Cancer Classification using CNV

Shawn Ng, Edison Wong, Renaud Chee and Abdullah Armain

Introduction

Cancer remains one of the leading causes of death worldwide, with distinct genomic characteristics across different cancer types. Accurate classification of cancers based on genomic data is vital for early detection and personalized treatment. This study focuses on classifying cancer types using copy number variation (CNV) profiles of human DNA, with the goal of categorizing each sample into one of the ten most prevalent cancer types, including breast, ovarian, and lung cancer.

Previous research has explored various machine learning techniques for cancer classification. Chaurasiya et al. demonstrated that Support Vector Machines (SVM) achieved high classification rates for breast cancer in their study, "Comparative Analysis of Machine Learning Algorithms in Breast Cancer Classification." Zelli et al., in "Classification of Tumor Types Using XGBoost Machine Learning Model: A Vector Space Transformation of Genomic Alterations," found that XGBoost outperformed other methods in distinguishing cancer types based on genomic bin features. Deep learning approaches, such as those explored in "An Investigation of Deep Neural Networks for Cancer Classification," have also been applied to genomic data, showing competitive performance but requiring large-scale training datasets.

However, Chaurasiya et al.'s focus on breast cancer may limit the applicability of their findings to other cancer types, as different cancers exhibit distinct genomic characteristics. Building on these studies, we implement and compare XGBoost, SVM, and TabNet to evaluate their effectiveness in cancer classification using CNV profiles. These models were chosen due to their strong performance in prior research and their ability to capture complex genomic patterns. By assessing their accuracy and classification performance, we aim to identify the most suitable model for genomic-based cancer classification.

Dataset

The dataset contains 2507 features, each representing a CNV value from a specific part of the genome. To better understand the dataset's distribution, we use UMAP (Uniform Manifold Approximation and Projection) to reduce the feature space to two dimensions. The resulting visualization allows us to observe clustering patterns among different cancer types. In Figure 1.1, where no feature selection has been applied, the cancer types exhibit significant overlap, with clusters blending into one another. This suggests that the raw feature space does not provide sufficient separation, making it harder for classification models to distinguish between cancer types.

In contrast, Figure 1.2, generated after applying Random Forest-based feature selection, reveals a much clearer structure. The clusters are more distinct, with better-defined boundaries and a wider spread across the UMAP space. This indicates that the most informative features contribute significantly to distinguishing between cancer types. The improved separation suggests that models trained on this reduced feature set are likely to achieve better accuracy while avoiding unnecessary complexity.

Given these observations, we apply feature selection or dimensionality reduction across all models to enhance performance. By reducing the number of features, we not only improve computational efficiency but also allow models to focus on the most relevant genomic signals, leading to more precise cancer classification.

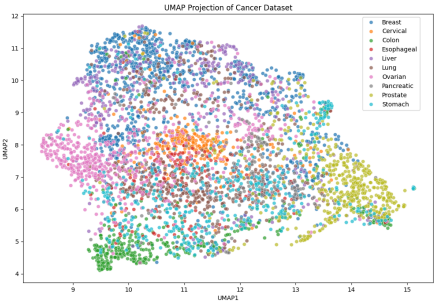

FIGURE 1.1 (UMAP of classes without feature selection)

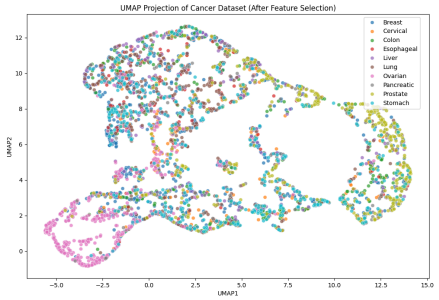

FIGURE 1.2 (UMAP of classes after Random Forest feature selection)

Class Imbalances

The genomic dataset exhibits a significant class imbalance, with breast cancer samples being overrepresented compared to pancreatic and esophageal cancers (Figure 1.3). This skewed distribution can lead to model bias, where predictions are more accurate for the breast cancer

class while underperforming for the minority classes. The imbalance also raises concerns about the model's ability to generalize effectively and perform well across all cancer types.

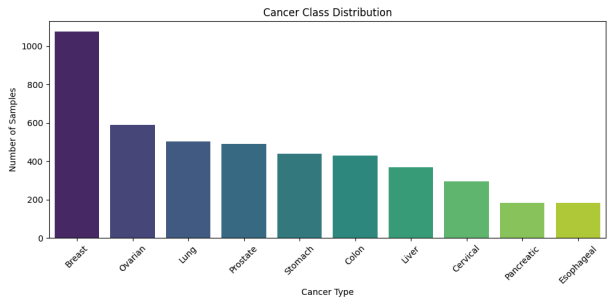

FIGURE 1.3 (distribution of cancer classes)

Methodology

To classify cancer types based on genomic data, specifically copy number variation (CNV) profiles, we employed three different machine learning models: Support Vector Machine (SVM), XGBoost, and TabNet. Each of these models has been chosen for their strong performance in classification tasks and their ability to handle high-dimensional, complex datasets like genomic data.

Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient boosting algorithm that builds an ensemble of decision trees to improve classification accuracy. Unlike traditional decision trees, which are prone to overfitting, XGBoost sequentially refines its predictions by learning from the mistakes of previous trees.

XGBoost formulates the classification problem as an ensemble of decision trees, where each tree corrects the errors made by the previous one. The objective is to minimize the loss function and achieve high accuracy in predicting the cancer types. The trees are built in an additive manner, where each subsequent tree attempts to correct the predictions of the previous trees, leading to a strong classifier.

SMOTE (Synthetic Minority Over-sampling Technique) was used to fix the class imbalance - it generated data to balance the classes. Feature selection was performed using a Random Forest model, such that only the top 100 features would be chosen to train the XGBoost model. The hyperparameters for XGBoost are then fine-tuned using a grid search approach with 5-fold cross-validation. The key hyperparameters optimized include learning rate, maximum tree depth, number of estimators, subsample ratio, and column sample per tree.

Tabnet

The second model is TabNet, a deep learning-based model that is particularly well-suited for tabular data like genomic data. TabNet utilizes attention mechanisms to process and select relevant features for classification. It learns a set of decision trees that capture complex relationships in the data, providing an efficient and interpretable model for high-dimensional data.

The preprocessing for TabNet only includes the use of SMOTE, as the neural network of TabNet is able to handle feature selection and does not require feature scaling, but struggles with class imbalances.

The hyperparameters of TabNet include learning rate, number of epochs, batch size, and virtual batch size. These hyperparameters are optimized using early stopping to prevent overfitting and ensure the model generalizes well to unseen data. The data was tested on a random sample of 20% of the dataset to keep it consistent with the other models, which use 5 fold Cross Validation for hyperparameter tuning.

Due to hardware limitations, we were unable to tune the hyperparameters of the TabNet model. As such, we capped the batch size and virtual batch size at 64, and kept the maximum epoch at 100.

Support Vector Machine (SVM)

The last selected model is the Support Vector Machine (SVM), a powerful classifier commonly used for high-dimensional datasets. SVM works by finding a hyperplane that best separates the different classes in the feature space.

Using SVM, we treat the problem as a multi-class classification problem where the goal is to find the optimal hyperplane that maximizes the margin between different cancer types. The preprocessing is done by first standardizing using the standard formula of $(X - \mu) / \sigma$.. For the model, we use a sigmoid kernel which is well-suited for this non-linearly separable data, enabling identification of relationships between features using neural network-like decision functions, also preventing overfitting which Radial Basis Functions are prone to with moderate class imbalances. The class imbalance is handled here using the parameter class_weight = 'balanced', which automatically adjusts weights inversely proportional to class frequencies.

The key hyperparameters of the SVM, such as C (penalty parameter) and gamma (kernel coefficient), are optimized through random-search CV for 8 iterations with cv=5, followed by an informed grid search CV for 16 parameter combinations with cv=3.

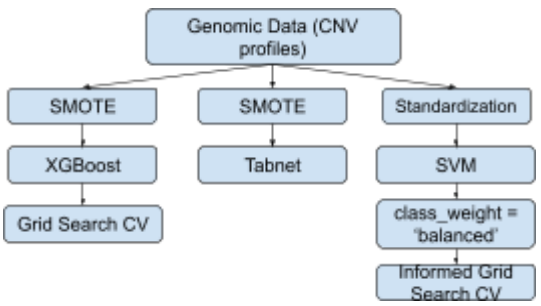

FIGURE 1.8 (pipelines of all 3 models)

Results

Extreme Gradient Boosting (XGBoost)

After hyperparameter tuning using GridSearch cross validation, we arrived at the best hyperparameters to be: colsample_bytree: 1.0, learning_rate: 0.1, max_depth: 7, n_estimators: 300, subsample: 0.8. XGBoost achieved an overall accuracy of 68.6%, with Top-2 and Top-3 accuracy of 82.1% and 88.4%, respectively.

XGBoost's strength lies in its ability to model interactions between features and handle noisy datasets through regularisation techniques. It performed best on ovarian (91.2%), prostate (81.3%), and colon (80.9%) cancers. To link it to real life medical use-cases, these particular cancer types probably possess highly distinctive CNV features.

However, XGBoost struggled with minority classes like esophageal (35.7%) and pancreatic (42.4%) cancers. Although SMOTE was utilised during data pre-processing to mitigate class imbalances, SMOTE-ing comes with some inherent limitations. Whilst SMOTE increases the number of training examples for underrepresented cancer types, it reinforces any underlying noise and class overlap present in the available data due to the nature of sample synthesis. Since XGBoost functions by partitioning the sample space via clear decision boundaries, it performs best when there are distinct class patterns. Thus, in the case where noise and overlaps are reinforced by SMOTE, a balanced set may not be sufficient to enhance performance, which may be a reason for XGBoost's underperformance with regards to Esophagus and Pancreatic cancer.

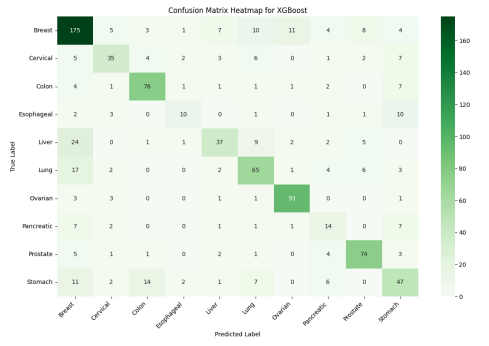

FIGURE 1.4 (XGBoost confusion matrix)

TABNET

Our TabNet model was trained on these hyperparameters: max_epochs: 100, patience: 10, batch_size: 64, virtual_batch_size: 64

TabNet achieved an overall accuracy of 76.5%, a Top-2 accuracy of 86.5%, and a Top-3 accuracy of 90.8% on the cancer classification task. It also exhibited a relatively small range of per-class accuracies (31.4%), indicating a stable and consistent performance across different cancer types. These results demonstrate TabNet's strong capacity to generalise across both easier and more difficult to identify cancer types.

This may be attributed to TabNet's key feature: its ability to dynamically select the most relevant features at each step through the implementation of a sparse attention mechanism. This essentially enables TabNet to focus on different subsets of features for different cancer types. Although this might intuitively sound redundant as feature selection had already been carried out earlier through a Random-forest algorithm, this is not the case. Whilst the earlier round of feature selection is done globally, TabNet carries out feature selection locally, on a per-instance basis, allowing for further fine tuning.

However, TabNet still faced challenges in classifying certain cancer types such as esophageal cancer (50.0%) and stomach cancer (48.9%). This suggests that even models with dynamic feature selection have intrinsic limits when the biological data itself is highly overlapping or lacks strong differentiating patterns.

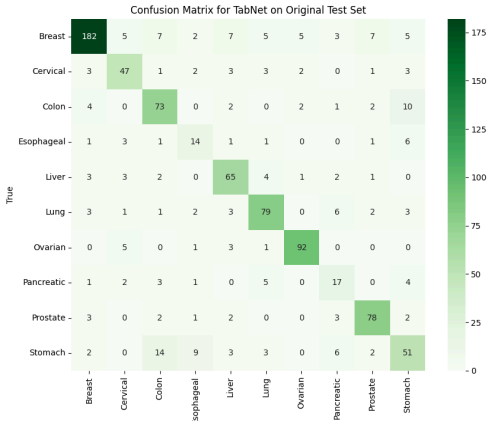

FIGURE 1.5 (TabNet confusion matrix)

SVM

Our SVM model was trained on these hyper parameters: C= 15, gamma= 0.0001, kernel= "sigmoid"

It achieved an overall accuracy of 72.4%, a Top-2 accuracy of 88.2%, and a Top-3 accuracy of 91.8%**.** Similar to our past 2 models, the SVM model performed particularly well on ovarian cancer (96.1%), prostate cancer (83.5%), and lung cancer (74.0%).

The use of kernel= "sigmoid" means that the model is able to generate non-linear boundaries to distinguish cancer types, essentially allowing it to more accurately draw partitions for each class. This is well suited for high-dimensional spaces such as genome CNV data.

The model exhibited weaker performance for stomach (52.2%) and colon (62.8%) cancer. This is likely due to the sigmoid kernel's inability to form sharp, localised boundaries as well as its sensitivity to the numerous, potentially noisy features in our dataset

One significant constraint of this model is interpretability. While linear SVMs can be interpreted by examining the feature weights, non-linear SVMs with kernel= "sigmoid" are less interpretable, as the decision function behaves like a single-layer perceptron, obscuring vital information such as contributing features and

boundary geometry. In clinical applications where explainability is essential, this limited interpretability could present challenges.

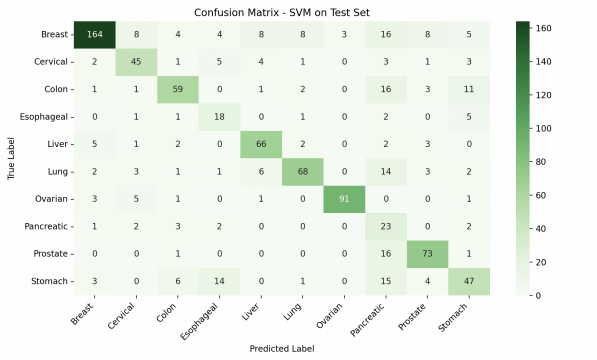

FIGURE 1.6 (SVM confusion matrix)

Comparative Analysis

Top-n \ Model	XGBoost	Tabnet	SVM
Top-1 accuracy	0.686	0.765	0.724
Top-2 accuracy	0.821	0.865	0.882
Top-3 accuracy	0.884	0.908	0.918

FIGURE 1.7 (Performance Metrics of the Models)

Comparing model performance, TabNet Outperforms XGBoost and SVM in overall Top-1 Accuracy, with 0.765, followed by SVM at 0.716 and then XGBoost at 0.686. However, SVM surprisingly yielded the highest Top-3 Accuracy at 0.916, hence it was extremely unlikely to misclassify the cancers within its top-3 predictions.

Trends within cancer classification rates in each class reveals that Ovarian cancer is the easiest to classify, with the lowest classification rate being TabNet at 0.902. However, Cervical, Liver and Stomach cancers have relatively low accuracy across all models, indicating that these cancers are the hardest to classify.

After analysis of the 3 cancers with the lowest classification rates (Cervical, Liver, Stomach), results reveal that the Cervical cancer is most misclassified as Breast Cancer (SVM: 8, TabNet: 3, XGBoost: 5), Liver Cancer is most misclassified as Breast Cancer (SVM: 5 TabNet: 3, XGBoost: 24), and Stomach Cancer is most misclassified as Colon Cancer (SVM: 6, TabNet: 14, XGBoost: 14)

These results concur with our UMAP projection of cancers - Cervical, Liver and Stomach are not separated away from other clusters, which makes it hardest for all the models to classify.

Cancer Type	TabNet	XGBoost	SVM
Cervical	0.277	0.426	0.308
Liver	0.207	0.526	0.185
Stomach	0.393	0.478	0.478

FIGURE 1.8 (Misclassification rates of difficult cancer types)

We also observed the performance variability of our models across all cancer types.

Model	Best Class Accuracy	Worst Class Accuracy	Range (Best - Worst)
XGBoost	91.2% (Ovarian)	35.7% (Esophageal)	55.50%
TabNet	80.3% (Liver)	48.9% (Stomach)	31.40%
SVM	89.2% (Ovarian)	52.2% (Stomach)	37.00%

FIGURE 1.9 (Performance variability for each model)

TabNet demonstrates the smallest range of 31.40%. This speaks to the model's greater scope for generalisation, ostensibly stemming from its ability to carry out instance-wise feature selection which would enable for a lower degree of performance fluctuation across the different cancer types.

Choice of Model

Noting that all three models have a low rate of classification for some cancers, we propose using TabNet as the main model of classification, but assisted by using the Top-3 accuracy of all three models, SVM, XGBoost and TabNet. In the classification of cancer, since biopsies are invasive, based on the top 3 accuracy of all 3 models, doctors can then test the most likely predictions.

In addition, as demonstrated in the results segment, TabNet's ability to generalise across both easier and harder to detect cancer classes also makes it the most efficient, resource-optimal choice for the main model for classification.

Moving forward, we believe that the key to addressing this lapse in model performance with regards to certain cancer types lies in feature engineering and expanding on domain-specific knowledge. Currently, the underperforming cancer types possess large feature overlaps that often prevent substantive splits. These issues will persist regardless of model type unless novel features are introduced.

APPENDIX

References

Chaurasiya, D. K.; Khamparia, A.; Pandey, B.; Gupta, D.; Khanna, A.; and Tiwari, S. 2020. Comparative Analysis of Machine Learning Algorithms in Breast Cancer Classification. *Procedia Computer Science* 167: 321–330. https://doi.org/10.1007/s11277-023-10438-9

Zelli, V.; Paci, P.; and Valerio, M. 2022. Classification of Tumor Types Using XGBoost Machine Learning Model: A Vector Space Transformation of Genomic Alterations. *Patterns* 3(2): 100470. https://doi.org/10.1186/s12967-023-04720-4

Rizwan, M.; Amin, M. B.; and Awan, M. J. 2020. An Investigation of Deep Neural Networks for Cancer Classification. *Procedia Computer Science* 178: 708–716. https://doi.org/10.1016/j.cmpb.2022.106951.

Member Contributions

Abdullah: Interpretation of results and model limitations (25%)
Edison: SVM implementation (25%)
Renaud: TabNet implementation (25%)
Shawn: XGBoost implementation (25%)
All team members contributed equally to the writing of the report