A Deep Learning Method Predicting Lung Adenocarcinoma Recurrence: Binary and Time-Based Approaches

By

Karma Hayek

BS, McGill University, 2023

Thesis

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in the Graduate Program of Biotechnology at Brown University

PROVIDENCE, RHODE ISLAND

May 2025

AUTHORIZATION TO LEND AND REPRODUCE THE THESIS

As the sole author of this thesis, I authorize Brown University to lend it to other institutions for the purpose of scholarly research.

Date 04/29/2025

Digitally signed by Karma Hayek
Date: 2025.04.29 16:19:19 -04'00'

Karma Hayek, Author

I further authorize Brown University to reproduce this thesis by photocopying or other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Date 04/29/2025

Digitally signed by Karma Hayek
Date: 2025.04.29 16:19:51 -04'00'

Karma Hayek, Author

This thesis by (Karma Hayek) is accepted in its present form by the Graduate Program in Biotechnology as satisfying the thesis requirements for the degree of Master of Science

Date_____ Signature: _____

Dr. Ece Uzun, Advisor

Date_____ Signature: _____

Dr. Jeremy Warner, Reader

Date_____ Signature: _____

Dr. Elizabeth Chen, Reader

Approved by the Graduate Council

Date_____ Signature: _____

Dr. Thomas Lewis, Dean of the Graduate School

Acknowledgments

First and foremost, I would like to thank the Brown Biotechnology Program for welcoming me with open arms and challenging me in the best ways possible. I am extremely grateful to have been able to model my course track according to my likings and to have had the chance to learn about what I love most on a daily basis.

Secondly, a heartfelt thank you to Dr. Ece Uzun, MS, PhD, FAMIA, who believed in me from day 1 and trusted that I could take on this project regardless of my purely biological background. Thank you for being a guiding voice along the way and for showing me that nothing is too difficult if you have the right mindset.

To the one and only (Uzun lab PhD student) Jessica Patricoski-Chavez, MS, thank you for always being one call away, for taking me under your wings, and for always finding a way to make me smile. Your energy is contagious and made my entire lab experience an absolute pleasure and great time. I can't wait to see you fly and accomplish great things.

Dr. Jeremy Warner, MS, MD, and Dr. Elizabeth Chen, PhD, thank you for your invaluable contributions to my thesis project and for supporting me along the way.

Dr. Jacquelyn Schell, PhD, you are the true engine and star behind this wonderful Master's program. Thank you for always being a listening ear and a guiding force. You made this experience an unforgettable and life-changing one.

Finally, to my family, I am forever grateful for your relentless support and motivation. Thank you for being the driving force behind everything that I do and for being by my side through thick and thin. My journey is yours just as much as it is mine. Merci à l'infini.

Table of Contents

Table of Figures

Table of Tables

Abstract of A Deep Learning Method Predicting Lung Adenocarcinoma Recurrence: Binary and Time-Based Approaches, by Karma Hayek, ScM, Brown University, May, 2025.

Despite advances in early detection and therapy, lung cancer is the leading cause of cancer death in the United States and accounts for approximately one-quarter of all cancer deaths. In early-stage non-small cell lung cancer (NSCLC), recurrence occurs in 30% to 55% of the patients, even after curative resection. In the event of recurrence, the 5-year survival after second operation decreases to 15%, indicating the poor prognosis with disease recurrence.. Lung adenocarcinoma (LUAD) patients account for a significant portion of recurred NSCLC cases, as LUAD is the most prevalent NSCLC subtype. In order to be able to offer more individualized treatment strategies for patients with LUAD, we developed two deep learning (DL) models tailored to predict binary (recurrence or no recurrence) and time-based (early vs. late recurrence i.e. before or after 12 months) LUAD recurrence by using clinical, mRNA, mutation, and methylation data.

To manage the high dimensionality of the multi-omics data, a modular feature selection pipeline was used, which integrated univariate filtering (ANOVA) and model-based methods (Logistic Regression, XGBoost, Random Forest) with hyperparameter optimization. The DL models were then built using Keras and benchmarked against conventional machine learning (ML) algorithms (K Neighbors Classifier, Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree).

The binary model had AUROC 0.898 and AUPRC 0.867 with perfect precision but with compromised sensitivity, reflecting high predictive specificity but possible under-recognition of recurrence. The time-based model had an AUROC of 0.916 and an AUPRC of 0.956, with the

detection of all true positives but over-estimation of false positives, indicating that further analysis is needed to obtain optimal model performance and generalizability. Further methods can be investigated in subsequent research to represent the complex, nonlinear relationships between omics layers (ex. graph neural networks, further hyperparameter tuning, etc.). Our method ultimately aims to investigate the usability of DL models to predict LUAD recurrence using clinical and genomic data from patients.

Chapter 1: Introduction

1.1 Lung cancer prevalence and subtypes

Despite advances in early detection and treatment, 234,580 new patients were diagnosed with lung cancer in 2024 in the United States (U.S), with a 25% 5-year survival rate (1). Lung cancer is the leading cause of cancer death in the U.S, accounting for around one quarter of all cancer-related deaths in the U.S, which is more than breast, prostate, and colorectal cancer combined. It is also increasing in never-smokers (1,2).

Lung cancers can be classified into several different subtypes based on histological, molecular, and clinical characteristics, which bear significant implications for prognosis and therapeutic strategies. The most common classifications of lung cancer are non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), with NSCLC making up about 85% of all cases (3). NSCLC is further divided into three major histological subtypes:

1. Lung Adenocarcinoma (LUAD): This is the most common subtype of NSCLC, which accounts for about 40% of all lung cancers (4). It usually develops in the outer parts of the lung and is the most common type among non-smokers (5).
2. Squamous cell carcinoma: Accounts for 25-30% of lung cancers and tends to occur, initially, near the center of the lung, adjacent to a bronchus (5).
3. Large cell carcinoma: Accounts for 5-10% of all lung cancers and also may arise in any portion of the lung (5).

LUAD, the most prevalent NSCLC subtype, is further classified based on histological characteristics and molecular attributes. According to the 2021 World Health Organization (WHO) classification, LUAD includes several histological subtypes, including minimally

invasive adenocarcinoma (nonmucinous and mucinous), invasive nonmucinous adenocarcinoma (with lepidic, acinar, papillary, micropapillary, and solid adenocarcinomas), invasive mucinous adenocarcinoma, mixed invasive mucinous and nonmucinous adenocarcinoma, colloid adenocarcinoma, fetal adenocarcinoma, and enteric type adenocarcinoma (6). Aside from these, adenocarcinoma not otherwise specified (NOS) is also a category for those cases that do not fit into any particular subtype. Molecular profiling, in contrast, has revealed a number of distinct lung adenocarcinoma (LUAD) subtypes, including terminal respiratory unit (TRU), proximal-inflammatory (PI), and proximal-proliferative (PP) subtypes (7,8).

The molecular classification of lung cancer is stratified by the expression of specific genetic alterations and biomarkers. For LUAD, common molecular alterations are mutations in *EGFR* and *KRAS*, as well as *ALK* rearrangements (8,9). All these molecular subtypes show important implications for targeted therapy and patient outcomes. Another important consideration in classification is staging with regard to tumor size, involvement of lymph nodes, and distant metastasis (TNM), as it deals with prognosis and treatment decision making (10,11). Such integration of histological, molecular, and clinical information allows a better understanding of the heterogeneity of lung cancers, which helps the clinician choose the most appropriate therapeutic strategies to improve patient outcomes (10). This thesis will specifically focus on LUAD.

1.2. LUAD treatments and outcomes

The landscape for treatment in LUAD has undergone many changes and still tends to be very dependent on the stage of disease and molecular characteristics. Despite these developments, the outcomes for advanced LUAD are still suboptimal.

For early-stage LUAD (stages I-IIIA), surgery usually is the cornerstone of treatment for early-stage LUAD, followed by adjuvant chemotherapy. The reported 5-year overall survival rate for patients with stage I NSCLC has been approximately 80%, though it is around 13 to 60% in stages II to IIIA (12).

In advanced or metastatic LUAD (stages IIIB-IV), systemic therapies become key to therapy:

1. Chemotherapy: Although the standard of care, platinum-based regimens have limited efficacy. The median overall survival for advanced lung cancer patients treated with chemotherapy alone is about 10.6 months (13).
2. Targeted therapies (e.g., EGFR or ALK inhibitors): Targeted therapies bring substantial benefits in selected cases with specific molecular alterations. These treatments are applicable only to a subset of patients.
3. Immunotherapy: Immunotherapies, including ICIs, have benefited a subset of patients. For example, the addition of pembrolizumab (Keytruda) to chemotherapy in metastatic non-squamous NSCLC achieved a five-year overall survival rate of 19.4%, as compared with 11.3% for chemotherapy alone (13).
4. Combination approaches: The combination of immunotherapy with chemotherapy has shown some promise. In the KEYNOTE-189 trial, the median overall survival for

pembrolizumab plus chemotherapy was 22.0 months versus 10.6 months for chemotherapy alone in metastatic non-squamous NSCLC (13).

In terms of treatment efficacy, a network meta-analysis reported that in the setting of advanced lung cancer, the highest response rates were: Targeted + Targeted therapy (82.9% cumulative probability), Chemo + Immuno (80.8%), and Targeted + Other Therapy (69.3%) (12). Moreover, the best disease-free control rates (DCR) were: Targeted + Others (93.4% cumulative probability), Chemo + Immuno (91.5%), and Targeted + Targeted Therapy (59.4%) (12).

The rates for 5-year relative survival from NSCLC vary greatly with the basis of the stage of diagnosis. In one study published in the *Journal of Thoracic Oncology*, 5-year relative survival rates for NSCLC is 60.1% for localized disease, 33.4% for regional disease, and 6.3% for distant disease (14). For advanced lung cancer patients, the median survival time was 13.46 months, the median disease-free survival time was 5.22 months, the response rate was 27.28%, and the DCR was 66.29% (12). Despite these advances, the prognosis for advanced LUAD is still unfavorable, and further research and development of novel therapeutic strategies are needed.

1.3. LUAD recurrence

In early NSCLC, 30% to 55% of patients develop recurrence despite curative resection, and the 5-year post-recurrence survival after repeat surgery is only 15% (15). Therefore, recurrence remains a challenge even in early-stage NSCLC. For LUAD specifically, recurrence rates vary from 13 to 23% in node-negative patients but tend to be higher in N1 patients (16).

There are three patterns of recurrence in LUAD: locoregional, distant, and both. The prognosis for recurrent LUAD is generally poor, with median PRS times ranging from 9 to 10.8 months (17,18).

However, a subset of patients can achieve long-term survival. A recent review of 635 patients who experienced postoperative recurrence of NSCLC reported an associated post-recurrence survival (PRS) rate of 13% at 5 years (18). Other determinants of improved PRS in the setting of LUAD include female sex, longer RFS, and locoregional-only recurrence (18). Of patients experiencing 5-year PRS, roughly 63% remained cancer-bearing largely treated with *EGFR*-TKIs whereas 37% remained cancer-controlled and within these individuals approximately 89% received upfront radical local treatment of their recurrence (18).

These findings are indicative of the heterogeneity in recurrence patterns and outcomes of early-stage patients with LUAD, thus mandating personalized strategies for follow-up and treatment. To improve overall treatment and survival outcomes, it is crucial to understand LUAD recurrence patterns. This may be done through Deep Learning (DL) methods, which have shown promising results in disease recurrence prediction (19).

1.4. Machine Learning and Deep Learning approaches in oncology and lung cancer

The technological advancement of computational tools, especially the advent of Artificial Intelligence or AI, has changed the face of cancer research. AI, comprising machine learning and deep learning, has found applications in cancer research and treatment areas, including but not limited to:

1. Detection: Google's LYNA model achieves 99.3% accuracy in identifying metastatic breast cancer, and even identifies micrometastases missed by pathologists (20).
2. Genomic Profiling: DeepVariant increases the accuracy of DNA sequencing, improving mutation detection (21).
3. Treatment Optimization: IBM Watson for Oncology (WFO) aligns patient data and clinical research to suggest individualized treatment (22).

AI refers to the simulation of human intelligence in machines programmed to think and learn like humans. ML, as a sub-branch of AI, consists of algorithms that let computers learn from big datasets without actually programming. DL is a more specific subset of ML and consists of neural networks with several layers that process complex data by feature extraction hierarchically.

AI-based methods have been used in the field of cancer studies:

1. Diagnosis: AI models are able to aid in the detection and classification of cancer. Researchers at Owkin, a company specializing in AI and ML for oncology, developed a model that predicts both the tumor-immune-stromal microenvironment status and treatment outcomes directly from medical images (23).
2. Prognosis: The Clinical Histopathology Imaging Evaluation Foundation (CHIEF) model predicts cancer prognosis based on histopathology images. CHIEF has shown strong performance in both cancer diagnosis and prognostic prediction. The DL model converts histopathological features into survival predictions and helps with personalized treatment

planning. Its high performance (outperforming other DL models by 36.1%) on heterogeneous cancer types highlights its ability to streamline clinical decision-making and potentially enhance patient outcomes (24).

3. Treatment Selection: AI is being used to develop new cancer treatments through novel approaches to drug discovery, drug repurposing, and predicting patient responses to treatment. NCI researchers have employed AI to help predict how immune cells called T cells will respond to tumors in a bid to improve immunotherapies (25).
4. Multi-omics Integration: AI models have been increasingly applied with multi-omics data for an integrative understanding of cancer biology. These techniques may provide a paradigm shift in cancer subtyping, risk stratification, prognostication, prediction, and clinical decision-making (26).
5. Precision Medicine: TME-NET, an interpretable deep neural network, was designed to forecast the response to immunotherapy across various cancers through the analysis of the tumor microenvironment. The model surpassed support vector machine and k-nearest neighbors in Area Under Curve (AUC) and accuracy when predicting the outcomes of treatment and also in elucidating the mechanism of immunotherapy. TME-NET can complement the current biomarkers through its interpretable predictions, thus improving personalized treatment strategies for cancer patients.

For lung cancer specifically, numerous ML and DL studies have been conducted, and have shown promising results (Table 1). The *IBPGNET*, proposed in 2024, adopted ML methods to integrate multi-omics data in a cohort of 505 patients with LUAD. An embedded pathway hierarchy relationship in the model realized promising performance, with an AUPR value of

0.790 in 5-fold cross-validation, AUC of 0.88, and accuracy of 0.82, which further demonstrates that the integration of biological knowledge is capable of enhancing the performance of a machine learning algorithm for the improvement of prediction results (27).

Additionally, the *Integrated Deep Learning Evaluation (IDLE)* study used a DL approach for analyzing the preoperative features of low-dose CT images in combination with histologic findings from a series of 182 patients who were stage IA NSCLC. A model was then developed and a 5-year AUC of 0.817 ± 0.037 obtained, proving the effectiveness of deep learning in useful information identification from medical imaging data for recurrence prediction (28) (Table 1).

The *Genomic Pathologic Annotated Risk Model*, developed in 2021, integrated demographic, imaging, pathologic, genomic, recurrence, and follow-up data in 426 patients with LUAD. The ML model outperformed the traditional TNM-based model with a concordance probability estimate of 0.73 compared to 0.61 for the TNM model. This shows that different types of data integration are helpful for increasing the risk stratification (29) (Table 1).

Finally, an immune gene-based ML model for recurrence prediction of LUAD was developed in 2023. In this analysis, it used an ensemble linear kernel support vector machine on the immune gene expression data and clinicopathologic factors of 41 early-stage lung cancer patients. Despite the relatively small size of the dataset, promising results were obtained in the validation set: accuracy of 0.90, sensitivity of 0.75, and specificity of 1.00 (30) (Table 1).

These studies collectively point towards the potential of ML and DL methods to improve the prediction of lung cancer recurrence. The models, applying various data types and the most recent analytical techniques, enable more accurate risk stratification and treatment strategies tailored to the individual patient with lung cancer. However, there is still a need for a more integrated approach that combines clinical information with genomic data to render an all-encompassing model with a more generalizable approach (Table 1).

Table 1. Non-exhaustive summary of ML/DL approaches in NSCLC research

Study	Methodology (ML or DL)	Dataset	Information Type	Performance
Interpretable Biological Pathway Graph Neural Networks (IBPGNET) (2024) (27)	ML (Graph Neural Networks)	Copy number variant (CNV) and somatic mutation data from the Xena TCGA Pan-Cancer web. 134 patients with recurrent LUAD 371 patients with non-recurrent LUAD	Pathway hierarchy relationships, multi-omics data	AUPR: 0.790 (in 5-fold cross-validation) AUC: 0.88 Accuracy: 0.82 F1 score: 0.68
Integrated Deep Learning Evaluation (IDLE) (2022) (28)	DL	182 stage IA NSCLC patients from the National Lung Screening Trial	Preoperative low-dose CT image features and histologic findings	5-year AUC of 0.817 ± 0.037
Genomic Pathologic Annotated Risk Model (2021) (29)	ML	426 patients with LUAD	Demographic, imaging, staging, pathologic, genomic, recurrence, and follow-up data	Concordance probability estimate of 0.73 vs 0.61 for TNM-based model
Immune gene-based ML model for LUAD recurrence prediction (2023) (30)	ML (Ensemble linear kernel support vector machine)	41 early-stage lung cancer patients (16 with recurrence, 25 without)	Immune gene expression data, clinicopathologic factors	Validation results: Accuracy: 0.90 Sensitivity 0.75 Specificity: 1 PPV: 1 NPV: 0.86

Chapter 2: Methods

2.1. Datasets

The Cancer Genome Atlas (TCGA) has molecularly classified approximately 10,000 tumors across 33 cancer types, including multiple types of lung cancer (31). To get an exhaustive dataset that accurately represents the LUAD tumors, we downloaded via cBioPortal (32) the clinical, mutation, methylation, and mRNA expression (RNA-seq z-scores) data from the LUAD PanCancer Atlas dataset (version 2023-08-13), which uses GDAC Firehose (version 2016-01-28) as its data source before transformations (33,34). We also gathered smoking history information from the same TCGA dataset, but downloaded from the Xena browser of the University of California Santa Cruz (UCSC) as this information was not available on cBioPortal (35).

2.1.1. Initial DL modeling using DFS-based labels

When first developing the models, Disease-Free Survival (DFS) status was used as the endpoint to define recurrence. Only patients with complete DFS, clinical, mRNA, mutation, and methylation data were included, leading to 196 samples in total (148 no recurrence, 25 early recurrence, and 23 late recurrence). Feature cleaning involved the use of a pairwise correlation threshold of 0.2, elimination of expression features with absolute median value > 3, and a variance threshold of 0.01. This resulted in a final dataset containing 348 mRNA features, 17 mutation features, and 30 methylation features. Individual DL models were trained on each modality from this DFS-based dataset and compared to baseline ML models (K Neighbors Classifier, Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree). Model performance for each modality is shown in Supplementary Figure S1. However, the

binary class imbalance and reduced sample size decreased the clinical utility of this end-point and made the models more prone to overfitting. Therefore, we discontinued this exploratory DFS-based pipeline in favor of a refined PFS-based approach, described in the sections that follow.

2.1.2. Clinical data pre-processing for final PFS-based modeling strategy

The TCGA dataset contained two kinds of separate clinical data: patient and sample data. The clinical dataset contained the patient-level attributes such as age, sex, tumor stage, and follow-up time. The sample dataset included attributes that are sample-specific: tumor type, tissue-related information. To start, preprocessing of the datasets was performed through the removal of irrelevant and redundant data. This involved dropping columns that are static or constant, like 'SUBTYPE', 'CANCER_TYPE_ACRONYM', and 'FORM_COMPLETION_DATE. Equally, columns holding null values for all entries were discarded. Patients whose data with fewer than six non-null values were removed. In respect to rows that had a lot of missing values, a threshold-based approach was taken in removing patients whose attribute non-null values were fewer than six.

The cleaned clinical dataset was then merged with the sample dataset based on PATIENT_ID. This ensured that patient-specific details, such as age, gender, and tumor stage, were matched with their corresponding sample. SAMPLE_ID was set as the index to ensure that each row corresponded to a unique sample while retaining vital patient-level features in the final dataset.

First, the data was merged — using SAMPLE_ID as a key — with supplemental information from Xena on smoking history, which defined smoking history in four categories: (1) never smoker in one's lifetime, (2) current smoker, (3) reformed smoker of more than 15 years, and (4) reformed smoker of 15 years or less. Rows with missing values in PFS_STATUS were dropped to preserve the integrity of the target variable for modeling work.

The final cleaned and integrated dataset included SAMPLE_ID as the index and retained 365 patients with 13 clinical features: tobacco_smoking_history, TUMOR_TYPE, TMB_NONSYNONYMOUS, AGE, SEX, AJCC_PATHOLOGIC_TUMOR_STAGE, DAYS_LAST_FOLLOWUP, PATH_M_STAGE, PATH_N_STAGE, PATH_T_STAGE, PFS_STATUS, PFS_MONTHS, and GENETIC_ANCESTRY_LABEL. The final dataset was exported into pickle (a python-specific serialization program) and CSV formats to be used in downstream DL and statistical analyses.

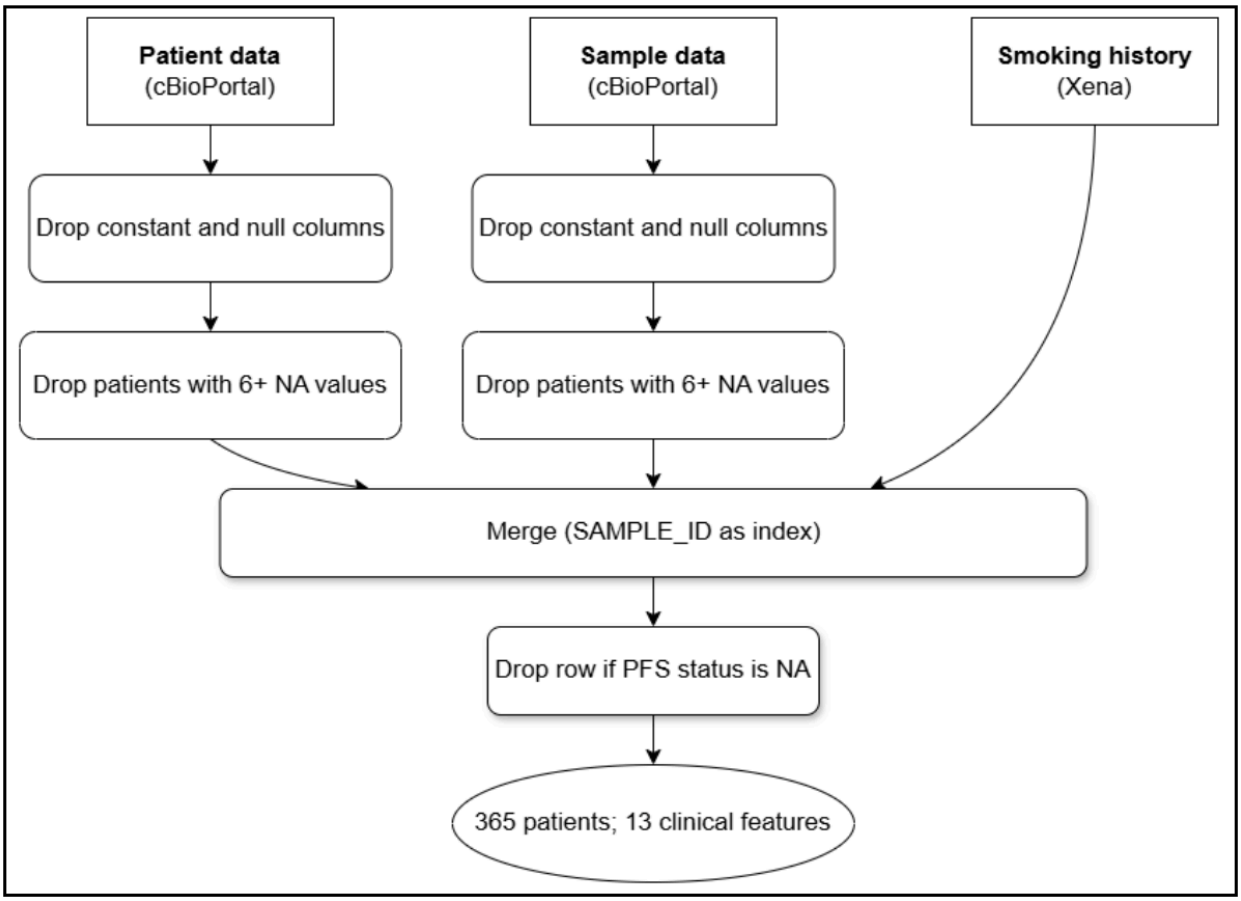


Figure 1. Formation of the clinical / sample dataset by integrating patient, sample, and smoking history data. Sample and patient-level clinical data was downloaded from cBioPortal, and smoking history was retrieved from UCSC Xena. After removal of constant and entirely missing columns, patients with more than six null values were filtered out. The datasets were merged on PATIENT_ID, indexed by SAMPLE_ID, and smoking history was incorporated. Samples with missing progression-free survival status (PFS_STATUS) were filtered out. The final dataset included 365 patients with LUAD and 13 clinical features for downstream analysis.

2.1.2 mRNA expression

The dataset containing mRNA expression (denoted in terms of z-scores) was derived from the raw TCGA RNA sequencing data (data_mrna_seq_v2_rsem_zscores_ref_all). The z-score denotes the relative expression for each gene with respect to the mean expression across samples; a positive value generally indicates increased expression, whereas a negative value

indicates decreased expression. Initially, this dataset consisted of 510 patient samples and 20,531 unique genes.

To facilitate consistency between the mRNA dataset and the corresponding clinical data, we used a clinical reference file, which had all the identifiers for patient samples and their corresponding PFS status (only_sampleID_and_PFSstatus.pkl). The clinical dataset included a total of 374 patients. All Patient IDs from the mRNA dataset not matching the clinical reference were excluded; this filtering step removed 139 patients. After the filtering, the mRNA dataset retained 371 patient samples.

Next, the dataset was cleaned of genes that did not have identifiers for Hugo_Symbol. Also, genes with missing values for more than 5 features were considered and removed. After this step, the total number of genes was decreased to 20,097. Finally, the index of the dataset was reset to correspond to the filtered rows only. To confirm the integrity of the cleaned dataset,we checked for any remaining null data; no further missing data was observed. The final mRNA expression dataset with 371 patients and 20,097 genes was saved in both pickle and CSV formats for downstream analysis (mrna_seqv2_zscores.pkl and mrna_seqv2_zscores.csv respectively).

This preprocessing step ensured that only high-quality, annotated gene expression data and clinically relevant patient samples were retained for further analyses. We then further processed the mRNA z-score dataset to facilitate downstream analysis. We looked into duplicate gene identifiers in the mRNA dataset. To negate this effect, we generated a combined identifier called Hugo_Entrez by merging the Hugo_Symbol and its corresponding Entrez_Gene_Id.

Duplicate pairs based on Hugo_Entrez were found; as the expression values for these duplicate rows were not the same, the suffixes _first and _second were appended to the respective rows to make them unique. This step made the gene identifiers unique while preserving the expression values for each entry.

After this cleaning, we dropped the Hugo_Symbol and Entrez_Gene_Id columns, only keeping the updated Hugo_Entrez identifier and the expression values. The data was transposed and SAMPLE_ID was set as the index to get the format required to facilitate the later combination with the clinical information. We then merged the processed mRNA expression data with a file containing the PFS_STATUS of each original patient. The patients that had mRNA expression data but that were missing progression-free survival status were removed, and a final sample size of 362 was adopted for further analysis, with 20,097 genes. We saved this dataset with mRNA z-scores and PFS status in pickle and CSV formats for further use (Figure 2).

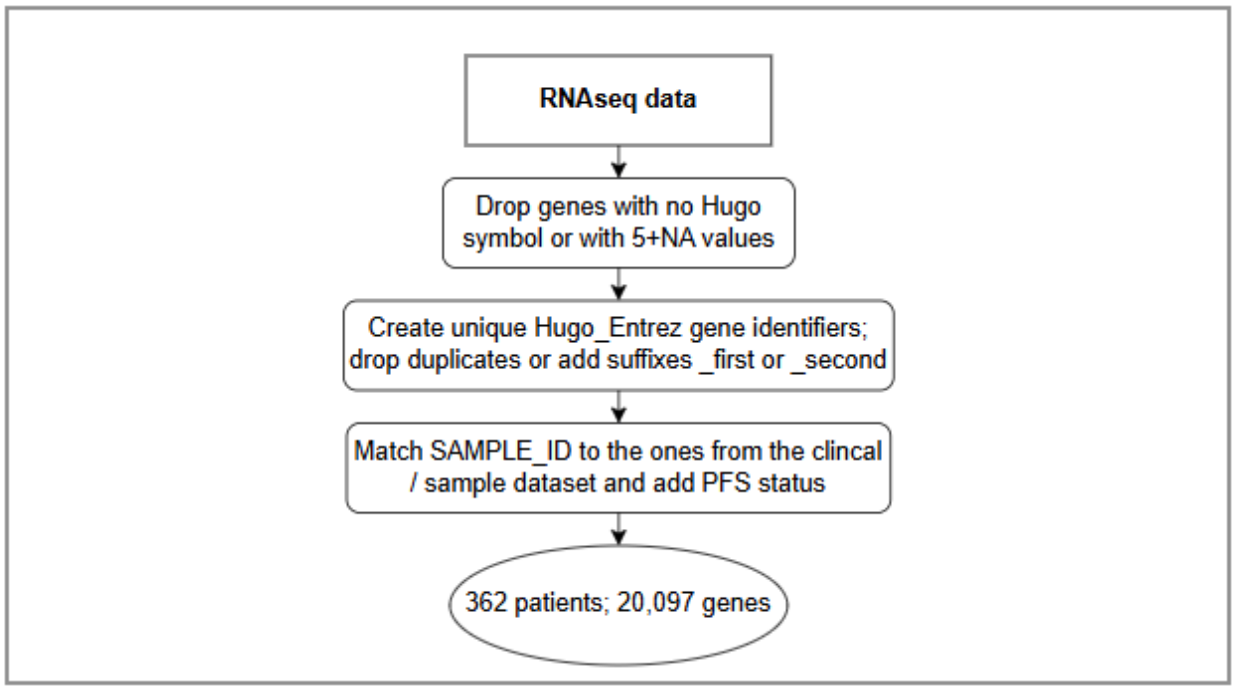


Figure 2. Filtering and pre-processing of RNA-seq z-score data before analysis. mRNA expression levels for the TCGA LUAD dataset were retained only for cases with corresponding clinical progression-free survival (PFS) data, narrowing the dataset from 510 to 371 patients. Genes with no identifiers or missing values for over five samples were filtered out, leaving 20,097 genes. To clarify gene names, a new unique identifier (Hugo_Entrez) was generated by concatenating Hugo_Symbol and Entrez_Gene_Id, and duplicates were prefixed with _first/_second when their expression differed. The original gene name columns were removed, data was transposed, and SAMPLE_ID was made index. Final filtration excluded samples lacking PFS status, retaining 362 patient samples and 20,097 genes for downstream analysis.

2.1.3 Mutation counts

The mutation information was pre-filtered to keep the non-synonymous mutations for the ML model. The original mutation file had 243,229 counts of mutations across 19,117 unique genes. Any rows that had missing values in crucial columns like Hugo_Symbol, Entrez_Gene_Id, Consequence, Variant_Classification, or Variant_Type were dropped. This resulted in a cleaned set of 243,285 mutations and 19,115 unique genes.

Then, the synonymous and RNA mutations were filtered out. So, from the Variant_Classification column, just the variants that were "Silent" or "RNA" only were removed. This eliminated 53,642 silent mutation and RNA-related variant records, and the dataset contained 189,563 mutations in 17,951 unique genes. Further filtration was done to keep only the non-synonymous mutations like "Missense_Mutation," "Nonsense_Mutation," "In_Frame_Del," "In_Frame_Ins," "Frame_Shift_Del," "Frame_Shift_Ins," "Translation_Start_Site," "Splice_Site," and "Splice_Region." There were 169,434 mutations in 17,316 unique genes after this step.

This filtered mutation information was then transformed and reorganized into the format of a mutation count matrix, with patient samples as rows and mutated genes as columns, with a "_MUT" suffix added to distinguish them. For example, the number "2" indicated 2 mutations in a gene for a patient, and "0" indicated no mutation. The processed mutation count matrix contained 17,316 gene columns and 370 sample ID rows.

To incorporate recurrence data, we joined the mutation sample IDs with sample / clinical dataset IDs and added the PFS_STATUS column alone (without the clinical information). Patient samples that contained mutation data but lacked PFS_STATUS values were discarded, reducing the patient samples to 361 but still maintaining 17,316 genes. We then pickled and stored the resulting dataframe in CSV format for downstream processing (Figure 3).

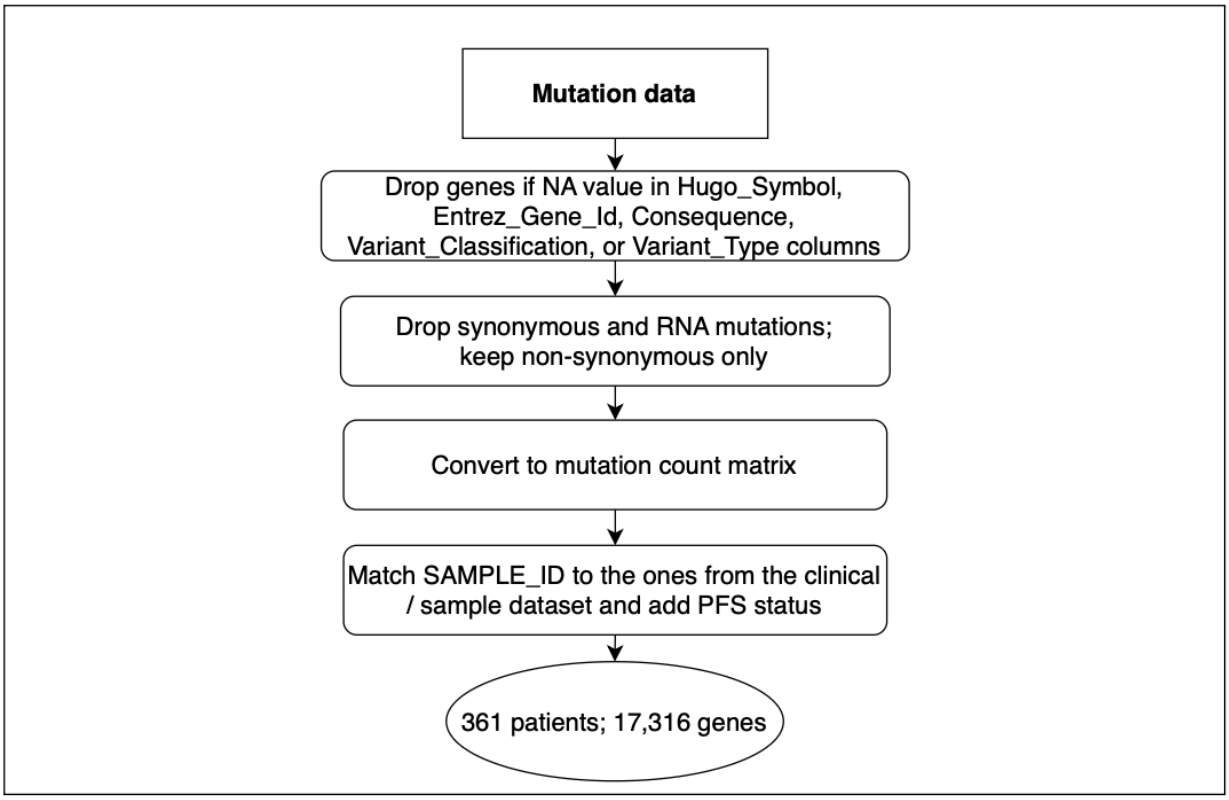


Figure 3. Mutation data pre-processing and transformation into a non-synonymous count matrix. Unfiltered raw mutation data from the TCGA were filtered down to only non-synonymous variants. RNA and silent mutations and entries without gene annotation or classification information present were filtered out, resulting in 169,434 mutations across 17,316 unique genes. These were transformed into a mutation count matrix with gene columns (with suffix "_MUT") and sample rows. The dataset was combined with PFS data, and the samples lacking PFS status were excluded from analysis. The finalized dataset included 361 LUAD samples and 17,316 gene mutation features for downstream analysis.

2.1.4 Methylation beta values

We extracted the input methylation dataset from the raw TCGA files, and this contained methylation beta values derived from Illumina HM450 methylation arrays. We then removed columns that did not give methylation beta value or gene name information, which were 'ENTITY_STABLE_ID', 'DESCRIPTION', and 'TRANSCRIPT_ID'. The 'NAME' column, which corresponded to gene IDs, was then set as the index for the individual methylation

features. At this stage, the matrix had a total of 562 samples (columns) and 22,601 methylation features (rows).

We then cleaned the data and prepared it for analysis by dropping missing values, which reduced the dataset to 19,705 methylation features. We then transposed this information into a matrix where samples would be the rows and the methylation features the columns. The names of the methylation features were changed by appending the suffix '_MET' to make them consistent in order to carry out the downstream analysis. The sample IDs were then set as the index to easily integrate them with clinical data.

We merged PFS status information into the methylation data using a left join operation to ensure that all methylation data for samples with available clinical labels were retained. After merging, the rows containing missing values were dropped to keep a complete and clean dataset. The final processed matrix consisted of 363 samples and 19,703 methylation features, with the 'PFS_STATUS' column included as the target variable for subsequent modeling. The processed dataset was exported in both pickle and csv formats (Figure 4).

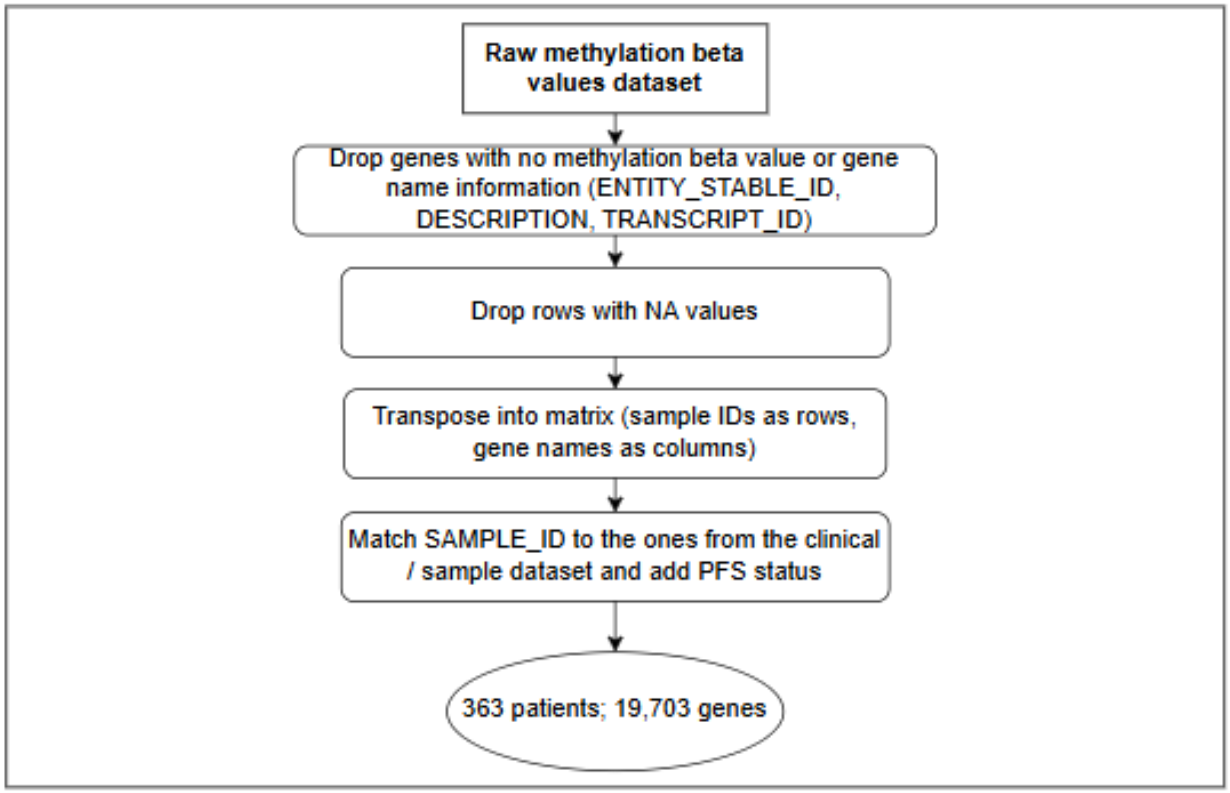


Figure 4. Methylation beta value data processing and formatting. TCGA LUAD sample methylation beta values derived from Illumina HM450 arrays were cleaned by removing columns with no gene information and setting gene identifiers as row indices. Features that had missing values were removed, then the data was transposed to form a matrix of 562 samples by 19,705 features (with an added "_MET" suffix for future use). SAMPLE_ID was made the index, and PFS status was combined by left join. Samples missing PFS information were discarded, leaving the final matrix consisting of 363 LUAD samples and 19,703 methylation features.

2.1.5. Formation of the combined dataset

2.1.5.1. STREAMLINE evaluation of individual omics modalities

After forming the individual omics datasets, we included a benchmarking step to ascertain the predictive potential of each modality in isolation. To this end, we employed STREAMLINE, an open-source AutoML pipeline developed by UrbsLab (https://github.com/UrbsLab/STREAMLINE), which automates preprocessing, feature selection, model training, and evaluation on high-dimensional datasets (36). The mRNA, mutation, and

methylation datasets (each labeled according to PFS-based binary recurrence status) previously described were inputted into STREAMLINE separately. No feature selection was done as STREAMLINE contains an autzomated feature selection step in its functionality.

While this benchmarking step provided insight into the predictive capacity of each data type, none of the modalities alone yielded sufficiently strong performance (Supplementary Figure S2). These findings informed our decision to merge the omics datasets into a combined multi-omics dataset to improve modeling performance, as described in the following section.

2.1.5.2. Construction of the binary combined dataset

The formation of the combined dataset integrated molecular and clinical data, ensuring the categorical clinical variables were numerically encoded for downstream machine learning tasks. This process was done through merging, cleaning, encoding, and exporting the dataset systematically.

The process started with the merging of molecular datasets. mRNA expression, mutation count, and methylation datasets were loaded. mRNA gene columns were renamed to include the suffix "_mRNA", for clarity, while the column "PFS_STATUS" was dropped from the mutation and methylation datasets since it was redundant. Then, an outer join on the column "SAMPLE_ID" was performed on the three molecular datasets to retain all available samples across the datasets. Next, clinical features were added using an inner join on SAMPLE_ID to ensure that the molecular and clinical data are aligned. Missing values were removed from the

resultant fully combined dataset, and the index was standardized to SAMPLE_ID. The resultant table consisted of 356 samples and 57,128 features (Figure 5).

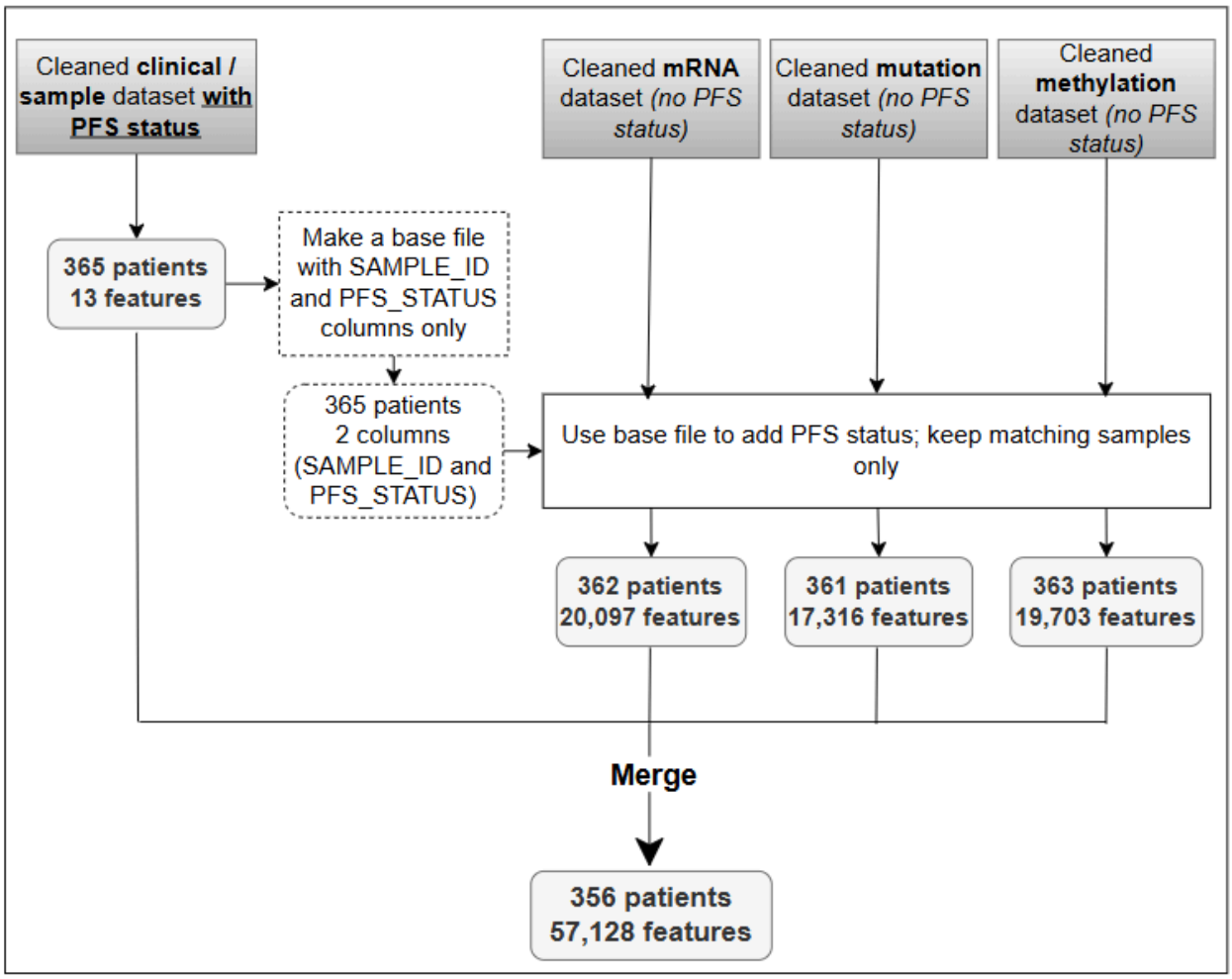


Figure 5. Integration of clinical and multi-omics data into a unified binary LUAD dataset for downstream feature selection and analysis. Clinical, mRNA expression, mutation count, and methylation datasets were filtered and cleaned individually. A core file containing only SAMPLE_ID and PFS_STATUS was taken from the clinical dataset and was used to align molecular data with PFS information. The datasets were then merged on SAMPLE_ID, which resulted in a combined dataset of 356 patients and 57,128 features. This binary-labeled dataset (based on PFS status) was subsequently used as input for further cleaning and feature selection before ML/DL analysis.

For the ML model, the categorical clinical variables were numerically encoded using label encoding. The encoding mappings for each clinical variable are as follows:

1. GENETIC_ANCESTRY_LABEL: Categories were encoded as AFR = 0, AFR_ADMIX = 1, EAS = 2, and EUR = 3.
2. PATH_N_STAGE: Tumor node stages were encoded as N0 = 0, N1 = 1, N2 = 2, N3 = 3, and NX = 4.
3. PATH_M_STAGE: Metastatic stages were encoded as M0 = 0, M1 = 1, M1A = 2, M1B = 3, and MX = 4.
4. PATH_T_STAGE: Tumor size and invasiveness stages were encoded as T1 = 0, T1A = 1, T1B = 2, T2 = 3, T2A = 4, T2B = 5, T3 = 6, T4 = 7, and TX = 8.
5. AJCC_PATHOLOGIC_TUMOR_STAGE: Tumor stages were encoded as STAGE I = 0, STAGE IA = 1, STAGE IB = 2, STAGE II = 3, STAGE IIA = 4, STAGE IIB = 5, STAGE III = 6, STAGE IIIA = 7, STAGE IIIB = 8, and STAGE IV = 9.
6. SEX: The binary gender variable was encoded as Female = 0 and Male = 1.
7. TUMOR_TYPE: The tumor subtypes were numerically encoded with Lung Adenocarcinoma (NOS) = 0, Lung Adenocarcinoma Mixed Subtype = 1, Mucinous Adenocarcinoma = 2, Lung Bronchioloalveolar Carcinoma, Mucinous = 3, and so on for all the remaining few rare types.

Once encoding was completed, the original categorical columns were dropped, leaving only their encoded counterparts. Clinical column names were standardized by appending

_clinical to their names for consistency. For example, "AGE" became "AGE_clinical", and "SEX_ENCODED" became "SEX_ENCODED_clinical".

This dataset was completed by removing the samples when PFS_STATUS equaled 0 and the time in DAYS_LAST_FOLLOWUP was less than six months. This was because LRR after surgery for NSCLC occurs in two main peaks, the first at 6–8 months and the second at 22–24 months (37). Therefore, if DAYS_LAST_FOLLOWUP was less than 6 months, then we did not have enough information to accurately know the patient's actual recurrence status. This left a total of 318 samples and 57,128 features, which formed an all-encompassing dataset with binary recurrence information. We saved it in CSV and pickle format, then further stored the individual data along with the PFS_STATUS data (clinical and PFS_STATUS, mRNA and PFS_STATUS, mutation and PFS_STATUS, and methylation and PFS_STATUS) (Figure 6).

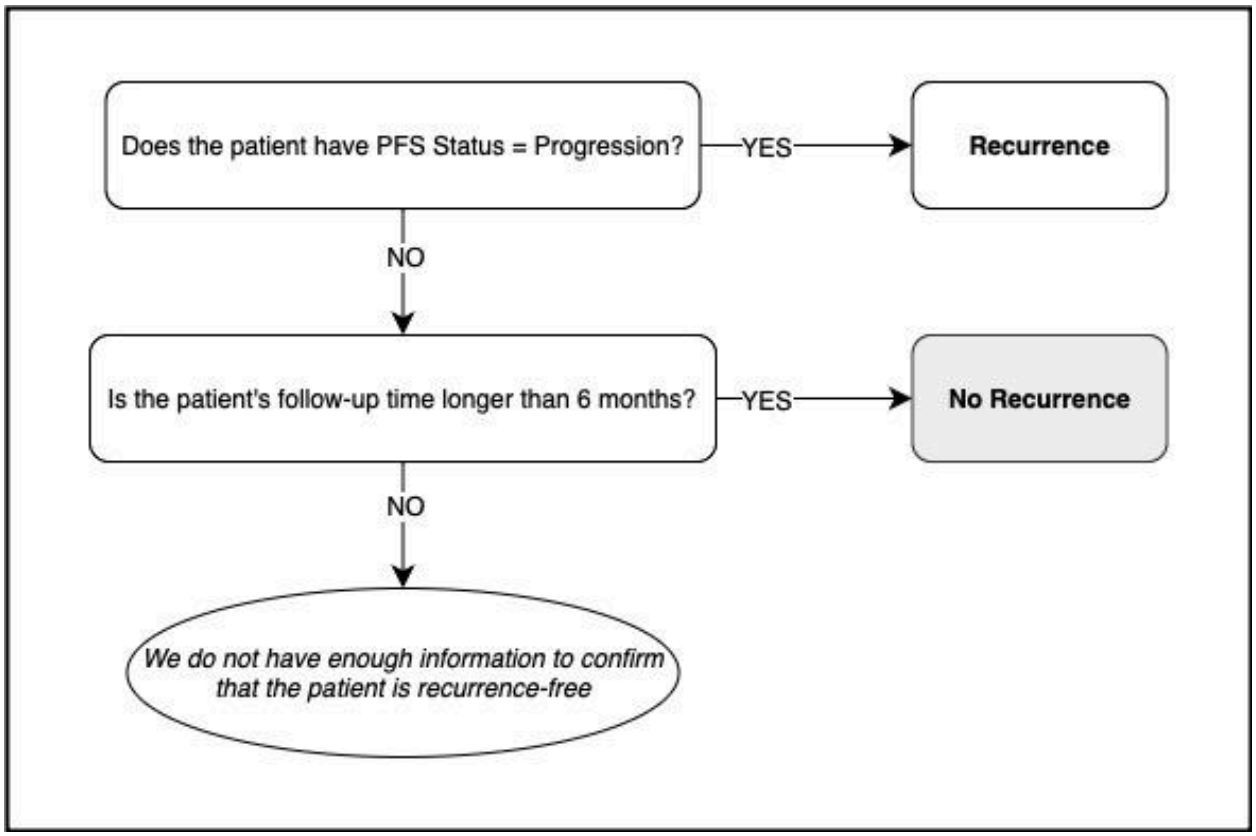


Figure 6. Refining the recurrence status of the patients. Patients with PFS_STATUS = 1 (indicating recurrence) were retained automatically. However, patients with PFS_STATUS = 0 and fewer than six months of follow-up (DAYS_LAST_FOLLOWUP) were dropped as we did not have enough information to confirm their recurrence-free status. The resulting binary-labeled dataset consisted of 318 patients and 57,128 features, meaning 36 patients were dropped.

The median and average PFS_MONTHS_clinical of the patients that had recurrent LUAD were 14.50 months and 16.45 months respectively (with a standard deviation of 12.27 months for the average recurrence time). These measures give a baseline understanding of the distribution of recurrence times present in the dataset.

2.1.5.3. Time-based combined dataset

In addition to looking at recurrence as a binary event, we further refined our dataset to show its time-sensitive nature. For NSCLC, post-surgery recurrence is higher during the first two

years after surgery, with many occuring within the first year (38). Moreover, Yun et. al showed that for NSCLC patients, the peak of the recurrence hazard was at about 12 months post-surgery, regardless of cancer stage (39). Therefore, we utilized a 12-month PFS cutoff to dichotomize patients into early and late recurrence groups.

The "PFS_MONTHS_clinical" was labeled as follows: patients with "PFS_MONTHS_clinical" less than 12 months were labeled as early recurrence (PFS_STATUS = 1), and those with "PFS_MONTHS_clinical" greater than or equal to 12 months were labeled as late recurrence (PFS_STATUS = 2). To focus only on distinguishing between early and late recurrence for model training and analysis, patients with no recurrence (PFS_STATUS = 0) were excluded from the dataset. This resulted in a transformed dataset intended for time-based recurrence prediction, retaining only patients with early/late recurrence events (Figure 7).

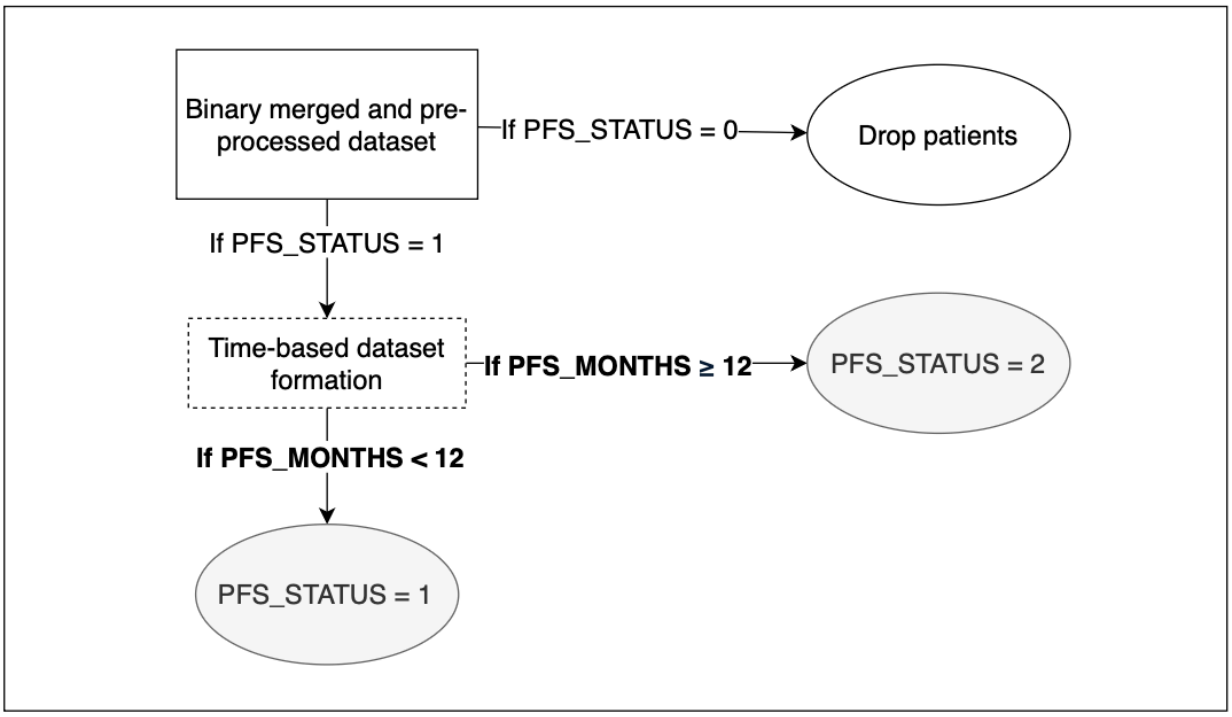


Figure 7. Time-based stratification of LUAD recurrence events using a 12-month PFS threshold. To analyze the recurrence time in LUAD, we only included the patients with disease progression (PFS_STATUS = 1), excluding non-recurrent cases (PFS_STATUS = 0) from the data. We then applied a 12-month progression-free survival (PFS) cutoff to split recurrent cases into two groups. Individuals who had PFS_MONTHS < 12 were labeled as early recurrence (PFS_STATUS = 1), and individuals who had PFS_MONTHS ≥ 12 were labeled as late recurrence (PFS_STATUS = 2). This refined, recurrence-only data was used to examine time-dependent recurrence in further downstream analyses.

In total, there were 115 patients with LUAD recurrence, of which 46 had early recurrence (within 12 months) and 69 had late recurrence (after 12 months). Subsequently, the "PFS_MONTHS" and "DAYS_LAST_FOLLOWUP" columns were dropped. This represented the final processed dataset containing early and late recurrence labels, along with all the individual feature subsets—clinical, mRNA, mutation, and methylation—saved in both csv and pickle formats to facilitate downstream feature selection and data processing.

2.1.6. Exclusion criteria

As noted earlier, in generating the binary and time-based datasets, specific criteria were placed for the exclusion of redundant, irrelevant, and poor-quality data in order to ensure the quality of the final analysis. The exclusions are detailed in Table 2).

Table 2. Exclusion criteria and reason for exclusion

Criterion	Reason for Exclusion
Columns with static / constant values	Did not provide informative variance
Columns with entirely missing values	Unusable columns as they lacked data
Patients with fewer than six non-null attributes	Too many missing values would lead to insufficient data for meaningful analysis
Patients missing PFS status	PFS status was our target variable and could not be missing
Samples without corresponding patient IDs in the clinical dataset	Ensured consistency between the clinical and molecular data
Genes missing Hugo symbol identifiers	Ensure gene recognition
Mutation entries missing key attributes (Hugo symbol, Entrez gene ID, Consequence, Variant classification, and Variant type)	Ensured all retained mutations had complete and valid information
Synonymous and RNA mutations	Kept only non-synonymous mutations
Methylation dataset columns without gene name or methylation beta values	Non-informative features and unrelated to methylation analysis
Samples missing methylation beta values	Ensured the integrity of feature-complete data
Patients with PFS_STATUS = 0 and DAYS_LAST_FOLLOWUP < 6 months	Lacked sufficient follow-up time to confirm recurrence

These exclusion criteria ensured that the final dataset was clean and contained high-quality, well annotated clinical and molecular data, facilitating robust ML/DL modeling, thereby leading to more accurate and precise results.

2.2. Data preprocessing

A custom feature elimination method was applied to reduce multicollinearity in the dataset and thus to make it more interpretable by finding and eliminating features highly correlated with each other. A threshold for the correlation was set to 0.95, such that every pair of features exhibiting a Pearson correlation coefficient above 0.95 (or below -0.95) was deemed redundant. In case of each pair of features that were correlated, one of the features was dropped based on its prevalence in the specified feature categories: clinical, expression, mutation, and methylation columns. It was supported by the Collinear Analysis class, whereby pairwise correlations were analyzed and collinear features were detected. The redundant features were then systematically dropped from the dataset to ensure that the final feature set retained the most important variables while dropping the ones that carried redundant information.

In the binary dataset, after data pre-processing, there were 318 patients remaining, for which the 10 clinical columns (tobacco_smoking_history, TUMOR_TYPE, TMB_NONSYNONYMOUS, AGE, SEX, AJCC_PATHOLOGIC_TUMOR_STAGE, PATH_M_STAGE, PATH_N_STAGE, PATH_T_STAGE, and GENETIC_ANCESTRY_LABEL) were kept the same. For the mRNA expression columns, 43 constant and 28 correlated columns were dropped, leaving 20,026 features. For the mutation counts, 3194 constant and 252 correlated columns were dropped, leaving 13,870 remaining

mutation features. Finally, for the methylation beta values, 0 constant and 15 correlated columns were dropped, keeping 12,290 remaining columns. This combined binary dataset then went through feature selection.

In the time-based dataset, the same 10 clinical columns were kept. After pre-processing, for the mRNA expression columns, 181 constant and 35 correlated columns were dropped, leaving 19,881 features. For the mutation counts, 8,241 constant and 919 correlated columns were dropped, leaving 8,156 remaining mutation features. Finally, for the methylation beta values, 0 constant and 16 correlated columns were dropped, keeping 12,289 remaining columns. We then applied feature selection to this time-based dataset.

2.3. Feature selection

Feature selection was performed in a structured manner to select the most relevant features across expression (ecols), mutation (mcols), methylation (metcols), and clinical (ccols) data categories. This would serve to reduce dimensionality while retaining important predictive variables for machine learning activities thereafter. Both univariate and model-based techniques were applied using well-known tools and classes from the scikit-learn library and the XGBoost framework.

Univariate feature selection was carried out using Analysis of Variance (ANOVA) F-score via the `SelectPercentile` method from scikit-learn. The features were ranked based on their statistical significance in terms of the F-statistic that measures the variance between groups defined by the target label (PFS_STATUS). Iteratively, the top 5%, 10%, 20%, and 30% of

features for each modality were chosen to determine those with the strongest association with the outcome.

Model-based feature selection was performed utilizing the “SelectFromModel” method from scikit-learn, which extracts important features based on weights or feature importance scores obtained from machine learning models. The feature ranking was obtained using five different models, namely, Logistic Regression (LR), Linear Support Vector Classifier (LinearSVC), eXtreme Gradient Boosting (XGBoost), ExtraTreesClassifier (ET), and RandomForestClassifier (RF).

Hyperparameter tuning was done using “RandomizedSearchCV” from scikit-learn in order to optimize model performance. The following hyperparameters were sampled during optimization:

- Logistic Regression (LR): Penalty (l1, l2), regularization strength (C: [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000]), solver (liblinear, saga), and maximum iterations (max_iter=3000).
- LinearSVC: C=[0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000], loss function of hinge or squared_hinge and tolerance by default 1e-5 (tol=1e-5).
- XGBoost (XGBClassifier) - number of estimators (n_estimators: [10, 100, 300, 500, 1000]), learning rate (learning_rate: [0.0001, 0.001, 0.01, 0.1, 0.2]), maximum tree depth (max_depth: [1, 5, 15, 21, 50, 100]) and gamma for regularization (gamma: [0, 0.1, 0.2, 0.3]).

- ExtraTreesClassifier (ET): Number of estimators (n_estimators: [10, 100, 300, 500, 1000]), maximum depth (max_depth: [1, 5, 15, 21, 50, 100, None]) and criterion (criterion: entropy).
- RandomForestClassifier (RF): Number of estimators (n_estimators: [10, 100, 300, 500, 1000]), maximum depth (max_depth: [1, 5, 15, 21, 50, 100]) and maximum features (max_features: sqrt, log2, None).

The feature selection was then performed through a dedicated function called "safe_feature_selection", which was used to perform feature selection and store the results in JSON files for reproducibility. This function could also load the selected features from the JSON files. This modular approach was set to ensure that selected features were extracted consistently across all modalities, while always keeping (and not selecting) the full set of clinical features (ccols).

2.4. DL model building and evaluating

The DL model was constructed using the TensorFlow/Keras framework, tailored to predict binary and time-based recurrence outcomes. The architecture and training process were systematically designed to balance simplicity and flexibility while ensuring optimal performance.

The model utilized the Keras Sequential API, starting with an input layer that matched the number of features in the training data (i.e., 1,636 features for the binary recurrence model and 1,253 for the time-based model). First, the hidden layer was formed by a tunable number of

neurons, referred to as layer1_nodes, with ReLU for activation. In some configurations, additional hidden layers with a specified number of neurons were added (other_nodes). Each hidden layer was followed by batch normalization and dropout layers (0.2 dropout rate) to help prevent overfitting and stabilize training. The final layer contains a single neuron in the output layer with a sigmoid activation function for predicting probabilities on a binary classification problem.

The model also incorporated configurable L1, L2, or elastic net regularization on the dense layers to prevent overfitting. Several configurations of regularization strength were tested, including L1 (e.g., alpha: 0.01), L2 (e.g., value: 0.05), and elastic net combinations (e.g., alpha: 0.005, l1_ratio: 0.3, among others). These were sampled to strike a balance between model complexity and generalizability. Learning rates were sampled across a range of values — learning_rate_configs: [0.001, 0.005, 0.008, 0.01, 0.015, 0.02] — to identify the best-performing rate.

The model was then compiled with the Adam optimizer, initialized with the sampled learning rate from each iteration. Binary cross-entropy was set as the loss function. Metrics for tracking were set to AUC and accuracy during training. Early stopping and ReduceLROnPlateau callbacks were utilized on the validation loss and validation AUC, respectively, so as not to let the model overfit and to converge as efficiently as possible. A custom AUROCCallback was implemented to save model weights corresponding to the best validation AUC during training.

The training was performed on a stratified dataset split into training, validation, and holdout sets. In the stratified split, 67% was partitioned to the training set, 18% to the validation set, and 15% to the holdout set. The model was trained with a batch size of 64 for 100 epochs, though this is often stopped much sooner because of early stopping. The holdout set was used only in the evaluation of the final model. The model architecture and the weights at the end of training were saved in a folder, model_architecture_path, for reproducibility and downstream uses. Saving the training configuration includes the optimizer, learning rate, regularization type, batch size, and number of epochs to JSON files, training_config_path, for traceability.

The final Keras models (one for binary recurrence, one for time-based recurrence) are the ones that performed best on the holdout set. They will be described in the results section.

We then evaluated each Keras model against traditional ML models in terms of AUPRC, sensitivity, specificity, accuracy, precision, recall, and F1 score. We used the same benchmark ML models as Yilmaz et. al (40), which were: i) Logistic Regression (LR) (41), ii) Support Vector Machine (SVM) (42), iii) Decision Tree (DT) (43), iv) Naive Bayes (NB) (44), v) Random Forest (RF) (45), vi) XGBoost (XGB) (46), vii) K Neighbors Classifier (KNN) (47). Next, we plotted the AUROC and AUPRC curves of each model tested and generated a SHapley Additive exPlanations (SHAP) summary plot to find the most important features for each of the two Keras models. The overview of our experimental approach is outlined in Figure 8.

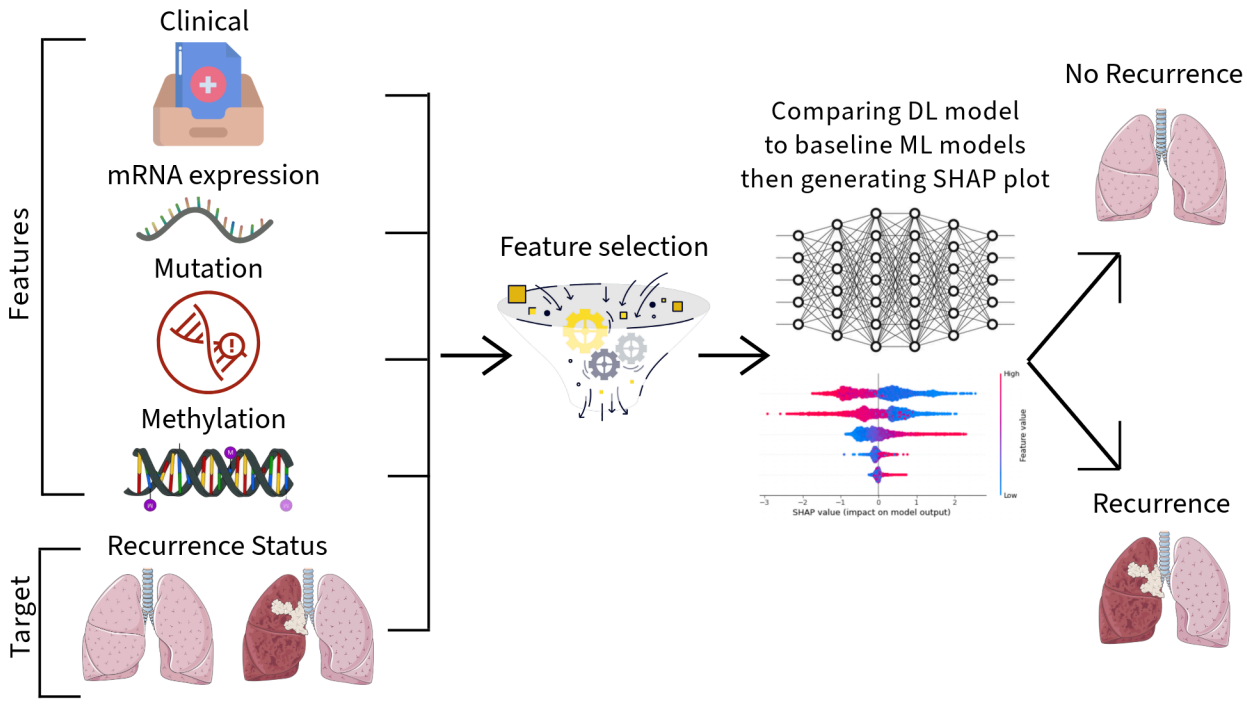


Figure 8. Overview of the experimental workflow for LUAD recurrence prediction. The figure illustrates the end-to-end modeling pipeline, from inputted combined clinical and multi-omics features (mRNA expression, mutation, and methylation) to recurrence status prediction. Feature selection was first performed for reducing dimensions. Training of the deep learning model and comparison with baseline machine learning models were then performed, ultimately classifying patients into no recurrence and recurrence groups. SHAP analysis explained the contribution of features to model predictions.

Chapter 3: Results

3.1. Binary prediction of LUAD recurrence

3.1.1. Binary feature selection results

In the binary recurrence model, feature selection was performed to identify the most relevant predictors across clinical, expression (ecols), mutation (mutcols), and methylation (metcols) data types. For the 318 patients studied (203 without recurrence, 115 with recurrence), clinical features were retained as such, making a total of 10 features. XGBoost was chosen as the feature selection modality for mRNA expression data, retaining 315 features. The mutation data feature selection was done through ANOVA with a threshold of 5%, and it returned 696 features. In a similar vein, methylation data were filtered through ANOVA at a threshold of 5% and returned 615 features. This procedure produced a final dataset containing a total of 1,636 features for the binary recurrence model. A complete list of the selected features for each modality is included in Appendix A. The full cohort description for the binary recurrence model can be found in Table 3.

Table 3. Binary dataset patient characteristics and clinical features

Metric	No Recurrence	Recurrence	Total
Number of Patients	203	115	318
Sex Distribution			
- Males (%)	44.8% (91)	47.8% (55)	45.9%
- Females (%)	55.2% (112)	52.2% (60)	54.1%
Age (Mean ± SD)	64.9 ± 10.2	65.4 ± 9.1	65.1 ± 9.9
Age (Median, in years)	65	66	65.5
Genetic Ancestry			
- European (%)	86.2% (175)	86.1% (99)	86.16%
- African (%)	8.9% (18)	6.1% (7)	7.86%
- Admixed African (%)	3.4% (7)	4.3% (5)	3.77%
- East Asian (%)	1.5% (3)	3.5% (4)	2.20%
Tumor Type (Most Frequent)	LUAD (NOS): 113	LUAD (NOS): 69	LUAD (NOS): 182
Tumor Stage			
- Stage 1 (%) (0, 1, 2 encoded)	69.0% (140)	53.0% (61)	63.2% (201)
- Stage 2 (%) (3, 4 encoded)	17.2% (35)	27.8% (32)	21.1% (67)
- Stage 3 (%) (5, 6 encoded)	10.8% (22)	13.0% (15)	11.6% (37)
- Stage 4 (%) (7 encoded)	3.0% (6)	6.1% (7)	4.1% (13)
Average Expression Z-Scores	-0.065	-0.074	-0.0695
Average Methylation Beta Values	0.34	0.34	0.34
Median Methylation Beta Value	0.21	0.20	0.205

3.1.2. Model structure and performance

For the binary recurrence model, the best Keras configuration was with 64 neurons in layer1_nodes, then 32 neurons in the rest (other_nodes). It also had l2 regularization at 0.05 and a learning rate of 0.005. We benchmarked the Keras model with the traditional ML models for LUAD recurrence prediction. Keras yielded the highest AUROC (0.90, p=0.011, DeLong's test for comparing the areas under two or more correlated AUROC curves (48)), which was significantly better than Logistic Regression (0.76), Support Vector Machine (0.72), Naïve Bayes

(0.69), Random Forest (0.64), XGBoost (0.59), Decision Tree (0.50), and K-Nearest Neighbors (0.48) (Figure 9). Similarly, in Precision-Recall, Keras scored the best AUPRC (0.86, p=0.022, bootstrap test (49)), followed by Logistic Regression (0.73), Naïve Bayes (0.68), SVM (0.57), XGBoost (0.55), Random Forest (0.43), Decision Tree (0.49), and KNN (0.31) (Figure 10). In terms of overall classification performance, looking at the other metrics evaluated enables us to have a better understanding of our DL model's performance. On one hand, the Keras model achieved an accuracy of 0.7708, showing good predictive capability. It also displayed perfect specificity (1.000) and precision (1.000), without any false positives. On the other hand, the model recorded a substantially low recall of 0.3529, which points to its failure to accurately detect a considerable portion of positive cases, a deficiency also shown in its lower F1 score of 0.5217 (Table 4).

Despite its class discrimination ability (AUROC: 0.90) and good handling of precision-recall trade-offs (AUPRC: 0.86), the Keras model's poor recall performance (0.3529) raises concern about its usefulness in the clinic. Although its perfect specificity ensures that non-recurrent cases are not misdiagnosed as recurrent, its high rate of false negatives (missed recurrence cases) is a major limitation in a clinical setting where the detection of recurrence is critical. The low F1 score of 0.5217 confirms that although the model has high precision, it is not dependable in correctly predicting positive cases (Table 4).

The performance of the Keras model demonstrates that recurrence prediction in LUAD is enhanced by deep learning's capacity to detect intricate patterns, yet recall gains are necessary prior to it being suitable for clinical decision-making. To negate these constraints, we need to

address hyperparameter optimization and/or perform class weight adjustment, all while preserving high specificity and accuracy (50).

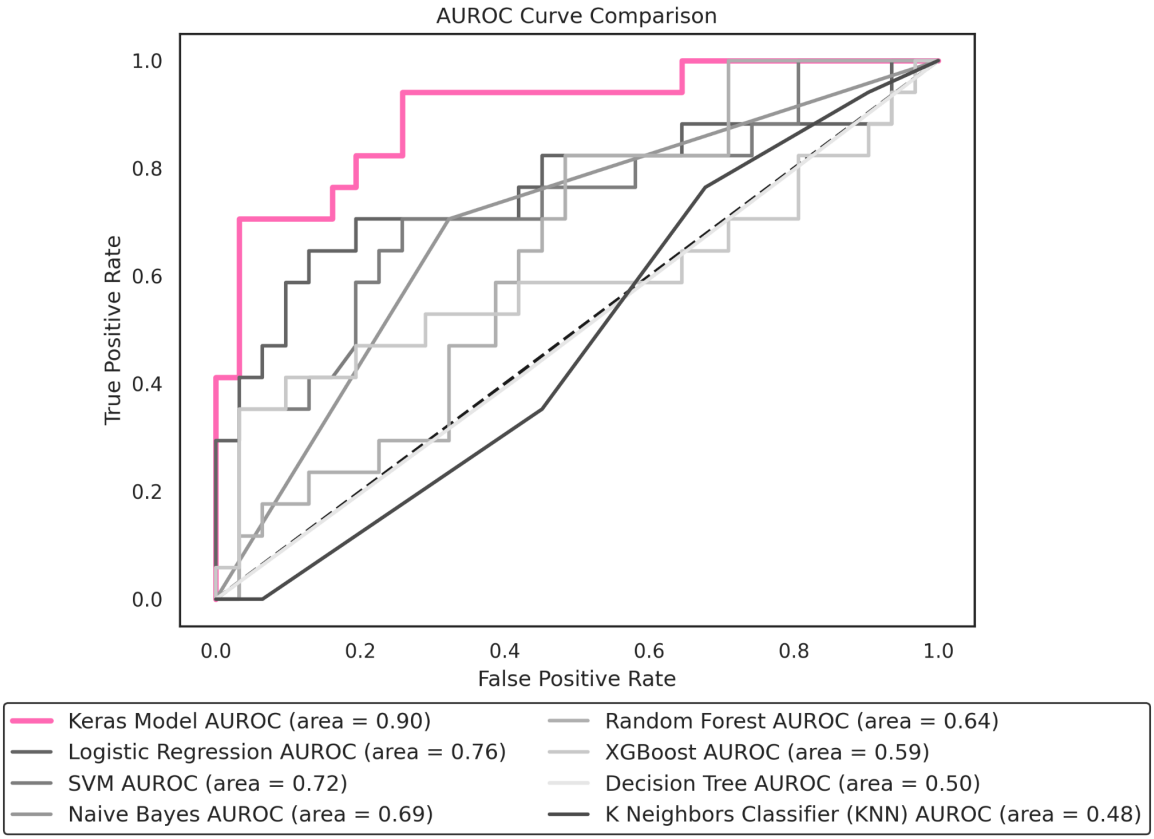


Figure 9. Comparison of AUROC of Keras with baseline ML models for predicting binary recurrence of LUAD. The Keras DL model outperformed all the traditional ML baselines (p = 0.001, DeLong's Test). Comparison AUROC values were as follows: Keras (0.90), Logistic Regression (0.76), Support Vector Machine (0.72), Naïve Bayes (0.69), Random Forest (0.64), XGBoost (0.59), Decision Tree (0.50), and K-Nearest Neighbors (0.48).

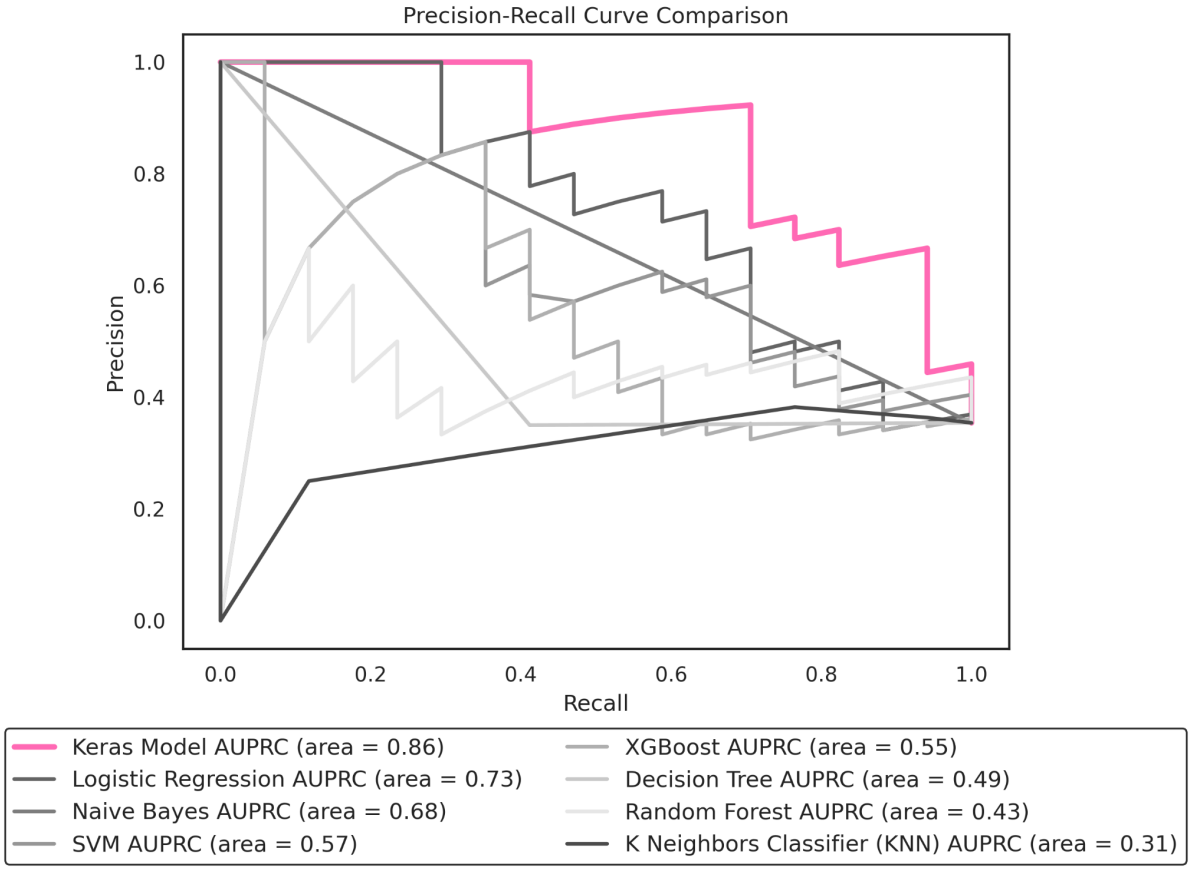


Figure 10. Comparison of AUPRC of Keras with baseline ML models for predicting binary recurrence of LUAD. The Keras DL model outperformed all the traditional ML baselines (p = 0.022, bootstrap test). Comparison AUROC values were as follows: Keras (0.86), Logistic Regression (0.73), Naïve Bayes (0.68), Support Vector Machine (0.57), XGBoost (0.55), Random Forest (0.43), Decision Tree (0.49), and K-Nearest Neighbors (0.31).

Table 4. Performance comparison of Keras and ML models on the binary combined dataset

Model	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
Keras	0.898	0.867	0.771	0.353	1.000	1.000	0.353	0.522
Logistic Regression	0.763	0.735	0.792	0.647	0.871	0.788	0.792	0.788
SVM	0.720	0.566	0.646	0.000	1.000	0.417	0.646	0.507
Naive Bayes	0.692	0.678	0.688	0.706	0.6777	0.715	0.688	0.694
Random Forest	0.636	0.431	0.646	0.118	0.935	0.603	0.646	0.567
XGBoost	0.592	0.564	0.729	0.353	0.935	0.734	0.729	0.698
Decision Tree	0.496	0.485	0.521	0.412	0.581	0.539	0.521	0.528
KNN	0.481	0.307	0.604	0.000	0.935	0.407	0.604	0.486

3.1.3. Feature importance

The SHAP summary and bar plot for feature importance in binary recurrence (Figure 11) showed Tumor Mutational Burden as the most influential feature, showing both positive and negative impacts on model predictions. Tumor Mutational Burden (TMB) has emerged as one of the crucial biomarkers in LUAD, as it clearly demonstrates a connection with outcomes of treatment and recurrence. In this case, TMB may be acting as a a proxy for smoking exposure, particularly in absence of pack-years data. In advanced lung adenocarcinoma, doubling smoking pack-years was associated with an increase in TMB, whereas doubling smoking-free months was associated with a decrease in TMB, after controlling for age, gender, and stage. (51,52).

CYP19A1 demonstrated significant associations with the development and progression of LUAD through multiple mechanisms. The gene encodes aromatase, a key enzyme in estrogen biosynthesis, and its polymorphisms can result in either increased or decreased aromatase activity, affecting LUAD risk (53). The SHAP summary plot (Figure 11) indicates that *CYP19A1* is the second most significant factor in forecasting binary LUAD recurrence, with its polymorphisms notably correlated with elevated lung cancer risk and *EGFR* mutations, particularly among female individuals who have never smoked (53). Specifically, the *CYP19A1* rs3764221 SNP has been prominently associated with the multicentric development of LUADs, where carriers of the less common (minor) allele had an increased risk of multicentric LUAD (OR = 1.72, P = 0.017). In addition, this variant was also associated with an increased risk of LUAD that presents in combination with atypical adenomatous hyperplasia (AAH) (OR = 1.74, P = 0.029) and components of bronchioloalveolar carcinoma (BAC) (OR = 1.41, P = 0.091). The underlying mechanism appears to involve differences in estrogen levels, as the variant was associated with higher circulating estradiol levels in postmenopausal women (54). Other variants of *CYP19A1* include rs28757157, which has been identified as a risk factor in lung cancer development, especially in specific subpopulations like females and smokers (OR = 1.30, 95% CI 1.03-1.64) (55). Although there are few direct studies on the role of *CYP19A1* in LUAD recurrence, its protein expression has been found to be positively correlated with PD-L1 expression in cancer tissues. Moreover, its inhibition downregulates PD-L1, IL-6, and TGF-β levels through GPR30-AKT signaling and hence may be involved in immune response and disease progression (56).

C1orf200 (alias for *PIK3CD-AS1*), ranked third in the SHAP summary plot, and is an antisense RNA of the *PIK3CD* gene. Even though there is a lack of direct research about *PIK3CD-AS1* in LUAD, its associated gene *PIK3CD* has been identified as a key factor in the development of LUAD through the *PI3K-AKT-mTOR* pathway. *PIK3CD* encodes the p110δ catalytic subunit of *PI3K* and was reported to be downregulated in LUAD tissues compared to normal samples (57). Interestingly, *PIK3CD-AS2*, another antisense RNA of *PIK3CD*, has been extensively studied in LUAD and shown to promote tumor progression. *PIK3CD-AS2* was overexpressed in LUAD tissues and associated with bigger tumor size, poor histological differentiation, and shorter overall survival (58,59). Mechanistically, *PIK3CD-AS2* exerts its function to promote cell proliferation and suppress apoptosis in p53 wild-type cells by binding to and stabilizing *Y-box binding protein 1 (YBX1)*, thereby repressing the p53 pathway (59). This *PIK3CD-AS2/YBX1/p5*3 signaling axis has been suggested as a possible therapeutic target, particularly for p53 wild-type patients with LUAD (59). Still, further research is warranted to pinpoint its accurate role and possible usefulness as a prognostic or therapeutic target in LUAD.

The SHAP summary plot shows that *Angiotensin-Converting Enzyme (ACE)* is the fourth most significant estimator for LUAD recurrence. *ACE* plays a central role in the renin-angiotensin system (RAS) since it takes part in the conversion of angiotensin I to angiotensin II — a peptide with pleiotropic activities that could induce proliferation in cancer cells. Quantitative measurements demonstrate that *ACE* activity in lung cancer tissues is approximately threefold lower (9.4 ± 3.2 mU/mg) compared with morphologically normal lung tissues (28.6 ± 14.6 mU/mg) (60). Interestingly, even though *ACE* levels are generally decreased in lung cancer tissues, increased angiotensin II production has been found in non-small cell lung

cancer tissues, which suggests a complex regulatory mechanism (61). Moreover, the role of *ACE* in LUAD has clinical implications, as *ACE* inhibitor administration may affect survival outcomes, though results in studies conducted regarding cancer progression were not always consonant, showing inconsistency in their results (61).

Finally, *SLC4A8* (Sodium Bicarbonate Cotransporter Family Member 8) is ranked fifth among the most influential features in LUAD recurrence prediction by SHAP summary plot. Even though the gene has been associated with a smoking initiation phenotype (62), its specific role in LUAD recurrence is yet to be determined, making it a novel gene requiring further studies. Tumor M stage does not show up in the top features, probably due to label imbalance and unequal representation of M stage, which was stratified by label.

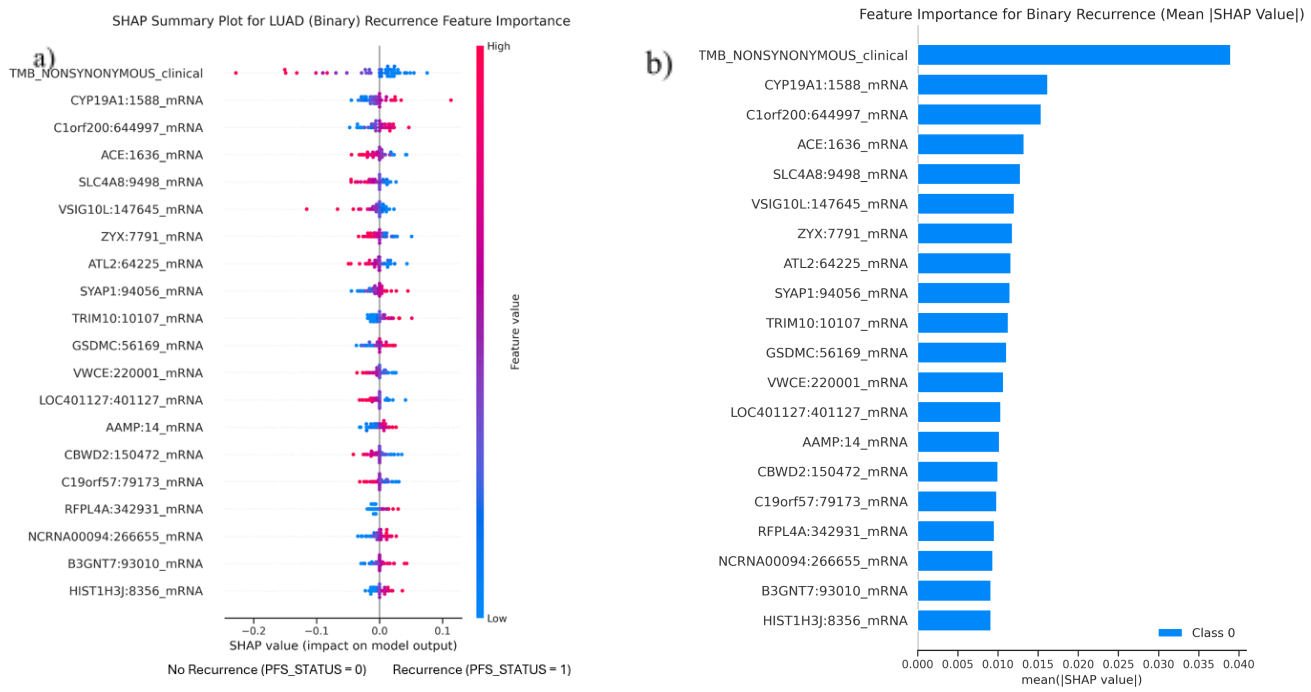


Figure 11. SHAP-based explanation of top features driving binary recurrence prediction. SHAP analysis was utilized to describe the individual feature contributions in the binary recurrence model. (a) SHAP summary plot showing both the direction and size of the effect of each feature on the prediction of the model, where color indicates feature value (low = blue, high = red). Positive SHAP values (right) are associated with higher probability of recurrence (PFS_STATUS = 1), and negative SHAP values (left) with no recurrence (PFS_STATUS = 0). (b) Bar plot ranking the features according to their mean absolute SHAP values, showing those most important in the classification of binary recurrence.

3.2. Time-based prediction of early vs late LUAD recurrence

3.2.1. Time-based feature selection results

Feature selection for the time-based recurrence model was done to capture those features that distinguished early and late recurrence. For the 115 patients with recurrence, clinical features were kept the same, hence contributing 10 features. For mRNA expression data, ANOVA was used at a threshold of 5%, and 994 features were retained. For the feature selection from the mutation data, XGBoost was used and resulted in 59 features. 190 features were selected from the methylation data using XGBoost. Thus, the final dataset for the time-based model had 1,253

features. Detailed lists of selected features for each modality are attached in Appendix B for reference and reproducibility. The full cohort description for the binary recurrence model can be found in Table 5.

Table 5. Time-based dataset patient characteristics and clinical features

Metric	Early Recurrence	Late Recurrence	Total
Number of Patients	46	69	115
Sex Distribution			
- Males (%)	56.5% (26)	42.0% (29)	47.8%
- Females (%)	43.5% (20)	58.0% (40)	52.2%
Age (Mean ± SD)	64.1 ± 9.2	66.2 ± 9.0	65.2 ± 9.1
Age (Median, in years)	66	67	66.5
Genetic Ancestry			
- European (%)	91.4% (42)	82.6% (57)	86.1%
- African (%)	4.3% (2)	7.2% (5)	6.1%
- Admixed African (%)	0.0% (0)	7.2% (5)	4.3%
- East Asian (%)	4.3% (2)	3.0% (2)	3.5%
Tumor Type (Most Frequent)	LUAD (NOS): 29	LUAD (NOS): 24	LUAD (NOS): 53
Tumor Stage			
- Stage 1 (%) (0, 1, 2 encoded)	43.5% (20)	59.4% (41)	53.0% (61)
- Stage 2 (%) (3, 4 encoded)	37.0% (17)	21.7% (15)	27.9% (32)
- Stage 3 (%) (5, 6 encoded)	19.5% (9)	8.8% (6)	13.0% (15)
- Stage 4 (%) (7 encoded)	0.0% (0)	10.1% (7)	6.1% (7)
Average Expression Z-Scores	-0.113	-0.165	-0.139
Average Methylation Beta Values	0.33	0.33	0.33
Median Methylation Beta Value	0.2	0.19	0.195

3.2.2. Model structure and performance

For time-based recurrence, the optimal Keras configuration was 64 nodes in the first layer (layer1_nodes) and 32 nodes in the rest (other_nodes). There was an l2 penalty at 0.008 and a learning rate of 0.001. Keras achieved the best AUROC (0.92, p=0.002, DeLong's test (48)), significantly outperforming Naïve Bayes (0.77), Decision Tree (0.62), Logistic Regression (0.62), SVM (0.57), XGBoost (0.46), Random Forest (0.44), and K-Nearest Neighbors (0.29) (Figure 12). Similarly, in Precision-Recall, Keras performed better with an AUPRC of 0.96 (p=0.04, bootstrap test (49)), followed by XGBoost (0.81), Naïve Bayes (0.80), Random Forest (0.76), Logistic Regression (0.75), SVM (0.72), Decision Tree (0.70), and KNN (0.70) (Figure 13). Considering other performance metrics provides a better understanding of the DL model's strengths and weaknesses. The Keras model has an accuracy of 0.83, precision of 0.786, and 1.000 recall, with an F1-score of 0.880. Although it was perfectly sensitive (1.000), with all positive cases correctly classified, its specificity (0.571) was lower, pointing to a high rate of false positives (Table 6).

Although our Keras model achieved high AUROC and AUPRC, indicating good class discrimination, its low specificity (0.571) indicates a bias for misclassifying negative cases as positive (high false positives). This kind of tradeoff guarantees coverage of all positive cases but is problematic in this case, where the exclusion of false positives is essential. In comparison, Naïve Bayes had better specificity (0.857) and precision (0.889), but was marked by low recall (0.727), with reduced capacity to identify all the positive cases. Logistic Regression and Decision Tree, among others, had moderate AUROC but were not capable of optimizing

precision, recall, and specificity. SVM, although having high sensitivity (0.909), was marked by low specificity (0.143), with an overabundance of false positives.

The performance of the Keras model highlights DL's power in detecting complex patterns in data, but the lower specificity also points to where refining is necessary before deployment in high-stakes decision-making. While the model is good at recall, hyperparameter tuning or class weight adjustment might be necessary in a way that will improve specificity without sacrificing predictive power. This balance will be crucial in fine-tuning the model's reliability for clinical use (63–65)).

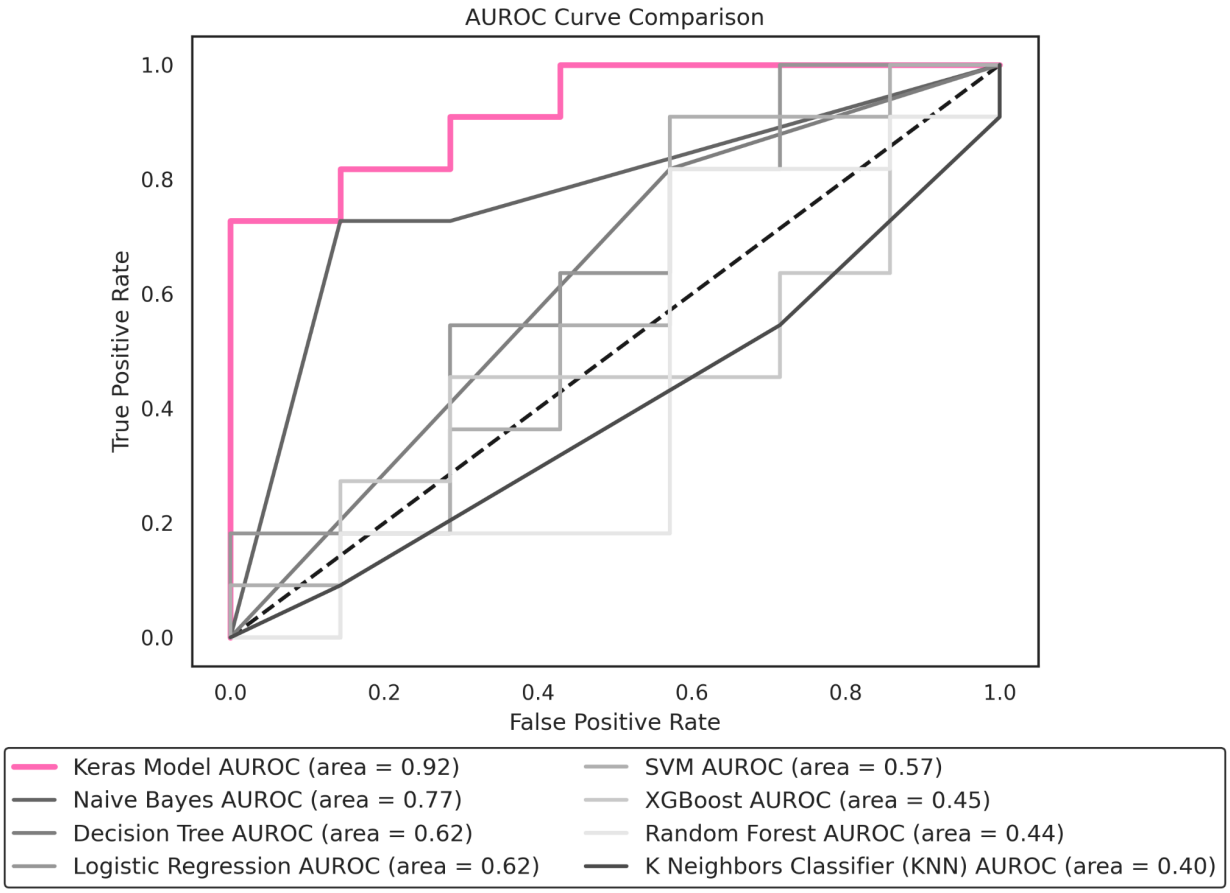


Figure 12. Comparison of AUROC of Keras with baseline ML models for predicting time-based recurrence of LUAD. The Keras DL model outperformed all the traditional ML baselines (p = 0.002, DeLong's Test). Comparison AUROC values were as follows: Keras (0.92), Naïve Bayes (0.77), Decision Tree (0.62), Logistic Regression (0.62), Support Vector Machine (0.57), XGBoost (0.45), Random Forest (0.44), and K-Nearest Neighbors (0.40).

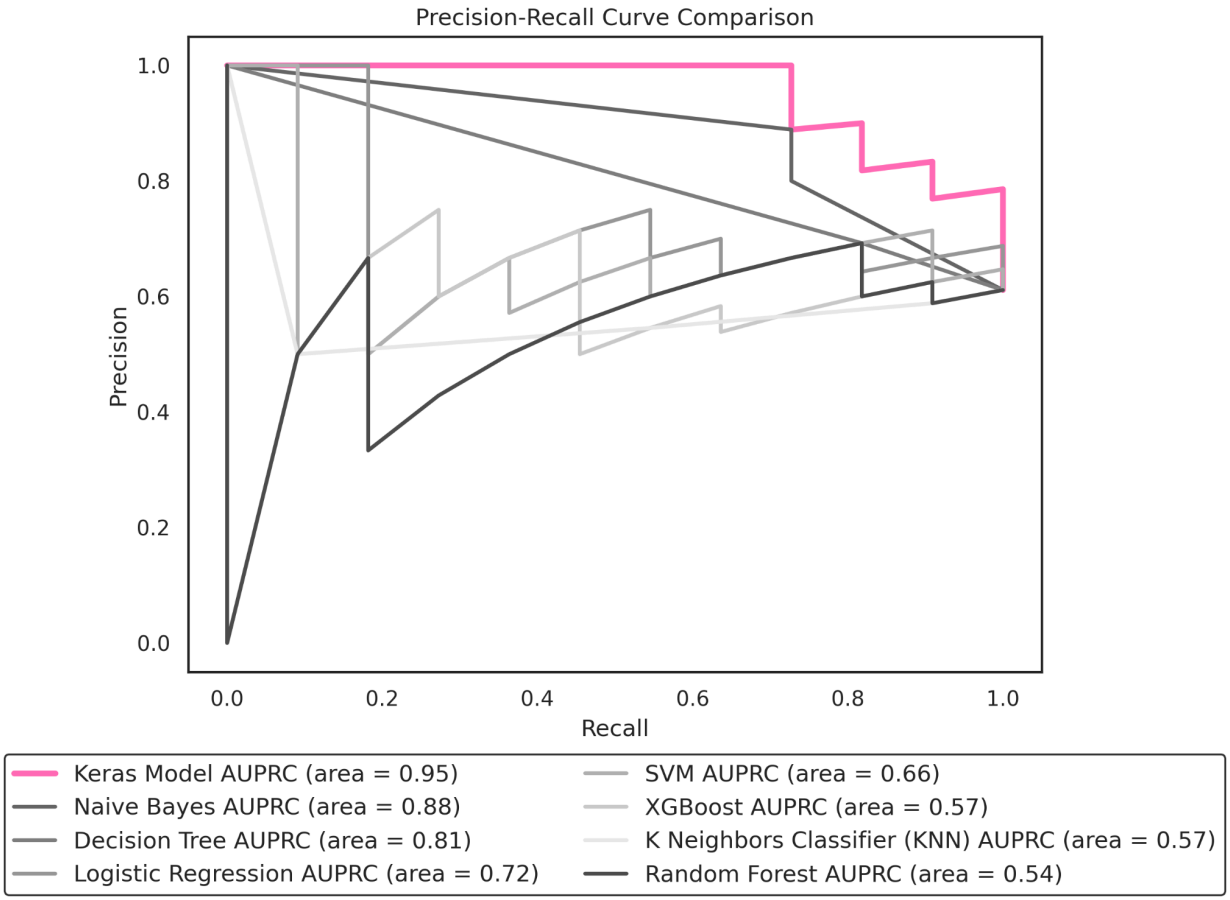


Figure 13. Comparison of AUPRC of Keras with baseline ML models for predicting time-based recurrence of LUAD. The Keras DL model outperformed all the traditional ML baselines (p = 0.04, bootstrap test). Comparison AUPRC values were as follows: Keras (0.95), Naïve Bayes (0.88), Decision Tree (0.81), Logistic Regression (0.72), Support Vector Machine (0.66), XGBoost (0.57), K-Nearest Neighbors (0.57), and Random Forest (0.54).

Table 6. Performance comparison of Keras and ML models on the time-based combined dataset

Model	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
Keras	0.916	0.956	0.833	1.00	0.571	0.786	1.000	0.880
Naive Bayes	0.773	0.879	0.778	0.727	0.857	0.802	0.778	0.781
Decision Tree	0.623	0.811	0.667	0.818	0.429	0.656	0.667	0.653
Logistic Regression	0.623	0.723	0.611	0.818	0.286	0.587	0.611	0.581
SVM	0.571	0.664	0.611	0.909	0.143	0.576	0.611	0.539
XGBoost	0.455	0.573	0.500	0.636	0.286	0.486	0.500	0.492
Random Forest	0.442	0.540	0.556	0.909	0.000	0.359	0.556	0.437
KNN	0.653	0.566	0.444	0.545	0.286	0.444	0.444	0.444

3.2.3. Feature importance

The SHAP summary plot provides the distribution of SHAP values for the most important features and shows their contribution to the model output for Class 0 (early recurrence) (Figure 14). TMB was, just like in the binary recurrence model, the most impactful feature; the wide range of SHAP values showed the dominant role it plays in driving predictions. High TMB values (red points) were strongly associated with positive Class 0 predictions, suggesting that patients with high TMB are more likely to experience early recurrence. Conversely, low TMB values (blue points) contributed negatively to Class 0, which suggests it is more predictive of Class 1, which is late recurrence Clinically, high TMB is associated with increased tumor

immunogenicity and better treatment responses in patients with LUAD, further reinforcing the importance of TMB in the SHAP results (66).

The second most important feature is age. Indeed, age stands as a critical factor in LUAD recurrence and treatment outcomes: elderly patients (≥70 years) experience higher recurrence risks and lower survival rates post-surgery, with 5-year overall survival at 52% compared to 67% in younger patients (67). Also, advanced age significantly influences treatment decisions, particularly regarding recurrence management; patients above 75 years have about half the chance of receiving active treatment compared with those below 55 years (OR 0.49 for locoregional recurrence and OR 0.42 for distant recurrence) (68). This confirms the significance of age on prognosis and treatment selection in LUAD cases, especially for the elderly.

The third most important feature, *AUH* (AU RNA binding methylglutaconyl-CoA hydratase) is an RNA-binding protein (RBPs) that is down-regulated in LUAD with a fold change of 0.829 ($p=1.46\times10^{-3}$) (69). Although some tumor suppressor genes are associated with AUH, its exact mechanism in the recurrence of LUAD is still unclear (69). With these uncertainties, no conclusion can be drawn at this point, and additional studies are needed to determine its role in LUAD recurrence.

Next, further investigation is needed for both *GJB7* and *UOX*, which rank as the fourth and fifth most important features, respectively. While *GJB7* is greatly downregulated in LUAD and belongs to the gap junction beta family of genes (that has been shown to have varying expression patterns in lung cancer), its specific role in LUAD recurrence remains unclear in

current literature (70,71). For *UOX* (Urate Oxidase), despite its appearance as a significant predictive feature in the SHAP analysis, there is limited direct evidence in the available literature explaining its specific mechanism in LUAD recurrence.

Finally, tumor M staging (from the TNM classification) is a key clinical feature and prognosis indicator, as shown on the SHAP plot. Patients in stage I LUAD with distant metastasis have a notably higher mortality rate, with 6.8% dying within 5 years of surgery compared to 2.6% for those with only locoregional recurrence (72). The presence of metastasis is of particular consequence in early-stage patients, where 30-50% experience metastatic recurrence after surgical treatment (73). Early-stage patients developing "blast metastasis" - defined as multiple distant site recurrences within one year of resection - show particularly poor survival outcomes, suggesting complex interactions between tumor immunoevasion and virulence factors rather than surgical failure (74). Consequently, understanding metastasis status is highly critical to treatment planning, risk stratification, and the determination of appropriate follow-up protocols among patients with LUAD.

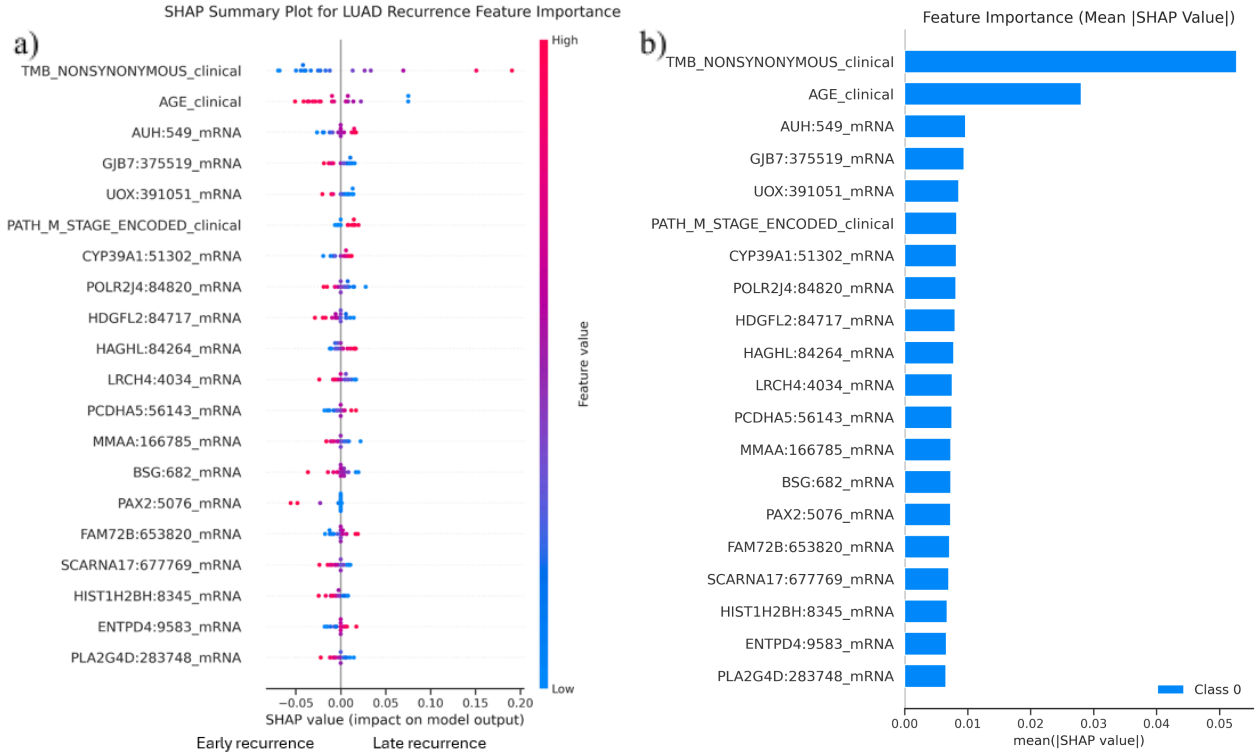


Figure 14. SHAP-based explanation of top features driving time-based recurrence prediction. SHAP analysis was utilized to describe the individual feature contributions in the time-based recurrence model. (a) SHAP summary plot showing both the direction and size of the effect of each feature on the prediction of the model, where color indicates feature value (low = blue, high = red). Positive SHAP values (right) are associated with higher probability of late recurrence, and negative SHAP values (left) with early recurrence. (b) Bar plot ranking the features according to their mean absolute SHAP values, showing those most important in the classification of time-based recurrence.

Chapter 4: Discussion and conclusions

LUAD recurrence remains a significant challenge despite advancements in diagnosis and treatments. This research studies the applications of ML and DL models to predict LUAD binary and time-based recurrence using clinical, genomic data including mRNA z-scores, mutation counts, and methylation data The use of mRNA z-scores enables a comprehensive profiling of gene expression in LUAD samples. Moreover, looking at (non-synonymous) mutation counts helps us understand which genes are mutated, and which of them have an impact on LUAD recurrence. Finally, methylation beta values give an insight on LUAD progression, with studies showing that hypermethylated tumor cells exhibit increased activity related to the hypoxia response, potentially contributing to tumor growth and metastasis (75). A multi-omic approach therefore seemed like the best method to create an all-encompassing dataset that accurately represents LUAD and its progression. However, a challenge of using multi-omics data is that it requires computationally expensive methods and that the high dimensionality of the datasets may lead to overfitting and poor generalizability of predictive models (76). This hurdle may be overcome by using feature selection to identify a subset of highly relevant features to LUAD recurrence.

Moreover, it is crucial to note that cancer recurrence is highly time-dependent. Indeed, some cancers - such as lung cancer - may recur within weeks, whereas others - such as breast cancer - may take years to recur (if they even recur) (77,78). Therefore, we not only looked at recurrence as a binary event, but also as a time-dependent phenomenon with LUAD recurrence

peaking at 12 months (39). This meant that we considered recurrences before the 12 months mark as early recurrences, and those after the 12 months mark as late recurrences.

Both our binary and time-dependent models showed modest prediction of LUAD recurrence, with AUROC scores of 0.90 and 0.92 respectively. However, it is only by looking at the remaining metrics that it is possible to better understand our model's performance and generalizability.

Finally, understanding the role of adjuvant therapy in modifying the risk of recurrence is essential. When it comes to adjuvant chemotherapy, the *LACE* meta-analysis showed that cisplatin-based adjuvant therapy results in a 5.4% 5-year overall survival benefit for patients with completely resected NSCLC, with the greatest benefit for those with stage II–IIIA disease (79). For adjuvant targeted therapy, The ADAURA trial showed that adjuvant osimertinib reduced the risk of recurrence or death by 80% in patients with resected stage IB–IIIA NSCLC with *EGFR* mutations (80). Finally, for adjuvant immunotherapy, the *IMpower010* trial showed that adjuvant atezolizumab after chemotherapy improved disease-free survival in resected stage II–IIIA NSCLC with PD-L1 ≥ 50% (81).

4.1. Performance analysis of the binary Keras model

4.1.1. The binary Keras model correctly identifies all true negatives but misses many true positives

The 0.898 AUROC of our binary Keras model shows that it has good discriminatory power and can reliably differentiate between positive and negative cases in most scenarios.

Similarly, Wang et al. (2023) trained a DL model on histopathological features after surgical resection of LUAD and achieved a similar AUROC of 0.89, showing the strong capability of DL in predicting recurrence (82). Our model's AUPRC of 0.867 indicates that it is able to keep a good balance between precision (correctly identifying positives) and recall (capturing positives). Also, the model's accuracy is 0.771, which means that it is correct 77.1% of the time. However, this can be misleading as the classes are slightly imbalanced (63.8% no recurrence, and 36.2% with recurrence) and the score may be favoring the majority class if our model predicts the majority group (no recurrence) most of the time, even if it misses some key positive cases. This is further demonstrated by the fact that the sensitivity and/or recall score is 0.353 (or 35.3%). This means that the model misses 64.7% of the true positives (TP), which entails a high rate of false negatives (FN), which is problematic. Contrarily, the specificity of 1.000 shows that the model correctly identifies all the true negatives (TN) and that there are no false positives (FP). Additionally, the precision of 1.000 asserts that every positive prediction made by the model is correct, even though it comes at the expense of missing many true positives (as indicated by the low sensitivity). The F1 score of 0.522 reflects the imbalance between the perfect precision and the low recall, highlighting the model's strength in being precise when predicting positives, but struggling to recall or identify a significant portion of the true positives. Therefore, while the model avoids false positives, it needs improvement in identifying more true positives to achieve a better balance.

4.1.2. Suggested improvements to the binary Keras model

In order to enhance the clinical applicability of our model, we may follow the following steps in further studies to optimize model performance. Firstly, we may proceed with a threshold

adjustment. Currently, the model outputs probabilities that reflect its confidence of the positive class prediction using a default threshold of 0.5 for classification. While efficient in many applications, this may not yield an optimal balance between false positives and false negatives. Therefore, a threshold adjustment would offer the model a way to favor recall and identify more positives, which is what is currently lacking. For instance, lowering the threshold (for example, to 0.3) would increase recall by capturing more positives. In the end, the best threshold should be determined empirically, based on the performance metrics tested across various thresholds and saved for later deployment to ensure consistent decision-making. Depending on the results of the threshold readjustment, we may secondly proceed with class imbalance handling through class weighting, which would enhance the model's ability to generalize across both classes, especially the minority class.

4.2. Performance analysis of the time-based Keras model

4.2.1. The time-based Keras model correctly identifies all true positives but lacks specificity, causing some false positives

The 0.916 AUROC of our time-based Keras model also demonstrates strong discriminatory power, indicating that the model can reliably differentiate between positive and negative cases. The AUPRC of 0.956 suggests excellent performance in balancing precision (correctly identifying positives) and recall (capturing as many positives as possible), even in the presence of moderate class imbalance (~40% early and ~60% late recurrence). Also, the model's accuracy of 0.833 indicates that it is correct 83.3% of the time. However, as previously mentioned, this metric alone can be misleading as it may favor the majority class.

On one hand, the sensitivity (or recall) of 1.000 shows that the model successfully identifies all TP cases, with no FNs. On the other hand, the specificity of 0.571 reveals that the model struggles with correctly identifying TNs, leading to a significant number of FPs. This indicates that while the model is highly sensitive to detecting positive cases, it probably does so by predicting 1 the majority of the time, therefore sacrificing the ability to confidently rule out negatives and leading to many FPs. The high rate of FP would be problematic in clinical applications, where a high rate of false positives may lead to unnecessary anxiety for patients, and would overburden healthcare resources with additional diagnostic tests and increased costs. In a more positive light, the precision of 0.786 indicates that 78.6% of the positive predictions are correct and are TPs. The F1 score of 0.880 reflects a strong balance between precision and recall, showcasing the model's overall effectiveness in identifying and predicting positive cases. While the model excels in capturing all TPs and maintains good precision, improving specificity would enhance its performance further by reducing FNs.

4.2.2. Suggested improvements to the time-based Keras model

In a similar way to the optimizations suggested for the binary Keras model, the time-based model may benefit from the following. Firstly, adjusting the threshold may lead to an optimal balance between sensitivity and specificity. The current default threshold of 0.5 may give a high sensitivity but lower specificity. Secondly, class imbalance—about 40% early versus about 60% late recurrence—could be handled with techniques such as class weighting to make the model generalize better over both classes and not ignore the minority class (early recurrence). Finally, refining the feature selection to identify predictors that would further differentiate true negatives from false positives could enhance specificity and result in a more clinically reliable

model. Such modifications would balance the strengths and limitations of the model and therefore make it more useful in clinical practice.

4.3. Strengths and pitfalls of feature selection

The feature selection process that was followed has some considerable advantages: it is a systematic and modular methodology that ensures coherence and reproducibility among all different modalities - expression, mutation, methylation, and clinical information. With the incorporated specific function “safe_feature_selection”, chosen features could be easily preserved and reintroduced for reducing inconsistencies in the analyses to be performed. Through dimensionality reduction, the dataset is simplified from thousands of features to a more tractable subset, thus relieving the curse of dimensionality while retaining important predictive variables. Moreover, the combination of univariate approaches, such as ANOVA, with model-based methods - Logistic Regression, XGBoost, and Random Forest - provides a synergistic viewpoint: ANOVA identifies statistically significant features by looking at group variance, while model-based methodologies use feature importance scores to highlight predictive importance. Furthermore, the use of hyperparameter optimization via RandomizedSearchCV ensures that the models used for feature selection are optimized for maximum performance, thus increasing the reliability of their output. Notably, clinical features were always retained, in recognition of their particular relevance to the domain and to preserve crucial high-level information. Furthermore, the process demonstrated adaptability by tailoring feature selection to different objectives, such as binary recurrence prediction and temporal modeling of recurrence, which exemplifies a design that is adaptable and purpose-driven.

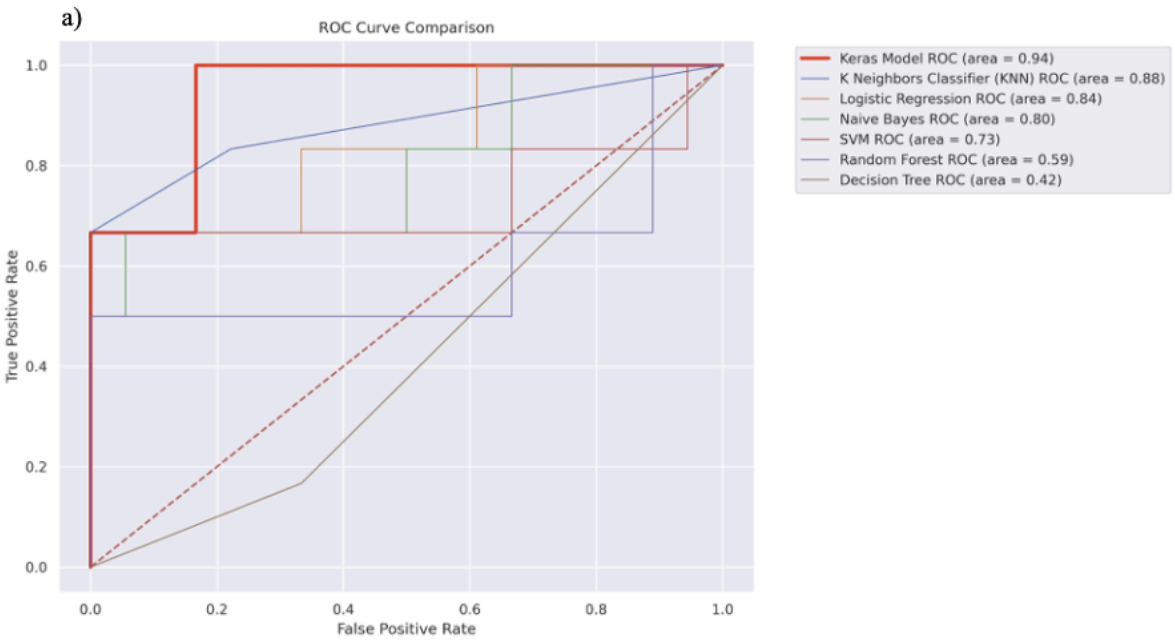
Although there were some advantages with the feature selection process, a few drawbacks can be seen. Even though there was some dimensionality reduction, the final datasets were still large, with 1,636 features in the binary model and 1,253 in the time-based model. This may pose a computational challenge and a barrier to interpretation. Furthermore, the reliance on ANOVA for some modalities and model-based methods for others leads to inconsistency in features across data types. ANOVA, though good for univariate selection, does not consider the correlation among features and tends to select redundant irrelevant features — even though we eliminated this problem a priori by data preprocessing. Also, the imbalanced distribution of the data (ex. 203 non-recurrent LUAD versus 115 recurrent LUAD) may be prejudicial in methods susceptible to class imbalance, like ANOVA. Moreover, model-based methods, such as XGBoost and Random Forest, are prone to overfitting, especially in small datasets like the time-based recurrence model with only 115 samples. This could compromise the generalizability of the selected features. This is further underlined by the large variability in the number of retained features across modalities, such as 315 mRNA features versus 696 mutation features in the binary model, which shows that the selection criteria are imbalanced and may overemphasize certain data types. Li et al. (2021) proposed a solution by integrating multi-omics data using graph convolutional networks (GCNs), constructing a patient similarity network for enhancing classification performance (83). Their graph-based learning solution effectively captures nonlinear interactions among genomic, transcriptomic, and proteomic features that could be overlooked by traditional feature selection methods. This enhancement would be useful in determining feature sets with high generalizability across various patient populations, thereby preventing overfitting without sacrificing predictive accuracy.

To conclude, this research shows that mRNA expression, mutation counts, methylation values, and clinical data can sufficiently guide LUAD recurrence prediction through ML and DL models. Both binary and time models had AUROCs over 0.90, highlighting a promising multi-omics strategy. However, both models had their respective pitfalls: the binary model was perfect in terms of precision but suffered from recall, while the time-based model captured all true positives at the expense of higher false-positive rates. Such compromises indicate requirements for threshold tuning of the models, class imbalance handling, and further feature selection optimization. Although feature selection reduced dimensionality and improved model generalizability, it is challenging to balance the heterogeneous data types while minimizing the risks of overfitting. As a potential future research direction, more sophisticated architectures (e.g., graph-based learning) can be investigated to better capture complicated relationships in the data. By surmounting these challenges, the field can move closer to creating clinically relevant tools, ultimately paving the way for earlier detection of recurrence and improved prognosis for patients with LUAD.

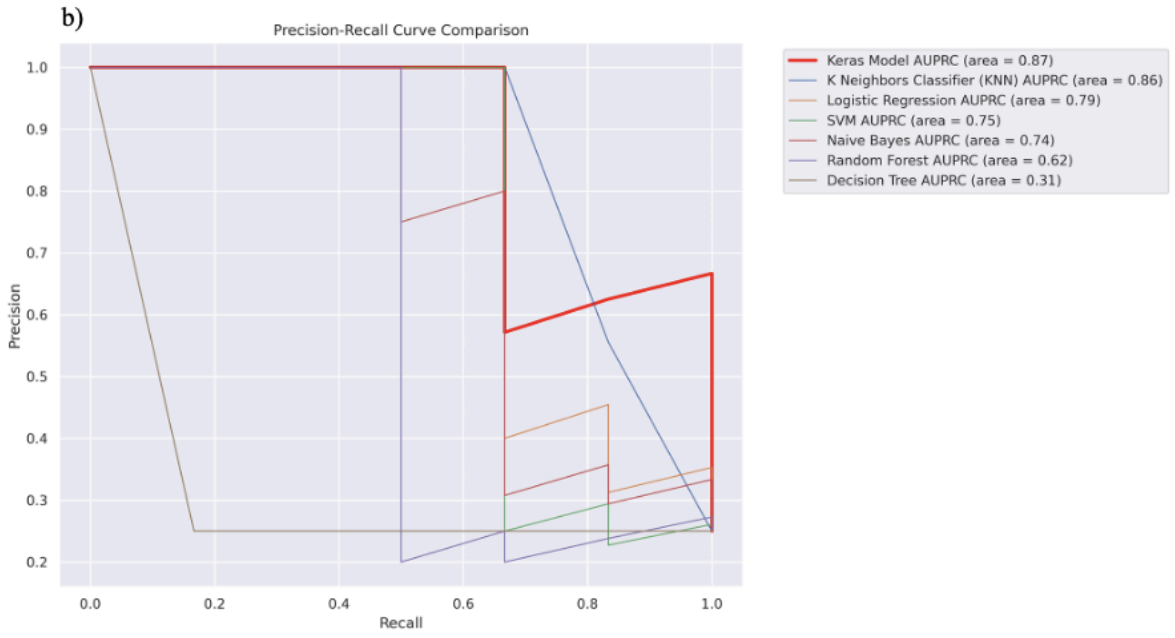
Supplementary Figures

---

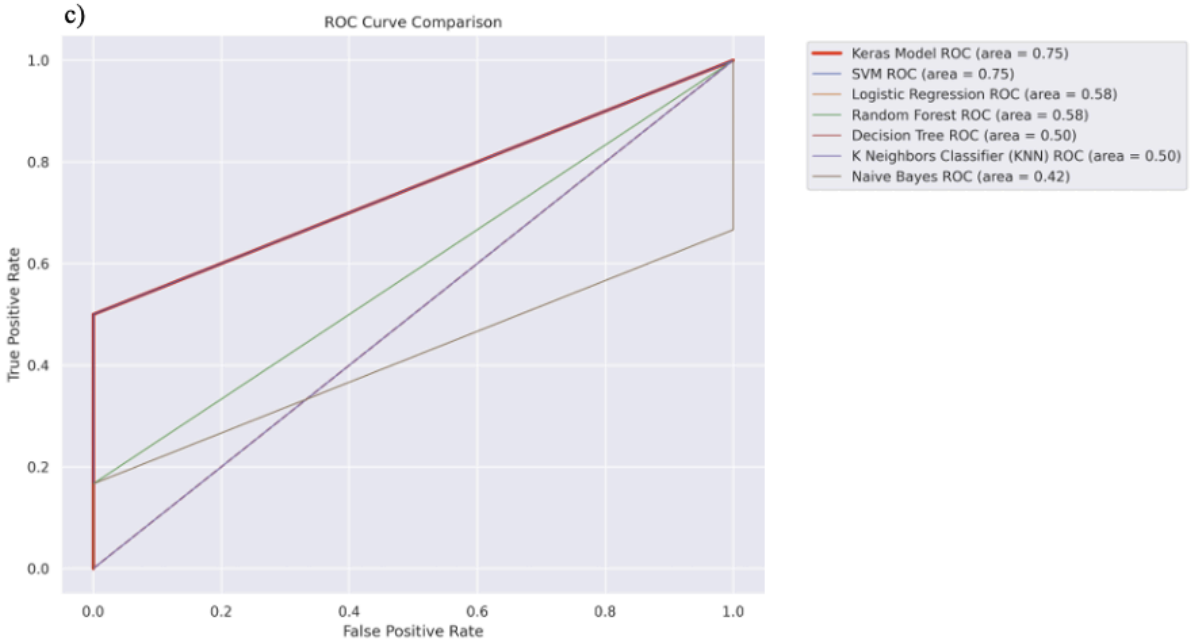
a)



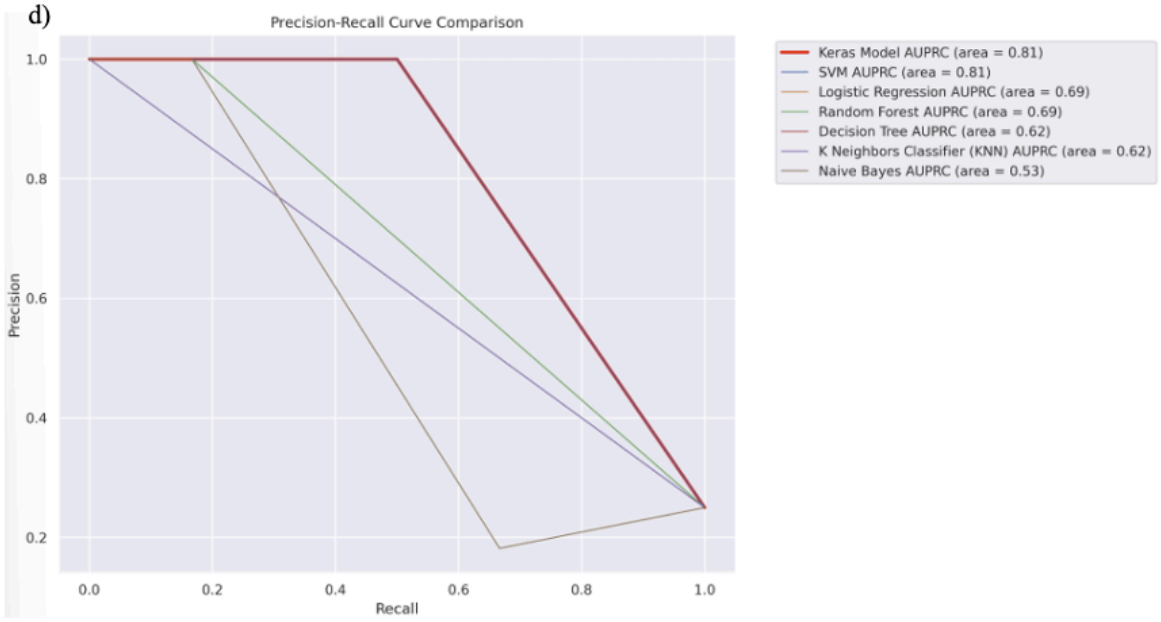
b)



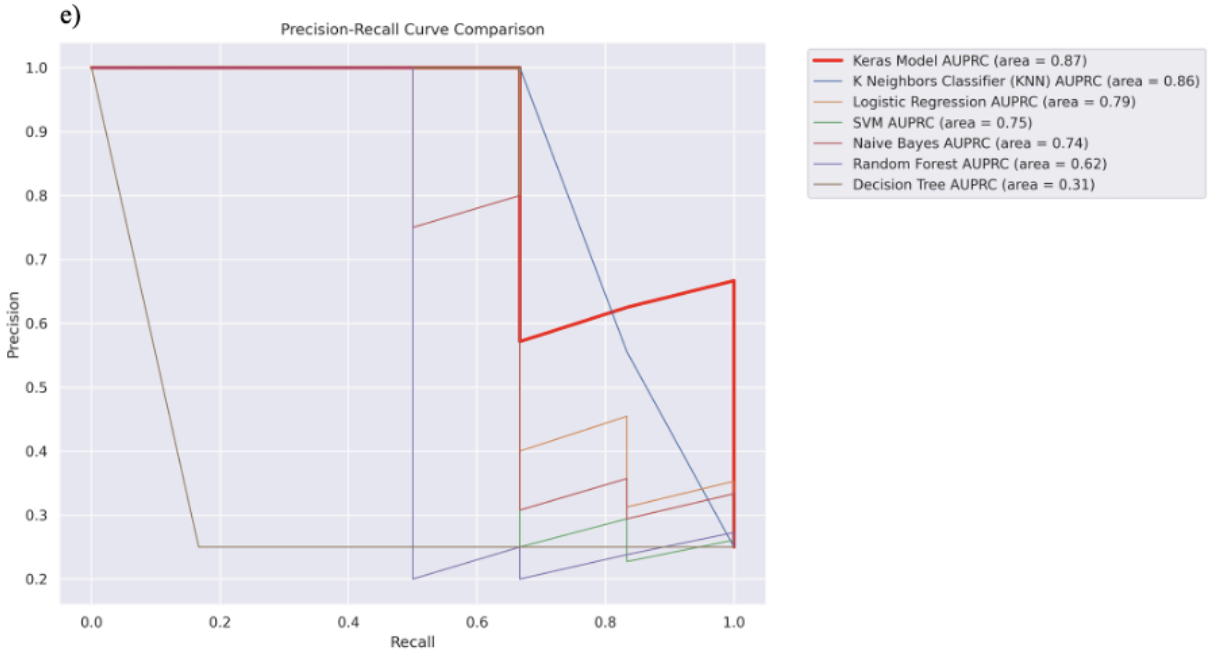
c)


ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Keras Model ROC (area = 0.75)
SVM ROC (area = 0.75)
Logistic Regression ROC (area = 0.58)
Random Forest ROC (area = 0.58)
Decision Tree ROC (area = 0.50)
K Neighbors Classifier (KNN) ROC (area = 0.50)
Naive Bayes ROC (area = 0.42)

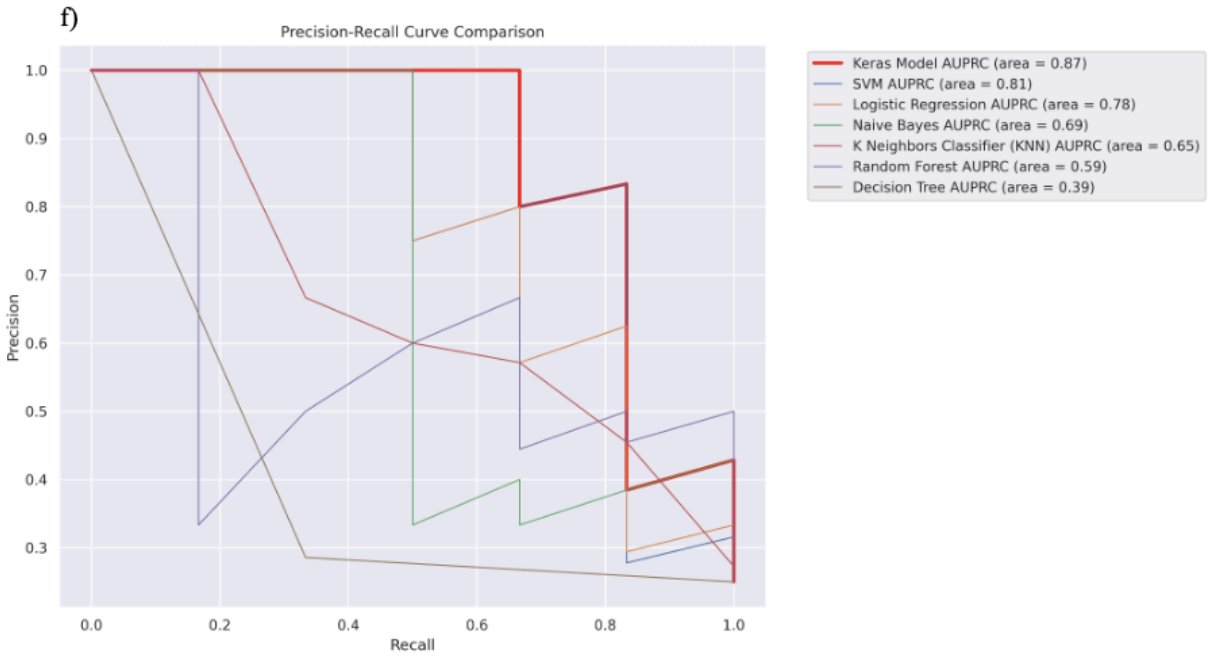
d)


Precision-Recall Curve Comparison
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Keras Model AUPRC (area = 0.81)
SVM AUPRC (area = 0.81)
Logistic Regression AUPRC (area = 0.69)
Random Forest AUPRC (area = 0.69)
Decision Tree AUPRC (area = 0.62)
K Neighbors Classifier (KNN) AUPRC (area = 0.62)
Naive Bayes AUPRC (area = 0.53)

e)


Precision-Recall Curve Comparison
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0
Keras Model AUPRC (area = 0.87)
K Neighbors Classifier (KNN) AUPRC (area = 0.86)
Logistic Regression AUPRC (area = 0.79)
SVM AUPRC (area = 0.75)
Naive Bayes AUPRC (area = 0.74)
Random Forest AUPRC (area = 0.62)
Decision Tree AUPRC (area = 0.31)

f)


Precision-Recall Curve Comparison
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0
Keras Model AUPRC (area = 0.87)
SVM AUPRC (area = 0.81)
Logistic Regression AUPRC (area = 0.78)
Naive Bayes AUPRC (area = 0.69)
K Neighbors Classifier (KNN) AUPRC (area = 0.65)
Random Forest AUPRC (area = 0.59)
Decision Tree AUPRC (area = 0.39)

1. Supplementary Figure S1. Comparisons of AUROC and AUPRC curves generated from individual modalities of the initial DFS-based dataset, showcasing individual Keras versus benchmark ML performance. One panel presents model performance per omics modality: (a, b) mRNA, (c, d) mutation, and (e, f) methylation. In both scenarios, a separate Keras deep learning model (red line) was also independently trained using the filtered DFS-labeled data and benchmarked against general ML baseline classifiers such as K-Nearest Neighbors (KNN), Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT). mRNA-based models (a, b) yielded superior AUROC (0.94) and AUPRC (0.87), whereas mutation (c, d) and methylation (e, f) models had more restricted performance. These outcomes, combined with the restricted DFS datasets and binary class imbalance, motivated a change of strategy to utilizing a PFS-based recurrence definition and constructing an integrated multi-omics modeling platform.

---

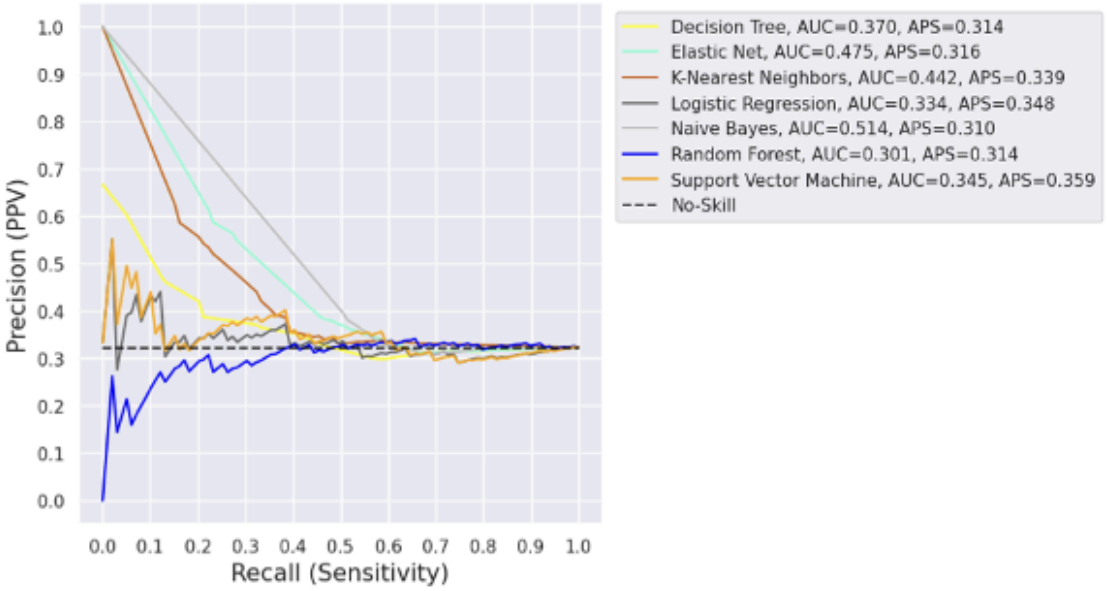
a)

ROC


True Positive Rate
False Positive Rate
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Decision Tree, AUC=0.457
Elastic Net, AUC=0.470
K-Nearest Neighbors, AUC=0.514
Logistic Regression, AUC=0.475
Naive Bayes, AUC=0.474
Random Forest, AUC=0.488
Support Vector Machine, AUC=0.491
No-Skill

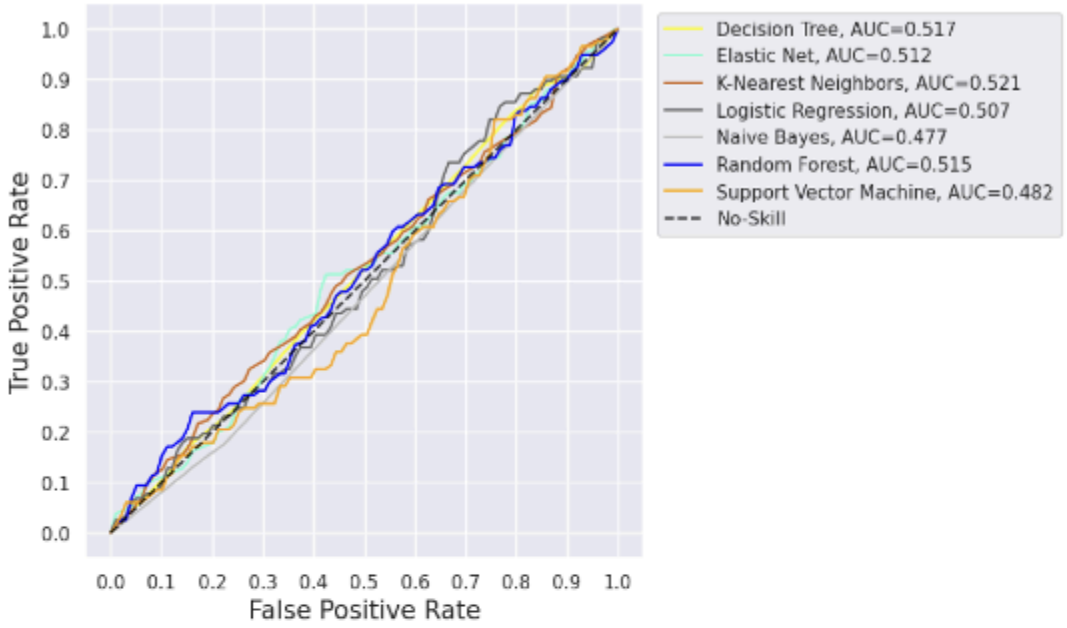
b)

PRC


Precision (PPV)
Recall (Sensitivity)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Decision Tree, AUC=0.370, APS=0.314
Elastic Net, AUC=0.475, APS=0.316
K-Nearest Neighbors, AUC=0.442, APS=0.339
Logistic Regression, AUC=0.334, APS=0.348
Naive Bayes, AUC=0.514, APS=0.310
Random Forest, AUC=0.301, APS=0.314
Support Vector Machine, AUC=0.345, APS=0.359
No-Skill

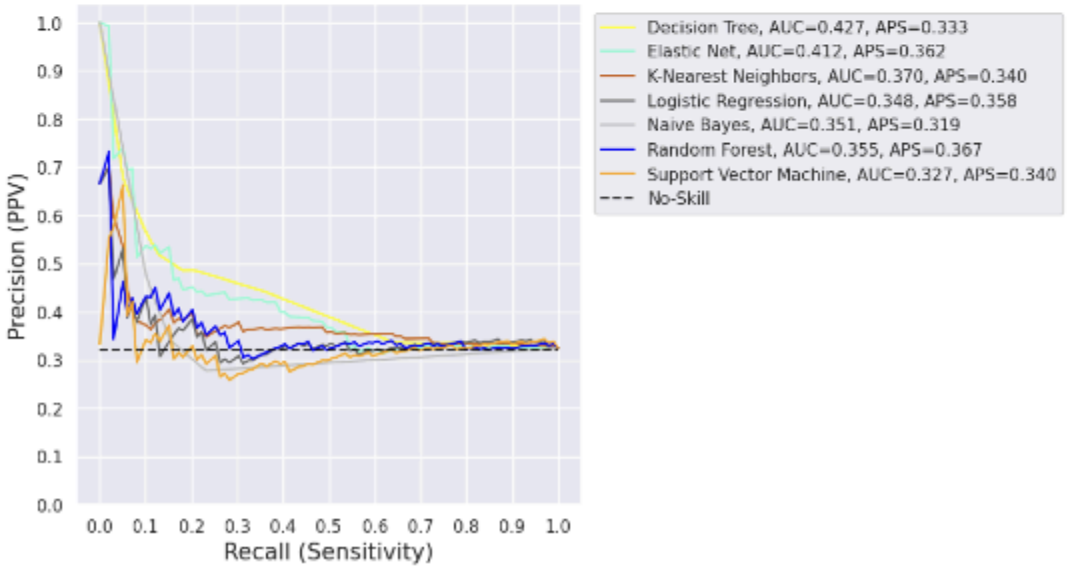
c)

ROC


True Positive Rate
False Positive Rate
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Decision Tree, AUC=0.517
Elastic Net, AUC=0.512
K-Nearest Neighbors, AUC=0.521
Logistic Regression, AUC=0.507
Naive Bayes, AUC=0.477
Random Forest, AUC=0.515
Support Vector Machine, AUC=0.482
No-Skill

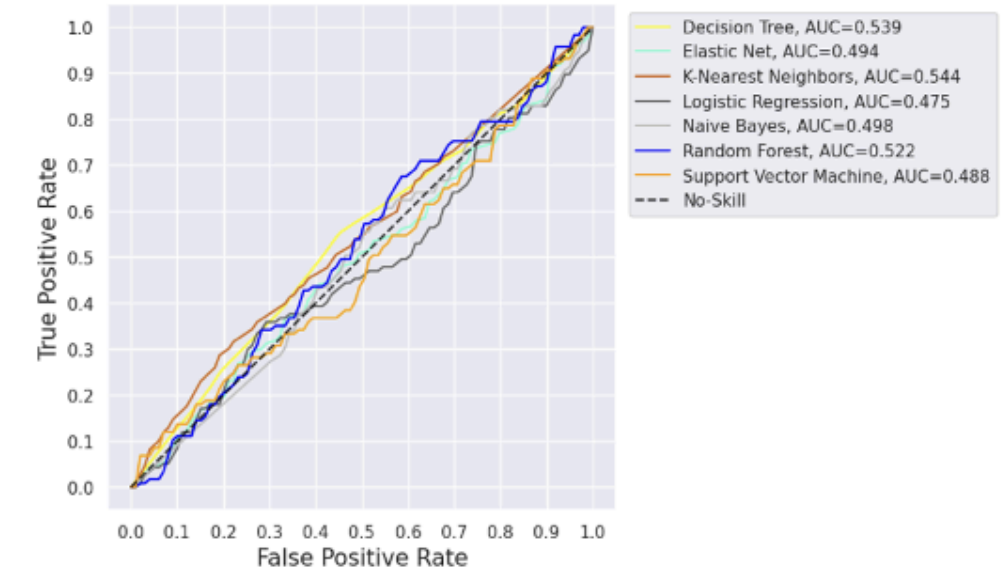
d)

PRC


Precision (PPV)
Recall (Sensitivity)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Decision Tree, AUC=0.427, APS=0.333
Elastic Net, AUC=0.412, APS=0.362
K-Nearest Neighbors, AUC=0.370, APS=0.340
Logistic Regression, AUC=0.348, APS=0.358
Naive Bayes, AUC=0.351, APS=0.319
Random Forest, AUC=0.355, APS=0.367
Support Vector Machine, AUC=0.327, APS=0.340
No-Skill

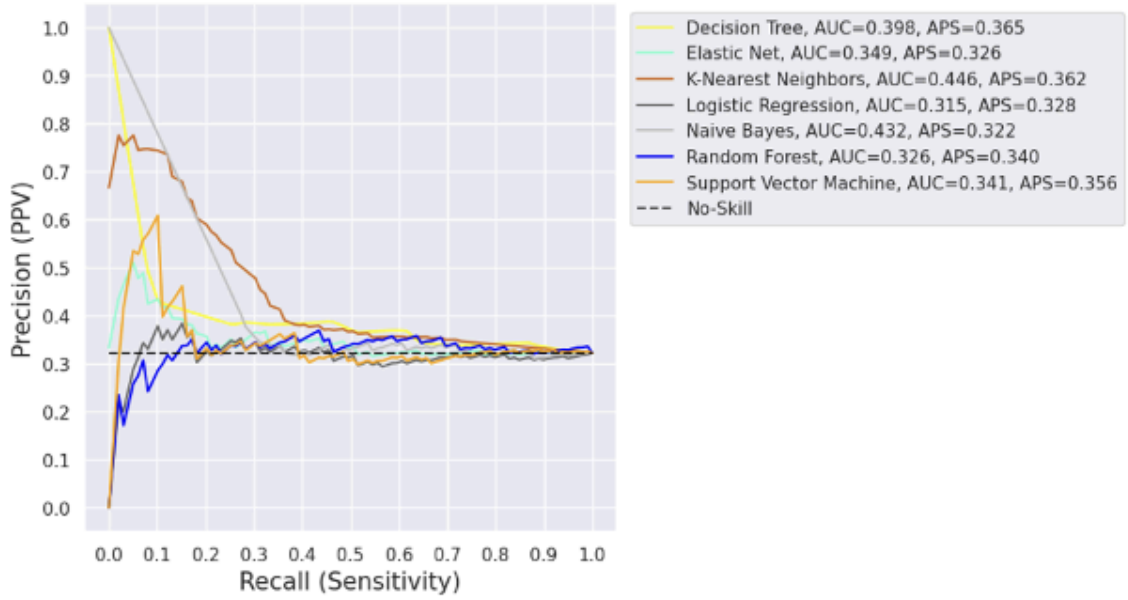
e)

ROC


True Positive Rate
False Positive Rate
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Decision Tree, AUC=0.539
Elastic Net, AUC=0.494
K-Nearest Neighbors, AUC=0.544
Logistic Regression, AUC=0.475
Naive Bayes, AUC=0.498
Random Forest, AUC=0.522
Support Vector Machine, AUC=0.488
No-Skill

f)

PRC


Precision (PPV)
Recall (Sensitivity)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Decision Tree, AUC=0.398, APS=0.365
Elastic Net, AUC=0.349, APS=0.326
K-Nearest Neighbors, AUC=0.446, APS=0.362
Logistic Regression, AUC=0.315, APS=0.328
Naive Bayes, AUC=0.432, APS=0.322
Random Forest, AUC=0.326, APS=0.340
Support Vector Machine, AUC=0.341, APS=0.356
No-Skill

2. Supplementary Figure S2. STREAMLINE-based ML AUROC and AUPRC curves of the pre-merge individual -omics datasets with binary PFS recurrence status. Every omics modality was inputted into STREAMLINE separately to evaluate its standalone predictive power before integrating the datasets. (a, b) mRNA-based modeling had the highest overall performance, with K-Nearest Neighbors possessing the highest AUROC (0.514) and Naïve Bayes the highest AUPRC (0.514). (c, d) Mutation models all had similar trends, with K-Nearest Neighbors again giving the best AUROC (0.521) and Decision Tree giving best results with AUPRC (0.427). (e, f) Methylation models all performed slightly better than the mutation models, with K-Nearest Neighbors giving the best AUROC (0.544) and Decision Tree giving best results with AUPRC (0.446). However, predictive performance for all modalities was poor, justifying the argument for the combination of omics datasets to improve model accuracy downstream. No-skill classifiers are included for comparison (dashed lines).

Appendix A: Selected binary features

Selected Expression Features: ['HMGB1P1:10357_mRNA', 'RNU12-2P:26823_mRNA', 'EZHIP:340602_mRNA', 'AACSL:729522_mRNA', 'AAMP:14_mRNA', 'ABCC10:89845_mRNA', 'ABCD1:215_mRNA', 'ABHD14A:25864_mRNA', 'ACD:65057_mRNA', 'ACE:1636_mRNA', 'ACSL1:2180_mRNA', 'ADAMTS8:11095_mRNA', 'ADAMTSL3:57188_mRNA', 'ADCK4:79934_mRNA', 'ADCY3:109_mRNA', 'ADPRHL1:113622_mRNA', 'AFP:174_mRNA', 'AIM1:202_mRNA', 'ALCAM:214_mRNA', 'ALPPL2:251_mRNA', 'AMY2B:280_mRNA', 'APH1B:83464_mRNA', 'API5:8539_mRNA', 'APPL2:55198_mRNA', 'ATL2:64225_mRNA', 'ATOH8:84913_mRNA', 'B3GNT7:93010_mRNA', 'BAAT:570_mRNA', 'BAK1:578_mRNA', 'BCL11A:53335_mRNA', 'BCL2L14:79370_mRNA', 'BCL6:604_mRNA', 'BEX5:340542_mRNA', 'BMP7:655_mRNA', 'BRF2:55290_mRNA', 'BRWD1:54014_mRNA', 'BZW2:28969_mRNA', 'C10orf11:83938_mRNA', 'C10orf46:143384_mRNA', 'C12orf68:387856_mRNA', 'C13orf29:283487_mRNA', 'C14orf143:90141_mRNA', 'C17orf60:284021_mRNA', 'C17orf63:55731_mRNA', 'C19orf10:56005_mRNA', 'C19orf26:255057_mRNA', 'C19orf43:79002_mRNA', 'C19orf57:79173_mRNA', 'C19orf73:55150_mRNA', 'C1orf128:57095_mRNA', 'C1orf200:644997_mRNA', 'C1orf31:388753_mRNA', 'C1orf83:127428_mRNA', 'C1QL3:389941_mRNA', 'C20orf70:140683_mRNA', 'C21orf56:84221_mRNA', 'C21orf67:84536_mRNA', 'C22orf41:644186_mRNA', 'C2CD4D:100191040_mRNA', 'C2orf55:343990_mRNA', 'C2orf56:55471_mRNA', 'C2orf58:285154_mRNA', 'C4orf33:132321_mRNA', 'C6orf26:401251_mRNA', 'C9orf116:138162_mRNA', 'C9orf125:84302_mRNA', 'C9orf21:195827_mRNA', 'CACNA1G:8913_mRNA', 'CALML4:91860_mRNA', 'CAPN1:823_mRNA', 'CAPS:828_mRNA', 'CASQ1:844_mRNA', 'CBFA2T3:863_mRNA', 'CBWD2:150472_mRNA', 'CCDC36:339834_mRNA', 'CD200R1:131450_mRNA', 'CDK10:8558_mRNA', 'CDK5:1020_mRNA', 'CDK5RAP1:51654_mRNA', 'CHCHD6:84303_mRNA', 'CLASP2:23122_mRNA', 'CLDN20:49861_mRNA', 'CNKSR3:154043_mRNA', 'COASY:80347_mRNA', 'CPA2:1358_mRNA', 'CPA4:51200_mRNA', 'CTDSPL2:51496_mRNA', 'CXCR2P1:3580_mRNA', 'CXorf57:55086_mRNA', 'CYBASC3:220002_mRNA', 'CYP19A1:1588_mRNA', 'CYP1B1:1545_mRNA', 'DAB1:1600_mRNA', 'DBNDD2:55861_mRNA', 'DEFB4A:1673_mRNA', 'DEM1:64789_mRNA', 'DHRS4:10901_mRNA', 'DIO1:1733_mRNA', 'DLGAP1:9229_mRNA', 'DMD:1756_mRNA', 'DNAJC1:64215_mRNA', 'DNAJC3:5611_mRNA', 'DNASE1:1773_mRNA', 'DPY19L1:23333_mRNA',

'DUS3L:56931_mRNA', 'DYDC2:84332_mRNA', 'E2F1:1869_mRNA', 'EGOT:100126791_mRNA', 'EGR1:1958_mRNA', 'EGR3:1960_mRNA', 'EIF2S1:1965_mRNA', 'EP400NL:347918_mRNA', 'FAM151B:167555_mRNA', 'FAM63A:55793_mRNA', 'FAM96B:51647_mRNA', 'FANCL:55120_mRNA', 'FBXO16:157574_mRNA', 'FBXO3:26273_mRNA', 'FFAR2:2867_mRNA', 'FGF1:2246_mRNA', 'LINC01634:100192420_mRNA', 'FOXK2:3607_mRNA', 'FSD1:79187_mRNA', 'GABBR2:9568_mRNA', 'GBP3:2635_mRNA', 'GDAP2:54834_mRNA', 'GDI1:2664_mRNA', 'GGT7:2686_mRNA', 'GLRB:2743_mRNA', 'GOLGA8C:729786_mRNA', 'GPC1:2817_mRNA', 'GPHA2:170589_mRNA', 'GPR31:2853_mRNA', 'GPR35:2859_mRNA', 'GRM1:2911_mRNA', 'GSDMC:56169_mRNA', 'GSTA4:2941_mRNA', 'HDAC10:83933_mRNA', 'HEATR3:55027_mRNA', 'HIST1H1E:3008_mRNA', 'HIST1H3J:8356_mRNA', 'HLA-DQB1:3119_mRNA', 'HLTF:6596_mRNA', 'HM13:81502_mRNA', 'HMHA1:23526_mRNA', 'HTA:283902_mRNA', 'HTATSF1:27336_mRNA', 'IFI27L2:83982_mRNA', 'IFITM5:387733_mRNA', 'IGF2BP2:10644_mRNA', 'INTS3:65123_mRNA', 'KCNMB1:3779_mRNA', 'KCTD10:83892_mRNA', 'KCTD15:79047_mRNA', 'KDM5C:8242_mRNA', 'KIAA0427:9811_mRNA', 'KIF26B:55083_mRNA', 'KIFC2:90990_mRNA', 'KLHDC9:126823_mRNA', 'KRT37:8688_mRNA', 'KRT8:3856_mRNA', 'LCTL:197021_mRNA', 'LDHAL6B:92483_mRNA', 'LIMA1:51474_mRNA', 'LNP1:348801_mRNA', 'LOC100192426:100192426_mRNA', 'LOC100240726:100240726_mRNA', 'LOC143188:143188_mRNA', 'LINC00667:339290_mRNA', 'LOC401127:401127_mRNA', 'SCGB1B2P:643719_mRNA', 'LINC00598:646982_mRNA', 'LOC728875:728875_mRNA', 'LRIT3:345193_mRNA', 'LRRC36:55282_mRNA', 'LYPD3:27076_mRNA', 'MAF1:84232_mRNA', 'MCART2:147407_mRNA', 'MGC23270:196872_mRNA', 'MLST8:64223_mRNA', 'MMP24:10893_mRNA', 'MORF4:10934_mRNA', 'MRPS24:64951_mRNA', 'MSX1:4487_mRNA', 'MTAP:4507_mRNA', 'MTMR7:9108_mRNA', 'MTRF1L:54516_mRNA', 'NAP1L2:4674_mRNA', 'NAP1L6:645996_mRNA', 'NCRNA00093:100188954_mRNA', 'NCRNA00094:266655_mRNA', 'NCRNA00152:112597_mRNA', 'NDUFB2:4708_mRNA', 'NEFL:4747_mRNA', 'NELL1:4745_mRNA', 'NGF:4803_mRNA', 'NKX2-8:26257_mRNA', 'NOTCH2NL:388677_mRNA', 'NRP1:8829_mRNA', 'NSFL1C:55968_mRNA', 'NSUN4:387338_mRNA', 'NTRK2:4915_mRNA', 'ONECUT2:9480_mRNA', 'OR4C6:219432_mRNA', 'ORMDL3:94103_mRNA', 'OTUB1:55611_mRNA', 'PAAF1:80227_mRNA', 'PAPOLG:64895_mRNA', 'PAPPA2:60676_mRNA', 'PDCD1LG2:80380_mRNA',

'PDGFD:80310_mRNA', 'PDXP:57026_mRNA', 'PFN4:375189_mRNA', 'PGBD5:79605_mRNA', 'PGM2L1:283209_mRNA', 'PHACTR3:116154_mRNA', 'PHF1:5252_mRNA', 'PIGT:51604_mRNA', 'PLA2G4A:5321_mRNA', 'PLEKHB2:55041_mRNA', 'PM20D1:148811_mRNA', 'PMM1:5372_mRNA', 'PNPLA5:150379_mRNA', 'PRICKLE3:4007_mRNA', 'PRKAA1:5562_mRNA', 'PRKAA2:5563_mRNA', 'PROS1:5627_mRNA', 'PRSS21:10942_mRNA', 'PRSS45:377047_mRNA', 'PSTK:118672_mRNA', 'PTH2R:5746_mRNA', 'PTK2B:2185_mRNA', 'PWWP2A:114825_mRNA', 'RAET1E:135250_mRNA', 'RAGE:5891_mRNA', 'RAX:30062_mRNA', 'RCAN2:10231_mRNA', 'RFPL4A:342931_mRNA', 'RIPK1:8737_mRNA', 'RNF24:11237_mRNA', 'RSAD2:91543_mRNA', 'RYR3:6263_mRNA', 'S100A1:6271_mRNA', 'SAP18:10284_mRNA', 'SBSN:374897_mRNA', 'SCAMP5:192683_mRNA', 'SCML1:6322_mRNA', 'SCRN1:9805_mRNA', 'SDHB:6390_mRNA', 'SECTM1:6398_mRNA', 'SELS:55829_mRNA', 'SERPINB6:5269_mRNA', 'SERPINE1:5054_mRNA', 'SLC16A13:201232_mRNA', 'SLC25A3:5250_mRNA', 'SLC2A8:29988_mRNA', 'SLC4A8:9498_mRNA', 'SLC5A8:160728_mRNA', 'SLC6A7:6534_mRNA', 'SLC9A6:10479_mRNA', 'SMPDL3B:27293_mRNA', 'SNAP47:116841_second_mRNA', 'SNHG4:724102_mRNA', 'SNORA21:619505_mRNA', 'SNX31:169166_mRNA', 'SPDYE1:285955_mRNA', 'SPOCK3:50859_mRNA', 'STK3:6788_mRNA', 'STXBP1:6812_mRNA', 'SUMO4:387082_mRNA', 'SYAP1:94056_mRNA', 'SYTL3:94120_mRNA', 'TATDN3:128387_mRNA', 'TBL1XR1:79718_mRNA', 'TCAM1P:146771_mRNA', 'TCAP:8557_mRNA', 'TFDP1:7027_mRNA', 'TIMM22:29928_mRNA', 'TINAG:27283_mRNA', 'TMEM123:114908_mRNA', 'TMEM161B:153396_mRNA', 'TMEM164:84187_mRNA', 'TMEM177:80775_mRNA', 'TMEM185B:79134_mRNA', 'TMOD1:7111_mRNA', 'TOE1:114034_mRNA', 'TOMM7:54543_mRNA', 'TPPP:11076_mRNA', 'TPSB2:64499_mRNA', 'TRIM10:10107_mRNA', 'TRO:7216_mRNA', 'TSG101:7251_mRNA', 'TTC9:23508_mRNA', 'TUBGCP3:10426_mRNA', 'TULP4:56995_mRNA', 'TXNDC5:81567_mRNA', 'UBE2MP1:606551_mRNA', 'UBE2Z:65264_mRNA', 'USP47:55031_mRNA', 'VSIG10L:147645_mRNA', 'VSNL1:7447_mRNA', 'VWCE:220001_mRNA', 'WDR16:146845_mRNA', 'WDR70:55100_mRNA', 'XPR1:9213_mRNA', 'YIPF1:54432_mRNA', 'ZADH2:284273_mRNA', 'ZNF138:7697_mRNA', 'ZNF394:84124_mRNA', 'ZNF414:84330_mRNA', 'ZNF593:51042_mRNA', 'ZNF614:80110_mRNA', 'ZNF630:57232_mRNA', 'ZNF83:55769_mRNA', 'ZYX:7791_mRNA']

Selected Mutation Features: ['PELO_MUT', 'ZCCHC8_MUT', 'GZMK_MUT', 'AHNAK_MUT', 'P2RY1_MUT', 'IGKV2D-29_MUT', 'PKN1_MUT', 'ZFC3H1_MUT', 'SLC9A8_MUT', 'GLI3_MUT', 'MYL3_MUT', 'SETDB2_MUT', 'BRAF_MUT', 'SERPINI2_MUT', 'MFRP_MUT', 'ZSCAN22_MUT', 'ANKRD39_MUT', 'XYLB_MUT', 'PHLDB1_MUT', 'EXO1_MUT', 'KNDC1_MUT', 'CASP9_MUT', 'PPTC7_MUT', 'ADAMTS14_MUT', 'TSC22D2_MUT', 'CAND1_MUT', 'ABCC8_MUT', 'CLLU1OS_MUT', 'MTG2_MUT', 'OR51L1_MUT', 'LRRC41_MUT', 'MAGEC3_MUT', 'ENTPD2_MUT', 'ALG9_MUT', 'CSH1_MUT', 'C14orf166B_MUT', 'ZNF223_MUT', 'OR10G9_MUT', 'LILRB1_MUT', 'VSNL1_MUT', 'ABCG4_MUT', 'MCTP2_MUT', 'SLC15A3_MUT', 'UTP18_MUT', 'FAM20A_MUT', 'LHX9_MUT', 'DAAM1_MUT', 'KCNV1_MUT', 'INSL5_MUT', 'FMNL3_MUT', 'PTP4A3_MUT', 'EDNRA_MUT', 'LEPRE1_MUT', 'DECR2_MUT', 'SUPT3H_MUT', 'TVP23C_MUT', 'TRABD2A_MUT', 'SPAG8_MUT', 'ADORA1_MUT', 'ILF2_MUT', 'MESP2_MUT', 'MFAP3L_MUT', 'STON2_MUT', 'CLCN3_MUT', 'PGK2_MUT', 'MRPL44_MUT', 'DNAJC10_MUT', 'MMP7_MUT', 'ACOT6_MUT', 'ALAS2_MUT', 'ZWINT_MUT', 'SLC26A3_MUT', 'FAM120A_MUT', 'OR2A2_MUT', 'MAP2K7_MUT', 'PYDC2_MUT', 'LRIT2_MUT', 'DOT1L_MUT', 'LINGO2_MUT', 'SLC32A1_MUT', 'WSCD2_MUT', 'PIGV_MUT', 'DSE_MUT', 'TMEM178A_MUT', 'TMEM168_MUT', 'MTMR8_MUT', 'KRTAP5-5_MUT', 'CTPS1_MUT', 'SNX24_MUT', 'C16orf46_MUT', 'DNAH17_MUT', 'MICAL2_MUT', 'CDKL2_MUT', 'AHSG_MUT', 'NFX1_MUT', 'SHPK_MUT', 'HTR3D_MUT', 'BANP_MUT', 'EIF6_MUT', 'RGS18_MUT', 'FLCN_MUT', 'MGAT1_MUT', 'VPS37A_MUT', 'ADAMTS5_MUT', 'KCTD3_MUT', 'PRMT10_MUT', 'DPYSL5_MUT', 'OR4X1_MUT', 'MRPL47_MUT', 'SKAP1_MUT', 'N4BP1_MUT', 'ZMYM1_MUT', 'TTC39A_MUT', 'HOMEZ_MUT', 'ZZEF1_MUT', 'BTBD10_MUT', 'HNRNPA2B1_MUT', 'GTPBP4_MUT', 'ERG_MUT', 'ALDH1L1_MUT', 'C14orf183_MUT', 'GDF2_MUT', 'OR4F15_MUT', 'ATP11C_MUT', 'IPO8_MUT', 'SIGLEC6_MUT', 'CCDC37_MUT', 'GAPVD1_MUT', 'PSIP1_MUT', 'OR4K2_MUT', 'ERC2_MUT', 'RAB30_MUT', 'DRD5_MUT', 'C16orf70_MUT', 'KREMEN2_MUT', 'POGK_MUT', 'SRSF2_MUT', 'ZNF862_MUT', 'ESPL1_MUT', 'ACYP2_MUT', 'ZNF408_MUT', 'SLC7A5_MUT', 'HR_MUT', 'WISP1_MUT', 'XXYLT1_MUT', 'TBC1D31_MUT', 'MTMR1_MUT', 'MAP3K6_MUT', 'COL4A3BP_MUT', 'ZNF512_MUT', 'UBA2_MUT', 'PIGS_MUT', 'CRYGA_MUT', 'PPP2R5B_MUT', 'MRPL35_MUT', 'KRT71_MUT', 'KL_MUT', 'AXDND1_MUT', 'TTF2_MUT',

'LPIN3_MUT', 'TFEB_MUT', 'NAA11_MUT', 'KCNAB2_MUT', 'WFS1_MUT', 'DSG2_MUT', 'PLXNB1_MUT', 'PCDH8_MUT', 'RFPL2_MUT', 'DNAJC8_MUT', 'C19orf66_MUT', 'MARCH6_MUT', 'PRICKLE1_MUT', 'CA9_MUT', 'UVRAG_MUT', 'TRAPPC8_MUT', 'CFHR3_MUT', 'FBXL2_MUT', 'IFNLR1_MUT', 'GPR124_MUT', 'TESC_MUT', 'FAM111A_MUT', 'MED23_MUT', 'CCDC136_MUT', 'SULT1C3_MUT', 'NUP155_MUT', 'TUBA1A_MUT', 'UBE2NL_MUT', 'LCLAT1_MUT', 'C1QTNF3_MUT', 'ATP2C1_MUT', 'IQCA1_MUT', 'CD300C_MUT', 'EVA1B_MUT', 'LCE3D_MUT', 'FAM84B_MUT', 'SOS1_MUT', 'C1orf87_MUT', 'C1orf63_MUT', 'NT5DC1_MUT', 'GLIPR2_MUT', 'OR8J3_MUT', 'LTBP4_MUT', 'ZNF282_MUT', 'CEP76_MUT', 'VHLL_MUT', 'PRAMEF17_MUT', 'SLC6A19_MUT', 'TMEM105_MUT', 'SBNO1_MUT', 'AZI1_MUT', 'HOXB3_MUT', 'CCDC14_MUT', 'LPAR3_MUT', 'OR1E1_MUT', 'GIMAP6_MUT', 'CALN1_MUT', 'ZNF432_MUT', 'HSD17B1_MUT', 'ZCCHC6_MUT', 'ENTHD2_MUT', 'DENND5A_MUT', 'CRISPLD2_MUT', 'IPO13_MUT', 'GUCY2D_MUT', 'BTK_MUT', 'B4GALNT2_MUT', 'TEX14_MUT', 'NGEF_MUT', 'RGS20_MUT', 'IFNGR1_MUT', 'ALDH5A1_MUT', 'LYST_MUT', 'PHC2_MUT', 'EIF4G3_MUT', 'KIN_MUT', 'LRRC37B_MUT', 'MYLK3_MUT', 'SLC26A11_MUT', 'HYOU1_MUT', 'NTSR2_MUT', 'ADAM28_MUT', 'SF3A1_MUT', 'ENOX2_MUT', 'ITGA2B_MUT', 'STOML3_MUT', 'PM20D2_MUT', 'ARRDC5_MUT', 'FLRT2_MUT', 'STK16_MUT', 'CCDC144A_MUT', 'ASIC1_MUT', 'MLH1_MUT', 'ATP13A3_MUT', 'EFCAB6_MUT', 'EBF2_MUT', 'MYOZ1_MUT', 'XAB2_MUT', 'JADE2_MUT', 'UPK3A_MUT', 'RORA_MUT', 'NES_MUT', 'CCDC80_MUT', 'ODF2L_MUT', 'AKR7A2_MUT', 'HAUS1_MUT', 'SIK1_MUT', 'ZDHHC4_MUT', 'E2F7_MUT', 'LHFPL3_MUT', 'IGDCC3_MUT', 'SLC6A13_MUT', 'DRAXIN_MUT', 'SLC6A7_MUT', 'MAGEB4_MUT', 'KRT16_MUT', 'MAP7D3_MUT', 'ULBP1_MUT', 'SGTB_MUT', 'ZNF438_MUT', 'MAGI2_MUT', 'SLFN12_MUT', 'PRICKLE2_MUT', 'POTEH_MUT', 'MRVI1_MUT', 'SLC22A8_MUT', 'ECE1_MUT', 'RDM1_MUT', 'VSIG4_MUT', 'EDEM3_MUT', 'MOGAT3_MUT', 'KCND1_MUT', 'C7orf34_MUT', 'SLCO1B7_MUT', 'ZNF404_MUT', 'ASNS_MUT', 'MST1R_MUT', 'FIGNL1_MUT', 'TLL2_MUT', 'ECHDC1_MUT', 'TFAP2B_MUT', 'LMNA_MUT', 'HAUS7_MUT', 'CLYBL_MUT', 'BRD1_MUT', 'TMEM215_MUT', 'SH3GL2_MUT', 'NUP188_MUT', 'AQP8_MUT', 'NANP_MUT', 'ACOX2_MUT', 'ZNF521_MUT', 'ERRFI1_MUT', 'TRAF3IP3_MUT', 'COL11A1_MUT', 'TMPRSS11F_MUT', 'B3GALNT2_MUT', 'ADAL_MUT', 'GJA10_MUT', 'DEPTOR_MUT', 'CYP46A1_MUT', 'FOXP3_MUT', 'ZNF407_MUT', 'SLC11A1_MUT', 'CIITA_MUT', 'PRPH_MUT',

'SMARCA5_MUT', 'KAZALD1_MUT', 'AFF3_MUT', 'ARMC9_MUT', 'ITGAX_MUT', 'PDIA6_MUT', 'DUS3L_MUT', 'GRM5_MUT', 'GPRIN3_MUT', 'CTNNA2_MUT', 'ITGA11_MUT', 'MAGT1_MUT', 'UTP14A_MUT', 'KRTAP19-3_MUT', 'AKR7A3_MUT', 'OR4D10_MUT', 'EIF4E_MUT', 'PPEF2_MUT', 'PRSS38_MUT', 'CLDN22_MUT', 'APBB2_MUT', 'FIGF_MUT', 'PPM1L_MUT', 'PELP1_MUT', 'FLYWCH1_MUT', 'TMEM81_MUT', 'MARCO_MUT', 'FUCA1_MUT', 'TOR1B_MUT', 'FFAR4_MUT', 'CD97_MUT', 'FBXL17_MUT', 'EPHA2_MUT', 'PDZRN4_MUT', 'SOAT2_MUT', 'GATAD1_MUT', 'CDHR1_MUT', 'KCNK17_MUT', 'EVPL_MUT', 'SLC24A4_MUT', 'VIM_MUT', 'FGL1_MUT', 'LAMP1_MUT', 'IGHV3-16_MUT', 'PKP2_MUT', 'ZC3H7B_MUT', 'SLC12A6_MUT', 'COG8_MUT', 'ASL_MUT', 'PCDHB3_MUT', 'SNRPN_MUT', 'GDPD5_MUT', 'MYH11_MUT', 'FARP1_MUT', 'KANSL1_MUT', 'FICD_MUT', 'CYP7A1_MUT', 'MAGEH1_MUT', 'TGM1_MUT', 'HFM1_MUT', 'CDC7_MUT', 'NCAPH2_MUT', 'BHLHB9_MUT', 'BRCA2_MUT', 'PTPRG_MUT', 'EPHA3_MUT', 'DAPK2_MUT', 'GPR84_MUT', 'LRRN4_MUT', 'AIFM2_MUT', 'APH1A_MUT', 'DHX30_MUT', 'GHR_MUT', 'PTHLH_MUT', 'SLC24A1_MUT', 'LCAT_MUT', 'NELFA_MUT', 'ARHGAP29_MUT', 'CALD1_MUT', 'KCNV2_MUT', 'ZNF775_MUT', 'TUBA1B_MUT', 'MED12_MUT', 'CHERP_MUT', 'OR2D2_MUT', 'CCSER1_MUT', 'SLITRK2_MUT', 'AAK1_MUT', 'NOX3_MUT', 'ARSJ_MUT', 'ARID3C_MUT', 'OLFM2_MUT', 'TBC1D19_MUT', 'H6PD_MUT', 'CASKIN1_MUT', 'PTPRT_MUT', 'DUSP22_MUT', 'OR11H6_MUT', 'GJC1_MUT', 'C9orf3_MUT', 'SPTBN2_MUT', 'NID1_MUT', 'PPP6C_MUT', 'ACER1_MUT', 'FYN_MUT', 'MAGEL2_MUT', 'UTP20_MUT', 'ADAM19_MUT', 'ZP2_MUT', 'C12orf5_MUT', 'MCHR1_MUT', 'PLEKHG1_MUT', 'LPO_MUT', 'WFDC12_MUT', 'PFKM_MUT', 'GPX7_MUT', 'RASA2_MUT', 'HIST1H2BH_MUT', 'RBMX_MUT', 'TSPAN8_MUT', 'ZNF558_MUT', 'NSRP1_MUT', 'ZNF703_MUT', 'TKTL1_MUT', 'PACRGL_MUT', 'AGBL1_MUT', 'VENTX_MUT', 'LRRIQ1_MUT', 'SETX_MUT', 'GBP4_MUT', 'ZNF318_MUT', 'ALDH1A3_MUT', 'LRRC32_MUT', 'UQCRB_MUT', 'TRAM2_MUT', 'CXXC1_MUT', 'IRX5_MUT', 'BLID_MUT', 'LIFR_MUT', 'CBX3_MUT', 'PIAS4_MUT', 'ZNF225_MUT', 'ADD1_MUT', 'SPTBN5_MUT', 'GPR152_MUT', 'WDR61_MUT', 'NPL_MUT', 'RXFP2_MUT', 'MMRN2_MUT', 'PCSK5_MUT', 'UGT2A2_MUT', 'PIDD_MUT', 'STK32B_MUT', 'BCL2L13_MUT', 'ZNF274_MUT', 'DYRK3_MUT', 'ZSWIM1_MUT', 'FBN1_MUT', 'MBD5_MUT', 'SLC35G3_MUT', 'PSMD1_MUT', 'DUSP21_MUT', 'CCDC120_MUT', 'SLC26A7_MUT', 'SAG_MUT', 'NLGN4X_MUT', 'DEFB112_MUT', 'MCM6_MUT',

'IL1RL2_MUT', 'RYR1_MUT', 'ACSM1_MUT', 'KRTAP5-3_MUT', 'GRK6_MUT', 'KIAA0907_MUT', 'KCNT1_MUT', 'OR12D2_MUT', 'FMN2_MUT', 'DLK1_MUT', 'ZNF484_MUT', 'C7orf31_MUT', 'TSNAXIP1_MUT', 'GPR160_MUT', 'RFFL_MUT', 'KCNMB4_MUT', 'DHRS12_MUT', 'TSPAN19_MUT', 'KIAA0368_MUT', 'CENPJ_MUT', 'TTC3_MUT', 'ARHGAP10_MUT', 'SOCS2_MUT', 'CHST1_MUT', 'HS3ST4_MUT', 'MAGI1_MUT', 'TMPRSS6_MUT', 'DDX20_MUT', 'S1PR4_MUT', 'SACS_MUT', 'MTMR11_MUT', 'VWA3B_MUT', 'CACNG4_MUT', 'GRM4_MUT', 'SEZ6L2_MUT', 'TMCO6_MUT', 'ARL2_MUT', 'EDAR_MUT', 'LMAN2_MUT', 'OR4S1_MUT', 'SAFB_MUT', 'ADAMTS3_MUT', 'KNOP1_MUT', 'OR14J1_MUT', 'PCDHB4_MUT', 'NPPA_MUT', 'ZNF778_MUT', 'RSRC2_MUT', 'PNPLA8_MUT', 'OTUD5_MUT', 'TRO_MUT', 'PFKP_MUT', 'QSOX1_MUT', 'PRKAR2A_MUT', 'KRTAP10-10_MUT', 'PDCD6IP_MUT', 'ALDH3B1_MUT', 'GRHL3_MUT', 'FOXL2_MUT', 'FBXO4_MUT', 'DENND4A_MUT', 'CACNA1A_MUT', 'PAX7_MUT', 'ALDH3B2_MUT', 'GPR83_MUT', 'ABCB4_MUT', 'CYB5D2_MUT', 'AIPL1_MUT', 'GAGE2B_MUT', 'GAS7_MUT', 'TMCO5A_MUT', 'DSG4_MUT', 'ABCF2_MUT', 'CYFIP2_MUT', 'ZNF169_MUT', 'SLC25A32_MUT', 'MTL5_MUT', 'ZNF227_MUT', 'GALNT6_MUT', 'KCNH1_MUT', 'IMP4_MUT', 'PLXNC1_MUT', 'OR5T3_MUT', 'SFPQ_MUT', 'KCTD19_MUT', 'FAM129A_MUT', 'AKAP11_MUT', 'KATNBL1_MUT', 'SLC5A5_MUT', 'PRB1_MUT', 'ABHD10_MUT', 'HDAC10_MUT', 'TRAJ23_MUT', 'ZNF23_MUT', 'TNRC6A_MUT', 'SCAF4_MUT', 'S100A12_MUT', 'USPL1_MUT', 'B4GALT2_MUT', 'TNFAIP2_MUT', 'CACNA1D_MUT', 'EWSR1_MUT', 'PCCA_MUT', 'ITGB5_MUT', 'RASGRF1_MUT', 'CRNN_MUT', 'DNAJB4_MUT', 'CSN1S1_MUT', 'RASL11A_MUT', 'DYNLRB1_MUT', 'NID2_MUT', 'CABP2_MUT', 'LRRC40_MUT', 'RGAG1_MUT', 'ZNF302_MUT', 'NSDHL_MUT', 'XRCC1_MUT', 'SLC39A12_MUT', 'GABRA1_MUT', 'FKBP2_MUT', 'AKNA_MUT', 'IL2RB_MUT', 'SH3TC2_MUT', 'KCMF1_MUT', 'CNOT4_MUT', 'PSG8_MUT', 'FTO_MUT', 'KMT2B_MUT', 'HOXB9_MUT', 'CACNA1G_MUT', 'TRAV22_MUT', 'TYW3_MUT', 'FZR1_MUT', 'IBSP_MUT', 'TRAF2_MUT', 'FMO3_MUT', 'GOT1L1_MUT', 'PNPLA5_MUT', 'LRRC1_MUT', 'ITGA6_MUT', 'NOTCH2NL_MUT', 'PIP4K2C_MUT', 'CTDSP2_MUT', 'POLR3C_MUT', 'BIRC6_MUT', 'IGLV3-1_MUT', 'GABRP_MUT', 'WDR78_MUT', 'XYLT2_MUT', 'MAPK1_MUT', 'MGA_MUT', 'GPR32_MUT', 'MMP26_MUT', 'CHAMP1_MUT', 'TRAV27_MUT', 'STXBP2_MUT', 'SLC4A3_MUT', 'FBXO41_MUT', 'OR8K3_MUT', 'WNK2_MUT', 'KTI12_MUT', 'COQ3_MUT', 'IVL_MUT', 'IMPG1_MUT', 'QSOX2_MUT', 'EGF_MUT',

'SEPT14_MUT', 'KRTAP19-4_MUT', 'ATP2B3_MUT', 'MBOAT2_MUT', 'ST18_MUT', 'NDUFA12_MUT', 'PITPNM2_MUT', 'ARHGAP44_MUT', 'SEPT3_MUT', 'ITGAM_MUT', 'PLA2G4C_MUT', 'WDFY1_MUT', 'RBBP6_MUT', 'DDX58_MUT', 'ATP12A_MUT', 'BRD8_MUT', 'C1R_MUT', 'MORC3_MUT', 'KRT33B_MUT', 'KCNK5_MUT', 'SLK_MUT', 'CNOT6_MUT', 'PRDM15_MUT', 'CYP4A11_MUT', 'ERBB4_MUT', 'CNTLN_MUT', 'LMO2_MUT', 'TRIM28_MUT', 'NMUR2_MUT', 'TRIM25_MUT', 'UBBP4_MUT', 'IGHV3-23_MUT', 'WDR65_MUT', 'SCAI_MUT', 'YWHAQ_MUT', 'COLEC10_MUT', 'MEMO1_MUT']

Selected Methylation Features: ['ELOVL5_MET', 'C19orf56;MORG1_MET', 'RB1_MET', 'ACRV1_MET', 'FASLG_MET', 'UBTD2_MET', 'ABCC3_MET', 'DGKH_MET', 'ANK1_MET', 'RARA_MET', 'C1orf59_MET', 'C18orf54_MET', 'TES_MET', 'PPEF1_MET', 'MOSC2_MET', 'ATG9A;ABCB6_MET', 'VCAN_MET', 'COBRA1_MET', 'RBM9_MET', 'HRASLS2_MET', 'ZNF784_MET', 'SEPHS1_MET', 'MAN1A1_MET', 'DOCK6;LOC55908_MET', 'ALDH1A2_MET', 'SLC4A5_MET', 'SRF_MET', 'CYFIP2_MET', 'TRH_MET', 'HERC3;NAP1L5_MET', 'GLMN;RPAP2_MET', 'HDAC10_MET', 'ZNF710_MET', 'ASNS_MET', 'FOXJ1_MET', 'NOP10_MET', 'HOXA9_MET', 'ZNF213_MET', 'TP53INP2_MET', 'ANKDD1A_MET', 'HSPA2_MET', 'PRAC;C17orf93_MET', 'RPH3A_MET', 'CTDP1_MET', 'HBEGF_MET', 'ZBED5_MET', 'MALL_MET', 'DCTN1_MET', 'JARID2_MET', 'ZBTB7B_MET', 'EPC1_MET', 'TARBP2;MAP3K12_MET', 'PKP2_MET', 'ANXA1_MET', 'CCDC7_MET', 'LSP1_MET', 'CACNA1D_MET', 'SKAP2_MET', 'CXCL1_MET', 'PCMT1_MET', 'SHANK2_MET', 'UNC13B_MET', 'SORT1_MET', 'SERPINA1_MET', 'OCLN_MET', 'FAM65A_MET', 'CCDC140;PAX3_MET', 'AP3D1_MET', 'MOSC1_MET', 'ZNF136_MET', 'USP14_MET', 'MMP9_MET', 'SEMG2_MET', 'ZBTB3_MET', 'PTBP1_MET', 'PRF1_MET', 'RAMP3_MET', 'C11orf65_MET', 'KIAA0125_MET', 'FAM186B_MET', 'HIST1H4I;HIST1H2BK_MET', 'TXK_MET', 'WDR82_MET', 'CYP24A1_MET', 'FLNC_MET', 'TMEM22_MET', 'ZNF432_MET', 'ARHGEF16_MET', 'KRT12_MET', 'SOX9_MET', 'FAM107B_MET', 'TBX1_MET', 'AP1G2_MET', 'SLC25A22_MET', 'CLIP1_MET', 'FOXA1_MET', 'FANCE_MET', 'SAV1_MET', 'TSSC1_MET', 'CCDC114_MET', 'TOR1AIP1_MET', 'ASCC1;C10orf104_MET', 'S100A9_MET', 'UBXN10_MET', 'MEGF8_MET', 'RAD54L_MET', 'MCAM_MET', 'C14orf4_MET', 'ZNF441_MET', 'USP44_MET', 'GJB6_MET', 'GYG2_MET', 'NTSR1_MET', 'LOC100233209;FAM113B_MET', 'KIR3DX1_MET', 'ZAK_MET', 'LOC100129794;SLC25A21_MET', 'GRTP1_MET', 'ICAM4_MET',

'CDH24_MET', 'LGALS7_MET', 'ZCCHC5_MET', 'OR10A4_MET', 'QRFP_MET', 'ALAS1_MET', 'C1QB_MET', 'FLVCR2_MET', 'GPR148_MET', 'C9orf171_MET', 'TCN2_MET', 'SPAG8_MET', 'CALCOCO2_MET', 'ASF1A_MET', 'CAPN7;DVWA_MET', 'CHST3_MET', 'ETV7_MET', 'PRIC285_MET', 'ZNF558_MET', 'PNKP_MET', 'VPS53_MET', 'GPS1;RFNG_MET', 'ZNF501_MET', 'STAP1_MET', 'ANO7_MET', 'PFN4;FAM228B_MET', 'GPR126_MET', 'GGT6_MET', 'SLC25A40;DBF4_MET', 'EIF4G1_MET', 'CD1B_MET', 'PROCA1_MET', 'PHF15_MET', 'RDBP;SKIV2L_MET', 'ACAD11;UBA5_MET', 'TMEM59;C1orf83_MET', 'RWDD2A;PGM3_MET', 'CAND1_MET', 'MCCC1_MET', 'C9orf131_MET', 'B4GALT4_MET', 'ETV4_MET', 'SLC6A4_MET', 'HOXC8_MET', 'TMEM45B_MET', 'PGF_MET', 'RAB39_MET', 'MIR548F5;MAB21L1;NBEA_MET', 'WDR90;RHOT2_MET', 'SYN1_MET', 'TTC26_MET', 'TM9SF1_MET', 'TMBIM6_MET', 'STOML3_MET', 'ZNF721;ABCA11P_MET', 'SP140_MET', 'C3orf52_MET', 'HOXC12_MET', 'CD69_MET', 'CXCL3_MET', 'SLFNL1_MET', 'PDLIM2_MET', 'CCL22_MET', 'SEPT5;GP1BB_MET', 'CGNL1_MET', 'LUC7L3_MET', 'C1orf150_MET', 'FYTTD1;KIAA0226_MET', 'PSME2;RNF31_MET', 'CHRDL1_MET', 'EID2_MET', 'PRELP_MET', 'SLC39A9;ERH_MET', 'CORO6_MET', 'UBE2D1_MET', 'CLDN3_MET', 'NFE2L2_MET', 'HP_MET', 'SPTLC2_MET', 'RIF1_MET', 'ST6GALNAC5_MET', 'FLJ45983;GATA3_MET', 'GLTP_MET', 'SLC22A2_MET', 'STIM2_MET', 'CELSR1_MET', 'KCNS1_MET', 'FAM129C_MET', 'PDCD5_MET', 'KRT23_MET', 'SORD_MET', 'NCOA6_MET', 'ZNF549_MET', 'GATA2_MET', 'LYZL6_MET', 'EIF4A3_MET', 'MLF1IP_MET', 'THOP1_MET', 'FAIM3_MET', 'NR0B2_MET', 'GNAQ_MET', 'TMEM126A_MET', 'TESC_MET', 'DCUN1D1_MET', 'IL24_MET', 'TRPM2_MET', 'ACSF2;CHAD_MET', 'CAB39_MET', 'LINC02914_MET', 'BUB1_MET', 'KRTAP11-1_MET', 'MTOR_MET', 'OAZ2_MET', 'NRAS;CSDE1_MET', 'CISH_MET', 'LOC144486;CCDC41_MET', 'SYNPO2L_MET', 'NXPH4_MET', 'KLHL23_MET', 'SLC22A3_MET', 'HOOK1_MET', 'SFXN1_MET', 'HYOU1_MET', 'ANXA9_MET', 'CD6_MET', 'PON1_MET', 'AP3M2_MET', 'ANGPTL6_MET', 'RGN_MET', 'BAK1_MET', 'CD3D_MET', 'ALDH3B1_MET', 'SYT17_MET', 'TOR1AIP2_MET', 'GPRC5C_MET', 'YIPF3;POLR1C_MET', 'IL13_MET', 'ESRRB_MET', 'OSBPL5_MET', 'IRAK3_MET', 'BACH2_MET', 'TFDP3_MET', 'DMBX1_MET', 'NNT_MET', 'C17orf57_MET', 'TTLL6_MET', 'ABHD8_MET', 'PDZD8_MET', 'G0S2_MET', 'TMEM139_MET', 'UCHL1_MET', 'BDKRB2_MET', 'SPARC_MET', 'TMCC2_MET', 'ZNF646;ZNF668_MET', 'FEM1C_MET', 'ZNF264_MET', 'KDM4A_MET', 'FHOD3_MET', 'RARRES1_MET', 'POM121L1P_MET', 'ALX3_MET',

'CMTM6_MET', 'CLIC5_MET', 'HELZ_MET', 'IFITM2_MET', 'HIST2H2AC;HIST2H2BE_MET', 'LPL_MET', 'GATC;TRIAP1_MET', 'TBC1D7_MET', 'APOE_MET', 'DMXL2_MET', 'SLC12A5_MET', 'ESPNL_MET', 'SLC39A4_MET', 'PPM1F_MET', 'RAD51L1_MET', 'C3orf14_MET', 'SH2D4A_MET', 'C9orf89_MET', 'POLS_MET', 'HPR_MET', 'CBS_MET', 'C1orf107_MET', 'SELE_MET', 'PLXNB1_MET', 'ABCD3_MET', 'ALOX15_MET', 'WFDC10B;WFDC13_MET', 'PRR5-ARHGAP8;ARHGAP8_MET', 'ZNF567_MET', 'WIPF1_MET', 'ZNF691_MET', 'SORL1_MET', 'EPS15_MET', 'BRAF_MET', 'RAB11FIP4_MET', 'TICAM1_MET', 'ZNF354A_MET', 'DYRK1B_MET', 'CCL5_MET', 'RASEF_MET', 'RAB6A_MET', 'LAPTM4A_MET', 'ALDH2_MET', 'CYP1A1_MET', 'KIAA0586;TIMM9_MET', 'FAM3C_MET', 'ZBTB7A_MET', 'C4orf7_MET', 'MMP14_MET', 'MRPL18;TCP1_MET', 'MIPOL1_MET', 'POU2AF1_MET', 'GPR89B_MET', 'PCDHB15_MET', 'BCAT1_MET', 'PHEX_MET', 'FAM57A_MET', 'PAPPA2_MET', 'FXYD2_MET', 'AIM2_MET', 'RHOBTB3_MET', 'SEMA3C_MET', 'ARAP3_MET', 'TCF3_MET', 'CCDC64_MET', 'IL18_MET', 'DNAJC14_MET', 'TARS2;ECM1_MET', 'HPX_MET', 'RNF113B;FARP1_MET', 'TNNI1_MET', 'SSH1_MET', 'MRPL35_MET', 'WBP2NL_MET', 'PLA2G2F_MET', 'DHRS11_MET', 'ERBB2_MET', 'AZGP1_MET', 'EFHD1_MET', 'CYTIP_MET', 'ADM_MET', 'TRIM42_MET', 'C3orf47;H1FX_MET', 'LOC145783;TCF12_MET', 'TNIP1_MET', 'PRG4;MIR548F1_MET', 'TUBGCP4;ZSCAN29_MET', 'SLC23A2_MET', 'MIR639;TECR_MET', 'SNORD101;RPS12_MET', 'STARD4_MET', 'CSHL1_MET', 'GRSF1_MET', 'ESPN_MET', 'ZNF566_MET', 'KCNK7_MET', 'ILF2_MET', 'PCMTD2_MET', 'MAEL_MET', 'ALDH16A1;PIH1D1_MET', 'KRT13_MET', 'PRR4;PRH1_MET', 'CDK8_MET', 'MGC29506_MET', 'CCR7_MET', 'NUDT16_MET', 'GPR55_MET', 'R3HDM1;ZRANB3_MET', 'RIOK1;CAGE1_MET', 'TIGIT_MET', 'JUP_MET', 'IL9_MET', 'GAS7_MET', 'C14orf138_MET', 'L3MBTL4_MET', 'RASGRP2_MET', 'MATN1_MET', 'ZNF546_MET', 'CDH16_MET', 'C14orf104_MET', 'ZNF230_MET', 'IRS2_MET', 'KLHDC7A_MET', 'TMOD1_MET', 'LPAR1_MET', 'ATP6V1C2_MET', 'HIST1H1A_MET', 'NAT15;ZNF597_MET', 'LRGUK_MET', 'HIST3H2A_MET', 'LYL1_MET', 'BRS3_MET', 'POLE4_MET', 'PDZK1IP1_MET', 'ATP5O_MET', 'GGT5_MET', 'TALDO1_MET', 'SIT1_MET', 'TGFBR1_MET', 'ZC3H15_MET', 'IAPP;SLCO1A2_MET', 'TUSC4;CYB561D2_MET', 'CCDC138_MET', 'SLC25A34_MET', 'ZNF71_MET', 'MTHFD2_MET', 'PRLHR_MET', 'ZNF660_MET', 'LTBR_MET', 'SCG5_MET', 'NOD1_MET', 'CD3G_MET', 'ABAT_MET', 'ACCS_MET', 'KIAA1199_MET', 'PRICKLE1_MET', 'ACIN1;C14orf119_MET',

'SF3B2_MET', 'WNT6_MET', 'MFAP1_MET', 'CSF2RB_MET', 'DENND3_MET', 'MAP1D_MET', 'SLC38A1_MET', 'TSEN34;MBOAT7_MET', 'WHSC2_MET', 'AHCYL2_MET', 'PTPN14_MET', 'UROD;HECTD3_MET', 'CYP4X1_MET', 'TMEM119_MET', 'ZDHHC8_MET', 'EDEM1_MET', 'FAM38A_MET', 'LIN54_MET', 'SOCS4;WDHD1_MET', 'TPH1_MET', 'SMC4;IFT80_MET', 'SLC25A15_MET', 'LCK_MET', 'GYPC_MET', 'MRPL45_MET', 'NINJ1_MET', 'MS4A3_MET', 'C10orf11_MET', 'FGFR3_MET', 'GLS_MET', 'SHBG_MET', 'AIRE_MET', 'RRP15_MET', 'GGT3P_MET', 'ST13;XPNPEP3_MET', 'NPC1L1_MET', 'EVL_MET', 'C15orf32_MET', 'IFI44L_MET', 'HS3ST1_MET', 'SERPINI1_MET', 'PYGL_MET', 'TNFRSF13B_MET', 'IL32_MET', 'ORAI3_MET', 'MARVELD3_MET', 'PRSS3_MET', 'ACAP3_MET', 'DDX27_MET', 'GPR160_MET', 'MYO1E;LDHAL6B_MET', 'ZNF498_MET', 'PDE6A_MET', 'RBM4B_MET', 'SOX21_MET', 'ZNF187_MET', 'GPR162_MET', 'FAM10A4_MET', 'C1orf220_MET', 'FBRS_MET', 'PYY2_MET', 'SLC34A2_MET', 'PBOV1;KIAA1244_MET', 'GGTLC1_MET', 'RGS3_MET', 'B3GALT6;SDF4_MET', 'IFNAR2_MET', 'DDO_MET', 'GPR124_MET', 'CORO2B_MET', 'RNF217_MET', 'ZNF645_MET', 'CLEC9A_MET', 'C5orf40;CYFIP2_MET', 'PPBP_MET', 'C17orf88_MET', 'CLDN11_MET', 'RHOQ_MET', 'HOXC13_MET', 'WIPI2_MET', 'MLX_MET', 'PTPRCAP_MET', 'CARTPT_MET', 'DPCR1_MET', 'PIK3CD_MET', 'TCIRG1_MET', 'GLIPR1L1_MET', 'AGR2_MET', 'HOXC9_MET', 'RGS7_MET', 'CD320_MET', 'CENPA_MET', 'CHRNB4_MET', 'KCTD4;GTF2F2_MET', 'BEX1_MET', 'MTHFD1L_MET', 'SLC22A13_MET', 'C13orf28_MET', 'HADH_MET', 'LCE2B_MET', 'PAK2_MET', 'FAM100A_MET', 'FGF6_MET', 'C10orf91_MET', 'ZNF197_MET', 'CLCC1_MET', 'ZADH2_MET', 'CD27;TAPBPL;LOC678655_MET', 'MS4A2_MET', 'SIAE;SPA17_MET', 'ANKRD12_MET', 'P4HA2_MET', 'RGS4_MET', 'C16orf67_MET', 'SPTA1_MET', 'DIRAS1_MET', 'BTN2A1_MET', 'HIVEP2_MET', 'UNC84A_MET', 'DCAF12L1_MET', 'BAG2_MET', 'RINL_MET', 'SETD2_MET', 'C16orf54_MET', 'SIK3_MET', 'GPR125_MET', 'CYP4B1_MET', 'GSTM3_MET', 'C15orf27_MET', 'C19orf2_MET', 'SMC1B;RIBC2_MET', 'ARID4B;GGPS1_MET', 'ELOF1_MET', 'GPRC5D_MET', 'ZFP28_MET', 'SLPI_MET', 'KLK12_MET', 'DFNB31_MET', 'NACC1;TRMT1_MET', 'CPEB4_MET', 'MAPK13_MET', 'TNS3_MET', 'GALNT5_MET', 'CD3E_MET', 'SERPINA7_MET', 'CD5_MET', 'ITFG1;PHKB_MET', 'AGAP1_MET', 'DLK2_MET', 'ZNF193_MET', 'APOC4_MET', 'MCOLN1;PNPLA6_MET', 'CFL2_MET', 'LMF2;NCAPH2_MET', 'ANKRD23_MET', 'FSCN1_MET', 'PLOD2_MET', 'SYT11_MET', 'TNK1_MET', 'CLPTM1L_MET',

'TMEM106C_MET', 'ISLR2;LOC283731_MET', 'ZNF536_MET', 'LRRC39_MET', 'RAB18_MET', 'PARVA_MET', 'SH2D1B_MET', 'PDE1A_MET', 'RPS11_MET', 'GPR132_MET', 'THOC1_MET', 'MAZ_MET', 'OR2L13_MET', 'C14orf39_MET', 'CHST12_MET', 'SPAG1_MET', 'CDC42EP1_MET']

Selected Clinical Features: ['tobacco_smoking_history_clinical', 'TMB_NONSYNONYMOUS_clinical', 'AGE_clinical', 'GENETIC_ANCESTRY_LABEL_ENCODED_clinical', 'PATH_T_STAGE_ENCODED_clinical', 'PATH_N_STAGE_ENCODED_clinical', 'PATH_M_STAGE_ENCODED_clinical', 'AJCC_PATHOLOGIC_TUMOR_STAGE_ENCODED_clinical', 'SEX_ENCODED_clinical', 'TUMOR_TYPE_ENCODED_clinical']

Appendix B: Selected time-based features

Selected Expression Features: ['ABHD13:84945_mRNA', 'ABHD8:79575_mRNA', 'ABI3BP:25890_mRNA', 'ACAT2:39_mRNA', 'ACER1:125981_mRNA', 'ACOT12:134526_mRNA', 'ACSM2B:348158_mRNA', 'ACSM3:6296_mRNA', 'ACSS2:55902_mRNA', 'ADAMTS4:9507_mRNA', 'ADAP1:11033_mRNA', 'ADD3:120_mRNA', 'ADM2:79924_mRNA', 'ADRBK1:156_mRNA', 'AGRN:375790_mRNA', 'AIDA:64853_mRNA', 'AKAP7:9465_mRNA', 'ALB:213_mRNA', 'ALDH1L2:160428_mRNA', 'AMELY:266_mRNA', 'AMIGO3:386724_mRNA', 'AMPD2:271_mRNA', 'ANGPTL4:51129_mRNA', 'ANKLE2:23141_mRNA', 'ANKRD13D:338692_mRNA', 'ANO8:57719_mRNA', 'ANXA8L1:728113_mRNA', 'AP1M1:8907_mRNA', 'AP1S3:130340_mRNA', 'AP2A1:160_mRNA', 'APOBEC3C:27350_mRNA', 'APOBEC3D:140564_mRNA', 'APOBEC3F:200316_mRNA', 'APOBEC3G:60489_mRNA', 'ARF1:375_mRNA', 'ARFGAP1:55738_mRNA', 'ARFIP1:27236_mRNA', 'ARHGAP22:58504_mRNA', 'ARHGAP4:393_mRNA', 'ARHGEF10:9639_mRNA', 'ARHGEF18:23370_mRNA', 'ARHGEF4:50649_mRNA', 'ARPP21:10777_mRNA', 'ASB4:51666_mRNA', 'ASB7:140460_mRNA', 'ASS1:445_mRNA', 'ATG3:64422_mRNA', 'ATP10D:57205_mRNA', 'ATP13A1:57130_mRNA', 'ATP6AP1:537_mRNA', 'ATP6V1H:51606_mRNA', 'AUH:549_mRNA', 'AURKAPS1:6791_mRNA', 'AZGP1:563_mRNA', 'B4GALT1:2683_mRNA', 'B4GALT2:8704_mRNA', 'BAI3:577_mRNA', 'BANF2:140836_mRNA', 'BCL3:602_mRNA', 'BCMO1:53630_mRNA', 'BET3L:100128327_mRNA', 'BLID:414899_mRNA', 'BMP1:649_mRNA', 'BMPR1B:658_mRNA', 'BOD1:91272_mRNA', 'BRPF1:7862_mRNA', 'BRS3:680_mRNA', 'BRWD1:54014_mRNA', 'BSG:682_mRNA', 'BTBD2:55643_mRNA', 'C11orf24:53838_mRNA', 'C11orf46:120534_mRNA', 'C11orf54:28970_mRNA', 'C11orf83:790955_mRNA', 'C12orf40:283461_mRNA', 'C12orf44:60673_mRNA', 'C13orf39:196541_mRNA', 'C14orf128:84837_mRNA', 'C15orf52:388115_mRNA', 'C16orf3:750_mRNA', 'C16orf52:730094_mRNA', 'C16orf5:29965_mRNA', 'C16orf57:79650_mRNA', 'C16orf63:123811_mRNA', 'C16orf7:9605_mRNA', 'C17orf107:100130311_mRNA', 'C17orf39:79018_mRNA', 'C17orf97:400566_mRNA', 'C19orf21:126353_mRNA', 'C19orf28:126321_mRNA', 'C19orf50:79036_mRNA', 'C19orf52:90580_mRNA', 'C19orf63:284361_mRNA', 'C19orf6:91304_mRNA', 'C19orf71:100128569_mRNA', 'C1orf203:84852_mRNA', 'C1QTNF9:338872_mRNA', 'C20orf132:140699_mRNA', 'C20orf135:140701_mRNA', 'C20orf152:140894_mRNA', 'C21orf34:388815_mRNA', 'C21orf84:114038_mRNA', 'C22orf33:339669_mRNA',

'C22orf43:51233_mRNA', 'C22orf9:23313_mRNA', 'C2CD4D:100191040_mRNA', 'C3orf17:25871_mRNA', 'C3orf23:285343_mRNA', 'C3orf49:132200_mRNA', 'C4orf32:132720_mRNA', 'C4orf33:132321_mRNA', 'C5orf41:153222_mRNA', 'C6orf126:389383_mRNA', 'C6orf204:387119_mRNA', 'C6orf211:79624_mRNA', 'C6orf57:135154_mRNA', 'C6orf72:116254_mRNA', 'C7orf27:221927_mRNA', 'C7orf28B:221960_mRNA', 'C7orf50:84310_mRNA', 'C8orf33:65265_mRNA', 'C8orf37:157657_mRNA', 'C8orf58:541565_mRNA', 'C8orf85:441376_mRNA', 'C9orf140:89958_mRNA', 'C9orf172:389813_mRNA', 'C9orf21:195827_mRNA', 'C9orf70:84850_mRNA', 'CACNB3:784_mRNA', 'CACNB4:785_mRNA', 'CADPS:8618_mRNA', 'CALCOCO2:10241_mRNA', 'CALM3:808_mRNA', 'CALR:811_mRNA', 'CAPN1:823_mRNA', 'CARM1:10498_mRNA', 'CASP12:120329_mRNA', 'CAT:847_mRNA', 'CBX4:8535_mRNA', 'CCDC117:150275_mRNA', 'CCDC123:84902_mRNA', 'CCDC124:115098_mRNA', 'CCDC150:284992_mRNA', 'CCDC152:100129792_mRNA', 'CCDC28A:25901_mRNA', 'CCDC53:51019_mRNA', 'CCDC61:729440_mRNA', 'CCDC78:124093_mRNA', 'CCDC79:283847_mRNA', 'CCDC9:26093_mRNA', 'CCL15-CCL14:348249_mRNA', 'CCL27:10850_mRNA', 'CCM2:83605_mRNA', 'CCND1:595_mRNA', 'CCNO:10309_mRNA', 'CD3EAP:10849_mRNA', 'CDC37:11140_mRNA', 'CDC42EP5:148170_mRNA', 'CDC5L:988_mRNA', 'CDCA7:83879_mRNA', 'CENPF:1063_mRNA', 'CEP170:9859_mRNA', 'CERCAM:51148_mRNA', 'CFL1:1072_mRNA', 'CHI3L2:1117_mRNA', 'CHMP7:91782_mRNA', 'CHPF2:54480_mRNA', 'CHPF:79586_mRNA', 'CHRDL1:91851_mRNA', 'CHRNE:1145_mRNA', 'CHST3:9469_mRNA', 'CIDEB:27141_mRNA', 'CLDN15:24146_mRNA', 'CLEC11A:6320_mRNA', 'CLEC18B:497190_mRNA', 'CLPTM1:1209_mRNA', 'CLPX:10845_mRNA', 'CLSTN3:9746_mRNA', 'CLYBL:171425_mRNA', 'CNOT3:4849_mRNA', 'COL1A1:1277_mRNA', 'COL5A3:50509_mRNA', 'COL6A1:1291_mRNA', 'COL9A1:1297_mRNA', 'COL9A2:1298_mRNA', 'COPE:11316_mRNA', 'CORO1B:57175_mRNA', 'COX11:1353_mRNA', 'COX19:90639_mRNA', 'COX6B2:125965_mRNA', 'CPA2:1358_mRNA', 'CPNE3:8895_mRNA', 'CPO:130749_mRNA', 'CPT1A:1374_mRNA', 'CPXM1:56265_mRNA', 'CRELD1:78987_mRNA', 'CRELD2:79174_mRNA', 'CRYZL1:9946_mRNA', 'CSNK1A1P:161635_mRNA', 'CSNK1E:1454_mRNA', 'CSNK1G2:1455_mRNA', 'CST9:128822_mRNA', 'CT47B1:643311_mRNA', 'CTDSPL2:51496_mRNA', 'CTNNA2:1496_mRNA', 'CTTN:2017_mRNA', 'CXCL6:6372_mRNA', 'CXorf57:55086_mRNA', 'CYCSP52:360155_mRNA', 'CYP2A6:1548_mRNA', 'CYP2A7:1549_mRNA',

'CYP39A1:51302_mRNA', 'CYP7A1:1581_mRNA', 'CYTH2:9266_mRNA', 'D4S234E:27065_mRNA', 'DAGLB:221955_mRNA', 'DAPK3:1613_mRNA', 'DAZAP1:26528_mRNA', 'DBN1:1627_mRNA', 'DBNDD1:79007_mRNA', 'DCAF10:79269_mRNA', 'DDA1:79016_mRNA', 'DDB2:1643_mRNA', 'DDN:23109_mRNA', 'DDX4:54514_mRNA', 'DEK:7913_mRNA', 'DGCR9:26220_mRNA', 'DHX34:9704_mRNA', 'DHX40:79665_mRNA', 'DIO1:1733_mRNA', 'DIRAS2:54769_mRNA', 'DIS3L:115752_mRNA', 'LOC93429:93429_mRNA', 'DMBT1:1755_mRNA', 'DMPK:1760_mRNA', 'DMRT2:10655_mRNA', 'DMWD:1762_mRNA', 'DNAJB11:51726_mRNA', 'DNAJB13:374407_mRNA', 'DNAJC18:202052_mRNA', 'DNMT1:1786_mRNA', 'DNTT:1791_mRNA', 'DOCK5:80005_mRNA', 'DOHH:83475_mRNA', 'DOK7:285489_mRNA', 'DOPEY1:23033_mRNA', 'DPP9:91039_mRNA', 'DRP2:1821_mRNA', 'DSCR3:10311_mRNA', 'DTNBP1:84062_mRNA', 'DVL1:1855_mRNA', 'DYNLL1:8655_mRNA', 'DYRK1B:9149_mRNA', 'ECHDC3:79746_mRNA', 'EDN3:1908_mRNA', 'EIF1B:10289_mRNA', 'EIF2AK1:27102_mRNA', 'EIF3B:8662_mRNA', 'ELF1:1997_mRNA', 'ELF2:1998_mRNA', 'ELL:8178_mRNA', 'EML5:161436_mRNA', 'ENTPD4:9583_mRNA', 'EPC2:26122_mRNA', 'EPM2A:7957_mRNA', 'EPN1:29924_mRNA', 'EPS15L1:58513_mRNA', 'EPS8L1:54869_mRNA', 'ERCC2:2068_mRNA', 'ERICH1:157697_mRNA', 'ESD:2098_mRNA', 'EXOC3L:283849_mRNA', 'F2RL1:2150_mRNA', 'F3:2152_mRNA', 'FADD:8772_mRNA', 'FAHD2A:51011_mRNA', 'FAM108B1:51104_mRNA', 'FAM122A:116224_mRNA', 'FAM122C:159091_mRNA', 'FAM138D:677784_mRNA', 'FAM154A:158297_mRNA', 'FAM160B2:64760_mRNA', 'FAM171A2:284069_mRNA', 'FAM178A:55719_mRNA', 'FAM196A:642938_mRNA', 'FAM198A:729085_mRNA', 'FAM27B:100133121_mRNA', 'FAM47A:158724_mRNA', 'FAM47B:170062_mRNA', 'FAM72B:653820_mRNA', 'FAM82A1:151393_mRNA', 'FAM83E:54854_mRNA', 'FAM8A1:51439_mRNA', 'FAM98A:25940_mRNA', 'FAM9B:171483_mRNA', 'FANCE:2178_mRNA', 'FASTK:10922_mRNA', 'FASTKD3:79072_mRNA', 'FBXL3:26224_mRNA', 'FBXO3:26273_mRNA', 'FBXO48:554251_mRNA', 'FEM1B:10116_mRNA', 'FGF19:9965_mRNA', 'FITM2:128486_mRNA', 'FIZ1:84922_mRNA', 'FKBP14:55033_mRNA', 'FKBP2:2286_mRNA', 'FKBP8:23770_mRNA', 'FKRP:79147_mRNA', 'FLJ16779:100192386_mRNA', 'QSOX1:200058_mRNA', 'FLJ25758:497049_mRNA', 'FLJ42393:401105_mRNA', 'FLNC:2318_mRNA', 'FMO5:2330_mRNA', 'FOXD3:27022_mRNA',

'FOXE3:2301_mRNA', 'FOXR2:139628_mRNA', 'FREM2:341640_mRNA', 'FRG2C:100288801_mRNA', 'FSCN1:6624_mRNA', 'FSTL3:10272_mRNA', 'FUT2:2524_mRNA', 'FUT3:2525_mRNA', 'FXC1:26515_mRNA', 'GAA:2548_mRNA', 'GAB1:2549_mRNA', 'GABPA:2551_mRNA', 'GABRP:2568_mRNA', 'GADD45B:4616_mRNA', 'GAK:2580_mRNA', 'GALNT2:2590_mRNA', 'GARS:2617_mRNA', 'GAS2:2620_mRNA', 'GAS2L1:10634_mRNA', 'GAST:2520_mRNA', 'GATAD2A:54815_mRNA', 'GDAP1L1:78997_mRNA', 'GDI1:2664_mRNA', 'GET4:51608_mRNA', 'GIPC1:10755_mRNA', 'GJB7:375519_mRNA', 'GJC3:349149_mRNA', 'GLT25D1:79709_mRNA', 'GLTSCR1:29998_mRNA', 'GLYATL2:219970_mRNA', 'GMIP:51291_mRNA', 'GMPPA:29926_mRNA', 'GMPR:2766_mRNA', 'GNB3:2784_mRNA', 'GNPNAT1:64841_mRNA', 'GNRHR:2798_mRNA', 'GOLGA3:2802_mRNA', 'GOLGA6D:653643_mRNA', 'GOLGA6L1:283767_mRNA', 'GOLGA6L6:727832_mRNA', 'GOLGA9P:283796_mRNA', 'GPI:2821_mRNA', 'GPLD1:2822_mRNA', 'GPR150:285601_mRNA', 'GPR172B:55065_mRNA', 'GPR174:84636_mRNA', 'GPR4:2828_mRNA', 'GPR77:27202_mRNA', 'GRAMD4:23151_mRNA', 'GRIK2:2898_mRNA', 'GSK3A:2931_mRNA', 'GTF2F1:2962_mRNA', 'GTF2IRD1:9569_mRNA', 'GTPBP3:84705_mRNA', 'GTPBP8:29083_mRNA', 'GUCY1A3:2982_mRNA', 'HADH:3033_mRNA', 'HAGHL:84264_mRNA', 'HARBI1:283254_mRNA', 'HBP1:26959_mRNA', 'HDGFL2:84717_mRNA', 'HEATR2:54919_mRNA', 'HELQ:113510_mRNA', 'HEPACAM:220296_mRNA', 'HES6:55502_mRNA', 'HIST1H2BH:8345_mRNA', 'HIST1H2BI:8346_mRNA', 'HIST1H2BM:8342_mRNA', 'HIST1H3C:8352_mRNA', 'HMG20A:10363_mRNA', 'HNF4G:3174_mRNA', 'HNRNPUL1:11100_mRNA', 'HNRPLL:92906_mRNA', 'HOMER3:9454_mRNA', 'HOXC5:3222_mRNA', 'HPSE2:60495_mRNA', 'HSD3B2:3284_mRNA', 'HSDL2:84263_mRNA', 'HSPA1A:3303_mRNA', 'HSPA7:3311_mRNA', 'HSPB3:8988_mRNA', 'HSPBP1:23640_mRNA', 'IBSP:3381_mRNA', 'ICA1:3382_mRNA', 'ID1:3397_mRNA', 'IFITM5:387733_mRNA', 'IGFL4:444882_mRNA', 'IKZF5:64376_mRNA', 'IL3:3562_mRNA', 'INSR:3643_mRNA', 'INTS1:26173_mRNA', 'IQCE:23288_mRNA', 'IQCF6:440956_mRNA', 'IQGAP2:10788_mRNA', 'IRF2BP1:26145_mRNA', 'IRF3:3661_mRNA', 'IRGC:56269_mRNA', 'ISOC2:79763_mRNA', 'JAG2:3714_mRNA', 'JUND:3727_mRNA', 'JUP:3728_mRNA', 'KALRN:8997_mRNA', 'KATNA1:11104_mRNA', 'KBTBD10:10324_mRNA', 'KBTBD5:131377_mRNA', 'KCNG4:93107_mRNA', 'KCNJ14:3770_mRNA', 'KCNJ3:3760_mRNA', 'KCNK16:83795_mRNA', 'KCNMB4:27345_mRNA',

'KCNS1:3787_mRNA', 'KCP:375616_mRNA', 'KDELR1:10945_mRNA', 'KDELR2:11014_mRNA', 'KDM4D:55693_mRNA', 'KDM6A:7403_mRNA', 'KIAA0415:9907_mRNA', 'KIAA0892:23383_mRNA', 'KIAA1586:57691_mRNA', 'KIAA1704:55425_mRNA', 'KIAA1967:57805_mRNA', 'KIFC3:3801_mRNA', 'KLC2:64837_mRNA', 'KLF16:83855_mRNA', 'KLHL14:57565_mRNA', 'KLHL17:339451_mRNA', 'KLHL23:151230_mRNA', 'KLHL26:55295_mRNA', 'KLK9:284366_mRNA', 'KLRG1:10219_mRNA', 'KRI1:65095_mRNA', 'KRT7:3855_mRNA', 'KRT80:144501_mRNA', 'KRTAP5-6:440023_mRNA', 'LAMA5:3911_mRNA', 'LAMB2:3913_mRNA', 'LASS4:79603_mRNA', 'LCE1B:353132_mRNA', 'LEFTY1:10637_mRNA', 'LENG9:94059_mRNA', 'LEPRE1:64175_mRNA', 'LEPREL2:10536_mRNA', 'LGALS3BP:3959_mRNA', 'LGR6:59352_mRNA', 'LIN52:91750_mRNA', 'LIN7A:8825_mRNA', 'LINGO1:84894_mRNA', 'LIPE:3991_mRNA', 'LMBRD1:55788_mRNA', 'LMNB2:84823_mRNA', 'C2CD4D-AS1:100132111_mRNA', 'TEKT4P2:100132288_mRNA', 'LOC100133469:100133469_mRNA', 'LINC00654:149837_mRNA', 'GNAO1-DT:283856_mRNA', 'TMEM235:283999_mRNA', 'LOC284100:284100_mRNA', 'LOC285696:285696_mRNA', 'LOC392196:392196_mRNA', 'LINC02817:400804_mRNA', 'LOC402377:402377_mRNA', 'LOC407835:407835_mRNA', 'PTGER4P2-CDK2AP2P2:442421_mRNA', 'SMG1P5:595101_mRNA', 'LOC653501:653501_mRNA', 'AFAP1-AS1:84740_mRNA', 'LONP1:9361_mRNA', 'LOXL2:4017_mRNA', 'LPA:4018_mRNA', 'LPAR2:9170_mRNA', 'LPPR4:9890_mRNA', 'LRCH2:57631_mRNA', 'LRCH4:4034_mRNA', 'LRFN4:78999_mRNA', 'LRPAP1:4043_mRNA', 'LRRC28:123355_mRNA', 'LRRC37A3:374819_mRNA', 'LSM6:11157_mRNA', 'LSR:51599_mRNA', 'LTA:4049_mRNA', 'LUZP2:338645_mRNA', 'LYSMD2:256586_mRNA', 'LZIC:84328_mRNA', 'MAD1L1:8379_mRNA', 'MAD2L1BP:9587_mRNA', 'MAFA:389692_mRNA', 'MAFK:7975_mRNA', 'MAGEC1:9947_mRNA', 'MAN2B1:4125_mRNA', 'MAN2B2:23324_mRNA', 'MAP1S:55201_mRNA', 'MAP2K2:5605_mRNA', 'MAP3K10:4294_mRNA', 'MAP4:4134_mRNA', 'MAP7D1:55700_mRNA', 'MAPK12:6300_mRNA', 'MAPK4:5596_mRNA', 'MARCH7:64844_mRNA', 'MARK4:57787_mRNA', 'MAT2B:27430_mRNA', 'MBOAT7:79143_mRNA', 'MCART6:401612_mRNA', 'MCOLN2:255231_mRNA', 'MCOLN3:55283_mRNA', 'MED18:54797_mRNA', 'MED26:9441_mRNA', 'MED30:90390_mRNA', 'MEOX1:4222_mRNA', 'METRNL:284207_mRNA', 'MEX3D:399664_mRNA', 'MFSD8:256471_mRNA', 'MGC12982:84793_mRNA', 'MGP:4256_mRNA',

'MGST2:4258_mRNA', 'MICALL2:79778_mRNA', 'MIER2:54531_mRNA', 'MINPP1:9562_mRNA', 'MITD1:129531_mRNA', 'MMAA:166785_mRNA', 'MOBKL2A:126308_mRNA', 'MOGS:7841_mRNA', 'MORC3:23515_mRNA', 'MOSC1:64757_mRNA', 'MPV17L2:84769_mRNA', 'MRPS18B:28973_mRNA', 'MSLN:10232_mRNA', 'MT1X:4501_mRNA', 'MT2A:4502_mRNA', 'MTMR10:54893_mRNA', 'MTMR12:54545_mRNA', 'MTRR:4552_mRNA', 'MTSS1L:92154_mRNA', 'MTTP:4547_mRNA', 'MUC20:200958_mRNA', 'MUC4:4585_mRNA', 'MUT:4594_mRNA', 'MUTED:63915_mRNA', 'MYEOV:26579_mRNA', 'MYH14:79784_mRNA', 'MYH16:84176_mRNA', 'MYH2:4620_mRNA', 'MYLK4:340156_mRNA', 'MYO9B:4650_mRNA', 'MYPOP:339344_mRNA', 'MYT1:4661_mRNA', 'N4BP2L1:90634_mRNA', 'NACAD:23148_mRNA', 'NACC1:112939_mRNA', 'NAV1:89796_mRNA', 'NCLN:56926_mRNA', 'NCRNA00051:619434_mRNA', 'NCRNA00158:54072_mRNA', 'NLGN2:57555_mRNA', 'NLRP8:126205_mRNA', 'NMBR:4829_mRNA', 'NME4:4833_mRNA', 'NMNAT1:64802_mRNA', 'NOC3L:64318_mRNA', 'NR1H3:10062_mRNA', 'NR2F6:2063_mRNA', 'NRN1:51299_mRNA', 'NUDT1:4521_mRNA', 'NUDT18:79873_mRNA', 'NUMBL:9253_mRNA', 'NUP153:9972_mRNA', 'NUP160:23279_mRNA', 'NUP43:348995_mRNA', 'NUP54:53371_mRNA', 'NUP62:23636_mRNA', 'OGDH:4967_mRNA', 'OLAH:55301_mRNA', 'OLFM1:10439_mRNA', 'OR10H2:26538_mRNA', 'OR10P1:121130_mRNA', 'OR13G1:441933_mRNA', 'OR13H1:347468_mRNA', 'OR1D4:8385_mRNA', 'OR2L13:284521_mRNA', 'OR4A16:81327_mRNA', 'OR4A47:403253_mRNA', 'OR51Q1:390061_mRNA', 'OR5B12:390191_mRNA', 'OR5P3:120066_mRNA', 'OS9:10956_mRNA', 'OSBPL5:114879_mRNA', 'OVGP1:5016_mRNA', 'OXCT1:5019_mRNA', 'OXR1:55074_mRNA', 'P2RY11:5032_mRNA', 'P4HA2:8974_mRNA', 'P4HB:5034_mRNA', 'PAFAH1B3:5050_mRNA', 'PAGE5:90737_mRNA', 'PAQR8:85315_mRNA', 'PARP11:57097_mRNA', 'PARS2:25973_mRNA', 'PAX1:5075_mRNA', 'PAX2:5076_mRNA', 'PAX6:5080_mRNA', 'PCCA:5095_mRNA', 'PCDH11X:27328_mRNA', 'PCDHA5:56143_mRNA', 'PCDHB13:56123_mRNA', 'PCDHB14:56122_mRNA', 'PCDHB17:54661_mRNA', 'PCDHB8:56128_mRNA', 'PCDHGA1:56114_mRNA', 'PCDHGA3:56112_mRNA', 'PCDHGA4:56111_mRNA', 'PCDHGA5:56110_mRNA', 'PCDHGB5:56101_mRNA', 'PCDHGC4:56098_mRNA', 'PCDHGC5:56097_mRNA', 'PCNP:57092_mRNA', 'PDE4A:5141_mRNA', 'PDIA4:9601_mRNA', 'PDIK1L:149420_mRNA', 'PDP2:57546_mRNA', 'PEX11A:8800_mRNA', 'PEX3:8504_mRNA', 'PF4V1:5197_mRNA', 'PFKP:5214_mRNA',

'PGA5:5222_mRNA', 'PHF20L1:51105_mRNA', 'PHKA1:5255_mRNA', 'PIGY:84992_mRNA', 'PIK3R2:5296_mRNA', 'PIP5K1B:8395_mRNA', 'PIP5K1C:23396_mRNA', 'PLA1A:51365_mRNA', 'PLA2G4A:5321_mRNA', 'PLA2G4D:283748_mRNA', 'PLCB3:5331_mRNA', 'PLD3:23646_mRNA', 'PLEKHF2:79666_mRNA', 'PLEKHG2:64857_mRNA', 'PLOD1:5351_mRNA', 'PLOD3:8985_mRNA', 'PLSCR4:57088_mRNA', 'PLXNA3:55558_mRNA', 'PLXNB2:23654_mRNA', 'PMFBP1:83449_mRNA', 'PMS2:5395_mRNA', 'PMS2CL:441194_mRNA', 'PNPLA4:8228_mRNA', 'PNRC2:55629_mRNA', 'POLD1:5424_mRNA', 'POLR1B:84172_mRNA', 'POLR2J3:548644_mRNA', 'POLR2J4:84820_mRNA', 'POLR3D:661_mRNA', 'POLRMT:5442_mRNA', 'PPAN:56342_mRNA', 'PPAPDC1A:196051_mRNA', 'PPDPF:79144_mRNA', 'PPIAL4E:730262_mRNA', 'PPIL4:85313_mRNA', 'PPP1R12C:54776_mRNA', 'PPP1R14B:26472_mRNA', 'PPP1R1C:151242_mRNA', 'PPP1R9B:84687_mRNA', 'PPP2R1A:5518_mRNA', 'PPP2R5B:5526_mRNA', 'PPP2R5C:5527_mRNA', 'PPP3R1:5534_mRNA', 'PRDM8:56978_mRNA', 'PRELID2:153768_mRNA', 'PRKCSH:5589_mRNA', 'PRKD2:25865_mRNA', 'PRKD3:23683_mRNA', 'PRLR:5618_mRNA', 'PRMT10:90826_mRNA', 'PRPF31:26121_mRNA', 'PRR12:57479_mRNA', 'PRR25:388199_mRNA', 'PSAPL1:768239_mRNA', 'PSAT1:29968_mRNA', 'PTAR1:375743_mRNA', 'PTK6:5753_mRNA', 'PTPRH:5794_mRNA', 'PTTG3P:26255_mRNA', 'PVR:5817_mRNA', 'PVRL2:5819_mRNA', 'PXDN:7837_mRNA', 'PXMP2:5827_mRNA', 'PYY2:23615_mRNA', 'QPCTL:54814_mRNA', 'QSOX1:5768_mRNA', 'R3HCC1:203069_mRNA', 'RAB11B:9230_mRNA', 'RAB3A:5864_mRNA', 'RAD21:5885_mRNA', 'RAI2:10742_mRNA', 'RANGAP1:5905_mRNA', 'RASGEF1B:153020_mRNA', 'RASSF1:11186_mRNA', 'RAVER1:125950_mRNA', 'RBM11:54033_mRNA', 'RBM16:22828_mRNA', 'RBM18:92400_mRNA', 'RBP5:83758_mRNA', 'RCAN2:10231_mRNA', 'RCHY1:25898_mRNA', 'RCN3:57333_mRNA', 'RDH13:112724_mRNA', 'REC8:9985_mRNA', 'REEP4:80346_mRNA', 'RELT:84957_mRNA', 'REV1:51455_mRNA', 'REXO1:57455_mRNA', 'RFX1:5989_mRNA', 'RFXANK:8625_mRNA', 'RFXAP:5994_mRNA', 'RG9MTD1:54931_mRNA', 'RGPD5:84220_mRNA', 'RIC3:79608_mRNA', 'RICTOR:253260_mRNA', 'RMND1:55005_mRNA', 'RNASE4:6038_mRNA', 'RNF115:27246_mRNA', 'RNF122:79845_mRNA', 'RNF138:51444_mRNA', 'RNF185:91445_mRNA', 'RNF187:149603_mRNA', 'RNF216L:441191_mRNA', 'RNFT2:84900_mRNA', 'RORA:6095_mRNA', 'RP1L1:94137_mRNA', 'RPS6KA4:8986_mRNA', 'RPS6KA5:9252_mRNA',

'RPS6KL1:83694_mRNA', 'RRBP1:6238_mRNA', 'RSBN1L:222194_mRNA', 'RUVBL2:10856_mRNA', 'RXFP3:51289_mRNA', 'S1PR2:9294_mRNA', 'SAFB2:9667_mRNA', 'SAMD1:90378_mRNA', 'SAPS1:22870_mRNA', 'SART1:9092_mRNA', 'SBNO2:22904_mRNA', 'SCAF1:58506_mRNA', 'SCAPER:49855_mRNA', 'SCARNA17:677769_mRNA', 'SCGBL:284402_mRNA', 'SCN11A:11280_mRNA', 'SCOC:60592_mRNA', 'SCT:6343_mRNA', 'SDK2:54549_mRNA', 'SEC13:6396_mRNA', 'SEMA3A:10371_mRNA', 'SEMA3B:7869_mRNA', 'SEMA4A:64218_mRNA', 'SEMA6B:10501_mRNA', 'SENP8:123228_mRNA', 'SEPP1:6414_mRNA', 'SERP2:387923_mRNA', 'SERPINC1:462_mRNA', 'SERPINH1:871_mRNA', 'SERTAD3:29946_mRNA', 'SEZ6L:23544_mRNA', 'SF4:57794_mRNA', 'SFRS13A:10772_mRNA', 'SH2B2:10603_mRNA', 'SH2D3A:10045_mRNA', 'SH3PXD2B:285590_mRNA', 'SHANK1:50944_mRNA', 'SHKBP1:92799_mRNA', 'SILV:6490_mRNA', 'SIX5:147912_mRNA', 'SLC11A1:6556_mRNA', 'SLC14A1:6563_mRNA', 'SLC16A11:162515_mRNA', 'SLC16A14:151473_mRNA', 'SLC17A1:6568_mRNA', 'SLC17A9:63910_mRNA', 'SLC22A20:440044_mRNA', 'SLC22A8:9376_mRNA', 'SLC23A3:151295_mRNA', 'SLC25A33:84275_mRNA', 'SLC25A41:284427_mRNA', 'SLC2A13:114134_mRNA', 'SLC2A4:6517_mRNA', 'SLC2A5:6518_mRNA', 'SLC2A7:155184_mRNA', 'SLC35A5:55032_mRNA', 'SLC35F5:80255_mRNA', 'SLC44A2:57153_mRNA', 'SLC4A3:6508_mRNA', 'SLC6A8:6535_mRNA', 'SMARCA4:6597_mRNA', 'SMARCA5:8467_mRNA', 'SMTN:6525_mRNA', 'SND1:27044_mRNA', 'SNORA12:677800_mRNA', 'SNORA16B:692157_mRNA', 'SNORA40:677822_mRNA', 'SOCS1:8651_mRNA', 'SP3:6670_mRNA', 'SPACA4:171169_mRNA', 'SPATS2:65244_mRNA', 'SPDYE5:442590_mRNA', 'SPPL2A:84888_mRNA', 'SPPL2B:56928_mRNA', 'SPTBN5:51332_mRNA', 'SRGAP1:57522_mRNA', 'SRGAP2:23380_mRNA', 'SRRM5:100170229_mRNA', 'SSBP4:170463_mRNA', 'SSH3:54961_mRNA', 'ST8SIA5:29906_mRNA', 'ST8SIA6:338596_mRNA', 'STAB1:23166_mRNA', 'STAG2:10735_mRNA', 'STBD1:8987_mRNA', 'STIP1:10963_mRNA', 'STK25:10494_mRNA', 'STRN4:29888_mRNA', 'STX17:55014_mRNA', 'STX1A:6804_mRNA', 'STXBP1:6812_mRNA', 'STXBP2:6813_mRNA', 'SUB1:10923_mRNA', 'SUPT5H:6829_mRNA', 'SYDE2:84144_mRNA', 'SYMPK:8189_mRNA', 'SYVN1:84447_mRNA', 'TAB2:23118_mRNA', 'TACC3:10460_mRNA', 'TAF15:8148_mRNA', 'TAF1B:9014_mRNA', 'TAF2:6873_mRNA', 'TAPBPL:55080_mRNA', 'TAS2R31:259290_mRNA', 'TATDN1:83940_mRNA', 'TBC1D15:64786_mRNA', 'TBC1D26:353149_mRNA', 'TBP:6908_mRNA',

'TBPL1:9519_mRNA', 'TCEANC:170082_mRNA', 'TCEB3B:51224_mRNA', 'TCF3:6929_mRNA', 'TCOF1:6949_mRNA', 'TDP2:51567_mRNA', 'TDRD7:23424_mRNA', 'TGDS:23483_mRNA', 'THAP5:168451_mRNA', 'THAP6:152815_mRNA', 'THAP9:79725_mRNA', 'THEM5:284486_mRNA', 'THOP1:7064_mRNA', 'TIFA:92610_mRNA', 'TM9SF4:9777_mRNA', 'TMC7:79905_mRNA', 'TMED9:54732_mRNA', 'TMEM132A:54972_mRNA', 'TMEM151A:256472_mRNA', 'TMEM154:201799_mRNA', 'TMEM161A:54929_mRNA', 'TMEM167A:153339_mRNA', 'TMEM184A:202915_mRNA', 'TMEM200C:645369_mRNA', 'TMEM206:55248_mRNA', 'TMEM213:155006_mRNA', 'TMEM214:54867_mRNA', 'TMEM81:388730_mRNA', 'TMEM88B:643965_mRNA', 'TMOD1:7111_mRNA', 'TNFRSF10B:8795_mRNA', 'TNFRSF18:8784_mRNA', 'TNFRSF25:8718_mRNA', 'TNKS1BP1:85456_mRNA', 'TNNI3K:51086_mRNA', 'TNRC18:84629_mRNA', 'TOM1L1:10040_mRNA', 'TP53BP2:7159_mRNA', 'TP53RK:112858_mRNA', 'TPD52L2:7165_mRNA', 'TPD52L3:89882_mRNA', 'TPH2:121278_mRNA', 'TPRG1:285386_mRNA', 'TREX2:11219_mRNA', 'TRIM11:81559_mRNA', 'TRIM38:10475_mRNA', 'TRIM46:80128_mRNA', 'TRIM50:135892_mRNA', 'TRIM65:201292_mRNA', 'TRIML1:339976_mRNA', 'TRIP10:9322_mRNA', 'TRPM6:140803_mRNA', 'TRPM7:54822_mRNA', 'TRPV1:7442_mRNA', 'TRY6:154754_mRNA', 'TSGA13:114960_mRNA', 'TSLP:85480_mRNA', 'TSTD1:100131187_mRNA', 'TTC21B:79809_mRNA', 'TTC9B:148014_mRNA', 'TUBB3:10381_mRNA', 'TUBGCP3:10426_mRNA', 'TYK2:7297_mRNA', 'TYRO3:7301_mRNA', 'U2AF2:11338_mRNA', 'UBE2MP1:606551_mRNA', 'UBE3A:7337_mRNA', 'UBIAD1:29914_mRNA', 'UBR1:197131_mRNA', 'UCN3:114131_mRNA', 'UNC13C:440279_mRNA', 'UOX:391051_mRNA', 'UPB1:51733_mRNA', 'UPF1:5976_mRNA', 'USP12:219333_mRNA', 'USP25:29761_mRNA', 'USP38:84640_mRNA', 'USP51:158880_mRNA', 'UTP23:84294_mRNA', 'UVRAG:7405_mRNA', 'VAC14:55697_mRNA', 'VASP:7408_mRNA', 'VGF:7425_mRNA', 'WDR11:55717_mRNA', 'WDR18:57418_mRNA', 'WDR26:80232_mRNA', 'WDR31:114987_mRNA', 'WFS1:7466_mRNA', 'WIPI2:26100_mRNA', 'WIZ:58525_mRNA', 'WTAP:9589_mRNA', 'XPO7:23039_mRNA', 'YIPF2:78992_mRNA', 'YJEFN3:374887_mRNA', 'YKT6:10652_mRNA', 'YPEL2:388403_mRNA', 'ZAK:51776_mRNA', 'ZBTB11:27107_mRNA', 'ZBTB45:84878_mRNA', 'ZBTB6:10773_mRNA', 'ZCCHC18:644353_mRNA', 'ZFP3:124961_mRNA', 'ZFP82:284406_mRNA', 'ZFR:51663_mRNA',

'ZFX:7543_mRNA', 'ZHX1:11244_mRNA', 'ZMIZ2:83637_mRNA', 'ZNF136:7695_mRNA', 'ZNF16:7564_mRNA', 'ZNF250:58500_mRNA', 'ZNF268:10795_mRNA', 'ZNF277:11179_mRNA', 'ZNF287:57336_mRNA', 'ZNF329:79673_mRNA', 'ZNF346:23567_mRNA', 'ZNF347:84671_mRNA', 'ZNF354B:117608_mRNA', 'ZNF442:79973_mRNA', 'ZNF451:26036_mRNA', 'ZNF469:84627_mRNA', 'ZNF510:22869_mRNA', 'ZNF526:116115_mRNA', 'ZNF540:163255_mRNA', 'ZNF568:374900_mRNA', 'ZNF574:64763_mRNA', 'ZNF582:147948_mRNA', 'ZNF597:146434_mRNA', 'ZNF643:65243_mRNA', 'ZNF658:26149_mRNA', 'ZNF673:55634_mRNA', 'ZNF679:168417_mRNA', 'ZNF704:619279_mRNA', 'ZNF763:284390_mRNA', 'ZNF787:126208_mRNA', 'ZNF815:401303_mRNA', 'ZNF828:283489_mRNA', 'ZNF843:283933_mRNA', 'ZNF99:7652_mRNA', 'ZP3:7784_mRNA', 'ZSCAN12:9753_mRNA']

Selected Mutation Features: ['C1orf173_MUT', 'ANK2_MUT', 'CACNA1E_MUT', 'SLC12A7_MUT', 'STK11_MUT', 'MYH2_MUT', 'ADAMTS5_MUT', 'MUC16_MUT', 'F5_MUT', 'ZFHX4_MUT', 'MUC17_MUT', 'CCDC147_MUT', 'OR2M3_MUT', 'PCDHB6_MUT', 'NRXN3_MUT', 'TTN_MUT', 'NAV3_MUT', 'DNAH1_MUT', 'FAT3_MUT', 'TSHZ3_MUT', 'HYDIN_MUT', 'ASTN1_MUT', 'NCKAP5_MUT', 'EYS_MUT', 'KRAS_MUT', 'DNAH7_MUT', 'SAMD9L_MUT', 'COL11A1_MUT', 'USH2A_MUT', 'PCDHB3_MUT', 'ATM_MUT', 'FHOD3_MUT', 'COL22A1_MUT', 'SYNE1_MUT', 'TP53_MUT', 'CDH10_MUT', 'ABCA6_MUT', 'KCNA4_MUT', 'OR8K5_MUT', 'AHNAK2_MUT', 'FMN2_MUT', 'OBSCN_MUT', 'EGFR_MUT', 'TENM1_MUT', 'CAMSAP2_MUT', 'RYR2_MUT', 'ZNF536_MUT', 'KCNH1_MUT', 'XIRP2_MUT', 'CSMD3_MUT', 'SPTA1_MUT', 'LILRB2_MUT', 'NBEA_MUT', 'ASXL3_MUT', 'BIRC6_MUT', 'FLG_MUT', 'ADAMTSL1_MUT', 'ASTN2_MUT', 'DNAH2_MUT']

Selected Methylation Features: ['ATP2A1_MET', 'MEOX2_MET', 'KIAA1409;COX8C_MET', 'TTC8_MET', 'PMM2;TMEM186_MET', 'nan_MET', 'QRFPR_MET', 'RHOC_MET', 'CCNB1IP1_MET', 'C9orf85;FAM108B1_MET', 'EPHA7_MET', 'CUL5_MET', 'GLRA1_MET', 'TULP2_MET', 'LOC100129726;ZFP36L2_MET', 'MPDU1_MET', 'MYH1_MET', 'MRPL52_MET', 'HTR2C_MET', 'SIL1_MET', 'TJAP1_MET', 'TBCK;AIMP1_MET', 'ACAA2;SCARNA17_MET', 'AMMECR1_MET', 'LTV1_MET', 'WDR37_MET', 'KLHL28;FAM179B_MET', 'TTLL4_MET', 'CYB5R2_MET', 'SLC24A3_MET', 'ZNF678_MET',

'PROKR1_MET', 'CYB5D2;ZZEF1_MET', 'TBC1D1;PTTG2_MET', 'T-SP1_MET', 'IL20_MET', 'TCP11_MET', 'LAT2_MET', 'SLC29A3_MET', 'GORAB_MET', 'SLC5A6;C2orf28_MET', 'SLC5A5_MET', 'ZIM2;PEG3_MET', 'KCTD8_MET', 'HSPA5_MET', 'RUNX2;SUPT3H_MET', 'SCARB1_MET', 'NEURL_MET', 'STK4_MET', 'SNX3_MET', 'WDR82_MET', 'MLNR_MET', 'PALB2;DCTN5_MET', 'GPR44_MET', 'CLIP1_MET', 'ASCC1;C10orf104_MET', 'ECHS1_MET', 'TFEB_MET', 'C14orf93_MET', 'VPS18_MET', 'ZNF227_MET', 'SPIB_MET', 'ZKSCAN3;ZNF323_MET', 'IGFBP7_MET', 'OR10A4_MET', 'RCN3_MET', 'RXFP3_MET', 'TSPYL4_MET', 'SPRR3_MET', 'GJB1;BCYRN1_MET', 'TMEM125_MET', 'TMEM106A_MET', 'OR7A17_MET', 'RPL7;RDH10_MET', 'TTC18_MET', 'STAT5B_MET', 'TMEM59;C1orf83_MET', 'PLAC2_MET', 'CDC7_MET', 'UGDH_MET', 'SAMD12_MET', 'SEMA4G_MET', 'CYP4F2_MET', 'LRSAM1;RPL12_MET', 'CLOCK_MET', 'MYH10_MET', 'HPN_MET', 'TMEM175;GAK_MET', 'RIF1_MET', 'SPIN2A_MET', 'DAZL_MET', 'STEAP3_MET', 'PAPOLA_MET', 'LPPR2_MET', 'SLC22A12_MET', 'G6PC2_MET', 'DAXX_MET', 'MORC2;TUG1_MET', 'NRTN_MET', 'ENTHD1_MET', 'VIP_MET', 'FUCA1_MET', 'TTR_MET', 'KDM4A_MET', 'SPRYD3_MET', 'FOXJ2_MET', 'POM121L1P_MET', 'STK33_MET', 'JPH1_MET', 'GABRG2_MET', 'XPNPEP1_MET', 'PSMD5;LOC253039_MET', 'APOL6_MET', 'ATXN3_MET', 'LRRN3;IMMP2L_MET', 'LAX1_MET', 'HMGCS2_MET', 'TMEM161B_MET', 'LPAR2_MET', 'BHMT_MET', 'TRNAU1AP_MET', 'PCDHB15_MET', 'HMGCS1_MET', 'FOLR2_MET', 'LMO3_MET', 'ACTR2_MET', 'RNF113B;FARP1_MET', 'TNNI1_MET', 'SLC36A4_MET', 'PRM1_MET', 'CCDC141_MET', 'HOXA1_MET', 'KTI12;TXNDC12_MET', 'TAAR6_MET', 'NEIL1_MET', 'TUBB3_MET', 'NME5_MET', 'HMGN1_MET', 'CDCA4_MET', 'BTRC_MET', 'PTH2_MET', 'NBPF15_MET', 'LSAMP_MET', 'PRKAB2_MET', 'CHCHD4;TMEM43_MET', 'SLC16A3_MET', 'RBBP7_MET', 'ZNF622_MET', 'LYL1_MET', 'CX3CL1_MET', 'NSUN4_MET', 'TMEM111_MET', 'MAP3K13_MET', 'MCM4;PRKDC_MET', 'TPM4_MET', 'XRCC5_MET', 'GRLF1_MET', 'TMEM50B_MET', 'ENO2_MET', 'PSG6_MET', 'TOM1L1_MET', 'TF_MET', 'P2RY13;MED12L_MET', 'MYH11_MET', 'AOC2_MET', 'HDAC2_MET', 'GPR160_MET', 'MMP19_MET', 'FAM10A4_MET', 'SLFN12_MET', 'SEC24C_MET', 'TRPM3_MET', 'ADA_MET', 'POR_MET', 'C20orf165_MET', 'ATP8B2;AQP10_MET', 'SH3YL1;ACP1_MET', 'CBLN2_MET', 'HIVEP2_MET', 'CLRN3_MET', 'NLGN4X_MET', 'ARL14_MET', 'KIAA0020_MET', 'C1orf126;TMEM51_MET', 'CXCR6;FYCO1_MET', 'S100A14_MET', 'LCTL_MET', 'MLL4_MET', 'STX8;WDR16_MET', 'SNX4_MET']

Selected Clinical Features: ['tobacco_smoking_history_clinical', 'TMB_NONSYNONYMOUS_clinical', 'AGE_clinical', 'GENETIC_ANCESTRY_LABEL_ENCODED_clinical', 'PATH_T_STAGE_ENCODED_clinical', 'PATH_N_STAGE_ENCODED_clinical', 'PATH_M_STAGE_ENCODED_clinical', 'AJCC_PATHOLOGIC_TUMOR_STAGE_ENCODED_clinical', 'SEX_ENCODED_clinical', 'TUMOR_TYPE_ENCODED_clinical']

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin. 2024;74(1):12–49.
2. Clark SB, Alsubait S. Non–small cell lung cancer. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 [cited 2024 Sep 20]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK562307/
3. Herbst RS, Heymach JV, Lippman SM. Lung Cancer. N Engl J Med. 2008 Sep 25;359(13):1367–80.
4. Alduais Y, Zhang H, Fan F, Chen J, Chen B. Non-small cell lung cancer (NSCLC): A review of risk factors, diagnosis, and treatment. Medicine (Baltimore). 2023 Feb 22;102(8):e32899.
5. Zappa C, Mousa SA. Non-small cell lung cancer: current treatment and future advances. Transl Lung Cancer Res [Internet]. 2016 Jun [cited 2024 Dec 26];5(3). Available from: https://tlcr.amegroups.org/article/view/8139
6. Nicholson AG, Tsao MS, Beasley MB, Borczuk AC, Brambilla E, Cooper WA, et al. The 2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015. J Thorac Oncol. 2022 Mar 1;17(3):362–87.
7. High-Resolution Profiling of Lung Adenocarcinoma Identifies Expression Subtypes with Specific Biomarkers and Clinically Relevant Vulnerabilities | Cancer Research | American Association for Cancer Research [Internet]. [cited 2024 Dec 26]. Available from: https://aacrjournals.org/cancerres/article/82/21/3917/709962/High-Resolution-Profiling-of-Lung-Adenocarcinoma
8. Ho KH, Huang TW, Liu AJ, Shih CM, Chen KC. Cancer Essential Genes Stratified Lung Adenocarcinoma Patients with Distinct Survival Outcomes and Identified a Subgroup from the Terminal Respiratory Unit Type with Different Proliferative Signatures in Multiple Cohorts. Cancers. 2021 Apr 28;13(9):2128.
9. Frontiers | Identification of lung adenocarcinoma subtypes and predictive signature for prognosis, immune features, and immunotherapy based on immune checkpoint genes [Internet]. [cited 2024 Dec 26]. Available from: https://www.frontiersin.org/journals/cell-and-developmental-biology/articles/10.3389/fcell.2023.1060086/full
10. Schneider F, Dacic S. Histopathologic and molecular approach to staging of multiple lung nodules. Transl Lung Cancer Res. 2017 Oct;6(5):540–9.
11. Yang K, Wu Y. A prognosis-related molecular subtype for early-stage non-small lung cell carcinoma by multi-omics integration analysis. BMC Cancer. 2021 Feb 6;21:128.
12. Comparing the Therapeutic Efficacies of Lung Cancer: Network Meta-Analysis Approaches - PMC [Internet]. [cited 2024 Dec 26]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9657796/
13. Five-Year Data for Merck's KEYTRUDA® (pembrolizumab) Plus Chemotherapy Showed Sustained Survival Benefit Versus Chemotherapy Alone in Two Studies for Metastatic Non-Small Cell Lung Cancer (NSCLC) - Merck.com [Internet]. [cited 2024 Dec 26]. Available from: https://www.merck.com/news/five-year-data-for-mercks-keytruda-pembrolizumab-plus-chemotherapy-showed-sustained-survival-benefit-versus-chemotherapy-alone-in-two-studies-for

-metastatic-non-small-cell-lung-cancer-nsclc/
14. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WEE, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol Off Publ Int Assoc Study Lung Cancer. 2016 Jan;11(1):39–51.
15. Uramoto H, Tanaka F. Recurrence after surgery in patients with NSCLC. Transl Lung Cancer Res. 2014 Aug;3(4):242–9.
16. Prognosis of recurrence after complete resection in early-stage lung adenocarcinoma based on molecular alterations: a systematic review and meta-analysis | Scientific Reports [Internet]. [cited 2024 Dec 26]. Available from: https://www.nature.com/articles/s41598-023-42851-2
17. Choi PJ, Jeong SS, Yoon SS. Prediction and prognostic factors of post-recurrence survival in recurred patients with early-stage NSCLC who underwent complete resection. J Thorac Dis. 2016 Jan;8(1):152–60.
18. Long-term survival outcome after postoperative recurrence of non-small-cell lung cancer: who is 'cured' from postoperative recurrence? | European Journal of Cardio-Thoracic Surgery | Oxford Academic [Internet]. [cited 2024 Dec 26]. Available from: https://academic.oup.com/ejcts/article/52/3/522/3805125?login=false
19. Deep learning in cancer diagnosis, prognosis and treatment selection | Genome Medicine | Full Text [Internet]. [cited 2024 Dec 26]. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-021-00968-x
20. Y L, T K, M N, Ge D, Jl S, A M, et al. Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. Arch Pathol Lab Med [Internet]. 2019 Jul [cited 2025 Mar 5];143(7). Available from: https://pubmed.ncbi.nlm.nih.gov/30295070/
21. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018 Nov;36(10):983–7.
22. Somashekhar SP, Sepúlveda MJ, Puglielli S, Norden AD, Shortliffe EH, Kumar CR, et al. Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann Oncol. 2018 Feb 1;29(2):418–23.
23. Biology-guided deep learning predicts prognosis and cancer immunotherapy response | Nature Communications [Internet]. [cited 2024 Dec 26]. Available from: https://www.nature.com/articles/s41467-023-40890-x
24. Wang X, Zhao J, Marostica E, Yuan W, Jin J, Zhang J, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. Nature. 2024 Oct;634(8035):970–8.
25. AI and Cancer - NCI [Internet]. [cited 2024 Dec 26]. Available from: https://www.cancer.gov/research/infrastructure/artificial-intelligence
26. Zhang B, Shi H, Wang H. Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. J Multidiscip Healthc. 2023 Jun 26;16:1779–91.
27. Xu Z, Liao H, Huang L, Chen Q, Lan W, Li S. IBPGNET: lung adenocarcinoma recurrence prediction based on neural network interpretability. Brief Bioinform. 2024 May 1;25(3):bbae080.
28. Huang P, Illei PB, Franklin W, Wu PH, Forde PM, Ashrafinia S, et al. Lung Cancer Recurrence Risk Prediction through Integrated Deep Learning Evaluation. Cancers. 2022

Aug 27;14(17):4150.
29. Jones GD, Brandt WS, Shen R, Sanchez-Vega F, Tan KS, Martin A, et al. A Genomic-Pathologic Annotated Risk Model to Predict Recurrence in Early-Stage Lung Adenocarcinoma. JAMA Surg. 2021 Feb 10;156(2):e205601.
30. Shen Y, Goparaju C, Yang Y, Babu BA, Gai W, Pass H, et al. Recurrence prediction of lung adenocarcinoma using an immune gene expression and clinical data trained and validated support vector machine classifier. Transl Lung Cancer Res. 2023 Oct 31;12(10):2055–67.
31. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018 Apr 5;173(2):291-304.e6.
32. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov. 2012 May 1;2(5):401–4.
33. Broad GDAC Firehose [Internet]. [cited 2024 Nov 15]. Available from: https://gdac.broadinstitute.org/
34. FireBrowse [Internet]. [cited 2024 Nov 15]. Available from: http://firebrowse.org/?cohort=LUAD
35. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020 Jun;38(6):675–8.
36. UrbsLab/STREAMLINE [Internet]. UrbsLab; 2025 [cited 2025 Mar 25]. Available from: https://github.com/UrbsLab/STREAMLINE
37. Mahvi DA, Liu R, Grinstaff MW, Colson YL, Raut CP. Local Cancer Recurrence: The Realities, Challenges, and Opportunities for New Therapies. CA Cancer J Clin. 2018 Nov;68(6):488–505.
38. Lou F, Sima CS, Rusch VW, Jones DR, Huang J. Differences in Patterns of Recurrence in Early-Stage Versus Locally Advanced Non-Small Cell Lung Cancer. Ann Thorac Surg. 2014 Nov;98(5):1755–61.
39. Yun JK, Lee GD, Choi S, Kim YH, Kim DK, Park SI, et al. Various recurrence dynamics for non-small cell lung cancer depending on pathological stage and histology after surgical resection. Transl Lung Cancer Res [Internet]. 2022 Jul [cited 2025 Jan 18];11(7). Available from: https://tlcr.amegroups.org/article/view/66365
40. Yilmaz A, Demircali AA, Kocaman S, Uvet H. Comparison of Deep Learning and Traditional Machine Learning Techniques for Classification of Pap Smear Images [Internet]. arXiv; 2020 [cited 2025 Mar 8]. Available from: http://arxiv.org/abs/2009.06366
41. Cox DR. The Regression Analysis of Binary Sequences. J R Stat Soc Ser B Methodol. 1958;20(2):215–42.
42. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep;20(3):273–97.
43. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification And Regression Trees [Internet]. 1st ed. Routledge; 2017 [cited 2025 Mar 8]. Available from: https://www.taylorfrancis.com/books/9781351460491
44. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn. 1992 Oct;9(4):309–47.
45. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32.
46. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the

22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. San Francisco California USA: ACM; 2016 [cited 2025 Mar 8]. p. 785–94. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785
47. Hart P. The condensed nearest neighbor rule (Corresp.). IEEE Trans Inf Theory. 1968 May;14(3):515–6.
48. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837–45.
49. Boyd K, Eng KH, Page CD. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In: Salinesi C, Norrie MC, Pastor Ó, editors. Advanced Information Systems Engineering [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013 [cited 2025 Mar 11]. p. 451–66. (Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al., editors. Lecture Notes in Computer Science; vol. 7908). Available from: http://link.springer.com/10.1007/978-3-642-40994-3_29
50. King G, Zeng L. Logistic Regression in Rare Events Data. Polit Anal. 2001 Jan;9(2):137–63.
51. Hsu Y, Chang Y, Chang G, Ho B, Yuan S, Li Y, et al. Tumor mutation burden and recurrent tumors in hereditary lung cancer. Cancer Med. 2019 Apr 2;8(5):2179–87.
52. Wang X, Ricciuti B, Nguyen T, Li X, Rabin MS, Awad MM, et al. Association between smoking history and tumor mutation burden in advanced non-small cell lung cancer. Cancer Res. 2021 Mar 2;81(8):2204–13. doi:10.1158/0008-5472.CAN-20-3991.
53. Ikeda K, Shiraishi K, Eguchi A, Osumi H, Matsuishi K, Matsubara E, et al. Association of a genetic variant of CYP19A1 with multicentric development of lung adenocarcinomas. Ann Surg Oncol. 2014 Mar;21(3):939–45.
54. Association of CYP19A1 polymorphisms with risks for atypical adenomatous hyperplasia and bronchioloalveolar carcinoma in the lungs | Carcinogenesis | Oxford Academic [Internet]. [cited 2024 Dec 4]. Available from: https://academic.oup.com/carcin/article-abstract/31/10/1794/2477200?login=false&redirectedFrom=fulltext
55. Zhang C, Cheng Y, Chen W, Li Q, Dai R, Wang Y, et al. Association of CYP19A1 rs28757157 polymorphism with lung cancer risk in the Chinese Han population. World J Surg Oncol. 2022 Dec 16;20(1):400.
56. Liu L, Mo M, Chen X, Chao D, Zhang Y, Chen X, et al. Targeting inhibition of prognosis-related lipid metabolism genes including CYP19A1 enhances immunotherapeutic response in colon cancer. J Exp Clin Cancer Res CR. 2023 Apr 13;42:85.
57. He W, Zhang H, Cheng H, Wen J, Li D. PIK3CD correlates with prognosis, epithelial–mesenchymal transition and tumor immune infiltration in breast carcinoma. Discov Oncol. 2023 Oct 20;14:187.
58. ResearchGate [Internet]. [cited 2024 Dec 5]. PIK3CD-AS2 is overexpressed in lung adenocarcinoma cancer (LUAD) and... Available from: https://www.researchgate.net/figure/PIK3CD-AS2-is-overexpressed-in-lung-adenocarcinoma-cancer-LUAD-and-associated-with_fig4_339892880
59. Zheng X, Zhang J, Fang T, Wang X, Wang S, Ma Z, et al. The long non-coding RNA PIK3CD-AS2 promotes lung adenocarcinoma progression via YBX1-mediated suppression of p53 pathway. Oncogenesis. 2020 Mar 12;9(3):34.
60. Tissue ACE phenotyping in lung cancer. PLOS ONE. 2019 Dec 26;14(12):e0226553.

61. Khosravian M, Momenzadeh M, Koosha F, Alimohammadi N, Kianpour N. Lung cancer risk and the inhibitors of angiotensin converting enzyme; an updated review on recent evidence. Immunopathol Persa. 2021 Aug 8;8(1):e19–e19.
62. Hegmann KT, Fraser AM, Keaney RP, Moser SE, Nilasena DS, Sedlars M, et al. The effect of age at smoking initiation on lung cancer risk. Epidemiol Camb Mass. 1993 Sep;4(5):444–8.
63. Lewis M, Elad G, Beladev M, Maor G, Radinsky K, Hermann D, et al. Comparison of deep learning with traditional models to predict preventable acute care use and spending among heart failure patients. Sci Rep. 2021 Jan 13;11(1):1164.
64. Didier AJ, Nigro A, Noori Z, Omballi MA, Pappada SM, Hamouda DM. Application of machine learning for lung cancer survival prognostication—A systematic review and meta-analysis. Front Artif Intell. 2024 Apr 5;7:1365777.
65. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning [Internet]. New York, NY: Springer; 2009 [cited 2025 Mar 9]. (Springer Series in Statistics). Available from: http://link.springer.com/10.1007/978-0-387-84858-7
66. Greillier L, Tomasini P, Barlesi F. The clinical utility of tumor mutational burden in non-small cell lung cancer. Transl Lung Cancer Res [Internet]. 2018 Dec [cited 2024 Dec 5];7(6). Available from: https://tlcr.amegroups.org/article/view/24870
67. Goodgame B, Viswanathan A, Zoole J, Gao F, Miller CR, Subramanian J, et al. Risk of recurrence of resected stage I non-small cell lung cancer in elderly patients as compared with younger patients. J Thorac Oncol Off Publ Int Assoc Study Lung Cancer. 2009 Nov;4(11):1370–4.
68. Wong ML, McMurry TL, Stukenborg GJ, Francescatti AB, Amato-Martz C, Schumacher JR, et al. Impact of age and comorbidity on treatment of non-small cell lung cancer recurrence following complete resection: A nationally representative cohort study. Lung Cancer Amst Neth. 2016 Dec;102:108–17.
69. Chiang HH, Ong CT, Chang CY, Wu KL, Wu YY, Lai JC, et al. Downregulated antisense lncRNA ENTPD3-AS1 contributes to the development of lung adenocarcinoma. Am J Cancer Res. 2024;14(2):854–68.
70. Qi J, Yin J, Ding G. A Connexin-Based Biomarker Model Applicable for Prognosis and Immune Landscape Assessment in Lung Adenocarcinoma. J Oncol. 2022 Oct 12;2022:9261339.
71. Aasen T, Sansano I, Montero MÁ, Romagosa C, Temprana-Salvador J, Martínez-Marti A, et al. Insight into the Role and Regulation of Gap Junction Genes in Lung Cancer and Identification of Nuclear Cx43 as a Putative Biomarker of Poor Prognosis. Cancers. 2019 Mar 6;11(3):320.
72. Jeong WG, Choi H, Chae KJ, Kim J. Prognosis and recurrence patterns in patients with early stage lung cancer: a multi-state model approach. Transl Lung Cancer Res. 2022 Jul;11(7):1279–91.
73. IBPGNET: lung adenocarcinoma recurrence prediction based on neural network interpretability | Briefings in Bioinformatics | Oxford Academic [Internet]. [cited 2024 Dec 5]. Available from: https://academic.oup.com/bib/article/25/3/bbae080/7638265?login=false
74. Ekeke CN, Mitchell C, Schuchert M, Dhupar R, Luketich JD, Okusanya OT. Early Distant Recurrence in Patients With Resected Stage I Lung Cancer: A Case Series of "Blast Metastasis." Clin Lung Cancer. 2021 Jan;22(1):e132–5.

75. Chen QF, Gao H, Pan QY, Wang YJ, Zhong XN. Analysis at the single-cell level indicates an important role of heterogeneous global DNA methylation status on the progression of lung adenocarcinoma. Sci Rep. 2021 Dec 2;11(1):23337.
76.Sharma A, Lysenko A, Jia S, Boroevich KA, Tsunoda T. Advances in AI and machine learning for predictive medicine. J Hum Genet. 2024 Oct;69(10):487–97.
77. Nors J, Iversen LH, Erichsen R, Gotschalck KA, Andersen CL. Incidence of Recurrence and Time to Recurrence in Stage I to III Colorectal Cancer: A Nationwide Danish Cohort Study. JAMA Oncol. 2024 Jan 1;10(1):54–62.
78. Pedersen RN, Esen BÖ, Mellemkjær L, Christiansen P, Ejlertsen B, Lash TL, et al. The Incidence of Breast Cancer Recurrence 10-32 Years After Primary Diagnosis. JNCI J Natl Cancer Inst. 2022 Mar 1;114(3):391–9.
79. Pignon JP, Tribodet H, Scagliotti GV, Douillard JY, Shepherd FA, Stephens RJ, et al. Lung adjuvant cisplatin evaluation: a pooled analysis by the LACE Collaborative Group. J Clin Oncol. 2008 Jul 20;26(21):3552–9. doi:10.1200/JCO.2007.13.9030
80. Hardenberg MC, Patel B, Matthews C, Califano R, Garcia Campelo R, Grohe C, et al. The value of disease-free survival and osimertinib in adjuvant non–small-cell lung cancer: an international Delphi consensus report. ESMO Open.2022 Sep 13;7(5):100572. doi:10.1016/j.esmoop.2022.100572.
81. Felip E, Altorki N, Zhou C, Csőszi T, Vynnychenko I, Goloborodko O, et al. Adjuvant atezolizumab after adjuvant chemotherapy in resected stage IB–IIIA non–small-cell lung cancer (IMpower010): a randomised, multicentre, open-label, phase 3 trial. Lancet. 2021 Oct 9;398(10308):1344–57. doi:10.1016/S0140-6736(21)02098-5.
82. Kim PJ, Hwang HS, Choi G, Sung HJ, Ahn B, Uh JS, et al. A new model using deep learning to predict recurrence after surgical resection of lung adenocarcinoma. Sci Rep. 2024 Mar 16;14(1):6366.
83. Li X, Ma J, Leng L, Han M, Li M, He F, et al. MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. Front Genet. 2022 Feb 2;13:806842.