Synergy between imputed genetic pathway and clinical information for predicting recurrence in early stage non-small cell lung cancer

Synergy between imputed genetic pathway and clinical information for predicting recurrence in early stage non-small cell lung cancer

Mohan Timilsina[a,*], Dirk Fey[b], Samuele Buosi[a], Adrianna Janik[e], Luca Costabello[e], Enric Carcereny[f], Delvys Rodrıguez Abreu[g], Manuel Cobo[h], Rafael López Castro[i], Reyes Bernabé[j], Pasquale Minervini[k], Maria Torrente[l], Mariano Provencio[l] and Vít Nováček[a,c,d]

[a]*Data Science Institute, Insight Centre for Data Analytics, National University of Ireland Galway, Ireland*
[b]*Systems Biology Ireland, University College Dublin, Ireland*
[e]*Accenture Labs, Dublin, Ireland*
[f]*Catalan Institute of Oncology, Hospital Universitari Germans Trias i Pujol, B-ARGO, IGTP, Badalona, Spain*
[g]*Hospital Universitario Insular de Gran Canaria, Las Palmas de Gran Canaria, Spain*
[h]*Medical Oncology Intercenter Unit. Regional and Virgen de la Victoria University Hospitals. IBIMA. Málaga. Spain*
[i]*Hospital Clínico Universitario de Valladolid,Spain*
[j]*Hospital Universitario Virgen del Rocio, Sevilla, Spain*
[k]*University College London, London, United Kingdom*
[l]*Medical Oncology Department, Hospital Universitario Puerta de Hierro Majadahonda, Madrid, Spain*
[c]*Faculty of Informatics, Masaryk University Brno, Czech Republic*
[d]*Masaryk Memorial Cancer Institute, Brno, Czech Republic*

ARTICLE INFO

Keywords:
regression
classification
imputation
recurrence
supervised
explanation

ABSTRACT

Objective: Lung cancer exhibits unpredictable recurrence in low-stage tumors and variable responses to different therapeutic interventions. Predicting relapse in early-stage lung cancer can facilitate precision medicine and improve patient survivability. While existing machine learning models rely on clinical data, incorporating genomic information could enhance their efficiency. This study aims to impute and integrate specific types of genomic data with clinical data to improve the accuracy of machine learning models for predicting relapse in early-stage, non-small cell lung cancer patients.

Methods: The study utilized a publicly available TCGA lung cancer cohort and imputed genetic pathway scores into the Spanish Lung Cancer Group (SLCG) data, specifically in 1348 early-stage patients. Initially, tumor recurrence was predicted without imputed pathway scores. Subsequently, the SLCG data were augmented with pathway scores imputed from TCGA. The integrative approach aimed to enhance relapse risk prediction performance.

Results: The integrative approach achieved improved relapse risk prediction with the following evaluation metrics: an area under the precision-recall curve (PR-AUC) score of 0.75, an area under the ROC (ROC-AUC) score of 0.80, an F1 score of 0.61, and a Precision of 0.80. The prediction explanation model SHAP (SHapley Additive exPlanations) was employed to explain the machine learning model's predictions.

Conclusion: We conclude that our explainable predictive model is a promising tool for oncologists that addresses an unmet clinical need of post-treatment patient stratification based on the relapse risk while also improving the predictive power by incorporating proxy genomic data not available for specific patients.

1. Introduction

Cancer is a widespread health complication globally and comprises various subtypes, making early identification of the cancer type and stage essential for effective treatment [1, 2]. Gene expression data and bioinformatics approaches have shown promise in stratifying cancer patients into high and low-risk cohorts [3]. Alterations in critical pathways

mohan.timilsina@insight-centre.org (M. Timilsina); dirk.fey@ucd.ie (D. Fey); samuele.buosi@insight-centre.org (S. Buosi); adrianna.janik@accenture.com (A. Janik); luca.costabello@accenture.com (L. Costabello); ecarcereny@iconcologia.net (E. Carcereny); drodabr@gobiernodecanarias.org (D.R. Abreu); manuelcobodols@yahoo.es (M. Cobo); rafalopezcastro@yahoo.es (R.L. Castro); reyesbernab@yahoo.es (R. Bernabé); p.minervini@ucl.ac.uk (P. Minervini); maria.torrente@salud.madrid.org (M. Torrente); mariano.provencio@salud.madrid.org (M. Provencio); vit.novacek@insight-centre.org (V. Nováček)

during cancer initiation and progression have been linked to clinical outcomes in patients [4, 5, 6, 7]. Therefore, recognizing the involved pathways and quantifying their deregulation is essential to understanding the malignancy process. Lung cancer, a disease with a poor prognosis, has several mediators of predominant tumor functions that have been identified through recent advances in understanding the molecular pathways behind the disease [8]. Pathway analysis allows scientists to examine genomics data systematically and gain insight into relevant problems. However, pathway screening is costly and time-consuming, making it challenging to perform in a clinical pathology laboratory [9]. A data-driven approach based on automatically derived latent patient similarities can partially overcome this difficulty by imputing pathway scores to patient records for whom they were not initially identified. This approach can augment machine learning models trained to predict the risk of relapse and ultimately increase their predictive performance.

Public data repositories such as The Cancer Genome Atlas (TCGA) provide molecular tumor maps collected from different laboratories for cancer-related research. Using such information for imputing pathway scores into clinical oncological data can provide an enhanced approach to predicting the risk of relapse in lung cancer patients. We implemented and validated this approach, imputing the pathway scores into clinical data by leveraging publicly available TCGA data and further evaluated the imputed pathway score by predicting relapse in early-stage lung cancer patients. To the best of our knowledge, such imputation of pathway score using TCGA data into clinical data was used for predicting relapse of early-stage lung cancer for the first time. Our experiments were conducted on the Spanish Lung Cancer Group (SLCG) dataset, and the results demonstrate the improved prediction performance of the imputed pathway scores compared to models trained with the clinical features alone.

The specific research questions investigated in this study are:

- Can we impute genetic pathway score by integrating publicly available genomic (TCGA) and medical data (SLCG)?
- How to ensure the quality of the pathway score imputation process?
- Can be augmenting imputed pathway scores with the patient's clinical information improve the accuracy of machine learning models aimed at relapse prediction for early stage patients?

From an Artificial Intelligence (AI) and Machine Learning (ML) perspective, our study does not introduce significant novelty. However, from a clinical and biomedical application point of view, our novelty lies in performing genomic pathway imputation from a publicly available patient cohort (TCGA) to a hospital lung cancer patient cohort. The quality of our pathway imputation is further validated by integrating pathway imputations with clinical features and predicting the tumor recurrence of the early stage lung cancer patient.

In conclusion, imputing pathway scores from publicly available TCGA lung cancer data into clinical data using a data-driven approach can improve the prediction performance of machine learning models aimed at relapse prediction for early-stage lung cancer patients. Our findings provide insights into the potential of using public data repositories to augment clinical data for improved patient outcomes in lung cancer treatment.

Contribution: Our work has two main contributions: first, we developed a method for pathway imputation, and second, we analyzed the imputation's impact on predicting tumour relapse. We trained a machine learning model using TCGA data and used it to impute pathways in hospital data. We evaluated the imputation's quality by predicting tumour relapse with clinical features. We also provided explanations for how the imputed pathway scores impact relapse prediction, which can aid clinicians in decision-making.

The method for pathway imputation developed in this study could help future researchers by providing a framework for integrating publicly available data with clinical data to improve the prediction performance of machine learning models aimed at relapse prediction for early-stage lung cancer patients. This approach could be applied to other types of cancer or diseases where publicly available data is abundant, but clinical data is limited. The findings of this study also provide insights into the potential of using public data repositories to augment clinical data for improved patient outcomes in early stage lung cancer treatment. Additionally, the explanations provided for how the imputed pathway scores impact relapse prediction could aid clinicians in decision-making by providing a better understanding of the underlying biological mechanisms involved in tumour relapse and potentially leading to more personalized treatment plans for patients.

1.1. Imputation of Pathway Scores

Imputation predicts missing values in data using available data and relationships within them [10]. Various algorithms [11, 12, 13] are used for imputation, and accurate imputation can save costs and assist physicians in effective prescriptions [14]. Feature selection methods like PCA increase imputation accuracy and decrease computational time [15]. However, feature selection considers only complete data, limiting their applicability to missing-value datasets without enough complete data samples [14]. Genetic information is case-specific and missing in clinical records, so imputation is used to replace missing values [16, 17]. Classical imputation methods like ZEROimpute, ROWimpute, and COLimpute do not account for correlations among genes [18]. K-Nearest Neighbor (KNN) [19], GMCimpute [20], and SLSSimpute [21] are examples of local approaches that utilize local correlation among genes. Matrix completion method is also employed for imputing missing values of gene expression data [22, 23]. SVD imputation [19] and Bayesian principal componenet analysis (BPCA) [24] are based on global correlation information derived from the entire data matrix but become less accurate when the genes exhibit dominant local similarity structures. Mean imputation is commonly used imputation technique but produces biased results [25]. Machine learning models are more sophisticated techniques that involve developing predictive approaches to handle missing values using labeled or unlabeled data [26]. KNN [27], SVM [28], Decision Tree [29], Random Forest Regression [30], XGBoost [31] and Artificial Neural Network (ANN) [32] are some of the mostly used supervised machine learning models for imputation purposes.

1.2. Recurrence Prediction in Lung Cancer

This work extends our previous work on recurrence prediction in lung cancer using machine learning models to explore baseline models for Non-small-cell lung cancer (NSCLC) patients [33, 34, 35]. The problem of predicting relapse can be formulated as a supervised learning problem, where demographic and clinical elements are used as input data and the output variable is the likelihood that the patient will relapse [36, 37]. Several machine learning algorithms have been explored, including penalized logistic regression, decision trees, random forests, multilayer perceptrons, and the Cox proportional hazards model [38]. While statistical algorithms, such as the Cox proportional hazards model [39], have been widely used, they may not accurately capture complex and potentially nonlinear relationships between features [40].

Several studies have utilized artificial neural networks (ANNs) to predict the survival of non-small cell lung cancer (NSCLC) patients using clinical and genomic data. Chen et al. [41] used ANNs with ten selected genes and clinical data, while Hanai et al. [42] integrated clinicopathological and immunohistochemical variables. Marchevsky et al. [43] used clinical-pathological and immunohistochemical variables to predict the survival of early-stage NSCLC patients, and Hsia et al. [44] predicted the survival time in advanced lung cancer patients using genetic polymorphism of specific genes and general patient data. Recently, Jones et al.[45] showed that genomic and clinicopathologic features predicted recurrence better than the TNM system. Chabon et al. [46] developed an ensemble classification framework for discriminating early-stage lung cancer patients. Researchers have proposed using deep learning algorithms, such as Multilayer Perceptrons, Recurrent Neural Networks, and Convolutional Neural Networks, in lung cancer diagnosis [47]. Although early results are encouraging, there is a need for further research as the field is rapidly evolving, and new knowledge is emerging in lung cancer biology and deep learning. Lai et al. [48] proposed an integrative deep neural network, but overfitting and lack of interpretability still remain challenge [49, 50].

2. Methods

2.1. Overview of pathway score imputation and relapse prediction

We took the following steps towards predicting relapse of the early-stage patient as shown in Figure 1. Each block of the pipeline is described as follows:

- **Step 1:** The main aim of this study is to predict the relapse of early-stage lung cancer patients. For this purpose, as an input, we have used TCGA molecular cancer genomic and SLCG EHR data.

- **Step 2:** We found the overlapping feature between the TCGA molecular cancer genomic and SLCG EHR data. We manually inspect the shared feature between the TCGA and SLCG datasets. We found five features that overlap between the two datasets. These features are as follows (i) TNM staging variables, (ii) Gender, (iii) Race, (iv) Age, and (v) Tumor stage. The TNM staging variables and Tumor stage variables are grouped in super categories [51] which allow bundling in similar categories. For example, if the patient has a T1a stage, it

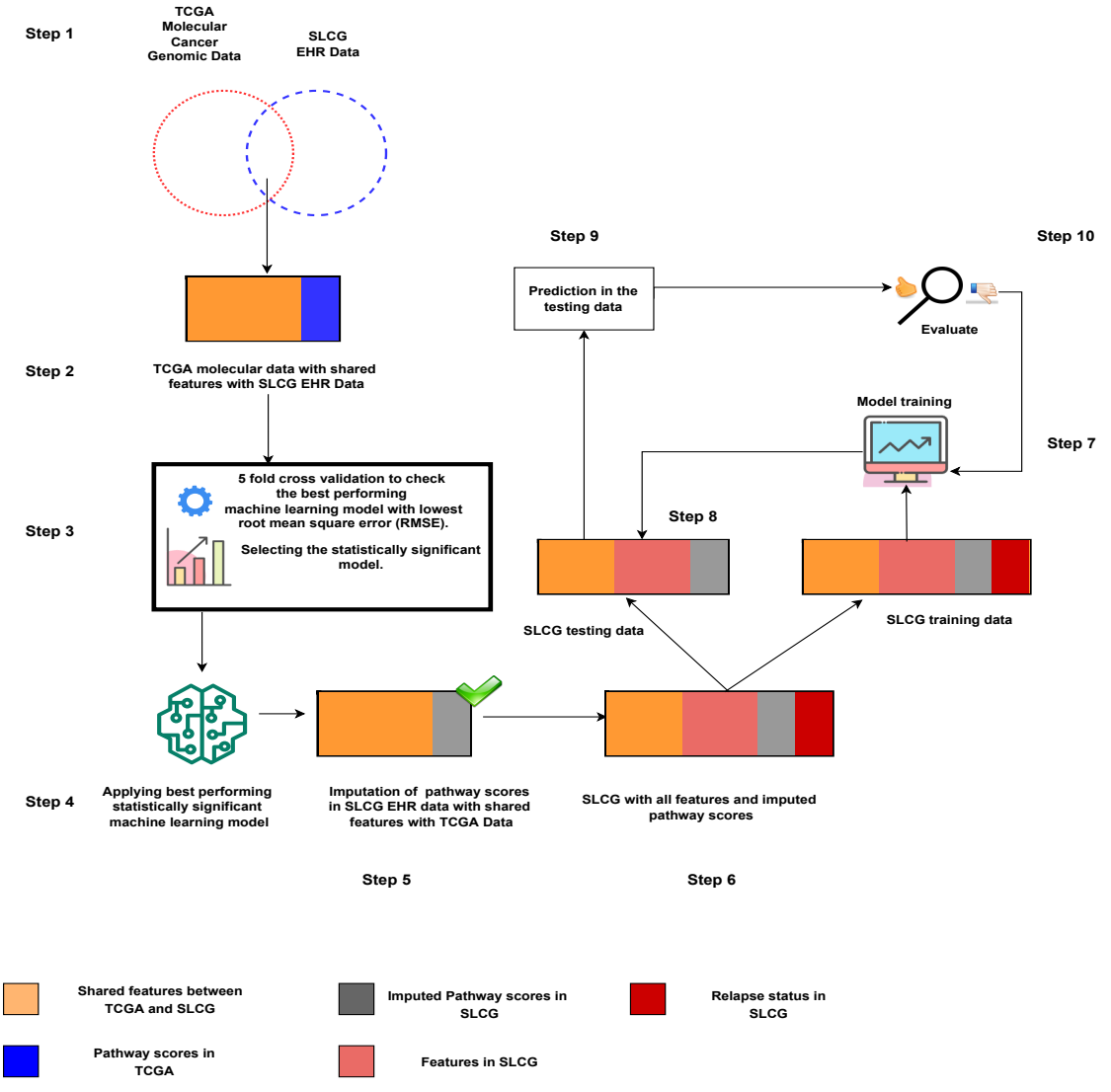


Figure 1: An illustration of the model training pipeline that we executed to apply the predictive models for relapse prediction using imputed pathway score.

is coded as T1 because it is the initial lung cancer stage. The categories for the TNM staging variable are shown in the supplementary material section 4. The overlapping features are represented by orange color code, and the pathway score in TCGA data is color-coded as blue in Figure 1.

- **Step 3:** Once we found the overlapping features, we trained several regression models on the TCGA data to find the best-performing regression models to predict pathway scores. A statistical test has been performed on them to find the significant model.
- **Step 4:** With the best performing significant model, we predicted the pathway scores score for the SLCG data. We called the predicted pathway score an imputed one for the SLCG dataset.
- **Step 5:** The gray color code represents the predicted pathway score in Figure 1.
- **Step 6:** Finally, we combined the other SLCG features shown in pink with the imputed pathway scores to train the classification models to predict the relapse. The relapse status is color-coded as red in the figure.

- **Step 7:** Data is splitted into train and testing set. Different classification model is trained in the training set.
- **Step 8:** Once the trained and optimized classification model is obtained, then it is applied in the testing set.
- **Step 9:** Once the model is applied in the testing set, we get the prediction for the testing set.
- **Step 10:** The prediction is then compared with the ground truth and the evaluation of the classification model is done using different evaluation metrics.

2.2. Imputation methods.

Chlioui et al. [52] demonstrated that traditional machine learning regression-based imputation techniques led to better classification results in cancer data. Therefore, we employed various regression-based methods to check the quality of the prediction of aneuploidy scores for TCGA. Once we finalized the best performing method for continuous variables in that dataset, we used that method to impute the aneuploidy score for the SLCG data. We use a set of eight off-the-shelf supervised machine learning models for continuous variable prediction: Dummy Regression (DR), Linear Regression (LR), Bayesian Regression (BR), Support Vector Regression (SVR), Random Forest Regression (RFR), Gradient Boosting Machine Regression (GBMR), K-Nearest Neighbor Regression (KNNR), and Multi-Layer Perceptron Regression or Neural Network Regression (NNR). A short description of each of these models and how they work is reported in supplementary material section 1.

2.3. Classification models.

Our task is to classify lung cancer patients into the "relapsing" and "non-relapsing" classes. It is a binary classification problem. Therefore, to tackle this problem, we have employed popular supervised machine learning classification algorithms: Support Vector Classification, Logistic Regression, Random Forest Classification, Gradient Boosting Machine Classifier, and Multi-Layer Perceptron Classifier. The regression variant of these algorithms is briefly described above. A short description of each of these models and how they work is reported in supplementary material section 2.

3. Data and experiment

3.1. TCGA Datasets

To identify pathways, publicly available data from The Cancer Genome Atlas were analyzed using PROGENy [53]. Lung small cell cancer data was part of the TCGA pan-cancer analysis project and is referred to as TCGA-LUSC-PAN-CAN-2018. The RNA-seq data v2 were downloaded using the cBioPortal for cancer genomics[1]. PROGENy computes a pathway activity score for 14 of cancer's most important signaling pathways. This computation is based on the patients' gene expression data and a linear model inferred from a large compendium of functional experiments in which each signaling pathway was perturbed/stimulated in functional pre-clinical cell-culture experiments [53]. The result is a table of patient-specific pathway activity scores for all 14 pathways and all patients. Similarly, for the aneuploidy score, we used the lung adenocarcinoma (LUAD) data from TCGA. LUAD is the most common histological subtype of lung cancer, with an average 5-year survival rate of 15%[54]. This data has recorded the aneuploidy score of the patient and is publicly available[2]. In addition, 501 patients have recorded aneuploidy and pathway score.

3.2. SLCG Datasets

This study uses electronic health records (EHRs) of lung cancer patients collected and stored by the SLCG [34, 35]. The cohort contains 1348 early-stage (stage I or II) NSCLC patients, where 491 (36.4%) had tumor relapse after successful treatment. The patient's average age in the dataset is 65.9 for those who relapsed. For those with disease-free survival, it is 65.7. In Table 1, we provide the summary of the patient cohort.

When training various machine learning models, we distinguish between the following types of features extracted from the EHRs:

- **General feature** contains the generic information about the patient, such as age, sex, race, smoking habit, and family cancer history.

[1] https://www.cbioportal.org

[2] https://www.cbioportal.org/study/summary?id=luad_tcga_pan_can_atlas_2018

Features		Relapse	Survival	Total
		491(36.4%)	**857 (63.6%)**	**1348 (100%)**
Age	Mean (Range)	65.9 (33-88)	65.7 (31 - 118)	65.7 (31-118)
Gender	Male	384 (38.0%)	626(62.0%)	1010(74.9%)
	Female	107 (31.7%))	231 (68.3%)	338 (25.1%)
Smoking	Current/Previous	436 (37.1%)	739 (62.9%)	1175 (87.2%)
	Non Smoker	55 (31.8%)	118 (68.2%)	173 (12.8%)
Cancer Stage	I	1 (100.0%)	0 (0.0%)	1 (0.0742%)
	IA	73 (28.7%)	181 (71.3%)	254 (18.8%)
	IA1	8 (32.0%)	17 (68.0%)	25 (1.85%)
	IA3	8 (17.4%)	38 (82.6%)	46 (3.41%)
	IA2	9 (13.6%)	57 (86.4%)	66 (4.9%)
	IIA	107 (45.9%)	126 (54.1%)	233 (17.3%)
	IB	154 (39.0%))	241 (61.0%)	395 (29.3%)
	IIB	131 (39.9%)	197 (60.1%)	328 (24.3%)
T stage	T2a	202 (40.6%)	296 (59.4%)	498 (36.9%)
	T1a	54 (32.3%))	113 (67.7%)	167 (12.4%)
	T2b	73 (42.0%)	101 (58.0%)	174 (12.9%)
	T1b	55 (25.7%)	159 (74.3%)	214 (15.9%)
	T3	81 (39.5%)	124 (60.5%)	205 (15.2%)
	T1c	14 (20.0%)	56 (80.0%)	70 (5.19%)
	Tx	12 (60.0%)	8 (40.0%)	20 (1.48%)
N stage	N0	387 (34.4%)	738 (65.6%)	1125 (83.5%)
	N1	104 (46.6%)	119 (53.4%)	223 (16.5%)
M stage	M0	491 (36.4%)	857 (63.6%)	1348 (100.0%)
Tumor size	Mean (Range)	36.0 (0.8-110.0)	33.2 (1.5-110.0)	34.6 (0.8-110.0)
Ecog status	0	243 (29.2%)	588 (70.8%)	831 (61.6%)
	1	216 (46.7%)	247 (53.3%)	463 (34.3%)
	2	26 (60.5%)	17 (39.5%)	43 (3.19%)
	3	5 (71.4%)	2 (28.6%)	7 (0.519%)
	4	1(100%)	0(0.0%)	1 (0.0742%)
Tumor differentiation	Non specified	137 (39.3%)	212 (60.7%)	349 (25.9%)
	Moderately	55 (26.8%)	150 (73.2%)	205 (15.2%)
	Poorly	44 (42.7%)	59 (57.3%)	103 (7.64%)
	Undifferentiated	3 (50.0%)	3 (50.0%)	6 (0.445%)
	Well	55 (43.0%)	73 (57.0%)	128 (9.5%)
Surgery		434 (32.2%)	835 (61.9%)	1269 (94.1%)
Chemotherapy		353 (26.2%)	305 (22.6%)	658 (48.8%)
Radiotherapy		38 (2.82%)	21 (1.56%)	59 (4.38%)

Table 1: A cohort analysis of the early stage patients which were utilized to build the datasets. Compare with Lung cancer in Spain: information from the Thoracic Tumors Registry (TTR study) for a set of NSCLC patients characteristics (Provencio et al. 2019 [55]) and [34, 35] for a cohort of all NSCLC patients in the dataset.

- **Diagnosis feature** details the tumor classification, histology at the time of diagnosis, and the symptoms and the patient history.
- **Treatment feature** includes the details of any chemotherapy, radiotherapy, or surgical procedures the patient underwent during their treatment.
- **All feature** includes the union of **General, Diagnosis** and **Treatment** features.

Labels. In this study, we do not focus on "when the relapse occurs" which is answered by survival analysis (time aware models). It is the case known as time to recurrence (TTR) which is defined as the time from date of curative surgery to the time of relapse. However, our focus is on whether "relapse happens or not". In order to fetch the data for the machine learning algorithms we label patients as positives if they either i) have a progression record with status "Progression"/"Relapse" or ii) their follow-up records include the status "Alive with disease"/"Dead" (with cause of

death "Lung cancer"). So, those patients who survived today after the surgery and or treatment but without relapse are negative cases but those patients who survived today after the treatment but relapse tomorrow fall into "Alive with disease (with the cause "Lung cancer")" and are positive cases.

3.3. Evaluation standard

- **RMSE:** We chose Root Mean Square Error (RMSE) metric to evaluate the regression model. It is because RMSE gives much more importance to large errors and is widely used in clinical applications to evaluate regression algorithms [56].
- **MAPE:** It is an evaluation metric for regression models which is a scale independent and expressed error as percentage.
- **MAE:** It is an error metric to evaluate regression model. It measures the average of all absolute difference between predicted and true values.
- **PR-AUC and ROC-AUC:** For the classification task, we used the average precision (i.e., the area under the precision-recall curve) (PR-AUC) and the area under the receiver operating characteristics curve (ROC-AUC). Higher values of PR-AUC and ROC-AUC mean better predictive performance across the board (i.e., regardless of specific decision thresholds optimized for precision or recall).

3.4. Hyperparameters Tuning

We have used grid search for tuning the hyperparameters. The hyperparameters and scoring function used by different regression and classification models is shown in the supplementary material section 3.
Source code: All the codes are written in Python 3.7 and are made publicly available[3].

3.5. Performance analysis and comparison of imputation method

We have 15 different pathway scores in our dataset. For each pathway score, we trained eight different regression-based machine learning models. The hyperparameters for these models are selected using a grid search procedure. For each model, we execute a set of 10 trials to find the optimal hyperparameter configuration, where each trial corresponds to a single hyperparameter configuration chosen from a predefined search space. We then select hyperparameters corresponding to the best performing configuration as the model best hyperparameters for each of the examined models.

The performance of each model is demonstrated in Figure 2. We can see that for the pathways *Aneuploidy, Trail, WNT, VEGF,* and *EGFR* the regression model predicted the average RMSE scores as less than 100 in comparison to other pathway scores. However, the model did not predict the lowest error for the rest of the pathway scores. Therefore, we included these four pathway scores in the relapse prediction. In the next experiment, we investigated whether including these features will boost the relapse prediction performance.

For *Aneuploidy* score, SVR performed better than other methods. Similarly, for *WNT, Trail* and *EGFR*, RFR has the lowest RMSE. On a similar note, for *VEGF*, KNNR has the lowest RMSE. However, these regression models have a marginal improvement over their competitors. Therefore, to identify the significant model, we applied the Wilcoxon signed rank test, a non-parametric test that does not rely on assumptions that data belongs to any particular distribution and is used for comparison between different machine learning models [57]. Thus we compute the p-value for the Wilcoxon signed-rank test between best performing models with the other compared regression methods. The result of the test can be found in supplementary material section 7. We can see that to predict *Aneuploidy* score SVR has significant difference with LR (p-value:0.001), GBMR (p-value:0.001) and NNR (p-value:0.013). For WNT, RFR significantly differs from KNNR (p-value 0.009). For VEGF, KNNR has a significant difference with LR (p-value 0.019). For Trail, RFR has a significant difference with GBMR (p-value 0.027). Similarly, for EGFR, RFR significantly differs from GBMR (p-value 0.001). Those model which does not have a significant difference from the best-performing model means that we have no evidence that the best-performing models are superior to the rest of the methods.

In terms of MAPE and MAE evaluation metric to predict pathway scores we see the same trend. For the lower value of RMSE, there is also a lower value for MAPE and MAE. For WNT we observe the tie between SVR and RFR with MAPE metric. Similarly, for EGFR we observe the tie between RFR and NNR with MAE metric.)

[3] https://github.com/timilsinamohan/SynergyIntegrationPathways

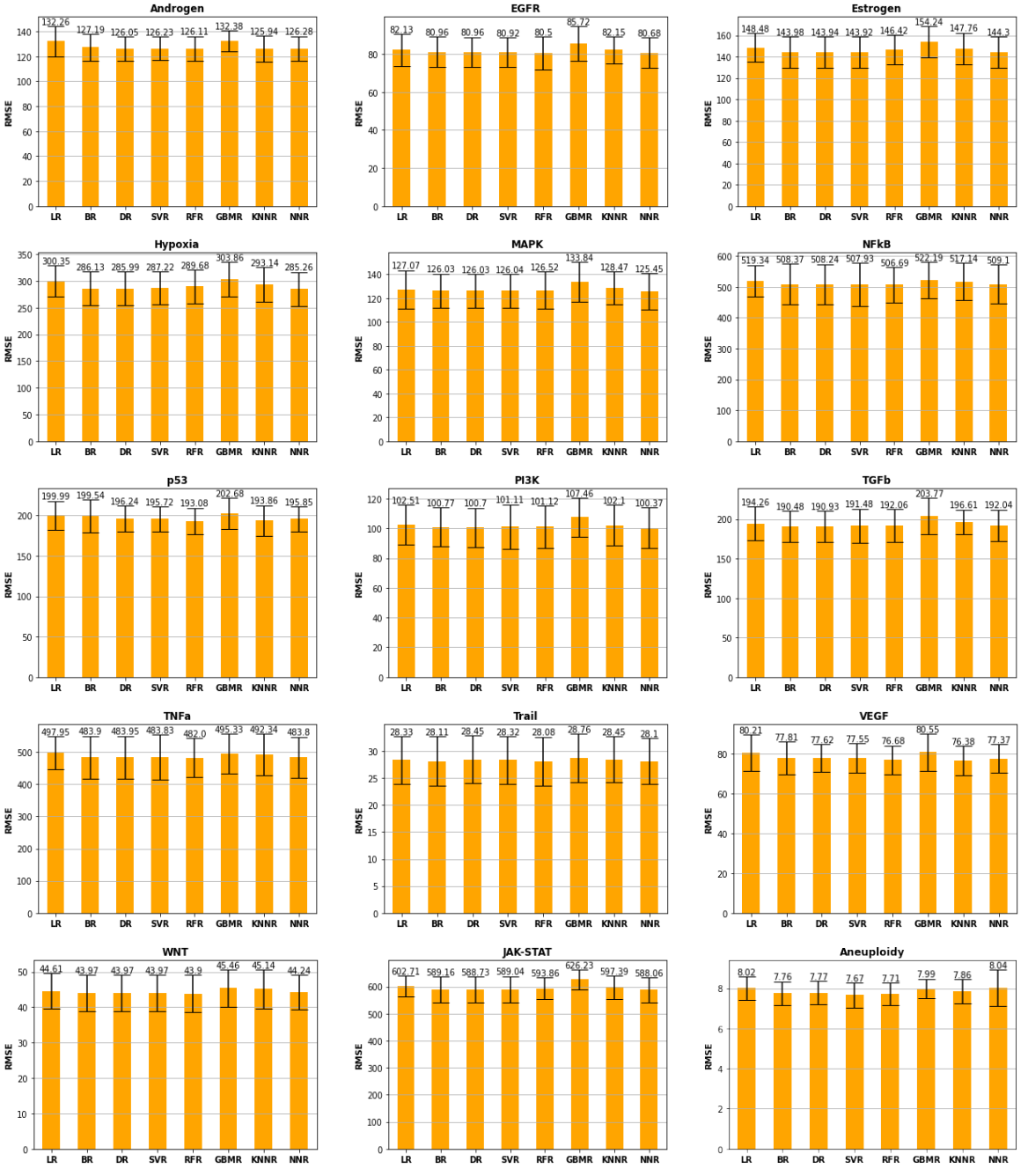


Figure 2: Bar chart is the Mean RMSE score of five-fold cross-validation to predict scores in the TCGA dataset. Each bar chart shows the mean RMSE score using different algorithms. The error bar is the standard deviation obtained from the five-fold cross-validation for the RMSE score.

3.6. Relapse Prediction

Table 2 and Table 3 demonstrates the results of our computational evaluation of the five examined models using a 5-fold cross-validation evaluation strategy. Using only all imputed features (Aneupoidy+WNT+VEGF+Trail+EGFR) to build a relapse prediction model, we observe the model has slightly better performance than the random model whose ROC_AUC = 0.5. However, including the imputed features with the clinical feature's performance of the model started to improve. We found that RF achieved the best result in ROC-AUC and PR-AUC using all clinical features,

Features	Evaluation Metric	SVC	LR	RF	GBM	NNC
All imputed Features (Aneuploidy + WNT +VEGF + Trail + EGFR)	Accuracy	0.64 ± 0.01	0.63 ± 0.01	0.62 ± 0.01	0.62 ± 0.02	0.63 ± 0.01
	Precision	0.2 ± 0.4	0.32 ± 0.2	0.31 ± 0.16	0.47 ± 0.08	0.21 ± 0.2
	Recall	0.01 ± 0.0	0.04 ± 0.04	0.06 ± 0.03	0.16 ± 0.07	0.03 ± 0.04
	F1	0.01 ± 0.01	0.08 ± 0.07	0.1 ± 0.05	0.22 ± 0.09	0.06 ± 0.07
	ROC_AUC	0.57 ± 0.04	0.58 ± 0.03	0.58 ± 0.03	0.58 ± 0.05	0.57 ± 0.03
	PR_AUC	0.41 ± 0.02	0.43 ± 0.04	0.42 ± 0.01	0.43 ± 0.05	0.44 ± 0.03
All clinical features	Accuracy	0.74 ± 0.02	0.74 ± 0.03	0.76 ± 0.03	0.74 ± 0.04	0.72 ± 0.03
	Precision	0.79 ± 0.08	0.75 ± 0.08	0.76 ± 0.06	0.71 ± 0.09	0.65 ± 0.06
	Recall	0.41 ± 0.07	0.46 ± 0.02	0.49 ± 0.06	0.49 ± 0.09	0.51 ± 0.08
	F1	0.53 ± 0.06	0.57 ± 0.03	0.59 ± 0.06	0.57 ± 0.07	0.57 ± 0.06
	ROC_AUC	0.78 ± 0.04	0.78 ± 0.04	0.79 ± 0.02	0.79 ± 0.03	0.76 ± 0.04
	PR_AUC	0.71 ± 0.05	0.71 ± 0.05	0.74 ± 0.05	0.73 ± 0.05	0.68 ± 0.05
All clinical feature + Aneuploidy	Accuracy	0.74 ± 0.03	0.75 ± 0.03	0.76 ± 0.02	0.75 ± 0.03	0.73 ± 0.03
	Precision	0.78 ± 0.08	0.74 ± 0.07	**0.81 ± 0.09**	0.77 ± 0.08	0.68 ± 0.07
	Recall	0.41 ± 0.07	0.50 ± 0.05	0.45 ± 0.04	0.48 ± 0.1	0.51 ± 0.07
	F1	0.54 ± 0.07	0.59 ± 0.04	0.58 ± 0.04	0.58 ± 0.08	0.58 ± 0.06
	ROC_AUC	0.77 ± 0.03	0.77 ± 0.04	0.79 ± 0.03	0.79 ± 0.03	0.78 ± 0.05
	PR_AUC	0.71 ± 0.04	0.7 ± 0.05	0.74 ± 0.05	0.73 ± 0.05	0.71 ± 0.06
All clinical feature + Aneuploidy +WNT	Accuracy	0.75 ± 0.03	0.74 ± 0.03	0.75 ± 0.03	0.75 ± 0.02	0.73 ± 0.03
	Precision	0.79 ± 0.07	0.72 ± 0.07	**0.81 ± 0.09**	0.79 ± 0.09	0.66 ± 0.07
	Recall	0.42 ± 0.06	0.47 ± 0.06	0.43 ± 0.05	0.46 ± 0.1	**0.55 ± 0.08**
	F1	0.55 ± 0.06	0.56 ± 0.05	0.56 ± 0.05	0.57 ± 0.07	0.6 ± 0.05
	ROC_AUC	0.78 ± 0.03	0.77 ± 0.04	**0.80 ± 0.04**	0.79 ± 0.03	0.76 ± 0.04
	PR_AUC	0.71 ± 0.04	0.69 ± 0.06	0.74 ± 0.05	0.73 ± 0.05	0.68 ± 0.04

Table 2: A Comparison between five supervised learning models on 5-fold cross validation to predict recurrence in NSCLC patients as a binary classification task using various features and evaluation metrics. Bold figures indicate the best performance or ties in the respective features and evaluation metric.

imputed Aneuploidy, WNT, VEGF, and Trail pathway scores. It achieved ROC-AUC and PR-AUC of 0.80 and 0.75, respectively. Similarly, trained with the same features, we found that the NNC has the best performance in Recall and F1 score of 0.55 and 0.61. We do not see the performance boost with all clinical features, imputed Aneuploidy, WNT, VEGF, and adding EGFR except for the marginal improvement over the Accuracy metric. One important thing to observe here is the difference between the predictive performance using all clinical features versus all clinical features with the imputed pathway scores. In Table 4, we demonstrated that all clinical features + Aneuploidy + WNT +VEGF + Trail achieved overall clinical features without imputed pathway scores.

In most cases, RF trained with All clinical features + Aneuploidy + WNT +VEGF +Trail has better performances than using only All clinical features. In terms of accuracy metric, we observe the ties. SVC trained with all clinical features with pathway score gives 1.26% in Precision, NNC gives 7.84% in Recall, and 3.38% in F1. Similarly, RF gives 1.26% in ROC_AUC and 1.35% in PR_AUC. It suggests that incorporating imputed pathway scores using publicly available genomic data leads to more accurate predictions. Even a slight improvement in prediction accuracy significantly impacts such a complex and life-threatening problem of relapse prediction in lung cancer. RF has slightly better performance in ROC_AUC and PR_AUC and ties with the Accuracy evaluation metric. Hence we checked the significance of the marginal improvement of ROC_AUC and PR_AUC using augmented pathway scores versus only clinical features.

The results of the 5 folds cross validation and the result of Wilcoxon signed rank test can be found in supplementary material section 6. We found that there is a significant difference in prediction performed by Random Forest augmenting pathway scores as features with respect to Random Forest using only clinical features in terms of ROC_AUC evaluation metric($p-value = 0.031$). However, there is no significant difference with respect to predictions performed by Random

Features	Evaluation Metric	SVC	LR	RF	GBM	NNC
All clinical feature + Aneuploidy +WNT+ VEGF	Accuracy	0.75 ± 0.03	0.74 ± 0.03	0.77 ± 0.03	0.75 ± 0.04	0.73 ± 0.03
	Precision	0.79 ± 0.06	0.73 ± 0.08	0.79 ± 0.07	0.79 ± 0.11	0.66 ± 0.04
	Recall	0.43 ± 0.07	0.48 ± 0.06	0.5 ± 0.09	0.44 ± 0.09	0.52 ± 0.09
	F1	0.55 ± 0.07	0.58 ± 0.06	0.61 ± 0.06	0.56 ± 0.08	0.58 ± 0.06
	ROC_AUC	0.77 ± 0.03	0.77 ± 0.04	0.79 ± 0.03	0.78 ± 0.03	0.77 ± 0.04
	PR_AUC	0.7 ± 0.03	0.70 ± 0.06	0.74 ± 0.05	0.73 ± 0.05	0.7 ± 0.04
All clinical feature + Aneuploidy+ WNT+ VEGF + Trail	Accuracy	0.75 ± 0.02	0.74 ± 0.03	0.76 ± 0.03	0.75 ± 0.04	0.75 ± 0.02
	Precision	0.8 ± 0.06	0.71 ± 0.06	0.78 ± 0.07	0.72 ± 0.1	0.69 ± 0.04
	Recall	0.43 ± 0.06	0.49 ± 0.05	0.49 ± 0.09	0.53 ± 0.08	**0.55 ± 0.06**
	F1	0.56 ± 0.06	0.58 ± 0.05	0.59 ± 0.07	0.6 ± 0.06	**0.61 ± 0.05**
	ROC_AUC	0.77 ± 0.03	0.76 ± 0.04	**0.80 ± 0.04**	0.78 ± 0.04	0.77 ± 0.04
	PR_AUC	0.71 ± 0.04	0.69 ± 0.05	**0.75 ± 0.05**	0.72 ± 0.06	0.72 ± 0.05
All clinical feature + Aneuploidy+ WNT+ VEGF + Trail + EGFR	Accuracy	0.75 ± 0.03	0.74 ± 0.03	**0.76 ± 0.03**	0.75 ± 0.03	0.72 ± 0.02
	Precision	0.79 ± 0.07	0.71 ± 0.07	0.79 ± 0.08	0.75 ± 0.08	0.65 ± 0.04
	Recall	0.42 ± 0.06	0.5 ± 0.08	0.48 ± 0.07	0.49 ± 0.03	0.52 ± 0.04
	F1	0.55 ± 0.06	0.58 ± 0.07	0.59 ± 0.06	0.59 ± 0.04	0.58 ± 0.04
	ROC_AUC	0.77 ± 0.04	0.76 ± 0.05	0.79 ± 0.03	0.78 ± 0.03	0.75 ± 0.02
	PR_AUC	0.71 ± 0.05	0.70 ± 0.06	0.74 ± 0.0	0.72 ± 0.04	0.68 ± 0.01

Table 3: A Comparison between five supervised learning models on 5-fold cross validation to predict recurrence in NSCLC patients as a binary classification task using various features and evaluation metrics. Bold figures indicate the best performance or ties in the respective features and evaluation metric.

Forest augmenting pathway scores as features with respect to Random Forest using only clinical features in terms of PR_AUC evaluation metric.

For the demonstration purpose, we have visualized a decision tree in the RF ensemble and provided the explanation which is reported in supplementary material section 5.

	Accuracy	Precision	Recall	F1	ROC_AUC	PR_AUC
All clinical features	0.76 (RF)	0.79 (SVC)	0.51 (NNC)	0.59 (RF)	0.79 (RF or GBM)	0.74 (RF)
All clinical feature + Aneuploidy+ WNT+ VEGF + Trail	0.76 (RF)	0.80 (SVC)	0.55 (NNC)	0.61 (NNC)	0.80 (RF)	0.75 (RF)
Relative improvement	0%	**1.26%**	**7.84%**	**3.38%**	**1.26%**	**1.35%**

Table 4: Summary of the best performing models across various evaluation metric. The last row presents relative improvement of using All clinical features + Aneuploidy + WNT +VEGF +Trail achieved over the all features without imputed pathway scores.

The most important features were ranked through Shapley Additive Explanations (SHAP) [58]. SHAP helps generate a summary plot for the model, demonstrating the order of feature importance and magnitude of each feature's contribution to making predictions. The bar plot in Figure 4 A shows those features in order of importance by summarising all the individual predictions to provide an overall picture of their impact; this is based on the mean absolute value of the SHAP values of each feature. Similarly, Figure 4 B shows the summary plots, each dot represents a patient in the dataset, and the color represents the feature value (red is high, blue is low). All feature values on the left of the vertical line at 0 have a negative impact on the predicted outcome, while those on the right have a positive impact. In the barplot, we can see that first-line intravenous chemotherapy is ranked highest. As the value for this feature increases, it positively affects the prediction of relapse. For the "Trail and aneuploidy imputed" feature, the

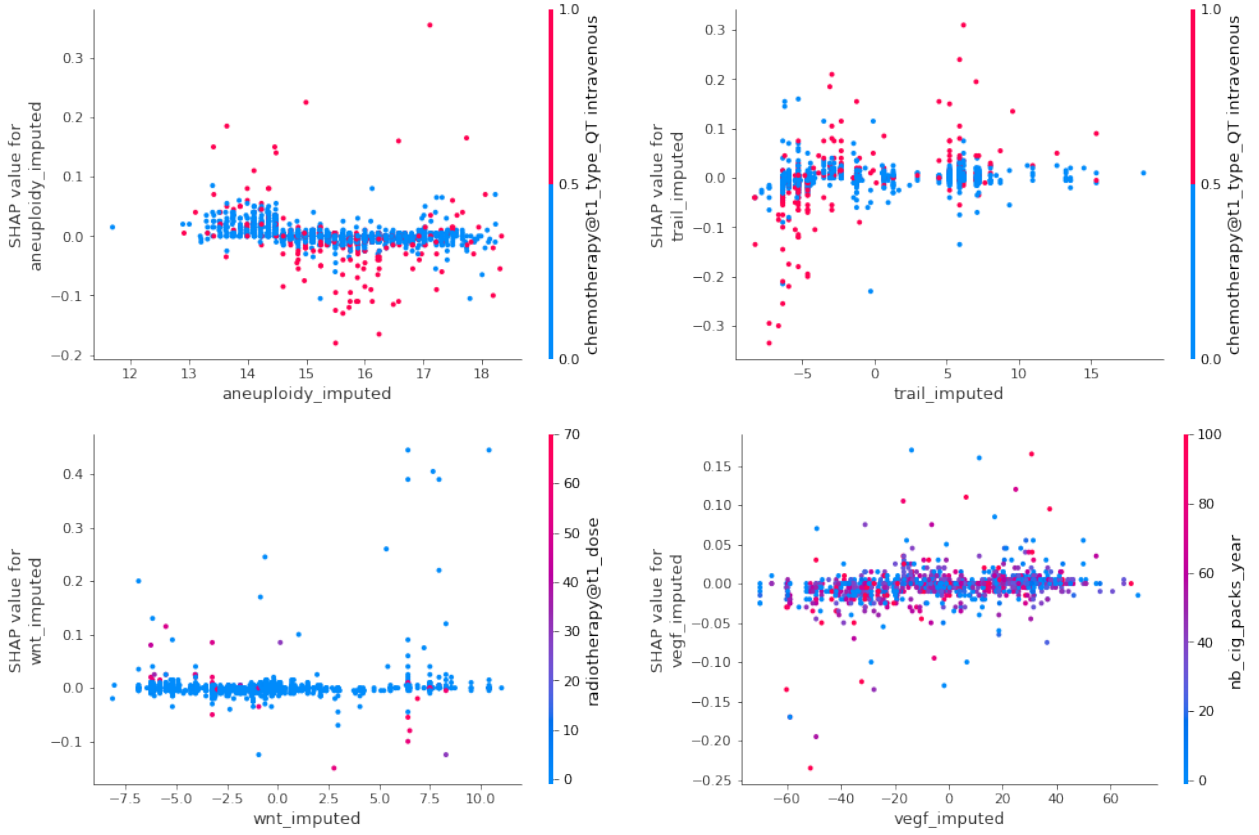
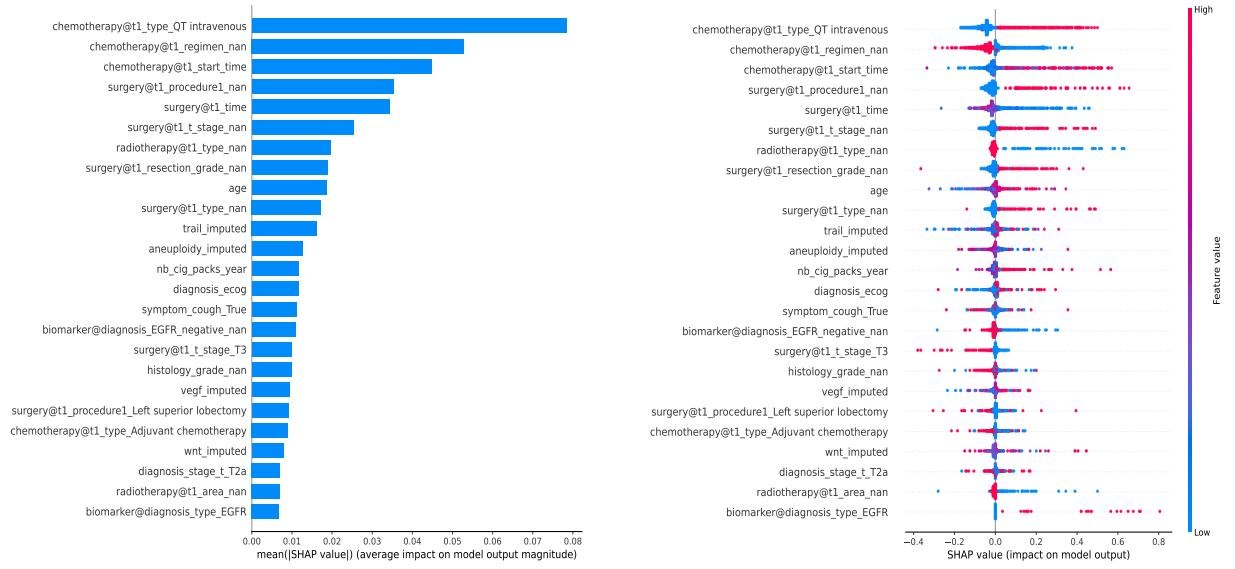


Figure 3: Dependency plot between imputed pathway scores and the interacting features.

RF model has ranked it 11^{th} and 12^{th} position, respectively, as shown in the barplot. VEGF is ranked 19^{th} and WNT imputed is ranked 22^{nd} important feature.

Similarly, in the summary plot, we can see that as the Trail imputed score increases, there is a high risk of relapse. Likewise, as VEGF score increases, there is also a high risk of relapse. However, with Aneuploidy and WNT, we can see red dots on both sides of the center. Therefore, these features are more complex and need further investigation. So, we further investigated the interaction of the imputed pathway scores using the dependency plot. A dependence plot is also one of the features provided by the SHAP package for capturing interaction effects. For Aneuploidy imputed, we can see from Figure 3 the interacting feature is chemotherapy@t1_type_ QT intravenous. Each dot in the dependency plot is a patient. The x-axis is the Aneuploidy imputed, and the y-axis is the SHAP value attributed to their chemotherapy@t1_type_ QT intravenous. Higher SHAP values represent a higher risk of relapse due to aneuploidy scores. Coloring each dot by the patient's chemotherapy@t1_type_ QT intravenous reveals that a high aneuploidy score is more concerning to the model for those few patients who are conducted with intravenous chemotherapy. In the case of Trail imputed, we also saw the interacting variable is chemotherapy@t1_type_ QT intravenous. As the Trail imputed score increases, the SHAP value also starts to increase, which means the risk of relapse also increases. A group of patients who are conducted with intravenous chemotherapy with low Trail imputed scores is relapse-free. In the case of wnt imputed, it interacts with radiotherapy doses. Only a few patients are at a high risk of relapse for the high wnt imputed score. A large group of patients with increased wnt imputed score and lower radiotherapy doses (less than or equal to 20) have low SHAP values with a low risk of relapse. Similarly, the VEGF imputed interacts with the number of cigarettes consumed by the patient annually. Therefore, those patients who are heavy smokers with high VEGF imputed scores are at high risk of relapse. However, there are also a few patients who are heavy smokers but with low VEGF scores at low risk of relapse.



(a) Bar Plot of the important features.

(b) Summary Plot of the features.

Figure 4: Relapse Prediction

4. Discussion

While pathway information can lead to more informed decisions in the context of lung cancer treatment, genetic testing for pathway screening is expensive and requires a longer waiting time. Patients may not be willing to do the extra tests and oncologists may not be able to wait for the result of genetic tests before making a decision on the next treatment steps. Therefore, a method like the one introduced in this paper—imputing such information into patient data using machine learning—may be a basis for a machine-aided supportive approach for the oncologist to understand a possible pathway profile of individual patients' disease without the need for practically infeasible generic screening.

The main goal of this work was to impute pathway scores and assess their usability in predicting relapse in early-stage non-small cell lung cancer patients. To that end, we used real EHR data on a comprehensive patient cohort from NSCLC patients selected from the Spanish Lung Cancer Group's Thoracic Tumor Registry (TTR). The integration of imputed pathway scores in the clinical data has boosted the prediction accuracy of predicting relapse in early-stage lung cancer. Our approach also enables rapid and easy experimentation with a broad range of possibly complementary models. One thing that is of note in this context is that for different pathways score predictions, there is not only one fixed model. For instance, in aneuploidy, SVR performed better. Likewise for Trail RFR regression model performed better than the regression model. Despite these differences, however, our experiments clearly demonstrated that the imputation provided a robust and sensitive approach to estimating missing data that improved the performance of relapse prediction models. We also showed how this can lead to interesting contextual findings via a prototype prediction explanation subsystem based on SHAP values that lets the users interpret the relapse prediction results and the specific contribution of the imputed pathway score.

Another important observation from our data driven approach is that the model picked surgery as the important feature. It is correct that surgical skills and techniques are central components of a surgeon's skill set and directly correlate with patient benefits. However, in our data we do not have any qualitative and quantitative information (kinematic data) about the surgeon's hands-on skills [59]. We have only the information about the surgery procedure, time, resection, type and stages. We can only speculate that the predictability of relapse prediction might be directly influenced by the surgeon's skill and knowledge but we do not have the evidence to prove this.

Predicting NSCLC relapse after surgery is crucial for bespoke monitoring plans and adjuvant therapies [60]. NSCLC which is considered to be a solid tumour also has a favourable impact from pathway therapy. It is due to the fact that pathway scores are the important tumour drivers. Therefore, our approach of imputing missing genetic pathway information in our hospital cohort patient aids to strengthen the relapse prediction. Our previous work [34] discussed baseline models that show clear promise for accurate lung cancer recurrence predictions, that are based not

on population statistics but on individual features of the patients. We now add imputed genetic pathway scores and explainable artificial intelligence methods to make sure clinicians make informed decisions when adopting the system in patient follow-ups. The brief example of the explainable system is reported in in the supplementary material section 8.

As missing data are unavoidable in clinical research, thus imputation techniques can help reducing bias and completeness of the medical data [61]. Our methods can also be generalized to other variants of the cancer such as blood [62],breast and colon [63] for genetic information imputation. We can impute the pathway scores from publicly available TCGA cohort and inference them in the missing hospital EHR data based on the shared features between TCGA and hospital patient.

The idea of imputing genomic data from public patient cohorts can be applied to various informatics-related problems, particularly in situations where the availability of genomic data is limited. For example, it can be used to develop predictive models for rare genetic diseases or to identify genetic variants that influence drug response in pharmacogenomics. This approach can also improve the accuracy of predictive models used in precision medicine by providing more personalized treatment options for patients. In infectious disease research, imputing genomic data can be used to identify genetic variants associated with disease susceptibility and develop targeted prevention and treatment strategies. Finally, in environmental health, imputing genomic data can help identify genetic variants that modify the effect of environmental exposures on health outcomes, leading to targeted prevention and intervention strategies. Overall, imputing genomic data from public patient cohorts is a promising approach that can improve our understanding of complex diseases and lead to more personalized and effective healthcare. By utilizing publicly available data repositories and integrating them with clinical data, this technique can be applied to a wide range of cancer types and genetic diseases. Not only that, but this approach can be extended to other data types, including proteomics, metabolomics, and epigenomics, to improve the accuracy of machine learning models aimed at predicting disease risk and treatment response. By imputing pathway scores and combining them with other clinical features, such as demographic data, medical history, and lifestyle factors, we can gain a more complete understanding of disease risk and treatment response. This can lead to the identification of novel biomarkers and therapeutic targets, ultimately helping us to improve patient outcomes and reduce the burden of these devastating diseases.

One limitation of this research is that it relies on imputing genetic pathway scores from a publicly available TCGA lung cancer cohort into similar records in the Spanish Lung Cancer Group (SLCG) data. The use of imputed data, rather than direct measurements, could potentially introduce errors and inaccuracies into the predictions by the machine learning models. Another limitation is that imputing genetic information may not fully capture the complexity of the patient's individual genetic profile. The imputed genetic information may only represent a subset of relevant genetic features, and there may be other genetic factors that are not captured by the imputation. The imputation may not capture the full range of genetic variation relevant to the disease. Some genetic variants may not be captured by the imputation method used, and these variants may be important in predicting disease outcomes.

Based on the results and findings of this study, there are several potential avenues for future work. One possibility is to extend the approach to other types of cancer and genetic diseases beyond lung cancer. This could involve leveraging publicly available data repositories and integrating them with clinical data to impute pathway scores and investigate their association with clinical outcomes. In addition, the approach presented in this study can also be used to identify novel biomarkers and therapeutic targets by identifying deregulated pathways and investigating their biological functions and interactions. By integrating pathway scores with other data sources and leveraging advanced analytical techniques, it may be possible to identify new targets for drug development and personalized treatment strategies.

In this study, our focus is not on validating imputed genetic pathway scores in the laboratory, rather it is on exploring the impact of imputation methods on models' performance in predicting tumour relapse, and more importantly, the impact of imputed scores and its explanations for the relapse prediction. For example, clinicians might less likely order genetic pathway tests for relapse free patients in comparison to those with risk factors of relapse. Thus, for relapse free patients, in this case, there might be more missing pathway scores. Using computationally cheap machine learning models we can impute such missing information which aid for enriching prediction models.

5. Conclusion

We presented the imputation of different pathway scores using the publicly available lung cancer genomic data from TCGA in the SLCG cohort to predict the relapse of the early-stage patient. Among 15 pathway scores, we found that Aneuploidy, WNT, VEGF, Trail, and EGFR scores are reasonably predicted by regression models using the shared

features between TCGA and SLCG. Incorporating these scores in the SLCG cohort along with general, diagnostic, and treatment features, we found a boost in the performance of classifier models for relapse prediction, particularly in Precision, Recall, F1, ROC_AUC, and PR_AUC metrics. One important observation here is that even though we see the boost in the performance, the difference with the classifier trained with all clinical features is not very high (cf Table 4). One reason might be the shared feature between the TCGA and SLCG. We only found (i) TNM staging variables, (ii) Gender, (iii) Race, (iv) Age, and (v) Tumor stage. Unfortunately, no molecular profile data of the patient is available in SLCG. This information could have further enriched the pathway imputation, and relapse prediction would have improved.

6. Statement of significance

Problem: The main problem addressed in this study is the lack of complete genomic data for all patients, which can limit the accuracy of predictive models for relapse in early stage non-small cell lung cancer. The study aims to address this issue by imputing genetic pathway scores from a publicly available TCGA lung cancer cohort into similar records in the Spanish Lung Cancer Group (SLCG) data, specifically in 1348 early-stage patients.

What is already known: The previous approaches used in studies that have explored the integration of genomic data with clinical information for predicting relapse in early stage non-small cell lung cancer. These approaches include gene expression profiling, whole genome sequencing, targeted sequencing, machine learning models, and biomarker identification.

What this paper adds: The study builds upon previous research that has shown the potential of integrating genomic data with clinical information to improve the accuracy of predictive models for relapse in early stage non-small cell lung cancer. The study demonstrates that imputing genetic pathway scores from a publicly available TCGA lung cancer cohort into similar records in the Spanish Lung Cancer Group (SLCG) data can improve the accuracy of machine learning models for predicting relapse. The findings have important implications for the development of personalized treatment plans and improved patient outcomes.

Funding

This work is funded by the Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 and the CLARIFY project funded by European Commission under the grant number 875160.

Declaration of Competing Interest

The authors declare that they have no known competing interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Mohan Timilsina: M.T conducted experiments, analysed the results and prepared the original draft.. **Dirk Fey:** D.F provided the guidance and feedback for the experiments and revised the draft.. **Samuele Buosi:** S.M revised the manuscript.. **Adrianna Janik:** A.Y provided the feedback on the experiments and revised the draft.. **Luca Costabello:** L.C revised the draft.. **Enric Carcereny:** E.C revised the draft.. **Delvys Rodrıguez Abreu:** D.R.A revised the draft.. **Manuel Cobo:** M.C revised the draft.. **Rafael López Castro:** R.L.C revised the draft.. **Reyes Bernabé:** R.B revised the draft. **Pasquale Minervini:** P.M revised the draft.. **Maria Torrente:** Ma.T revised the draft.. **Mariano Provencio:** M.P revised the draft.. **Vít Nováček:** V.N provided the guidance and extensively reviewed the manuscript .

Acknowledgements

We would like to thank Unit for Information Mining and Retrieval (UIMR) at Data Science Institute, Insight Center for Data Analytics for providing us the computational infrastructure.

References

[1] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[2] Mohan Timilsina, Meera Tandan, and Vít Nováček. Machine learning approaches for predicting the onset time of the adverse drug events in oncology. *Machine Learning with Applications*, page 100367, 2022.
[3] André CA Nascimento, Ricardo BC Prudêncio, Marcilio CP De Souto, and Ivan G Costa. Mining rules for the automatic selection process of clustering methods applied to cancer gene expression data. In *International Conference on Artificial Neural Networks*, pages 20–29. Springer, 2009.
[4] Jian Ren, Kubra Karagoz, Michael Gatza, David J Foran, and Xin Qi. Differentiation among prostate cancer patients with gleason score of 7 using histopathology whole-slide image and genomic data. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, page 1057904. International Society for Optics and Photonics, 2018.
[5] Yotam Drier, Michal Sheffer, and Eytan Domany. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, 110(16):6388–6393, 2013.
[6] Andrea H Bild, Guang Yao, Jeffrey T Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary-Beth Joshi, David Harpole, Johnathan M Lancaster, Andrew Berchuck, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(7074):353–357, 2006.
[7] Elke K Markert, Hideaki Mizuno, Alexei Vazquez, and Arnold J Levine. Molecular classification of prostate cancer using curated expression signatures. *Proceedings of the National Academy of Sciences*, 108(52):21276–21281, 2011.
[8] Michalis Alevizakos, Serafim Kaltsas, and Konstantinos N Syrigos. The vegf pathway in lung cancer. *Cancer chemotherapy and pharmacology*, 72(6):1169–1181, 2013.
[9] Luis A Diaz Jr. The current clinical value of genomic instability. In *Seminars in cancer biology*, volume 15, pages 67–71. Elsevier, 2005.
[10] Chengsheng Mao, Liang Yao, and Yuan Luo. Medgcn: Medication recommendation and lab test imputation via graph convolutional networks. *Journal of Biomedical Informatics*, 127:104000, 2022.
[11] Akbar K Waljee, Ashin Mukherjee, Amit G Singal, Yiwei Zhang, Jeffrey Warren, Ulysses Balis, Jorge Marrero, Ji Zhu, and Peter DR Higgins. Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, 3(8):e002847, 2013.
[12] Yuan Luo, Peter Szolovits, Anand S Dighe, and Jason M Baron. Using machine learning to predict laboratory test results. *American journal of clinical pathology*, 145(6):778–788, 2016.
[13] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.
[14] Shatha Awawdeh, Hossam Faris, and Hazem Hiary. Evoimputer: An evolutionary approach for missing data imputation and feature selection in the context of supervised learning. *Knowledge-Based Systems*, 236:107734, 2022.
[15] Chia-Hui Liu, Chih-Fong Tsai, Kuen-Liang Sue, and Min-Wei Huang. The feature selection effect on missing value imputation of medical datasets. *Applied Sciences*, 10(7):2344, 2020.
[16] Allison W Kurian, Kent A Griffith, Ann S Hamilton, Kevin C Ward, Monica Morrow, Steven J Katz, and Reshma Jagsi. Genetic testing and counseling among patients with newly diagnosed breast cancer. *Jama*, 317(5):531–534, 2017.
[17] Türem Delikurt, Graham R Williamson, Violetta Anastasiadou, and Heather Skirton. A systematic review of factors that act as barriers to patient referral to genetic services. *European Journal of Human Genetics*, 23(6):739–745, 2015.
[18] Ash A Alizadeh, Michael B Eisen, R Eric Davis, Chi Ma, Izidore S Lossos, Andreas Rosenwald, Jennifer C Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.
[19] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.
[20] Ming Ouyang, William J Welsh, and Panos Georgopoulos. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, 20(6):917–923, 2004.
[21] Xiaobai Zhang, Xiaofeng Song, Huinan Wang, and Huanping Zhang. Sequential local least squares imputation estimating missing value of microarray data. *Computers in biology and medicine*, 38(10):1112–1120, 2008.
[22] Arnav Kapur, Kshitij Marwah, and Gil Alterovitz. Gene expression prediction using low-rank matrix completion. *BMC bioinformatics*, 17(1):1–13, 2016.
[23] Mohan Timilsina, Haixuan Yang, Ratnesh Sahay, and Dietrich Rebholz-Schuhmann. Predicting links between tumor samples and genes using 2-layered graph based diffusion approach. *BMC bioinformatics*, 20(1):1–20, 2019.
[24] Shigeyuki Oba, Masa-aki Sato, Ichiro Takemasa, Morito Monden, Ken-ichi Matsubara, and Shin Ishii. A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.
[25] Aisyah Mat Jasin, Daniel Neagu, and Attila Csenki. The wild bootstrap resampling in regression imputation algorithm with a gaussian mixture model. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 218–230. Springer, 2018.
[26] Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. A survey on missing data in machine learning. *Journal of Big Data*, 8(1):1–37, 2021.
[27] Jesus Maillo, Sergio Ramírez, Isaac Triguero, and Francisco Herrera. knn-is: An iterative spark-based design of the k-nearest neighbors classifier for big data. *Knowledge-Based Systems*, 117:3–15, 2017.
[28] Feng Honghai, Chen Guoshun, Yin Cheng, Yang Bingru, and Chen Yumei. A svm regression based approach to filling in missing values. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 581–587. Springer, 2005.
[29] Bhekisipho Twala. An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, 23(5):373–405, 2009.
[30] Fei Tang and Hemant Ishwaran. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.
[31] G Madhu, B Lalith Bharadwaj, G Nagachandrika, and K Sai Vardhan. A novel algorithm for missing data imputation on machine learning. In *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 173–177. IEEE, 2019.
[32] Simon Lebech Cichosz, Morten Hasselstrøm Jensen, and Ole Hejlesen. Short-term prediction of future continuous glucose monitoring readings in type 1 diabetes: Development and validation of a neural network regression model. *International Journal of Medical Informatics*,

151:104472, 2021.
[33] Mohan Timilsina, Samuele Bousi, Dirk Fey, Adrianna Janik, Maria Torrente, Mariano Provencio, Alberto Bermúdez, Enric Carcereny, Luca Costabello, Delvys Abreu, Manuel Cobo, Rafael Castro, Reyes Bernabé, Maria Guirado, Pasquale Minervini, and Vít Novácek. Integration of clinical information and imputed aneuploidy scores to enhance relapse prediction in early stage lung cancer patients. In *AMIA 2022, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 5-9, 2022*. AMIA, 2022.
[34] Sameh K Mohamed, Brian Walsh, Mohan Timilsina, Maria Torrente, Fabio Franco, Mariano Provencio, Adrianna Janik, Luca Costabello, Pasquale Minervini, Pontus Stenetorp, et al. On predicting recurrence in early stage non-small cell lung cancer. In *AMIA Annual Symposium Proceedings*, volume 2021, page 853. American Medical Informatics Association, 2021.
[35] Adrianna Janik, Maria Torrente, Luca Costabello, Virginia Calvo, Brian Walsh, Carlos Camps, Sameh K Mohamed, Ana L Ortega, Vít Nováček, Bartomeu Massutí, et al. Machine learning-assisted recurrence prediction for early-stage non-small-cell lung cancer patients. *arXiv preprint arXiv:2211.09856*, 2022.
[36] Ewout W Steyerberg et al. *Clinical prediction models*. Springer, 2019.
[37] Aaron N Richter and Taghi M Khoshgoftaar. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial intelligence in medicine*, 90:1–14, 2018.
[38] Kien Wei Siah, Sean Khozin, Chi Heem Wong, and Andrew W Lo. Machine-learning and stochastic tumor growth models for predicting outcomes in patients with advanced non–small-cell lung cancer. *JCO clinical cancer informatics*, 1:1–11, 2019.
[39] Yue-Hua Zhang, Yuquan Lu, Hong Lu, and Yue-Min Zhou. Development of a survival prognostic model for non-small cell lung cancer. *Frontiers in Oncology*, 10:362, 2020.
[40] Hiroko K Solvang, Ole Christian Lingjærde, Arnoldo Frigessi, Anne-Lise Børresen-Dale, and Vessela N Kristensen. Linear and non-linear dependencies between copy number aberrations and mrna expression reveal distinct molecular pathways in breast cancer. *BMC bioinformatics*, 12(1):1–12, 2011.
[41] Yen-Chen Chen, Wan-Chi Ke, and Hung-Wen Chiu. Risk classification of cancer survival using ann with gene expression data from multiple laboratories. *Computers in biology and medicine*, 48:1–7, 2014.
[42] Taizo Hanai, Yasushi Yatabe, Yusuke Nakayama, Takashi Takahashi, Hiroyuki Honda, Tetsuya Mitsudomi, and Takeshi Kobayashi. Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression. *Cancer science*, 94(5):473–477, 2003.
[43] Alberto M Marchevsky, Sachin Patel, Karen J Wiley, Mark A Stephenson, Margaret Gondo, Richard W Brown, Eunhee S Yi, William F Benedict, Rose C Anton, and Philip T Cagle. Artificial neural networks and logistic regression as tools for prediction of survival in patients with stages i and ii non-small cell lung cancer. *Modern Pathology: an Official Journal of the United States and Canadian Academy of Pathology, Inc*, 11(7):618–625, 1998.
[44] Te-Chun Hsia, Hung-Chih Chiang, David Chiang, Liang-Wen Hang, Fuu-Jen Tsai, and Wen-Chi Chen. Prediction of survival in surgical unresectable lung cancer by artificial neural networks including genetic polymorphisms and clinical parameters. *Journal of clinical laboratory analysis*, 17(6):229–234, 2003.
[45] Gregory D Jones, Whitney S Brandt, Ronglai Shen, Francisco Sanchez-Vega, Kay See Tan, Axel Martin, Jian Zhou, Michael Berger, David B Solit, Nikolaus Schultz, et al. A genomic-pathologic annotated risk model to predict recurrence in early-stage lung adenocarcinoma. *JAMA surgery*, 156(2):e205601–e205601, 2021.
[46] Jacob J Chabon, Emily G Hamilton, David M Kurtz, Mohammad S Esfahani, Everett J Moding, Henning Stehr, Joseph Schroers-Martin, Barzin Y Nabet, Binbin Chen, Aadel A Chaudhuri, et al. Integrating genomic features for non-invasive early lung cancer detection. *Nature*, 580(7802):245–251, 2020.
[47] Raquel Dias and Ali Torkamani. Artificial intelligence in clinical and genomic diagnostics. *Genome medicine*, 11(1):1–12, 2019.
[48] Yu-Heng Lai, Wei-Ning Chen, Te-Cheng Hsu, Che Lin, Yu Tsao, and Semon Wu. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific reports*, 10(1):1–11, 2020.
[49] Hongguang Fu, Yanbing Liang, Xiuqin Zhong, ZhiLing Pan, Lei Huang, HaiLin Zhang, Yang Xu, Wei Zhou, and Zhong Liu. Codon optimization with deep learning to enhance protein expression. *Scientific reports*, 10(1):1–9, 2020.
[50] Khoa A Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D Williams, John V Pearson, and Nicola Waddell. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, 13(1):1–17, 2021.
[51] Timo M Deist, Frank JWM Dankers, Priyanka Ojha, M Scott Marshall, Tomas Janssen, Corinne Faivre-Finn, Carlotta Masciocchi, Vincenzo Valentini, Jiazhou Wang, Jiayan Chen, et al. Distributed learning on 20 000+ lung cancer patients–the personal health train. *Radiotherapy and Oncology*, 144:189–200, 2020.
[52] Imane Chlioui, Ibtissam Abnane, and Ali Idri. Comparing statistical and machine learning imputation techniques in breast cancer classification. In *International Conference on Computational Science and Its Applications*, pages 61–76. Springer, 2020.
[53] Michael Schubert, Bertram Klinger, Martina Klünemann, Anja Sieber, Florian Uhlitz, Sascha Sauer, Mathew J Garnett, Nils Blüthgen, and Julio Saez-Rodriguez. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature communications*, 9(1):1–11, 2018.
[54] Joshua D Campbell, Anton Alexandrov, Jaegil Kim, Jeremiah Wala, Alice H Berger, Chandra Sekhar Pedamallu, Sachet A Shukla, Guangwu Guo, Angela N Brooks, Bradley A Murray, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature genetics*, 48(6):607–616, 2016.
[55] Mariano Provencio, Enric Carcereny, Delvys Rodríguez-Abreu, Rafael López-Castro, María Guirado, Carlos Camps, Joaquim Bosch-Barrera, Rosario García-Campelo, Ana Laura Ortega-Granados, José Luis González-Larriba, et al. Lung cancer in spain: information from the thoracic tumors registry (ttr study). *Translational lung cancer research*, 8(4):461, 2019.
[56] James A Bartholomai and Hermann B Frieboes. Lung cancer survival prediction via machine learning regression, classification, and statistical techniques. In *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 632–637. IEEE, 2018.
[57] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[58] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[59] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Evaluating surgical skills from kinematic data using convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, pages 214–221. Springer, 2018.

[60] Xiangxue Wang, Andrew Janowczyk, Yu Zhou, Rajat Thawani, Pingfu Fu, Kurt Schalper, Vamsidhar Velcheti, and Anant Madabhushi. Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital h&e images. *Scientific reports*, 7(1):13543, 2017.

[61] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

[62] Akanksha Farswan, Anubha Gupta, Ritu Gupta, and Gurvinder Kaur. Imputation of gene expression data in blood cancer and its significance in inferring biological pathways. *Frontiers in oncology*, 9:1442, 2020.

[63] Mala Pande, Aron Joon, Abenaa M Brewster, Wei V Chen, John L Hopper, Cathy Eng, Sanjay Shete, Graham Casey, Fredrick Schumacher, Yi Lin, et al. Genetic susceptibility markers for a breast-colorectal cancer phenotype: Exploratory results from genome-wide association studies. *Plos one*, 13(4):e0196245, 2018.