Tumor-Only Mutation Calling with Supervised Learning

Author: Aram Adamyan
BS in Data Science
American University of Armenia

Supervisor: Boris Shpak
BostonGene Technologies

Abstract – Identifying somatic mutations in tumor genomes is crucial for cancer diagnosis, treatment, and research. Traditional mutation-calling approaches typically require paired samples of tumor and normal tissues to differentiate between the somatic and germline variants. This paper presents a novel supervised learning approach for mutation calling that relies solely on tumor genome data, eliminating the need for normal sample sequencing which reduces cost and saves time. Here we use LightGBM model to perform the task of classifying the mutation types in tumor samples using a combination of genomic features. Our method involves extensive training using a dataset of known mutations from cancer patients, which allows the model to learn the distinguishing characteristics of somatic and germline variants. In This paper you will see how different approaches and designs have improved the performance. We also illustrate the robustness of our approach across various cancer types and calculating tumor-mutational burden (TMB). This innovation not only reduces the sample requirements and cost of genomic analysis but also has significant implications for precision oncology, where rapid and accurate mutation profiling is critical for personalized treatment strategies.

***Index Terms*—Tumor-Only Mutation calling, Tumor mutation burden, Light Gradient Boosting Machine, Classification, Somatic, Germline.**

I. Introduction

Variant/Mutation calling is the process of identifying mutations in the DNA sequence. In our case, we consider single nucleotide polymorphisms (SNPs) which include insertions, deletions, and substitutions. Variant calling is essential in cancer genomics, where it is used to identify mutations driving cancer development. This information is crucial for understanding cancer pathways, developing targeted therapies, and personalizing cancer treatment by improving drug efficacy and reducing side effects. Humans share more than 99% of their genome sequence indicating a high degree of genetic similarity among individuals. However, each individual has an estimated 3 to 4 million single nucleotide variants (SNVs) that contribute to genetic diversity. This difference arises from two main sources. It could be inherited from our parents through the germ cells, and is the reason why one person is different from one another. This type of difference is called germline variants. Secondly, it could arise within the cell and tissues throughout individual's lifetime due to environmental factors or other cellular processes, which we don't transmit to the next generation, and is called Somatic variants. These somatic mutations lead to the development of tumors and identifying them makes it a useful target for personalized cancer therapies.

The traditional approach for finding these somatic mutations include extracting the DNA of Normal (non-cancerous source) and Tumor cells of the patient and sequencing it. We then run a mutation caller (Strelka, Mutect, GATK etc.) which detect the differences of the nucleotides between the normal and the tumor genome. These differences are considered the somatic mutations. Moreover, to identify the germline variants the mutation caller compares the tumor and a universal human reference genome and gets the differences. from that it filters out the somatic variants that we found earlier, and we will be left with the germline mutations. The mutation caller also identifies false positives which are the nucleotides that are not mutations, but it has identified as one.

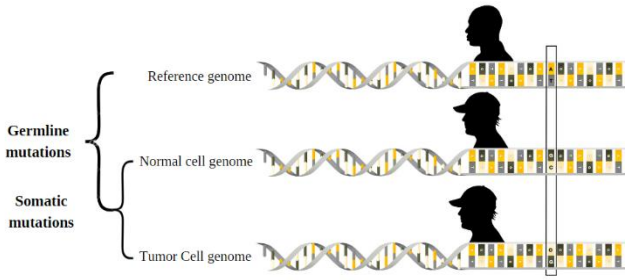

Figure 1. Paired tumor-normal mutation calling

II. RESULTS AND DISCUSSION

With our approach we have created a model that could identify and classify the mutation as germline, somatic or false positive (FP). The data is in a tabular format which initially had 27 columns (features), 36,482,012 rows where each row is a single mutation, covered 2020 patients that had been diagnosed with 119 different cancers. The features are information that describe the mutation such as Tumor_VAF, t_depth, POPAF, purity of the sequence and

etc. We processed the data by joining the tables of different features, cleaning duplicates, filtering out bad samples, standard-scaling and one-hot encoding the corresponding columns and fixing data related issues. After that phase we were left with 29,449,541 rows (mutations) and 92 columns which were ready to go into the model.
For the training approach we used Leave-one-patient-out-cross-validation (LOPOCV), Leave-one-group-of-patient-out-cross-validation (LOGOPOCV) and Leave-one-diagnosis-out-cross-validation (LODOCV). We have chosen these approaches so that we don't split mutations of same patients in the different sets, which is more realistic to the problem that it will solve in real clinical case.

Class	Counts	Percentage (%)
Germline	17317262	58.80
FP	11898876	40.40
Somatic	233403	0.79

Figure 2. There is a huge imbalance within the distribution of the classes.

The main evaluation metric chosen for the model was Precision recall area under the curve score (PR AUC). This metric is particularly useful here as we are dealing with a highly imbalanced dataset. And our focus is on the performance of positive class (somatic) predictions. We also track other metrics such as the f1 score, MCC, and classification report and plot the confusion matrix.
Finally, to check the effect on clinical use cases, we compare the value of the calculated Biomarkers, specifically the TMB (Tumor Mutation Burden) of the predicted and the actual mutations.

For our first iteration we performed a binary classification using two models, LGBM and logistic regression, by dropping the false positives from the dataset and training only somatic vs germline using LOPOCV. The reason for that is to compare the two models with a smaller amount of data and compare their performances to proceed with the best model and perform hyperparameter tuning.

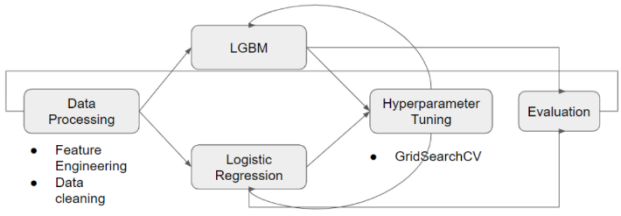

Figure 3. Architecture of the binary model

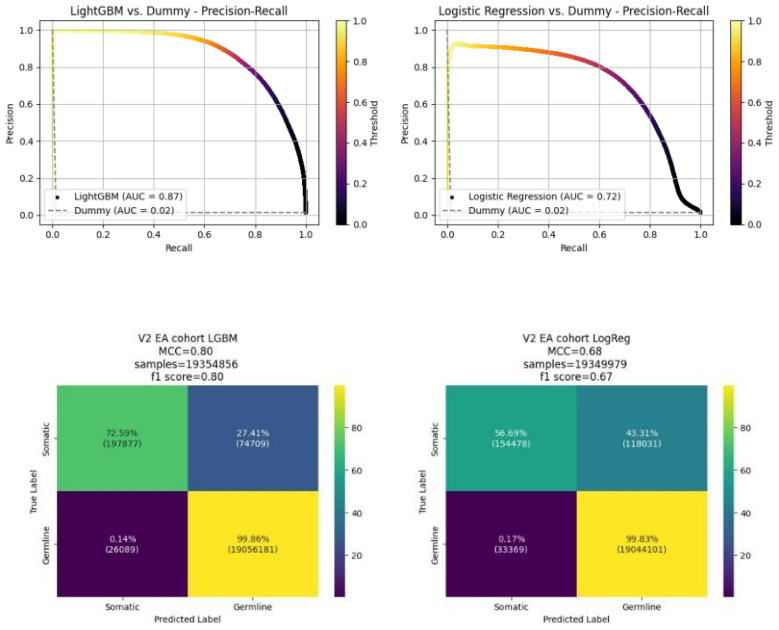

Figure 4. Performance of the binary model. The LGBM gave a result of 0.87 PR AUC for the somatic class and the log reg performed with 0.72 PR AUC on the positive class.

These are the hyperparameters that was used for the further model: **LGBMClassifier(learning_rate=0.1, max_depth=9, random_state=80, objective='multiclass', num_class=3, bagging_seed=13, bagging_fraction=0.8, feature_fraction=0.8, bagging_freq=5, reg_alpha=0.1, reg_lambda=0.1)**

After getting the optimal hyperparameters for the model we performed a multiclass classification on the whole data. This gave a result of 0.74.

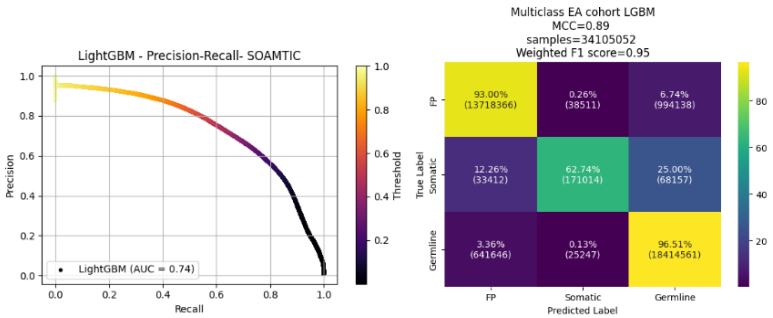

Figure 5. Performance of the multiclass classification

After some exploration and analysis, we came up with a better performing model that improved the PR AUC with an absolute value of 5%.

The architecture includes a two-step classification model. The first step of the LGBM model performs a binary classification by first identifying the FP vs (somatic and germline) which we will call TP. And after getting the TP with a very high accuracy we filter out the FP from the original dataset and run a multiclass LGBM model.

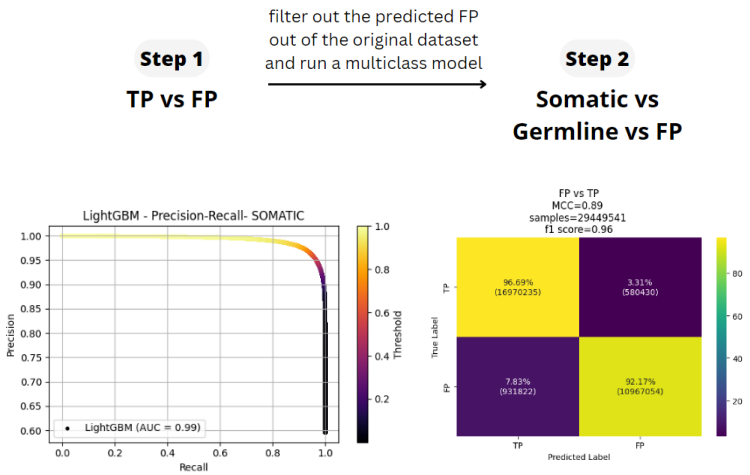

Figure 6. The performance of the binary model in Step 1. We observe a high precision and recall for predicting the false positives from the true positives using the LGBM model with the same hyperparameters.

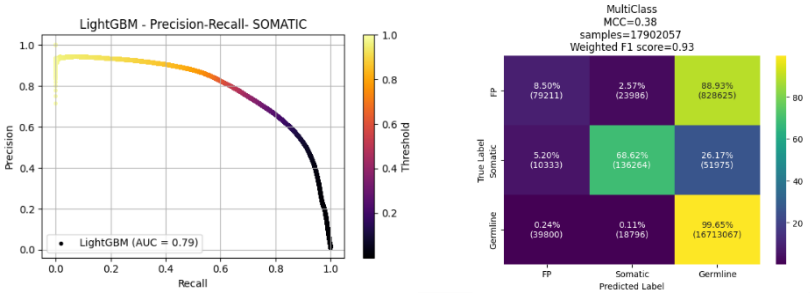

Figure 7: Two-step multiclass model evaluation. Step 2.

After this model we have tried stacking model for every diagnosis type on top of this base model hoping to improve the performance, but it didn't have any improvement. So, we did a deep dive to see how it is performing for each category.

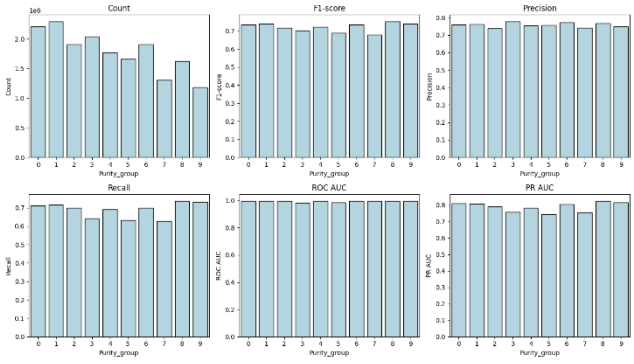

Figure 8: Performance on different binarized purity groups. So purity has values from 0 to1. So, group 0 are the mutations with purity [0, 0.1]. Group 1 are the mutations with purity from (0.1, 0.2] and so on until group 9 (0.9,1].

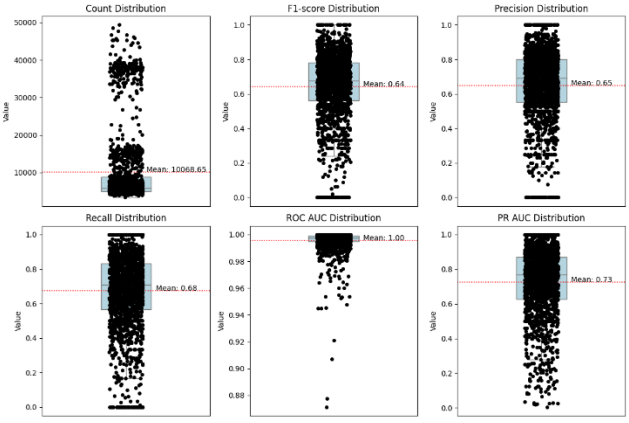

Figure 9: Performance of each patient. Each datapoint is the metric of the model for that patient. Count shows the number of mutations each patient has.

Looking at the PR AUC distribution of patients we came up with an idea of training the model on the good samples which we defined as the patients that have a PR AUC score higher than 0.65. We trained on the good samples but predicted on the whole samples using LODOCV. This improved our PR AUC results even further with an absolute value of 3%.

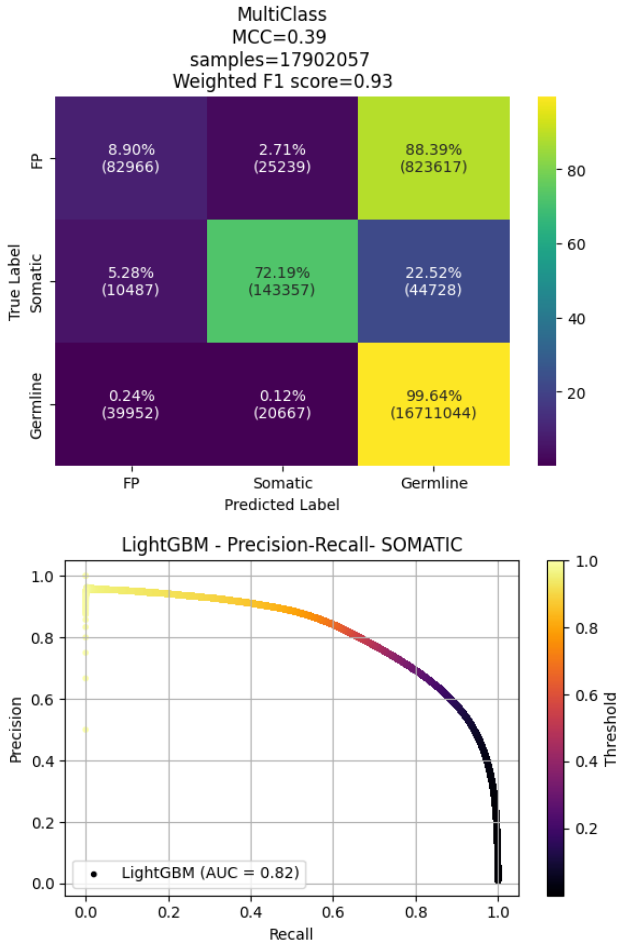

Figure 10: Two Step Multiclass Classification LODOCV (good samples PR AUC>= 0.65)

After getting a satisfactory result we also calculated the Tumor mutation Burden (TMB) using both predicted somatics and the actual labels. TMB is defined as the number of somatic mutations per megabase, and it is an important biomarker for cancer treatment.

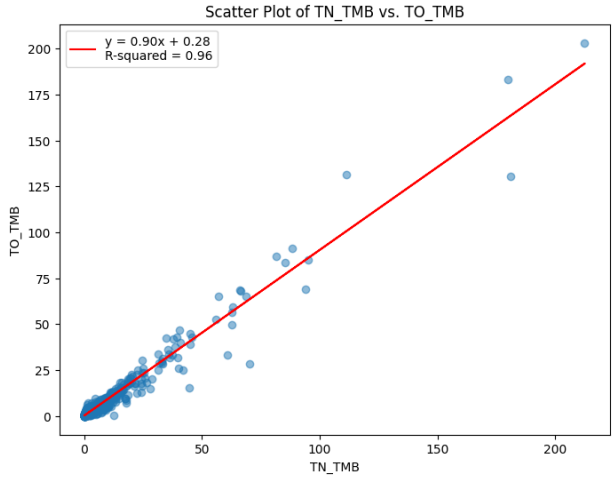

Figure 11: TMB values of each patient are calculated using the predicted values and the original values and the relationship is shown here. As is visible we have achieved a very high R2 score that indicates a very good performing model that can be used for clinical purposes.

III. Conclusion and References

The development of this supervised learning approach for tumor-only mutation has shown an increased performance compared to previous models. Using the LightGBM model, this study has achieved notable success in identifying somatic and germline variants within tumor genome data, eliminating the need for traditional paired samples. The two-step multiclass classification model demonstrated robustness across different cancer types and in the calculation of Tumor Mutation Burden (TMB), a key biomarker for precision oncology. The use of multiple cross-validation strategies, like Leave-One-Patient-Out Cross-Validation (LOPOCV), ensured rigorous testing and minimized bias. Ultimately, the paper shows that with this approach, significant progress can be made in clinical settings, potentially leading to faster, more accurate mutation profiling that underpins personalized treatment strategies. Future research may focus on improving the model by trying deep learning methods, exploring additional metrics, and calculating other biomarkers to increase the validity in precision oncology.

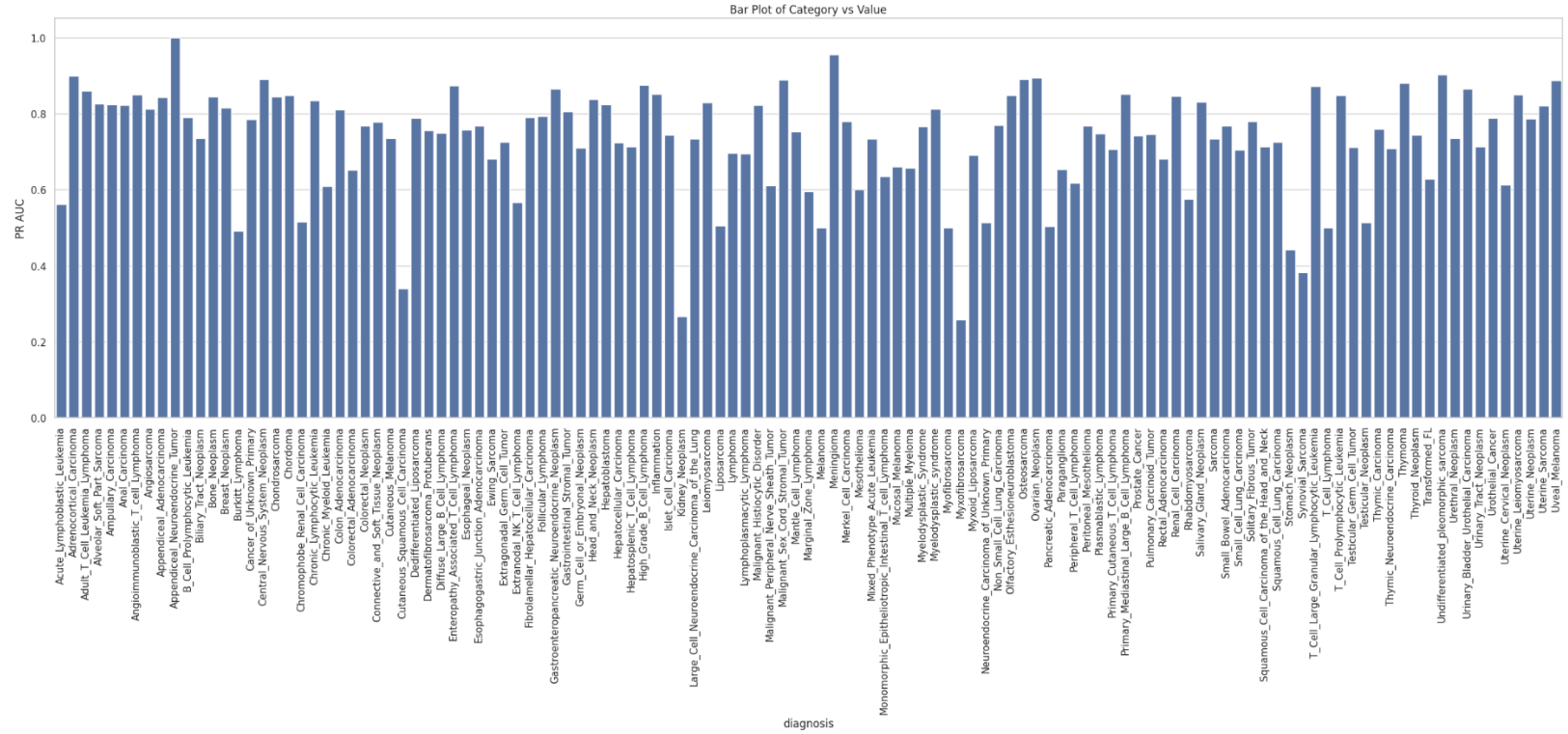

Figure 12: This graph shows the PRAUC values per diagnosis which proves that the model is quite robust on every cancer type.

REFERENCES

[1] Kim, S., Scheffler, K., Halpern, A.L. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591–594 (2018). https://doi.org/10.1038/s41592-018-0051-x

[2] McLaughlin, R.T., Asthana, M., Di Meo, M. *et al.* Fast, accurate, and racially unbiased pan-cancer tumor-only variant calling with tabular machine learning. npj Precis. Onc. 7, 4 (2023). https://doi.org/10.1038/s41698-022-00340-1

[3] Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics (Oxford, England), 28(14), 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

[4] Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., & Saunders, C. T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nature methods, 15(8), 591–594. https://doi.org/10.1038/s41592-018-0051-x

[5] Sha, D., Jin, Z., Budczies, J., Kluck, K., Stenzinger, A., & Sinicrope, F. A. (2020). Tumor Mutational Burden as a Predictive Biomarker in Solid Tumors. Cancer discovery, 10(12), 1808–1825. https://doi.org/10.1158/2159-8290.CD-20-0522