AI-based Classification of Histopathology with Multi-Omic Data, for Improved Risk Stratification in Medulloblastoma.

Sai Anugna Sureddi[1,2.], Dr. Ruth Cranston[2.], Prof. Steven Clifford[2.], Dr. Ed Schwalbe[1,2.]

1.School of Natural Sciences , Northumbria University; 2.Department of Translational and Clinical Research, Newcastle University.

INTRODUCTION

Medulloblastoma (MB) is a malignant (Grade IV) tumour occurring **in the cerebellum** and is **a leading cause of cancer-related death in children.**

2nd most commonly diagnosed Central Nervous System (CNS) tumour in children less than 15 years of age (~13.5%)

MB is **very heterogenous:** WHO-recognises 4 principal molecular groups **(WNT,SHH,Group3&4)** & 3 major histopathological types-**Classic (CLA), Large Cell/Anaplastic (LCA), Desmoplastic/Nodular(DN)**

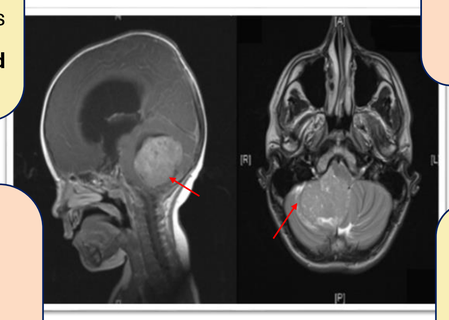
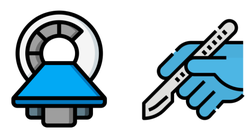


Figure1:Magnetic Resonance Imaging **(MRI) Scan** of head with Sagittal (Left) and Horizontal (Right) views. The **red arrows** show the **Medulloblastoma (MB) tumour in the posterior fossa,** (DeSouza et al., 2014).

In 2021, WHO recognised **subgroups of MB molecular groups** with distinct molecular features and clinical behaviours



After the tumour has been detected, it is surgically resected.

H&E (Haematoxylin and Eosin) Staining and Immunohistochemistry.

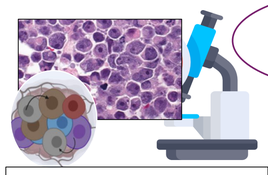
Patient directed risk stratification

Standard Risk(~80% survival) lower doses of Craniospinal Irradiation(CSI)

High Risk (60% survival) high-dose CSI

Infants (aged under 3) - chemotherapy without CSI, sparing developing brain

THE RESEARCH PROBLEM & APPROACH



I think it's a **Classic** MB tumour!

I think it has a few **anaplastic** cells

Intra (molecular and sub-clonal level) tumour heterogeneity

Inter-observer Variability

Potential Misdiagnosis

Undertreatment of High-risk patients – Increased Risk of Relapse

Overtreatment of Standard risk patients – Treatment-based late effects

I CAN HELP with my ability to rapidly analyse high-dimensional data and provide PREDICTIONS

You can TRAIN me with OMIC DATA for me to LEARN THE PATTERNS and DERIVE THE HISTOLOGY that may been seen in the patient!

Thanks! That would help us **PERSONALISE THE TREATMENT** based on **patients' individual and UNIQUE RISK FEATURES!**

AIMS

AIM: To develop an omic-based **AI pipeline** to **aid diagnosis and prognosis** of heterogeneous **Medulloblastoma tumours**.

OBJECTIVES:

- Establish a **training cohort** with archival and clinical trial patient samples with histopathology annotation labelled through **Central Pathology Review (CPR)**, a **consensus histology call** arrived upon by **experienced neuropathologists**.
- Train **Machine Learning (ML)** models on **methylation array data for classification of histopathological type** in **MB**.
- **Validate models on unseen test cohorts, assess prediction probabilities and recall**.

RESEARCH METHODOLOGY AND RESULTS

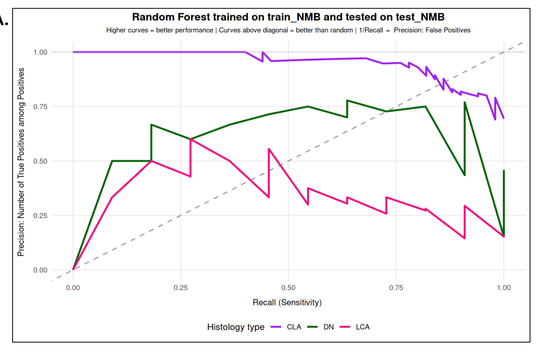
Let's start with **methylation array data** that has been established as a **gold standard for classification of MB molecular groups. 'NMB'** is an **in-house cohort** with archival and clinical trial samples while **'Cavalli'** is an **external MB cohort** published by Cavalli *et al.*, 2017.

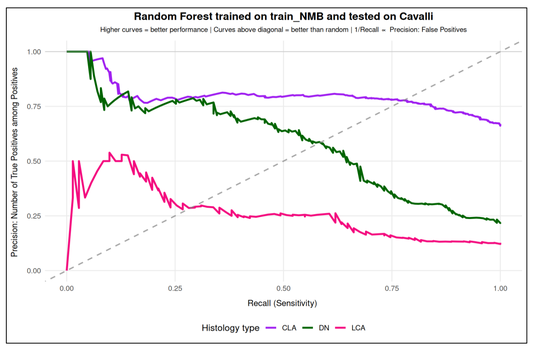
Cohort	Total number of patient samples	After filtering out samples with poor quality and missing data	Type of Histology Labels
'NMB'	510	**430**	Central Pathology Review
'Cavalli'	763	**585**	Institutional Pathology Calls

Copy! I'll use **'NMB' as the training cohort** since labels were assigned by **CPR**, making my learnt histology type omic patterns **reliable**! Along with **testing** the model on **'Cavalli'**, I'll do the **80:20 split of 'NMB'**, **training** the model on **80%** of 'NMB' **(train_NMB)** and using the remaining **20%** as a CPR-labelled independent **test set (test_NMB)**.

Ok, let's use a **Random Forest** (RF) model with 5-fold cross validation for you to be trained as it has **previously** been successfully **implemented** for **classification of MB molecular groups**!

In **test_NMB**, The RF seemed to classify most of the **CLAs correctly followed by DN with few False Positives (FPs)** and with the **LCA class** having **more FPs** [Fig.2A] **.In 'Cavalli',** there are **FPs in all the histology classes** [Fig.2B]





Sensitivity: True positives (TP)/[TP+ FN (False Negatives)]

Precision: TP/[TP+ FP(False Positives)]

Figure 2A and 2B: **Random Forest (RF)** model **trained on train_NMB with** an **imbalance, is biased to** the **class** with **majority** of the instances, **Classic/CLA (purple curve)** and displays **limited specificity in the minority classes Desmoplastic/DN(green curve) and Large Cell Anaplastic/LCA(pink curve).**

I think due to **class imbalance**, the model is **biased to the majority class, CLA**. Let's try **balancing** the dataset with **SMOTE** (Synthetic Minority Oversampling TEchnique) which would generate **synthetic samples** by **interpolating between 2 nearest minority instances** within **k-means clusters**. This would give us sufficient samples to work with!

RESEARCH METHODOLOGY AND RESULTS

Sure, the RF model **trained on an augmented cohort that included synthetic data** performed similarly **to the one trained on an unbalanced set, still not able to classify** the histology classes **confidently** [Fig. 3A&3B].

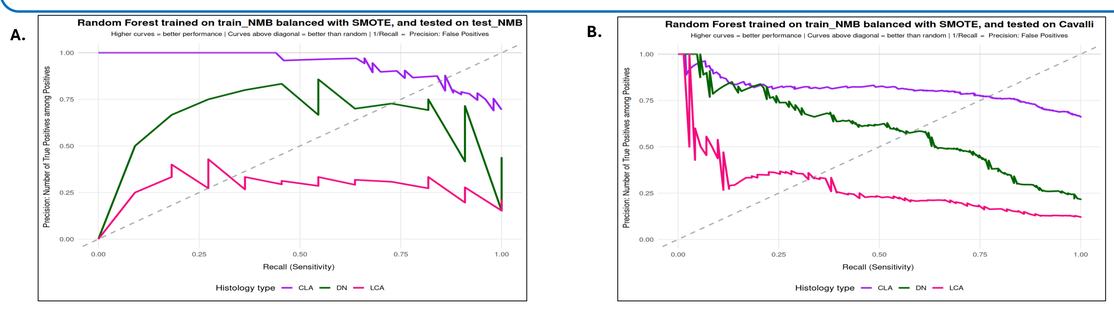


Figure 3A and 3B: **Balancing Train_NMB** with synthetic samples generated by **SMOTE(Synthetic Minority Oversampling TEchnique) before training** the Random Forest (**RF**) model **did not make the model better,** it was **still biased to the majority class**, **Classic/CLA** ,with **limited specificity** in the minority classes, **Desmoplastic/DN** and **Large Cell Anaplastic/LCA,** with **skewed probability scores of assignment** to these classes by the RF model.

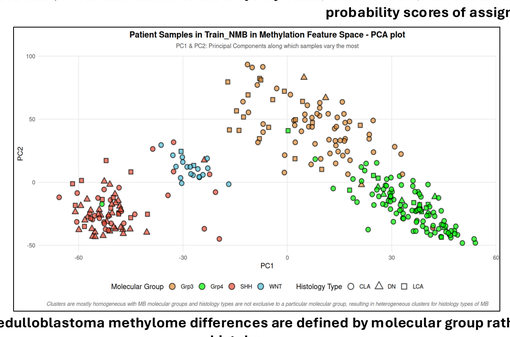


Figure 4: Medulloblastoma methylome differences are defined by molecular group rather than histology.

Maybe the **synthetic samples** are **not** really **balancing**! **SMOTE** is being **confounded by molecular type.** Let's say when it's trying to **interpolate between 2 DN samples,** it **might** end up **creating a CLA** sample **instead,** making the **data noisy** [Fig.4] and the **model** further **overfitted to CLA!**

Even if I were to **stratify** the dataset by **molecular types before** applying **SMOTE**, I would have **fewer samples** in some of the **groups** which might lead to **unequal representation across molecular groups**! How about we try **under-sampling CLA** samples in **train_NMB** instead?

Yes. This time, let's also **calibrate the output probabilities** with **Multinomial Regression** since **raw** output **probability scores of** the **RF** model **might not reflect the true likelihood** of the **predicted class** labels !

Ok, along with **improving recall** in minority classes **DN** and **LCA**, calibration also seemed to **stabilise probabilities** of assignment in **Cavalli** (**Relatively smoother P/R curves**; Fig3B vs Fig 5B). However, we still have **False Positives in LCA** (in **test_NMB and Cavalli**) and in **DN and CLA** (in **Cavalli**) [Fig.5].

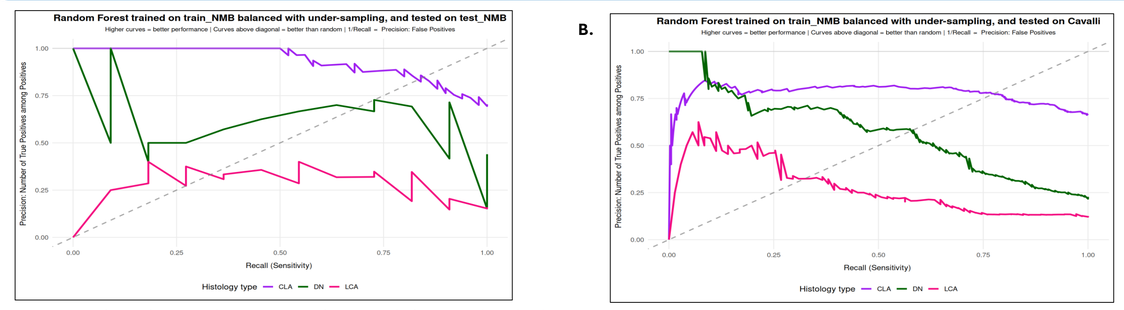


Figure 5A and 5B:**Balancing Train_NMB** with **under-sampling the majority class, Classic/ CLA**, made the model **detect true positives** in the **minority classes ,Desmoplastic/DN** and **Large Cell Anaplastic/LCA effectively** while it still **misclassified CLA samples as DN/LCA** leading these to these classes to have a **False Positive Rate.**

Aside from LCA FPs, the model seems to be **good in classifying DN samples correctly,** because **most** of them **belong to the molecular group SHH,** making the **model confident** about **SHH** samples **being DN** [Fig 6]. However, **in Cavalli**, the model has **True Negatives and False positives across all classes**, do you think the **model is "overfitted" to NMB ?**

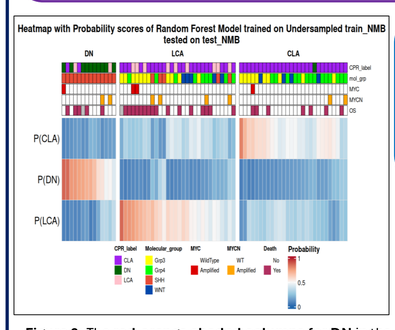


Figure 6: The **red-orange shaded columns for DN** in the heatmap indicate that the **model is relatively confident (high P(DN))** about labelling **SHH-MB (**Sonic Hedgehog Medulloblastoma)samples as **Desmoplastic/DN histology.**

From the **heatmap,** the **FPs in LCA** have **P(CLA) and P(LCA) close to each other,** do you think instead of prediction by highest probability score, we can **determine class-wise thresholds with Youden's J Statistic** and **apply** them in a **step-wise manner** to be able to **predict them correctly?**

Maybe not, because the **survival outcomes** of **histology** groups in **Cavalli, predicted by the model trained on NMB, match** with **consensus** (as **in NMB**) (DN-good prognosis; LCA-poor prognosis) [Fig7&8], suggesting that the **predictions might be biologically meaningful.**

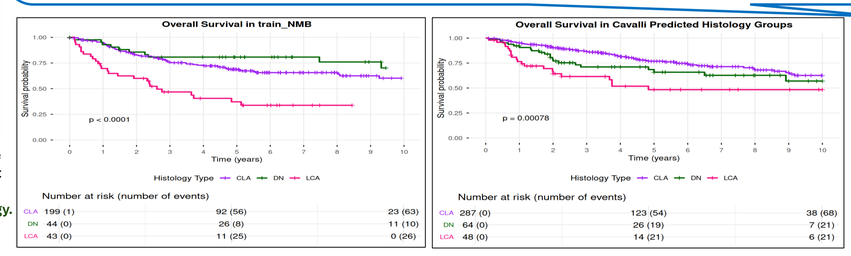


Figure 7: The **Overall Survival** in **train_NMB histology groups** represent the **consensus prognostic outcomes** seen in patients with the respective histopathology **(CLA and DN-good prognosis; LCA-poor prognosis)**

Figure 8: The **Overall Survival** in **histology groups of 'Cavalli'** as **predicted by the Random Forest model, match** the **expected consensus** prognosis indicating that **RF 'learnt' meaningful class specific representations.**

With class thresholds applied in a sequential order, CLA samples previously **misclassified as LCA** could **pass** Youden's **CLA threshold (for correct prediction),** but some **LCA** samples were **misclassified as CLA** now [Fig 9], **costing** us the **high true positive rate we had in LCA !**

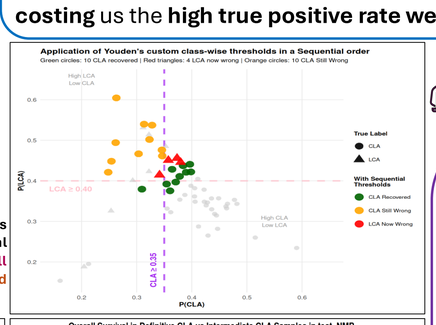


Figure 9: The **dashed lines are the class-wise thresholds determined by Youden's J statistic.** By **sequential application** of **Classic/CLA** and **Large Cell Anaplastic/LCA thresholds,** few of the **CLA misclassified as LCA were now correctly classified as CLA.**

Maybe **FP LCA samples have both CLA and LCA features** making them "**intermediate**" samples that the **model is "uncertain" about** (|P(CLA)-P(LCA)| <= 0.1 & low P(DN)) [Fig.10]. We **could say** they are **transitional** and **may not be** entirely **different** as there's **no difference in prognosis of "confident" and "intermediate" samples** [Fig.11].

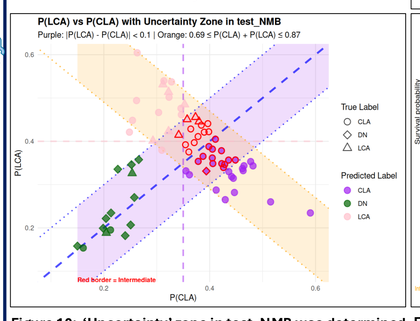


Figure 10: 'Uncertainty' zone in test_NMB was determined with 'intermediate' (CLA/LCA) samples.

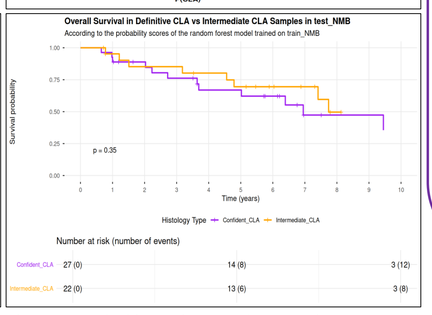


Figure 11:No significant difference in the **survival outcomes** in the **Classic/CLA Intermediate samples** and the **CLA** samples that the **model** is **confident** about.

DISCUSSION

- "**Intermediate**" CLA/LCA **might be indicative of Classic and Large Cell/Anaplastic histology types existing on a continuum**. Clinically, this is a recognised phenomenon.
- Annotation of **these samples with % of anaplasia and other clinico-molecular features** would help further **characterise, define and classify** them with **confidence**.

REFERENCES

1.DeSouza, R.-M., Jones, B.R.T., Lowis, S.P. and Kurian, K.M. (2014). Pediatric Medulloblastoma - Update on Molecular Classification Driving Targeted Therapies. *Frontiers in Oncology*, 4. doi:https://doi.org/10.3389/fonc.2014.00176.

2.Cavalli, F.M.G., Remke, M., Rampasek, L., Peacock, J., Shih, D.J.H., Luu, B., Garzia, L., Torchia, J., Nor, C., Morrissy, A.S., Agnihotri, S., Thompson, Y.Y., Kuzan-Fischer, C.M., Farooq, H., Isaev, K., Daniels, C., Cho, B.-K., Kim, S.-K. Wang, K.-C. and Lee, J.Y. (2017). Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell*, 31(6), pp.737-754.e6. doi:https://doi.org/10.1016/j.ccell.2017.05.005.