review of modality fusion techniques													
Initial processing of data modalities													
Modeling and training of learning models based on each data modality													
First test of late fusion of data modalities													
Review and application of GANs													
Late fusion of data modalities													
Early fusion of data modalities													
Evaluation of data fusion strategies													
Writing the 1st article													
Writing the 2nd article													
Identification of problems and opportunities for model improvement													
Design of modality fusion strategies to improve and correct detected problems													
Evaluación y validación													
Writing the thesis													

Objetive 1 | Objetive 2 | Objetive 3 | Objetive 4 | Activity completed | Activity to be carried out

References

Instituto Nacional del Cáncer. (s.f.). Sistema linfático. Instituto Nacional del Cáncer. https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/sistema-linfatico
American Cancer Society. (s.f.). Señales y síntomas del linfoma no Hodgkin. American Cancer Society. https://www.cancer.org/es/cancer/tipos/linfoma-no-hodgkin/deteccion-diagnostico-clasificacion-por-etapas/senales-sintomas.html
StatPearls Authors. (2023). Lymphoma. En StatPearls. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK560826/
Häggström, H., Leithner, D., Alvén, J., Campanella, G., Abusamra, M., Zhang, H., ... Mayerhoefer, M. E. (2024). Deep learning for [18F] fluorodeoxyglucose-PET-CT classification in patients with lymphoma: a dual-centre retrospective analysis. The Lancet Digital Health, 6(2), e114–e125.
Naji, H., Hahn, P., Pisula, J. I., Ugliano, S., Simon, A., Büttner, R., & Bozek, K. (2025). Deep learning-based interpretable prediction of recurrence of diffuse large B-cell lymphoma. BJC Reports, 3(1), 34.
Cairns, J., Frood, R., Patel, C., & Scarsbrook, A. (2025). The role of AI in lymphoma: An update. Seminars in Nuclear Medicine, 55(3), 377-386.
Lipkova, J., Chen, R. J., Chen, B., Lu, M. Y., Barbieri, M., Shao, D., ... & Mahmood, F. (2022). Artificial intelligence for multimodal data integration in oncology.