Machine Learning Classification of Thermal Liquid Biopsy

December 6, 2023

0.1 Introduction

Thermal liquid biopsy (TLB) is a growing field of biochemistry that holds potential to be used as a tool to diagnose and monitor disease. Thermograms, which are the results of TLB, differ based on the health status of a patient, and can be used to train classification models to identify an illness. This report details the work that has been done with Lung Cancer (LC) Thermograms during the Fall 2023 semester. The goal of this work was to use machine learning to train classification models to identify LC diagnosis and stage using TLB. First, feature selection was used to identify any bias in the thermograms, then random forests were cross-validated to train classification models to identify cancer type and stage. Table 1 shows the LC types and frequencies included in the dataset.

Figure 1 shows the median thermogram for each LC type. It can be seen that the median curves present differently for each type of LC. Despite the differences in the median curves, variation in individual samples is high. Figure 2 shows the median thermogram curve, as well as the 5th and 95th quantiles observed. The variation of individual samples within each group suggests that classification using thermograms will be difficult.

0.2 Thermogram Feature Importance

Due to the way TLB is collected, identifying where the tails of a thermogram begin involves human input. As a result, evaluating any bias that exists in thermogram tails needed to be evaluated. This was done by selecting the two largest classes in the dataset, Control and Adenocarcinoma, and training a random forest to classify the type with and without the thermogram tails. Feature importance for each random forest was collected, and compared to determine if the presence of the tails were biasing a thermogram classification.

Figure 3 shows the mean feature importance of each temperature, including the tails (45-90°C), for 1000 bootstrapped train test splits. Figure 4 shows the truncated thermogram temperatures, without the tails (51-83°C), for 1000 bootstrapped train test splits. Feature importance in thermograms that contain the tails are concentrated in the ends, whereas when the tail are truncated, feature importance becomes more distributed across the entirety of the thermogram. This indicates that the tails are biased heavily influencing the outcome of a classification model. In order to prevent the tails from influencing models, all models presented in this report will contain temperatures 51-83°C.

0.3 Results of Pairwise Classifications

For all LC types with 10 or more thermograms, pairwise classifications were conducted. 1,000 iterations of bootstrap cross-validation (BSCV) was utilized to evaluate random forest model per-

formance. For each bootstrapped sample, a grid of hyper parameter combinations was searched to find the best version. The parameters that were evaluated were the number of trees, the number of features, and tree depth. The hyper parameter states were grouped together, and the combination with the highest balanced accuracy was selected as the final model for each combination. Table 2 shows the results of each cancer type combination that were classified. Balanced accuracy was used as the primary metric for evaluating performance because it accounts for class imbalance. The balanced accuracy ranges from 0.51 - 0.67. The poor model performance for each cancer pair is not surprising given the large amount of thermogram overlap between cancer type.

0.4 Results of Stage Prediction

Thermograms were also used to predict lung cancer stage. Early vs late stages were predicted , as well as stages 1-4 for each cancer type for each cancer type with 10 or more samples. Table 3 and 4 shows the groups and counts for each stage. Figure 5 shows the median thermograms for each cancer and stage. Due to limited sample sizes, cancer types with thermograms that are particularly distinct, such as Large Cell, were not included in classification pairs.

Table 4 indicates that thermograms cannot be used to predict early vs late stage within individual cancer types, and also cannot be used to predict early vs late across all groups due to the low balanced accuarcies observed. Again, this is not surprising given how similar each cancer stage is to the other.

Finally, a multiclassifier was used to predict stages 1-4. BSCV was utilized to evaluate model performance, and a grid search for hyper parameters was used to find the optimal model. Table 5 shows the sample size for each group and stage. Table 6 shows that LC thermograms are not useful in classification problems with respect to stage. With balanced accuracies between 0.329 and 0.367, the models did not perform well predicting cancer stages. Classifiers were trained on lung cancer thermograms to identify cancer type, early/late stage, and stages 1-4 for each cancer type.

0.5 Conclusion

Despite initial visualization, analysis of LC thermograms indicate that they contain bias, and do not perform well in a classification problem. When evaluating feature importance in classifiers trained with and without tails, models that included the tails had high feature importance values concentrated at the ends the thermograms, and models trained without tail had feature importance values that were distributed across thermograms. Additionally, cross-validated random forests did not perform well classifying lung cancer type or stage, indicating that LC thermograms are not distinct enough to be used in classification problems.

Table 1: Number of Lung Cancer Samples by Type

	Cancer Type	Count
0	Adenocarcinoma	72
1	Control	51
2	Squamous	46
3	SCLC	16
4	NOS	8
5	Large cell	6

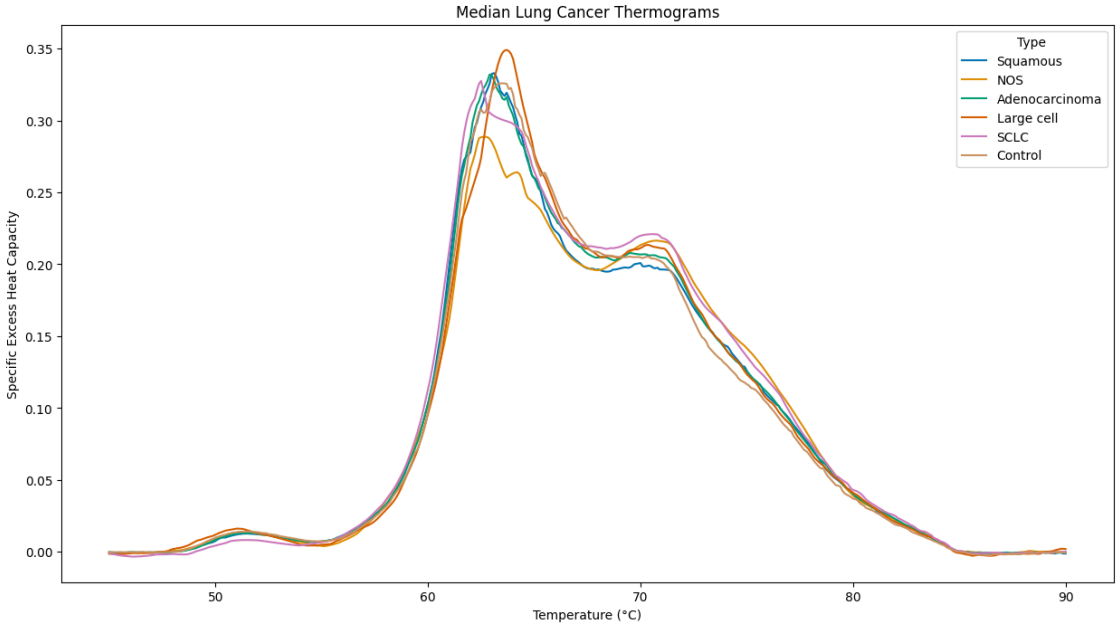

Figure 1: Median Lung Cancer Thermograms by type.

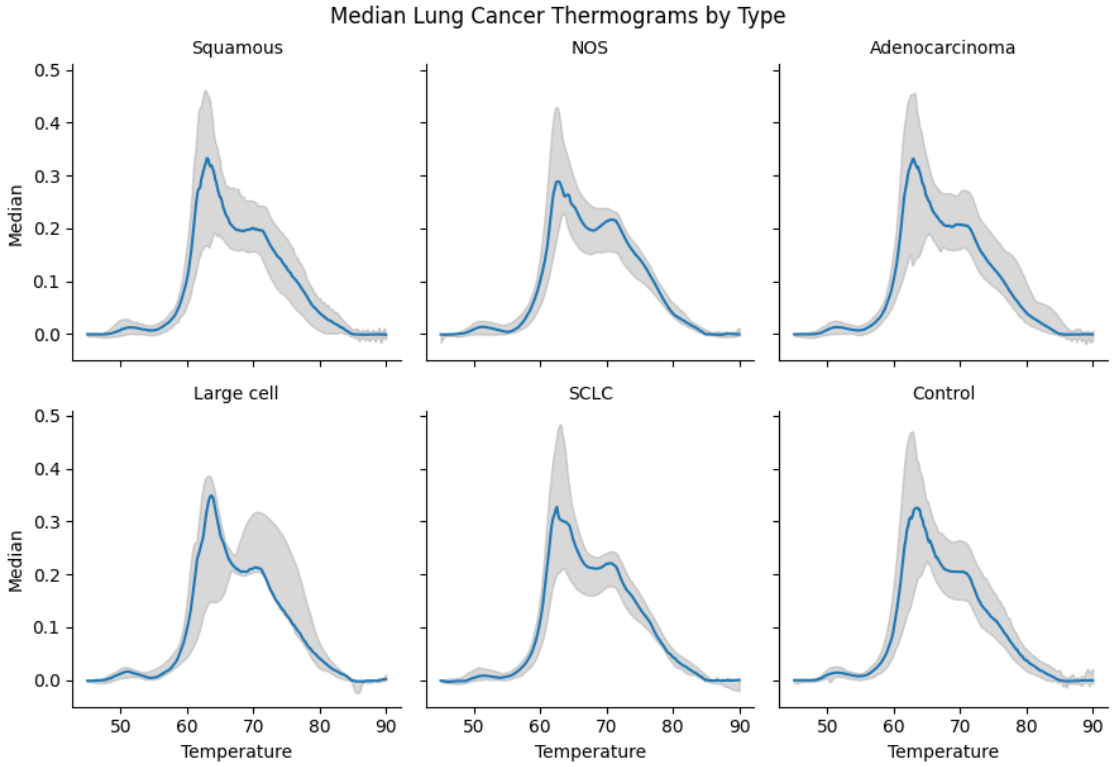

Figure 2: Lung Cancer Thermograms by Type with Quantile Ribbons

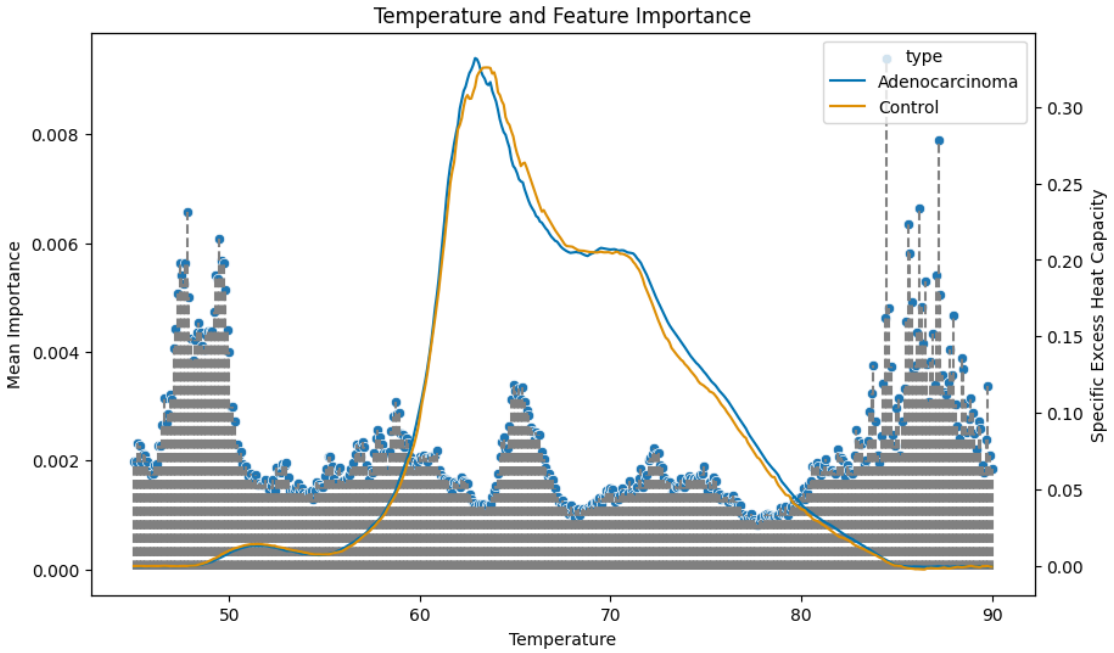

Figure 3: Feature importance for classifying with tails.

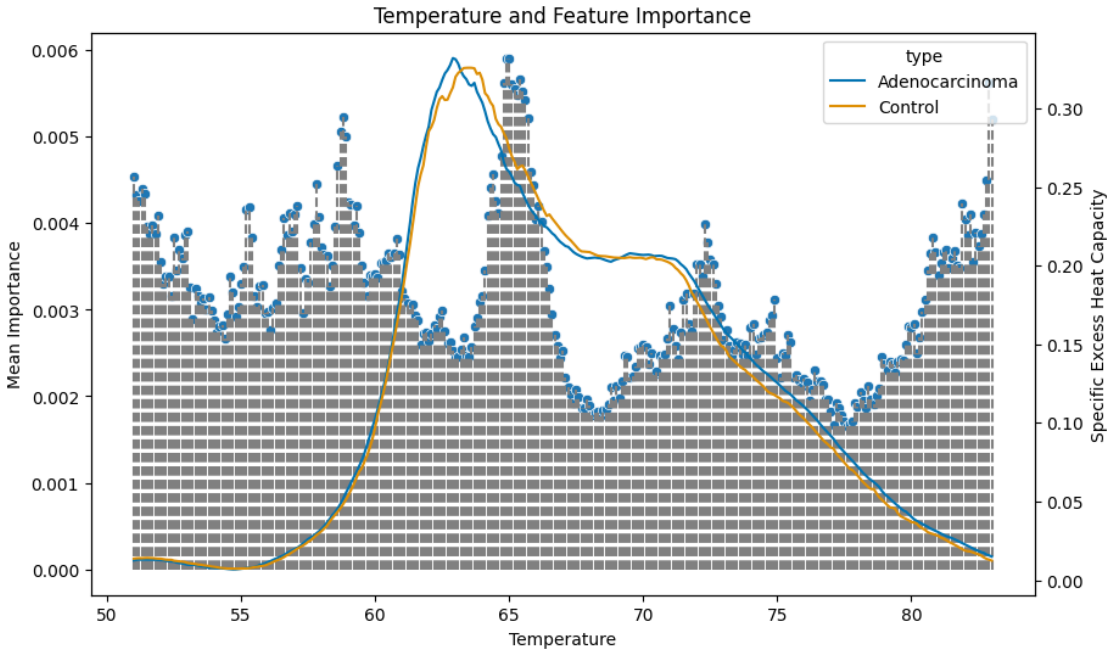

Figure 4: Feature importance for classifying without tails.

Table 2: Classification Results per Cancer Pair

	Pair	Weighted Accuracy	AUC
5	Control - SCLC	0.677083	0.799829
4	Squamous - Control	0.588179	0.622351
2	Adeno - Control	0.575928	0.619652
3	Squamous - SCLC	0.541017	0.598070
0	Adeno - SCLC	0.530405	0.593018
1	Adeno - Squamous	0.513628	0.524180

Table 3: Classification Results per Cancer Pair

		count
Diagnosis	Stage	
AC	Early	36
	Late	34
Large	Early	4
	Late	2
Mix	Early	1
	Late	1
NOS	Late	6
	Early	2
SCC	Early	23
	Late	23
SCLC	Early	11
	Late	5

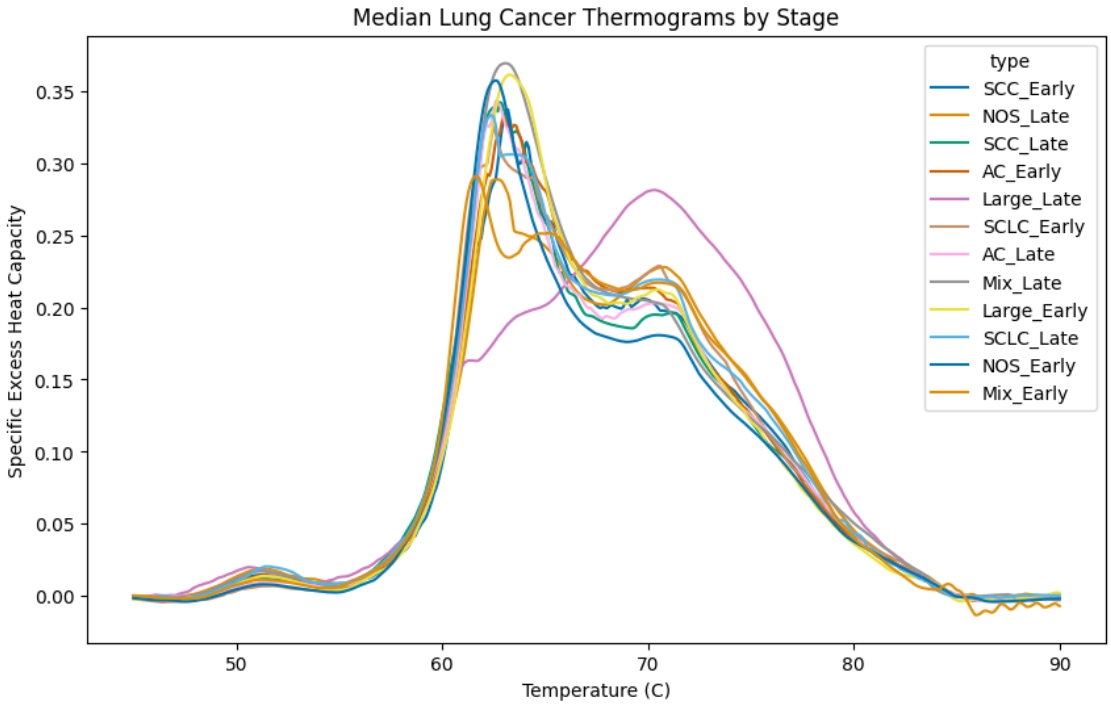

Figure 5: Median Lung Cancer Thermograms by Stage.

Table 4: Classification Results per Cancer Type for Early vs Late Stage

	Type	Weighted Accuracy	AUC
0	AC	0.550638	0.584997
1	SCC	0.508604	0.505465
2	SCLC	0.597812	0.684095
3	All	0.544999	0.561948

Table 5: Stages for Each Cancer Type

Diagnosis	Current Clinical Stage	count
AC	1	16
	2	10
	3	15
	4	29
Large	1	2
	2	2
	4	2
Mix	3	1
	4	1
NOS	1	1
	3	1
	4	6
SCC	1	7
	2	8
	3	16
	4	15

Table 6: Classification Results per Cancer Stage 1-4

	Type	Weighted Accuracy	AUC
0	AC	0.367951	0.624821
1	SCC	0.329758	0.605949
2	All	0.339149	0.622328