IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 9, 2021, accepted July 29, 2021, date of publication August 17, 2021, date of current version August 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104820

A Multi-Learning Training Approach for Distinguishing Low and High Risk Cancer Patients

LUCAS VENEZIAN POVOA[1,2,3,4], **URIEL CAIRÊ BALAN CALVI**[2], **ANA CAROLINA LORENA**[2], **CARLOS HENRIQUE COSTA RIBEIRO**[1,2], **AND ISRAEL TOJAL DA SILVA**[3]

[1] Bio-Engineering Laboratory, Aeronautics Institute of Technology (ITA), São José dos Campos 12228-900, Brazil
[2] Computer Science Division, Aeronautics Institute of Technology (ITA), São José dos Campos 12228-900, Brazil
[3] Laboratory of Bioinformatics and Computational Biology, A. C. Camargo Cancer Center (ACCCC), São Paulo 01508-010, Brazil
[4] Computer Science Department, Federal Institute for Education, Science, and Technology of São Paulo (IFSP), Jacareí 12322-030, Brazil

Corresponding author: Carlos Henrique Costa Ribeiro (carlos@ita.br)

The work of Lucas Venezian Povoa was supported by the A. C. Camargo Cancer Center under Research Grant ACCCC-ITA-LVP: 201803. The work of Carlos Henrique Costa Ribeiro was supported by CNPq under Research Grant 303093/2016-1. The work of Israel Tojal da Silva was supported by FAPESP under Research Grant 15/19324-6.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT All cancers are caused by changes in the DNA within cells that occur over the course of an individual's lifetime. These mutations confer extensive genetic and phenotype variations within individuals, making the identification of appropriate treatments hard and costly. Moreover, cancer datasets are usually highly sparse due to the presence of few samples and many input features, making it difficult to design accurate predictors to classify patients into risk groups. Here, we report on the Multi Learning Training (MuLT) algorithm, which employs supervised, unsupervised, and self-supervised learning methods in order to take advantage of the interplay of clinical and molecular features for distinguishing low and high risk cancer patients. Our solution is evaluated using three independent and public cancer data sets considering three different performance aspects, through 5-fold cross-validation experiments. MuLT outranks other methods achieving AUCs between 0.65 and 0.77 and mean squared errors smaller than 0.24, while reducing classification complexity. These findings confirm the benefits of combining different learning algorithms and of coupling molecular and clinical data for supporting clinical decision making in Oncology.

INDEX TERMS Machine learning, cancer, molecular bio-markers, composable approach, clinical and molecular data, cancer risk prediction.

I. INTRODUCTION

Engineering, physical sciences, and oncology integration have been providing several significant contributions to cancer research over the past fifty years. The convergence of these disciplines driven by Machine Learning (ML) algorithms could lead to new computational models for modeling complex cancer systems, helping to increase the effectiveness of treatments, reducing costs, and saving lives [1]. This is a very important fact, once cancer contributes to a large number of deaths world wide, fatally reaching more than 4 million people by 2018 [2] with a projection of 19.8 million new cases per year for 2025 [3].

All cancer are caused by changes in the DNA within cells, with high genetic and phenotypic variation. This prevents the identification of appropriate treatments, which may become more complex and costly, whilst less effective [4], [5]. Although the development of novel cancer treatments improved survival over the years, only a subset of patients benefits from them, which motivates the study of new treatment response predictors to aid cancer diagnosis and risk evaluation [6]. Herewith, the identification and collection of relevant markers is a fundamental step to create new effective and efficient treatment protocols. Currently, the biotech industry is providing a full sequence of human genome and this technology has improved at a faster pace than Moore's law. These facts have direct impact on the possibility to create personalized treatments based on a set of highly informative predictors [7].

As a consequence of those developments, large volumes of cancer data have been gathered, at both molecular and clinical levels. Nonetheless, these data sets often pose several statistical and computational challenges [8]. Among them, one can mention: high sparsity due to low number of patients

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

and high number of molecular features, presence of noise and inconsistencies, unbalanced number of samples related to each cancer event (e.g., high risk group), and no clear relationship of the input variables with treatment outcomes.

ML algorithms could be roughly defined as computational methods to optimize a performance metric (e.g., Mean Squared Error (MSE)) based on past stored data representing a specific set of tasks (e.g., risk classification, salary estimation).

Size and quality of data are recognized as critically important to create useful ML-based models [9]. ML is being widely used to drive cancer studies, with successful applications to different cancer types [10], [11]. These methods allow to automatically extract general patterns hidden in a given data set related to cancer diagnosis and prognosis. Whilst modern ML-based predictors need to be trained on hundred of thousands samples, in general molecular cancer data sets contain only few hundreds samples [12]. This requires an additional care in the development of such models, in order to guarantee that the final predictors obtained effectively generalize to new data samples.

All these facts define a challenge on defining and implementing algorithms to classify patients into risk groups, which usually is an information derived from features like overall survival time, progression-free survival (PFS), or vital status. ML algorithms could be an important tool as part of clinical decision-making processes, helping to indicate effective treatments in order to save lives, or at least to provide better end-life quality.

In this paper, we propose the Multi Learning Training (MuLT) algorithm for classifying cancer patients as low or high risk. MuLT leverages from multiple data preprocessing and ML algorithms, combining supervised, unsupervised and self-supervised methods in order to create more accurate predictors. It takes clinical and molecular data as input considering an interplay of these data into the same model. Therefore, all public data sets used here are composed of both types of data. Clinical data usually represent information about demography, pharmacology, toxicology, safety efficacy, and disease data (e.g., age, cancer stage, beta 2, days to disease progression), while molecular data provide information about gene expression levels.

We evaluate our approach via cross-validation experiments on three public cancer data sets. As baselines, we compared our solution to other popular ML models used in the literature for cancer risk prediction, namely K-Nearest Neighbors (KNN) [13], Light Gradient Boosting Machine (LightGBM), Multilayer Perceptron (MLP) [14], and Support Vector Machines (SVM) [15]. We defined risk categories (i.e, low and high risk) based on overall survival, PFS, or vital status features as available in each data set. Our experimental results allowed us to conclude that, by combining different learning methods to generate new data representations and by extracting latent data information embedding molecular, clinical, and treatment data, it is possible to reach more precise patient risk predictions. In fact, MuLT performance was analyzed using data sets of different types of cancer, reaching better results than all baseline methods employed in our experiments.

To ensure the reproducibility of our results, we publicly share the source code used in our experiments at https://github.com/lucasvenez/mult. It not only includes the code for the proposed approach and the baselines, but also the code for data download and pre-processing, parameter optimization, and evaluation.

The remainder of this paper is organized as follows. Section II presents some preliminary definitions, and reviews existing work and previous technical approaches. Section III describes details of the MuLT approach. Section IV details the experimental design. Section V presents and discusses our results. Finally, Section VI brings our final remarks and point outs future work directions.

II. BACKGROUND AND RELATED WORK

This section summarizes the mathematical notations employed in this work, presents fundamental definitions of molecular genetics and data derived from such area, and describes some related work that encompasses ML techniques applied to cancer data.

Regarding the mathematical notation employed, Table 1 summarizes the main symbols used to designate the data items, operations and sets, among others.

TABLE 1. Mathematical notation summary.

Symbols	Meanings
$\mathbf{x}$, $\mathbf{X}$, A	vector, matrix, and set
$\boldsymbol{X}$	random variable
$\lvert A\rvert$	number of elements in A
$A \cup B$	union of A and B
$A \backslash B$	set difference between A and B
$a \leftarrow b$	assigns b to a
$\overline{A}$	average of A
$\sigma(A)$	standard deviation of A
$A^{(t)}$	A at iteration t
$\lvert\lvert\mathbf{x}\rvert\rvert$	length of $\mathbf{x}$
$d(\mathbf{x}, \mathbf{y})$	Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$
log	logarithm base 2
$\mathrm{relu}(x)$	rectifier activation function $\max(0, x)$
$\mathrm{sigmoid}(x)$	sigmoid activation function $(1 + e^{-x})^{-1}$
$\lfloor x \rfloor$	floor of x
$(\mathbf{X}\lvert\mathbf{Y})$	column concatenation of $\mathbf{X}$ and $\mathbf{Y}$
$r(\mathbf{X})$	number of rows of $\mathbf{X}$
$\mathbf{X}^T$	transpose of $\mathbf{X}$
$\boldsymbol{X} \sim \mathcal{N}(\mu, \sigma^2)$	$\boldsymbol{X}$ is normally distributed

A. MOLECULAR GENETICS

Oncogene generation is driven by series of mutations in genes that implies in changes of cell functions. These mutations occur within the coding sequence of genes and can lead to abnormal proteins being produced [16]. In addition, these mutations can accelerate growth and induce uncontrolled cell divisions that can result in anomalous chromosome segregation and changes in chromosome number [17].

As genes encode proteins and proteins dictate cell function, the monitoring of these dynamics may lead to a better understanding of the regulatory events involved both in health and disease conditions, including cancer treatment. Thus, assessment of molecular mechanisms underlying cellular events coupled with environment, and lifestyle for each person is a pivotal approach called precision medicine.

Guided by this knowledge, several researches have been carried out to extract information from genomic data. These studies and analyzes improved the understanding of genetic networks in the development of human diseases and allowed us to discover new genes associated with these diseases [18].

A common way to inspect genomic data is using gene expression profiles. Gene expression profiling measures the expression level of mRNAs in a cell population at a certain time [19]. The expression profile is often represented by array-based techniques and provides a high-throughput approach to analyze thousands of genes simultaneously. In this format, it can be correlated to pathological diagnosis, clinical outcomes, or even therapeutic response [20].

A gene expression data set may be considered an $n \times l$ matrix M, where each row $M_{i.}$ represents expression levels of a set of n genes ($G = g_1, \ldots., g_n$), the columns $M_{.j}$ represent expression profiles of a set of l samples $S = (s_1, \ldots, s_l)$ at different conditions, and each element $m_{i,j}$ is the expression level of gene g_i $(1 \leq i \leq n)$ on sample s_j $(1 \leq j \leq l)$.

In this work we use public gene expression levels together with clinical markers as input for ML models in order to classify cancer patients as low or high risk.

B. RELATED WORK

There are several researches applying ML algorithms in cancer prognosis prediction or to identify new characteristics that can be used to understand such a complex disease. The general idea is to induce predictive models from collections of patients' data with known outcomes, which can then be used in the analysis of new patients' data. Here we present a compendium of work related to this research.

A novel simulation-based approach to identify subgroups of Multiple Myeloma patients that might benefit from a treatment of interest is presented in [6]. The study considers both bortezomib [21] and lenalidomide [22] treatments.

They consider 10,581 gene features as input to ML models using techniques such as random forests [23] and support vector machines (SVM) [15]. No clinical features are considered. All data were collected from the MMRF Research Gate [24] and Gene Expression Omnibus (GEO) Database [25]. Experimental results evidence that their solution can be helpful to estimate whether a patient will benefit or not from a given treatment. Authors also stated that predicting non treatment benefit has equal importance to predicting treatment benefit, and both can give useful information towards personalized medicine.

A novel transfer learning-based process based on ML techniques such as SVM, XGBoost [26] and Deep Boosting [27] aiming to classify patients of Multiple Myeloma (MM), triple-negative breast cancer, and breast cancer as treatment sensitive or resistant is proposed in [28]. The study was limited to bortezomib treatment. The data sets were obtained from the Affymetrix Human Genome U133A Array platform and GEO database. The authors successfully used gene expression data to generate predictions of drug response, and reported a superior performance of their approach compared to baseline approaches.

The Quadratic Phenotypic Optimization Platform (QPOP) is presented in [29], which aims to optimize treatment combinations to effectively treat bortezomib-resistant MM. Authors state that bortezomib is present in 58% of clinically used therapies, having shown overall response rates as high as 93% in newly diagnosed patients. QPOP results present an R^2 of 0.803 in the regression problem of predicting effective drug combinations that optimized treatment efficacy.

In [30] a study with 1,181 MM patients is reported. Authors explore the relationship among different abnormalities. Their results provide evidences that only a set of abnormalities classified as high-risk do not contain all necessary information to predict Multiple Myeloma prognostic.

Evidences of correlation between a mutation called KRAS and the colorectal patient resistance to cetuximab [31] treatment are presented in [32]. That study was conducted with thirty patients. Advances from this study are presented in [33], which explores molecular features to predict treatment outcome. The authors state that, through the application of molecular analyzes, they found several new markers associated with the prognosis and the treatment outcome.

A risk predictor for breast cancer based on Cox model [34] and genetic variables is presented in [35]. The study considered more than 700 patients and used breast cancer sub-types (luminal A, luminal B, HER2-enriched, and basal-like) as an important feature to train the model. It concludes the sub-type model predicted neoadjuvant chemotherapy efficacy with a negative predictive value for pathologic complete response of 97%. Furthermore, the sub-type feature added significant prognostic and predictive information.

The study presented in [36] describes a novel *in silico* screening process based on Association Rule Mining (ARM). It identifies molecular markers as candidate drivers of treatment response, rather than clinical data, such as characteristics of tumors. The authors argue that clinical data have reached their limit and are insufficient to guide therapeutic solutions or predict therapy response. The tests performed in this study demonstrated that the association rule mining process was able to highlight relevant molecular correspondences to evaluate responsiveness to certain drugs.

In order to improve accuracy on risk prediction, MuLT combines clinical markers and gene expression levels to compose the patient description. Moreover, new representations of gene expression levels are created based on unsupervised and self-supervised learning algorithms to find latent predictive information. These data are then used to distinguish low and high risk patients. Once MuLT defines treatments as input, final predictors could also be used to simulate optimal treatment in a personalized manner.

It is worth noting that MuLT can also be beneficial in other application domains which require combining multiple data sources. This is the case of credit risk assessment problems, for instance, where the work [37] proposed a dynamic model combining the use of customer profile data with politic-economic factors.

III. PROPOSED APPROACH

The MuLT approach proposed in this paper is composed of tree main modules: feature selection, feature extraction and patient risk prediction. It takes clinical, molecular and treatment data as input.

We will present an overview on MuLT in Section III-A. In Section III-B, we detail the feature selection module. In Section III-C, we provide information about each part of the feature extraction module. Finally, in Section III-D we describe the patient risk prediction module.

A. OVERVIEW

The MuLT approach has the following characteristics: i) it takes multiple types of data available from the patients as input, allowing to take into account different data representations; ii) it avoids noisy and irrelevant features by employing a feature selection step; iii) it finds latent information on data by feature engineering using unsupervised learning methods; iv) it creates a more robust representation of the data using an autoencoder; and v) it uses highly accurate ensemble predictors. Figure 1 shows an overview of the MuLT approach and components.

Briefly, MuLT takes clinical data (e.g., age, race, stage, transplant), molecular data (e.g., gene expression levels), and treatment as input. Then, it starts by executing the Clinical Feature Selection (CFS) and Molecular Feature Selection (MFS) components from the Feature Selection module, which are responsible for filtering noisy and irrelevant input features. A Cross Feature Selection (XFS) algorithm is used in this filtering (see details in the next section). Using the subset of molecular features previously chosen, the Feature Extraction module performs Patient Clustering (PC), Gene Clustering (GC), and Gene Denoising (GD), which aim to create gene expressed-based signatures able to improve prediction accuracy. Finally, the Ensemble Predictor (EP) component classifies cancer patients into risk groups via a supervised ensemble learning method. All features previously selected and created are used as input to this classifier, which outputs a continuous value between 0 and 1, namely Low Risk Score (LRS), where low values for the LRS indicate high risk patients, whilst high values are attributed to patients with low risk.

From the architectural perspective, we define MuLT as a composable approach that can be extended to embrace different data types (e.g., images), to extract specialized information from raw data or to perform some training or inference regularization or normalization. Next all modules and components of MuLT are described in more details.

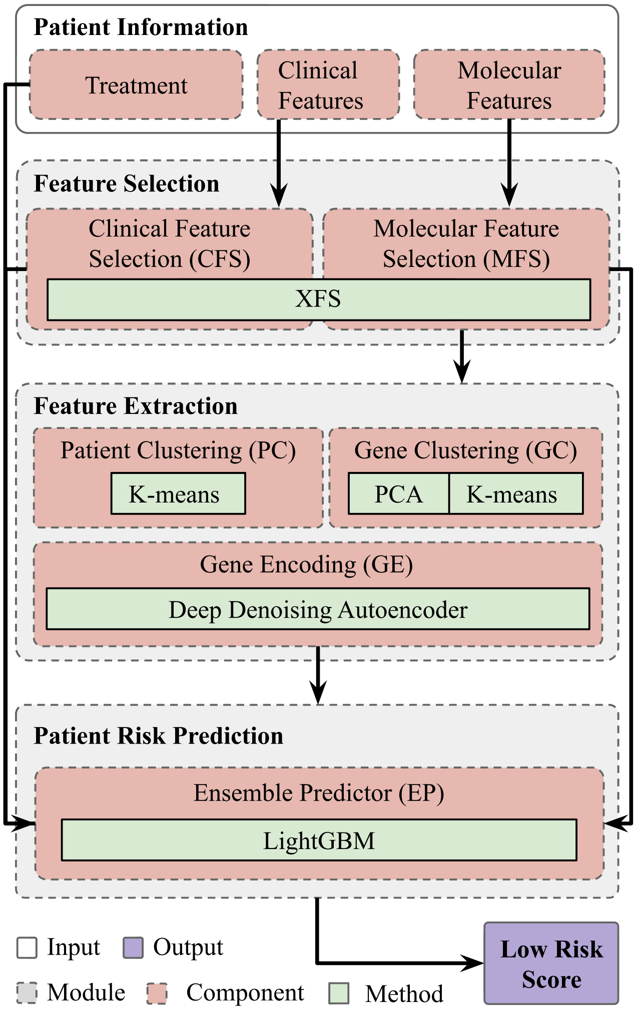


FIGURE 1. A MuLT instance, which uses clinical, molecular, and treatment input features, performs two independent feature selection steps, three automatic feature extraction steps and then predicts cancer patient risk using a LightGBM ensemble.

B. FEATURE SELECTION

In order to design our feature selection module, we established three main requirements: i) choose features that are helpful to discriminate low and high risk patients; ii) avoid features with redundant information; and iii) avoid features that only segregate a particular set of patients (i.e, biased features).

Let $\mathbf{X}$ be a matrix composed of features at columns and patients at rows, where each element x_{ij} describes the value of a feature M associated with a patient p. Let $\mathbf{y}$ be a vector associating a risk class (i.e., either low or high) to a patient p. Algorithm 1 receives as input $\mathbf{X}$ and $\mathbf{y}$, and generates two independent sets of patients (i.e., rows of $\mathbf{X}$) based on clinical outcome $\mathbf{y}$. The ComputeOrRetrieveKS function takes values

Algorithm 1 Pseudocode of the Feature Selection Algorithm

```
function FeatureSelection($\mathbf{X}, \mathbf{y}, \alpha, \beta$)
    $EF \leftarrow \{\}$                      ▷ Set of excluded features
    $AF \leftarrow \{\}$                      ▷ Set of analyzed features
    $SF \leftarrow$ Names($\mathbf{X}$)       ▷ Set of selected features
    for all $i \in SF$ do
        $AF \leftarrow AF \cup \{i\}$
        $p_i \leftarrow$ ComputeOrRetrieveKS($\mathbf{X}_{*i}, \mathbf{y}$)
        if $p_i > \alpha$ then
            $EF \leftarrow EF \cup \{i\}$
        else
            for all $j \in SF \setminus AF$ do
                $p_j \leftarrow$ ComputeOrRetrieveKS($\mathbf{X}_{*j}, \mathbf{y}$)
                if $p_j > \alpha$ then
                    $EF \leftarrow EF \cup \{j\}$
                else
                    $pc \leftarrow$ Pearson($\mathbf{X}_{*i}, \mathbf{X}_{*j}$)
                    if $pc > \beta$ then
                        if $p_j \geq p_i$ then
                            $EF \leftarrow EF \cup \{j\}$
                        else
                            $EF \leftarrow EF \cup \{i\}$
    $S \leftarrow SF \setminus EF$
    $\mathbf{X}' \leftarrow \mathbf{X}_{*S}$
    return $\mathbf{X}'$
```

of a feature $\mathbf{X}_{*i}$ of all patients, splits them into low and high risk groups described by vector $\mathbf{y}$ and computes the p-value of a Kolmogorov–Smirnov (KS) test [38]. For each feature $\mathbf{X}_{*i}$, the Algorithm 1 computes the KS test in order to quantify the distance between the empirical distribution functions of low and high risk groups assuming a significance level $\alpha = 0.05$. Features with p-value greater than α are excluded. After that, Algorithm 1 computes a pairwise linear correlation between the remainder features, excluding a feature $\mathbf{X}_{*i}$ if it has a Pearson correlation coefficient greater than a threshold $\beta = 0.75$ with any feature $\mathbf{X}_{*j}$ ($\forall i \neq j$) with smaller p-value. Then, it outputs a matrix $\mathbf{X}'$ associating patients to selected features.

Algorithm 1 addresses the requirements i and ii, and in order to address requirement iii we split our data set rows in three similar parts balancing the number of low and high risk patients in each part. Then, we apply our algorithm to each combination of data subsets, returning the intersection between results. We called that final algorithm XFS.

XFS is independently applied over clinical features (CFS component) and molecular features (MFS component). CFS and MFS return matrix $\mathbf{X}'_{\mathbf{c}}$ associating patients to selected clinical features, and matrix $\mathbf{X}'_{\mathbf{g}}$ associating patients to selected molecular features, respectively.

C. FEATURE EXTRACTION

In order to create a better representation [39] of cancer patients, we use self-supervised and unsupervised ML methods to find latent information in the raw molecular data. In the following sections, we detail the Patient Clustering (PC), Gene Clustering (GC) and Gene Encoding (GE) components.

1) PATIENT CLUSTERING

Previous studies [40], [41] have shown that ability to classify patients based on gene expression profile stratifies those with different outcomes. Thus, we proposed a PC Component to detect clusters of patients with similar molecular profile by using the k-means clustering algorithm [42]. k-means is a partitional clustering algorithm, whose purpose is to split a set of samples into k non-empty disjoint sets $S = \{S_1, S_2, \ldots, S_k\}$, that is, all $S_i \neq \emptyset$ and $S_i \cap S_j = \emptyset$, for $i \neq j$.

PC takes as input the matrix $\mathbf{X_g}$ and applies k-means for partitioning the patient data into varying numbers of clusters $k = 2, 3, \ldots, 50$. Equation 1 presents the objective function of k-means, which tries to cluster points such that the distance of each element to the centroid of its nearest cluster S_i is minimized.

$$J(X_g) = \arg\min_{S} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} d(\mathbf{x}_j, \mu_\mathbf{i}) \tag{1}$$

where μ_i is the centroid of cluster S_i and $\mathbf{x}_j$ is an element from cluster S_i. From a random set of k centroids, the k-means algorithm greedily assigns elements to their nearest centroids, which can then be updated. The cluster assignments and centroids' update steps are repeated until there are no changes in the clusters or a given maximum number of iterations is reached. The assignment of a patient $\mathbf{x}_j$ into a cluster S_i at a given iteration t can be described by Equation 2.

$$S_i^{(t)} = \left\{\mathbf{x}_j : d(\mathbf{x}_j, \mu_i^{(t)}) \leq d(\mathbf{x}_j, \mu_l^{(t)})\ \forall l \neq i\right\} \tag{2}$$

In order to define the number of clusters of patients, that is, the k value in k-means, we use the silhouette metric (sil) [43], defined by Equation 3.

$$sil(\mathbf{x}) = \begin{cases} (b(\mathbf{x}) - a(\mathbf{x}))/\max(a(\mathbf{x}), b(\mathbf{x})), & \text{if } |S_\mathbf{x}| > 1 \\ 0, & \text{if } |S_\mathbf{x}| = 1 \end{cases} \tag{3}$$

where $S_\mathbf{x}$ represents the cluster assigned to $\mathbf{x}$,

$$a(\mathbf{x}) = \frac{1}{|S_\mathbf{x}| - 1} \sum_{\mathbf{y} \in S_\mathbf{x}, \mathbf{x} \neq \mathbf{y}} d(\mathbf{x}, \mathbf{y})$$

and

$$b(\mathbf{x}) = \min_{\mathbf{z} \neq \mathbf{x}} \frac{1}{|S_\mathbf{z}| - 1} \sum_{\mathbf{y} \in S_\mathbf{z}} d(\mathbf{x}, \mathbf{y}), \quad \forall\, S_\mathbf{x} \neq S_\mathbf{z}$$

Taking the silhouette coefficient, the number of clusters k is optimized according to Equation 4, which selects a k value that provides high average silhouette and is penalized by standard deviation in order to better balance the variations and sizes of the clusters.

$$\arg\max_{k} \frac{\overline{sm(S)} - \sigma(sm(S))}{\sigma(cs(S)) + 1} \tag{4}$$

where $sm(S)$ and $cs(S)$ give, respectively, the silhouette metric of all samples in the partition S and the number of elements in each cluster $S_i \in S$.

Finally, the PC component outputs a new matrix $\mathbf{L}$ with the Euclidean distances between each patient and each of the clusters' centroids in the chosen partition S.

2) GENE CLUSTERING

The GC component aims to detect pairwise relationships among genes who are co-expressed into cancer-relevant signaling pathways [44]. Consecutively, GC has the purpose of defining gene clusters and generating an integrated measure from them.

GC takes as input $\mathbf{X_g}^T$, that is, the transpose of $\mathbf{X_g}$, and first applies Principal Component Analysis (PCA) [45] to reduce data dimensionality and obtain a linear combination of the patient features which represent better the data variance. Let $r(\cdot)$ be a function that retrieves the number of rows of a matrix, only the first $l = \max(2, \lfloor 0.1r(\mathbf{X_g})\rfloor)$ principal components are kept, as defined by Equation 5.

$$\mathbf{x}_i = \mathbf{X_g}' - \sum_{j=1}^{i-1} \mathbf{X_g}'\mathbf{w}_j\mathbf{w}_j^T \tag{5}$$

where $\mathbf{w}_j = \arg\max_{||\mathbf{w}||=1} d(\mathbf{x}_j, \mathbf{w})$ and $\mathbf{x}_i$ is a principal component [39], [45].

Taking the reduced data set formed by the first l principal components, GC clusters the genes using the same k-means-based procedure employed by the PC component. Finally, GC returns a matrix $\mathbf{E} = (e_{ij}) \in \mathbb{R}^{n \times k}$ associating each patient i to the average of expression level $h(\cdot)$ of genes in a cluster C_j, where n is the number of patients (sample size), k is the number of clusters, and

$$e_{ij} = \frac{1}{|C_j|}\sum_{s \in C_j} h\big(\mathbf{X_g}_{*s}\big).$$

3) GENE ENCODING

The GE component aims to create a noise-resilient representation of the molecular features [46]. This background noise arises during the gene expression quantification possibly obtained under sample preparation or in the sequencing step.

GE takes $\mathbf{X_g}'$ as input and it is founded on a Deep Denoising Autoencoder (DDA) [47], a specialized dense neural network based on self-supervised learning and composed of an input layer, encoder and decoder layers and an output layer. The encoder layer transforms a corrupted vector $\mathbf{x}'$ into a hidden representation $\mathbf{y}$ and a decoder layer maps back $\mathbf{y}$ to a reconstructed vector $\mathbf{z} \sim \mathbf{x}'$.

Our DDA architecture serializes an input layer, three encoder layers, two decoder layers, and one output layer. Encoder and decoder layers have $\lfloor 0.5m\rfloor$, $\lfloor 0.4m\rfloor$, $\lfloor 0.3m\rfloor$, $\lfloor 0.4m\rfloor$, and $\lfloor 0.5m\rfloor$ units, respectively, where m is the number of attributes in $\mathbf{X_g}'$. It is optimized by minimizing the Mean Squared Error (MSE) between $\mathbf{z}$ and uncorrupted data $\mathbf{x}$ regularized by Tikhonov regularization [14] defined as $\lambda\mathbf{W}^T\mathbf{W}$, where $\lambda = 0.001$ is the regularization factor, and $\mathbf{W}$ is the weight matrix of the whole DDA network. $\mathbf{x}'$ is corrupted by adding a value from a random variable $R \sim \mathcal{N}(0, 1)$ with a probability of 0.75. DDA is trained by AdaGrad algorithm [48] with a learning rate of 0.001 and a batch size of 250 samples. For each learning iteration an R is regenerated.

Finally, GE outputs a matrix $\mathbf{G_e}$ associating patients to encoded values generated by the deeper encoder layer in DDA computed from uncorrupted $\mathbf{X_g}'$.

D. PATIENT RISK PREDICTION

The final module of MuLT, namely PRP, takes as input the concatenation $\mathbf{X} = (\mathbf{X_g}' \mid \mathbf{X_c}' \mid \mathbf{L} \mid \mathbf{E} \mid \mathbf{G_e} \mid \mathbf{T})$.

LightGBMs [49] embedded to the Bootstrap Aggregating (bagging) [50] meta-algorithm are used to model the individual patient LRS. The training is composed of two parts. First, the training data is split into two folds and a hyper-parameter optimization using the Bayesian optimization (BO) algorithm [51] is applied to define model parameter values to improve generalization and accuracy. One fold is used to train the model, and another to estimate the log loss defined by Equation 6.

$$L(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \tag{6}$$

where $\mathbf{y}$ is a vector of expected risk classes and $\hat{\mathbf{y}}$ is the estimated one.

The hyper-parameter optimization process returns the LightGBM parameters with the minimum average log loss in 50 independent iterations. Taking the optimized parameters, the training data set is then split into three folds. An independent TS predictor is created based on each pairwise fold. Training is stopped after one iteration without log loss improvement, or after 100 iterations. Final TS score is defined via the average of TS scores computed by each predictor. It is important to note that all the processes previously described are applied only on the data available for training the models.

Finally, PRP outputs a matrix $\mathbf{Y}$ associating patients to their LRS.

IV. EXPERIMENTAL EVALUATION

We conducted a series of experiments in order to analyze our approach in different cancer data sets and to compare it with existing ML classification methods.

A. DATA SETS

Experiments were performed using four independent and public cancer data sets gathered from the GEO database, an international repository that freely distributes genomic data sets. GEO is supported by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) [25]. Table 2 presents the cancer type, number of samples, number of molecular features (i.e., genes),

TABLE 2. Summarized description of the cancer data sets used to driven experiments. All data sets are publicly available at GEO Database via Dataset ID.

Metadata	GSE135820	GSE68465	GSE94873
Cancer type	Tubo-Ovarian	Lung	Melanoma
Gender (F)	100.00%	39.17%	49.55%
Samples	3798	442	720
Molecular features	513	22,283	169
Clinical features	15	10	4
Missing values	0.00%	2.96%	0.21%
Overlapping classes	22.64%	40.49%	35.92%
Reference	[52]	[53]	[54]

number of clinical features (e.g., age, gender, treatment, progression-free survival), percentage of missing values, percentage of female patients, and main references of each selected data set.

GSE135820 is composed by patients with high-grade serous ovarian tube cancer (HGSOC), and was originally used to develop and validate the gene expression sub-type predictor called PrOTYPE [52]. The original data set contains 4,077 samples, with 16 clinical features and 513 genes. Patients have an average Overall Survival (OS) of 1,618 days and an average age of 60.3 (range 21-93) years. There are 1,314 samples marked as alive, 2,668 as dead and 95 as undefined. We cleaned it by removing samples with missing Vital Status (VS), age, or cancer stage. We removed features that promote feature leakage [55], have more than 5% of missing values or that have no predictive information (e.g., hospital information, year of diagnosis). After these adjustments, we ended up with a data set composed by 3,798 samples, and no missing values. Risk groups were defined based on OS and VS. We defined patients as low risk (1) if $OS \geq \overline{OS}$ and as high risk (0) otherwise.

GSE68465 contains data about lung adenocarcinoma, the most common type of lung cancer. These data were originally used on a study of survival prediction in lung adenocarcinoma [54]. It has 462 samples, with 16 clinical features, and 22,283 gene features. There are 220 patients marked as female, and 19 as undefined. The average age is 64.4 (range: 33-87) years and 64.9% of all patients have a positive smoking history. We removed samples with missing VS or months to first progression, and deleted the first progression or relapse, and months to last clinical assessment to avoid data leakage issues. Patients were marked as low risk (1) if *months to last contact or death* is greater than or equal to its mean and as high risk otherwise.

GSE94873 contains data about advanced melanoma patients. It was used to identify blood-based features that can predict clinical response and one year survival on patients treated with Tremelimumab [53]. This data set is composed by 720 samples, with 7 clinical features and 169 genes. There are 438 male patients. The average age is 54.9 (range: 18-90) years, and there are 398 patients marked as dead and 322 as alive. We removed tissue and immunotherapy response features, once the former has constant value and the latter promotes feature leakage. Once this data set does not contain a survival time-based feature, patients were marked as low risk (1) if VS equals to alive and high risk (0) otherwise.

We stress that there is no universal boundary to categorize low and high risk cancer patients, mainly because cancer is a heterogeneous disease consisting of many different subtypes, and low and high risk categorization can vary as a function of the research or clinical decision-making process objectives. Our experiments considered boundaries from averages of time-based features or vital status features in order to address the fundamental concept that longer survival time is associated with lower risk factors.

Finally, in order to ensure reproducibility, we implemented a set of algorithms that download a data set directly from GEO Database and structure it into two tables, one for clinical and other for molecular data.

B. EVALUATION METHODOLOGY

In order to directly compare MuLT performance, we defined four baseline approaches, which are a MuLT variant by removing the feature extraction module and replacing EP component by either KNN, LightGBM, MLP, and SVM methods preceded by a hyper-parameter optimization. In addition, we perform experiments using the 5-fold cross validation methodology [56]. Table 3 presents all

TABLE 3. Complete list of hyper-parameters optimized by BO algorithm in our experiments described per method.

Method	Hyper-parameter
KNN	Number of neighbors
	Weight function used in prediction
	Leaf size
	Power parameter for the Minkowski metric
LightGBM	Max number of leaves in one tree
	Weight of labels with positive class
	Min number of data in one leaf
	Num. of data that sampled to construct bins
	Max number of bins that feature will be bucketed in
	Min sum hessian in one leaf
	Perc. of selected data without resampling
	Perc. of features on each tree to be selected
	Perc. of features on each tree node to be selected
MLP	Num. of neurons in the hidden layer
	Learning rate schedule for weight updates
	Initial learning rate
	Maximum number of iterations
	Tolerance for the optimization
SVM	Regularization parameter
	Kernel coefficient
	Degree of the polynomial kernel function
	Kernel type to be used in the algorithm (e.g., poly, rbf)

hyper-parameters per method optimized by the BO algorithm in our experiments.

We perform an independent experiment for each data set, which starts by splitting entire data set into five disjoint parts. The splitting process shuffle samples and generates parts with similar numbers of patients per risk group. Consecutively, an experiment consists of five rounds, each of which uses four parts as training data and one part as testing data. Each round employs a different part as testing data. Final results of an experiment are described by the average performance metrics across the rounds.

We evaluated and compare our approach by three different perspectives. Classification performance driven by Area Under the receiver operating characteristic Curve (AUC) [57] metric, residual analysis [58], and classification complexity reduction [59] obtained by MuLT when compared to raw data only. We chose these perspectives because accuracy and residuals together can provide a detailed understanding of generalizability and bias avoidance while classification complexity reduction analysis can provide a clear view of the benefits of combining different feature types.

All experiments were conducted into a server with 32 GB of RAM, 12 i7 cores of 2.20 GHz each, and a GPU NVIDIA GeForce GTX 1050 Ti with 4 GB GDDR5 of dedicated memory and 768 cuda cores.

V. RESULTS AND DISCUSSION

In this section, we present the main results and discuss our findings.

A. CLASSIFICATION PERFORMANCE

Figure 2 presents the classification performance obtained by the different algorithms in terms of AUC metric. The reported values correspond to the average of the measures obtained for five independent CV rounds, according to the evaluation methodology described in Section IV-B.

In this comparison, MuLT outperforms the other baseline algorithms on the three analyzed data sets. The LightGBM method is the second best. The observed difference between MuLT and LightGBM is higher for the GSE68465 data set. This can be explained by the facts that GSE68465 has a smaller sample size, more missing values, and more overlapping classes as described in Table 2. As result, algorithms that do not generate a more resilient representation of data have a higher tendency to incorrectly classify patient risk for more complex data sets. Considering absolute numbers, the higher MuLT predictive performance observed in our experiments provides the correct risk classification for more than 150 cancer patients, and the MuLT time consumption was 2.3 times higher in the worst case.

We can furthermore observe that ensemble-based approaches (i.e., LightGBM, MuLT) were able to provide better accuracies than non-ensemble baselines for patient risk classification. Other aspect that we can observe in Table 4 is the number of selected features for each data set. Once again GSE68465 has a latent characteristic compared to the other data sets, containing a higher number of selected features. This fact impacts in a dimensionality challenge for the classification task, which could indicates another important MuLT capability.

TABLE 4. Number of selected features in each data set. The counts encompass the union of molecular and clinical features among CV rounds.

Data set	Molecular features	Clinical features
GSE135820	260	3
GSE68465	367	1
GSE94873	69	0

Finally, a direct performance comparison with related algorithms [6], [28] was not applicable, as it requires to be reproducible and evaluated under the same conditions of our proposed algorithm (i.e., data set composition, response variables, data distributions, train-validation-test partition). However, when comparing MuLT architecture to these algorithms we can highlight its composability, which makes MuLT easier to be extended with new functionalities and modules focused on representation transformation based on specific characteristics of each feature type and via different learning methods, thus taking advantage of potential hidden information in the raw data.

B. RESIDUAL ANALYSIS

In our experiments, binary classes were estimated using a continuous value between 0 and 1. Based on these predictions, we computed residuals $e = y - \hat{y}$, and the Mean Squared Error defined by MSE $= 1/n \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where e is the residual from an inference, y is the actual patient risk (i.e., 0 for high and 1 for low risk class), $\hat{y}$ is the estimated class, and n is the sample size.

Figure 3 presents the residuals associated with each method applied in each data set coloring high and low risk points differently. It also presents the associated MSE and a linear regression curve of residuals estimated from LRS with its coefficient of determination $R^2 = 1 - \sum(e - \bar{e})^2 / \sum(e - \hat{e})^2$, where $\bar{e}$ is the average residual, and $\hat{e}$ is the residual estimated from LRS.

R^2 represents a goodness-of-fit, measuring it by values between 0 and 1, where 0 represents a random model and 1 a perfect one. MuLT performed better than all methods for GSE135820 and GSE68465 data sets and performed second-best for GSE94873 where SVM take the best R^2. We can observe that residuals are not easy to predict in all data sets. For a linear regression, it is expected that the variable to be predicted follows a normal distribution, and for residuals it is expected mean zero once it is not desirable big difference between expected and predicted values. But in our results ks tests give p-values equals zero when testing $e \sim \mathcal{N}(0, 1)$ for all methods and data sets.

In a complementary perspective, we can observe that MuLT generated smaller MSE when compared to all other methods in studied data sets. This observation was verified by

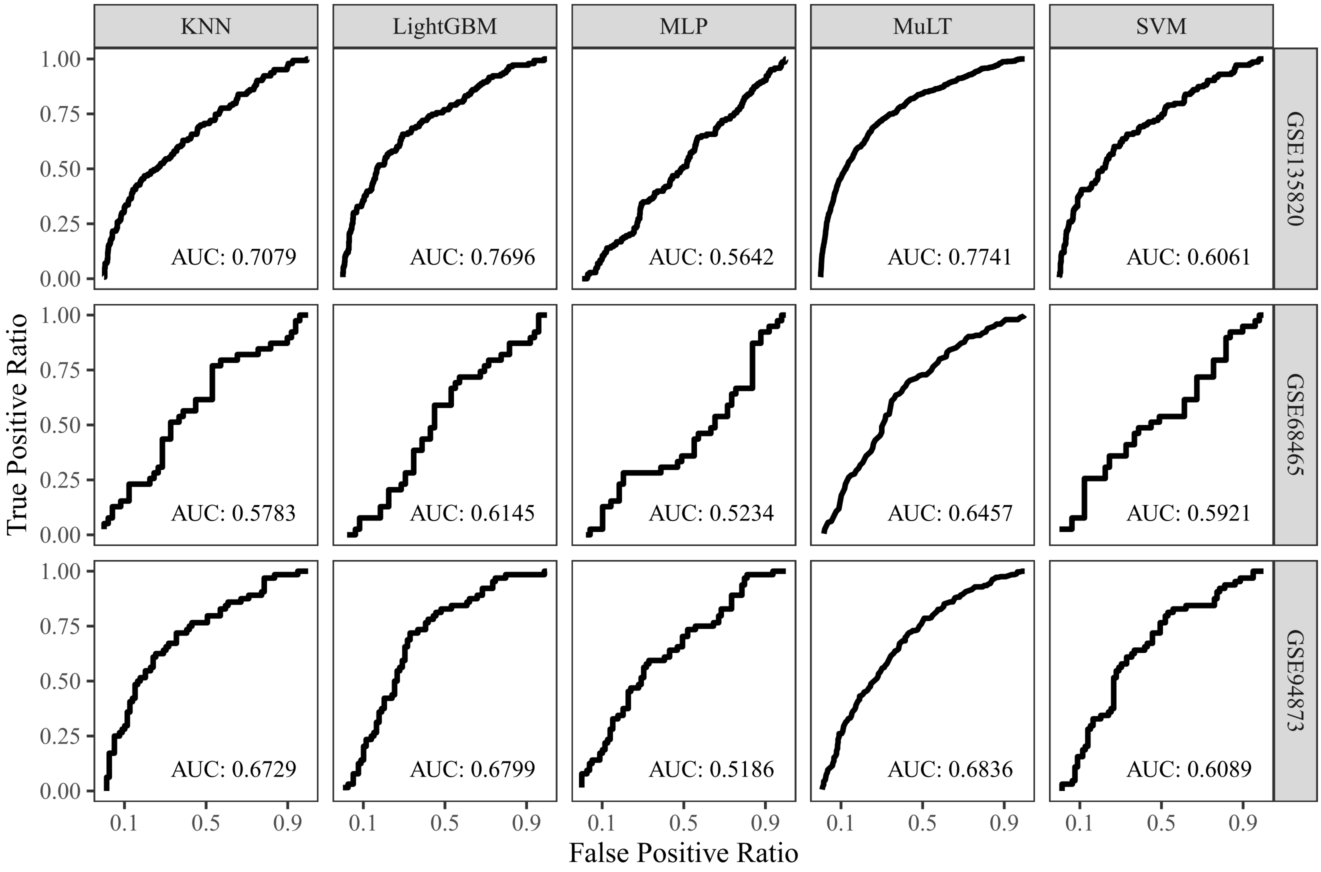


FIGURE 2. AUC resulting from each method applied to the three cancer data sets used in this study.

testing the alternative hypothesis of MuLT's MSE being less than MSE generated by other methods. The t-test provides a p-value lower than 3×10^{-3}, which is a strong evidence that our alternative hypothesis is true when assuming a significance level of 5%.

These results reinforce the predictive performance gains obtained by MuLT, mainly in the GSE68465 data set that has higher classification complexity in terms of sample size, amount of missing data, and overlapping of the classes (see Table 2).

C. CLASSIFICATION COMPLEXITY ANALYSIS

Once MuLT builds different data representations for a given data set, we evaluate in this section how this impacts the overall classification complexity of the underlying cancer risk prediction problem. For such, we employ a set of measures devoted to estimate the hardness level of a classification problem based on data characteristics, aka data complexity measures [59]. A set of feature-based complexity measures was extracted for the original (raw) data sets and their MuLT processed counterparts. The objective is to compare the hardness level required to solve the classification problem before and after the multiple data representations from MuLT are extracted.

The classification complexity measures employed, implemented in a tool named pyMFE [60], are:

- **Fisher's discriminant ratio (ft_f1)**: measures the overlap between the values of the features in the different classes. Lower ft_f1 values are obtained for simpler data sets, in which the individual input features are able to discriminate the classes well, considering the usage of hyper-planes perpendicular to the features' axes to separate the classes.
- **Directional-vector Fisher's Discriminant Ratio (ft_f1v)**: it is similar to ft_f1, but seeks for a vector which can separate the classes after the examples have been projected into it. With this procedure, the hyper-planes used to separate the data can be oblique in relation to the features' axis. Low ft_f1v values are obtained when a linear hyper-plane is able to separate the data. In this case, the classification problem can be considered simpler compared to the need of a non-linear decision boundary.
- **Volume of Overlapping Region (ft_f2)**: this measure computes the overlap of the distributions of the features values within the classes, taking the minimum and maximum values they assume for observations of different classes. The higher the ft_f2 value, the greater the amount of overlap between the classes considering the features' values and, consequently, the greater the classification complexity of the problem regarding the volume of overlapping aspect.

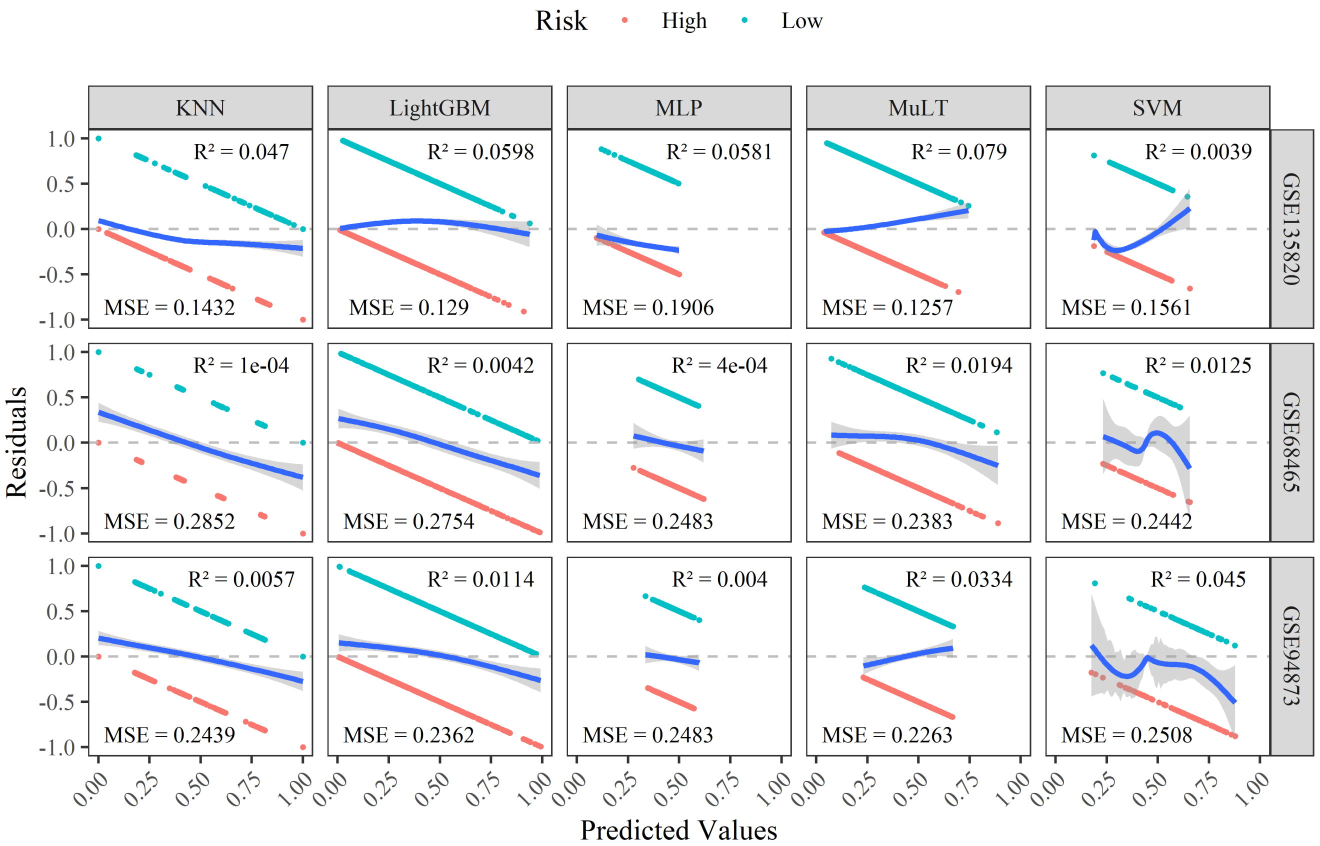


FIGURE 3. Residuals associated to predicted values for each association between method and data set.

- **Maximum Individual Feature Efficiency (ft_f3)**: this measure estimates the individual efficiency of each feature in separating the classes. Low values are obtained for problems where few examples lie in an overlapping region of the classes for at least one input feature.

Considering the results shown in Figure 4, we can see that MuLT data representation has lead to feature spaces with reduced overlapping between the classes as compared to the usage of the original data, for all data sets. The lower the overlapping of the classes, the easier is their separation. This is a clear advantage of the MuLT representation compared to the most common direct usage of the raw data.

VI. FINAL REMARKS AND FUTURE WORK

In this final section, we first present final remarks of our work and then give directions on future work.

A. FINAL REMARKS

We have proposed a novel Multi-Learning Training Approach named MuLT for distinguishing low and high risk cancer patients, based on clinical and molecular features. MuLT is composed by different algorithms in order to improve data representation and to find hidden information. Each component has an specific goal based on data characteristics, taking into account patient genetic profile, discovery of relevant groups of genes, amount of uncertainty in the data capture process, and high dimensionality. From the architecture perspective, our approach enables an easy combination of methods used in each component and an extensible structure which

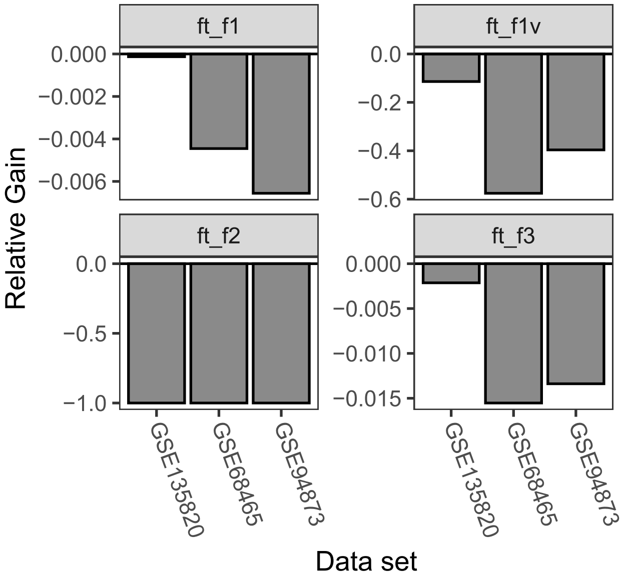


FIGURE 4. Measures of classification complexity reduction obtained by MuLT. These values were computed by $-(1 - ft_{MuLT}/ft_{raw})$, where ft_{raw} is the metric computed from raw data, and ft_{MuLT} is the metric computed from raw data combined with MuLT feature extraction.

can be complemented based on new technical or biological requirements.

Our experiments considered three different and independent public data sets, and evaluated and compared MuLT along three different perspectives: classification performance, residual analyses, and complexity reduction. We observed that MuLT had the best classification performance, being particularly advantageous when the data set has a high volume of missing data, high overlap among risk classes, and less available information. MuLT also gets a lower average error and its inner modules are capable of generating new features from raw data that reduced classification complexity w.r.t. feature overlapping.

B. FUTURE WORK

Our plan for future work focus on extending MuLT by adding medical images as input data, creating specific modules to treat that kind of data. We are also interested in two important aspects of our approach. First, we want to get more insight and knowledge on the gain obtained by each module individually, identifying local and global gains in the training process, overfitting reduction and bias reduction. This is motivated by the fact that in incipient studies we identified that MuLT could avoid bias in particular genetic profile, clinical conditions, or treatments. Finally, we foresee a potential utilization of MuLT in simulation processes that could be helpful to driven laboratory researches or clinical decision making protocols in a real medical setting related to cancer treatment.

ACKNOWLEDGMENT

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

[1] M. J. Mitchell, R. K. Jain, and R. Langer, "Engineering and physical sciences in oncology: Challenges and opportunities," *Nature Rev. Cancer*, vol. 17, no. 11, pp. 659–675, Nov. 2017, doi: 10.1038/nrc.2017.83.

[2] The Union for International Cancer Control. (2018). *New Global Cancer Data: GLOBOCAN 2018*. [Online]. Available: https://www.uicc.org/new-global-cancer-data-globocan-2018

[3] K. I. Block *et al.*, "Designing a broad-spectrum integrative approach for cancer prevention and treatment," *Seminars Cancer Biol.*, vol. 35, pp. S276–S304, Dec. 2015, doi: 10.1016/j.semcancer.2015.09.007.

[4] R. A. Rosales, R. D. Drummond, R. Valieris, E. Dias-Neto, and I. T. da Silva, "SigneR: An empirical Bayesian approach to mutational signature discovery," *Bioinformatics*, vol. 33, no. 1, pp. 8–16, Jan. 2017.

[5] R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton, "The causes and consequences of genetic heterogeneity in cancer evolution," *Nature*, vol. 501, no. 7467, pp. 338–345, Sep. 2013.

[6] J. Ubels, P. Sonneveld, E. H. van Beers, A. Broijl, M. H. van Vliet, and J. de Ridder, "Predicting treatment benefit in multiple myeloma through simulation of alternative treatment effects," *Nature Commun.*, vol. 9, no. 1, pp. 1–10, Dec. 2018, doi: 10.1038/s41467-018-05348-5.

[7] Singularity University, Moffett Field, CA, USA. (2018). *Exponential Trends in Healthcare: A Singularity University Industry Insights Report*. [Online]. Available: https://bit.ly/2Ti2ZIA

[8] A. C. Lorena, I. G. Costa, N. Spolaôr, and M. C. P. de Souto, "Analysis of complexity indices for classification problems: Cancer gene expression data," *Neurocomputing*, vol. 75, no. 1, pp. 33–42, Jan. 2012.

[9] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[10] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Informat.*, vol. 2, pp. 1–19, Jan. 2006.

[11] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, no. 1, pp. 8–17, 2015.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[13] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, Aug. 1992.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[15] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, vol. 10, pp. 1–11, Mar. 1999.

[16] T. Schrader, "Mutagens," in *Encyclopedia of Food and Health*, B. Caballero, P. M. Finglas, and F. Toldrá, Eds. Oxford, U.K.: Academic, 2016, pp. 20–28. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123849472004761

[17] M. A. Vasef, A. Auerbach, D. R. Czuchlewski, T. Bocklage, D. Chabot-Richards, N. Aguilera, and K. H. Karner, Eds., "Gene mutations," in *Diagnostic Pathology: Molecular Oncology* (Diagnostic Pathology). Philadelphia, PA, USA: Elsevier, 2016, pp. 1-10–1-13. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323376785500104

[18] J.-S. Lee, "Exploring cancer genomic data from the cancer genome atlas project," *BMB Rep.*, vol. 49, no. 11, pp. 607–611, Nov. 2016. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27530686

[19] P. Muller and R. T. Cate, "Oligoarticular and polyarticular juvenile idiopathic arthritis," in *Pediatrics in Systemic Autoimmune Diseases* (Handbook of Systemic Autoimmune Diseases), vol. 11, R. Cimaz and T. Lehman, Eds. Amsterdam, The Netherlands: Elsevier, 2016, ch. 1, pp. 1–30. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444635969000013

[20] D. E. Hansel, C. Magi-Galluzzi, and M. Zhou, "Molecular genitourinary pathology," in *Cell and Tissue Based Molecular Pathology*, R. R. Tubbs and M. H. Stoler, Eds. Philadelphia, PA, USA: Churchill Livingstone, 2009, ch. 28, pp. 379–392. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978044306901750033X

[21] C. Green, J. Bryant, A. Takeda, K. Cooper, A. Clegg, A. Smith, and M. Stephens, "Bortezomib for the treatment of multiple myeloma patients," *Health Technol. Assessment*, vol. 13, no. 1, Jun. 2009, Art. no. 5, doi: 10.3310/hta13suppl1-05.

[22] X. Armoiry, G. Aulagner, and T. Facon, "Lenalidomide in the treatment of multiple myeloma: A review," *J. Clin. Pharmacy Therapeutics*, vol. 33, no. 3, pp. 219–226, Jun. 2008, doi: 10.1111/j.1365-2710.2008.00920.x.

[23] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[24] Multiple Myeloma Research Foundation. (2019). *MMRF Researcher Gateway*. [Online]. Available: https://research.themmrf.org/

[25] E. Clough and T. Barrett, "The gene expression omnibus database," in *Statistical Genomics* (Methods in Molecular Biology), vol. 1418. New York, NY, USA: Humana Press, 2016, pp. 93–110, doi: 10.1007/978-1-4939-3578-9_5.

[26] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[27] C. Cortes, M. Mohri, and U. Syed, "Deep boosting," in *Proc. 31st Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 32, no. 2, E. P. Xing and T. Jebara, Eds., Beijing, China, Jun. 2014, pp. 1179–1187. [Online]. Available: http://proceedings.mlr.press/v32/cortesb14.html

[28] T. Turki and J. T. L. Wang, "Clinical intelligence: New machine learning techniques for predicting clinical drug response," *Comput. Biol. Med.*, vol. 107, pp. 302–322, Apr. 2019, doi: 10.1016/j.compbiomed.2018.12.017.

[29] M. B. M. A. Rashid, T. B. Toh, L. Hooi, A. Silva, Y. Zhang, P. F. Tan, A. L. Teh, N. Karnani, S. Jha, C.-M. Ho, W. J. Chng, D. Ho, and E. K.-H. Chow, "Optimizing drug combinations against multiple myeloma using a quadratic phenotypic optimization platform (QPOP)," *Sci. Transl. Med.*, vol. 10, no. 453, Aug. 2018, Art. no. eaan0941.

[30] M. Binder, S. V. Rajkumar, R. P. Ketterling, P. T. Greipp, A. Dispenzieri, M. Q. Lacy, M. A. Gertz, F. K. Buadi, S. R. Hayman, Y. L. Hwa, S. R. Zeldenrust, J. A. Lust, S. J. Russell, N. Leung, P. Kapoor, R. S. Go, W. I. Gonsalves, R. A. Kyle, and S. K. Kumar, "Prognostic implications of abnormalities of chromosome 13 and the presence of multiple cytogenetic high-risk abnormalities in newly diagnosed multiple myeloma," *Blood Cancer J.*, vol. 7, no. 9, p. e600, Sep. 2017.

[31] B. Vincenzi, G. Schiavon, M. Silletta, D. Santini, and G. Tonini, "The biological properties of cetuximab," *Crit. Rev. Oncol./Hematol.*, vol. 68, no. 2, pp. 93–106, Nov. 2008, doi: 10.1016/j.critrevonc.2008.07.006.

[32] A. Lièvre, J.-B. Bachet, D. Le Corre, V. Boige, B. Landi, J.-F. Emile, J.-F. Côté, G. Tomasic, C. Penna, M. Ducreux, P. Rougier, F. Penault-Llorca, and P. Laurent-Puig, "KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer," *Cancer Res.*, vol. 66, no. 8, pp. 3992–3995, Apr. 2006.

[33] A. Walther, E. Johnstone, C. Swanton, R. Midgley, I. Tomlinson, and D. Kerr, "Genetic prognostic and predictive markers in colorectal cancer," *Nature Rev. Cancer*, vol. 9, no. 7, pp. 489–499, Jul. 2009.

[34] T. M. Therneau and P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model*. New York, NY, USA: Springer, 2000.

[35] J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J. Clin. Oncol.*, vol. 27, pp. 1160–1167, Mar. 2009.

[36] K. Vougas *et al.*, "Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining," *Pharmacol. Therapeutics*, vol. 203, Nov. 2019, Art. no. 107395.

[37] S. Moradi and F. M. Rafiei, "A dynamic credit risk assessment model with data mining techniques: Evidence from Iranian banks," *Financial Innov.*, vol. 5, no. 1, pp. 1–27, Mar. 2019, doi: 10.1186/s40854-019-0121-9.

[38] G. Marsaglia, W. Tsang, and J. Wang, "Evaluating Kolmogorov's distribution," *J. Statist. Softw.*, vol. 8, no. 18, pp. 1–4, Nov. 2003.

[39] S. Haykin, *Neural Networks and Learning Machines* (Neural Networks and Learning Machines), no. 10. Upper Saddle River, NJ, USA: Prentice-Hall, 2009. [Online]. Available: https://books.google.com.br/books?id=K7P36lKzI_QC

[40] H. Yin, C. Zhang, X. Gou, W. He, and D. Gan, "Identification of a 13-mRNA signature for predicting disease progression and prognosis in patients with bladder cancer," *Oncol. Rep.*, vol. 43, pp. 379–394, Dec. 2019, doi: 10.3892/or.2019.7429.

[41] J. Zhu, L. Muskhelishvili, W. Tong, J. Borlak, and M. Chen, "Cancer genomics predicts disease relapse and therapeutic response to neoadjuvant chemotherapy of hormone sensitive breast cancers," *Sci. Rep.*, vol. 10, no. 1, May 2020, Art. no. 8188, doi: 10.1038/s41598-020-65055-4.

[42] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.

[43] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.

[44] M. H. Kabir, R. Patrick, J. W. K. Ho, and M. D. O'Connor, "Identification of active signaling pathways by integrating gene expression and protein interaction data," *BMC Syst. Biol.*, vol. 12, no. S9, pp. 77–87, Dec. 2018.

[45] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 2002, doi: 10.1007/b98835.

[46] S. Clancy and W. Brown, "Translation: DNA to MRNA to protein," *Nature Educ.*, vol. 101, no. 1, p. 1, 2008. [Online]. Available: https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/

[47] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*. New York, NY, USA: Association for Computing Machinery, 2008, pp. 1096–1103, doi: 10.1145/1390156.1390294.

[48] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Jul. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2021068

[49] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2017, pp. 3147–3155.

[50] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/bf00058655.

[51] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 2951–2959. [Online]. Available: https://bit.ly/2HG5RIg

[52] A. Talhouk *et al.*, "Development and validation of the gene expression predictor of high-grade serous ovarian carcinoma molecular SubTYPE (PrOTYPE)," *Clin. Cancer Res.*, vol. 26, no. 20, pp. 5411–5423, Jun. 2020, doi: 10.1158/1078-0432.ccr-20-0103.

[53] P. Friedlander, K. Wassmann, A. M. Christenfeld, D. Fisher, C. Kyi, J. M. Kirkwood, N. Bhardwaj, and W. K. Oh, "Whole-blood RNA transcript-based models can predict clinical response in two large independent clinical studies of patients with advanced melanoma treated with the checkpoint inhibitor, tremelimumab," *J. ImmunoTherapy Cancer*, vol. 5, no. 1, Aug. 2017, Art. no. 67, doi: 10.1186/s40425-017-0272-z.

[54] K. Shedden *et al.*, "Gene expression–based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study," *Nature Med.*, vol. 14, no. 8, pp. 822–827, Aug. 2008, doi: 10.1038/nm.1790.

[55] S. Kaufman, S. Rosset, and C. Perlich, "Leakage in data mining," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 556–563, doi: 10.1145/2020408.2020496.

[56] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 2. San Francisco, CA, USA: Morgan Kaufmann, 1995, pp. 1137–1143.

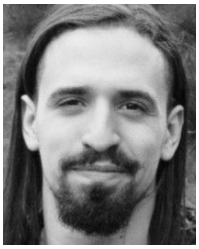
[57] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY, USA: Springer, 2001.

[58] D. R. Cox and E. J. Snell, "A general definition of residuals," *J. Roy. Stat. Soc., B, Methodol.*, vol. 30, no. 2, pp. 248–265, Jul. 1968, doi: 10.1111/j.2517-6161.1968.tb00724.x.

[59] A. C. Lorena, L. P. F. Garcia, J. Lehmann, M. C. P. Souto, and T. K. Ho, "How complex is your classification problem?" *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–34, Oct. 2019.

[60] E. Alcobaça, F. Siqueira, A. Rivolli, L. P. F. Garcia, J. T. Oliva, and A. C. P. L. F. de Carvalho, "MFE: Towards reproducible meta-feature extraction," *J. Mach. Learn. Res.*, vol. 21, no. 111, pp. 1–5, 2020. [Online]. Available: http://jmlr.org/papers/v21/19-348.html

LUCAS VENEZIAN POVOA was born in Ipaussu, São Paulo, Brazil, in 1990. He received the B.S. degree in information system from the Faculdade de Tecnologia (FATEC), Ourinhos, São Paulo, in 2012, and the M.S. degree in computer science from the Universidade Federal de São Carlos (UFSCar), São Carlos, São Paulo, in 2014. He is currently pursuing the Ph.D. degree in electronic engineering and computing with the Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, São Paulo. He is also an Associate Professor with the Instituto Federal de Educação Ciência e Tecnologia de São Paulo (IFSP), Jacareí, São Paulo. His research interests include artificial intelligence and distributed systems.



URIEL CAIRÊ BALAN CALVI was born in Americana, São Paulo, Brazil, in 1997. He received the Technologist degree in systems analysis and development from the Instituto Federal de Educação Ciência e Tecnologia de São Paulo (IFSP), Caraguatatuba, São Paulo, in 2018. He is currently pursuing the M.Sc. degree in electronic engineering and computing with the Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, São Paulo. His research interests include artificial intelligence and machine learning.

ANA CAROLINA LORENA received the B.S. and Ph.D. degrees in computer science from the Universidade de São Paulo, São Carlos, Brazil, in 2001 and 2006, respectively. From 2007 to 2012, she was an Assistant Professor with the Universidade Federal do ABC, Santo André, Brazil. From 2012 to 2018, she was an Associate Professor with the Universidade Federal de São Paulo, São José dos Campos, Brazil. Since 2018, she has been an Associate Professor with the Instituto Tecnológico de Aeronáutica, São José dos Campos. She has 48 articles in peer-reviewed academic journals and 75 complete papers in conferences. Her research interests include artificial intelligence, machine learning, data mining, and data science. She is also a CNPq research productivity fellow in Brazil.

CARLOS HENRIQUE COSTA RIBEIRO received the B.S. degree in communications engineering from the Instituto Militar de Engenharia, Rio de Janeiro, Brazil, in 1990, and the Ph.D. degree in electrical engineering from Imperial College London, U.K., in 1998. Since 1999, he has been with the Instituto Tecnológico de Aeronáutica, Brazil, where he currently holds a Full Professor position at the Computer Science Division. He has 28 articles in peer-reviewed academic journals and 126 papers in conferences, in the areas of artificial intelligence, machine learning, complex networks, robotics, and engineering education.

ISRAEL TOJAL DA SILVA received the B.S. degree in computer science from the Universidade Paulista, Ribeirão Preto, Brazil, in 2001, and the Ph.D. degree in sciences from the Universidade de São Paulo, Ribeirão Preto, in 2009. In 2012, he started his postdoctoral training at The Rockefeller University, New York City. He was a Research Associate with The Rockefeller University, from 2014 to 2015, where he continued as a member of the Adjunct Faculty, from 2015 to 2019. Since 2015, he has been the Head of the Laboratory of Bioinformatics and Computational Biology, A. C. Camargo Cancer Center, São Paulo, Brazil. He has 56 articles in peer-reviewed academic journals. His main research interests include genomics, bioinformatics, and artificial intelligence.