Model-Agnostic Interpretation of Cancer Classification with Multi-Platform Genomic Data

Olatunji Oni
School of Computational Science and Engineering
McMaster University
Hamilton, Ontario, Canada
onioa@mcmaster.ca

Sanzheng Qiao
Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada
qiao@mcmaster.ca

ABSTRACT

Machine learning models are often criticised for being black-boxes. Recent work in this field has aimed to address this criticism by developing methods to explain the underlying behaviour of machine learning models. These explanations are designed to help the end-user interpret how the models input features are used to make a prediction. Here, we present an extension to one such method, referred to as local interpretable model-agnostic explanations, to interpret multimodal tumor type classification from multi-platform genomic data. We propose a framework for transparent biomedical machine learning by leveraging interpretable dimensionality reduction to facilitate gene-wise explanations ~~for~~ the model behaviour. Experimental results using RNA-seq expression and single nucleotide variation (SNV) data from eight cancer types uncovered the models use of clinically relevant genes for cancer cell stratification. We demonstrate that model-agnostic explanations can provide valuable information to a clinician or scientist when predictive ability and interpretability are of absolute importance.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**; • **Applied computing** → **Health informatics**.

KEYWORDS

Model interpretation, Machine learning, Information fusion, Cancer detection

ACM Reference Format:
Olatunji Oni and Sanzheng Qiao. 2019. Model-Agnostic Interpretation of Cancer Classification with Multi-Platform Genomic Data. In *ACM-BCB '19: 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, September 07–10, 2019, Niagara, NY*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Recent advances in biotechnology have enabled a multidimensional approach for exploring human disease. New high-throughput technologies can quantify and characterize the biomolecules that define the architecture, behaviour, and dynamics of a biological system. Research in the last decade has introduced a multifaceted exploration of cancer biology at an unprecedented scale [15]. Cancer research projects are characterizing the genome, epigenome, and transcriptome to capture the complexity and phenotypic heterogeneity of cancer cells [8]. The prevalence of rich cellular descriptions of cancer cells has encouraged the adoption of advanced predictive analytics in clinical decision ~~support~~. Clinicians are increasingly adopting coupled frameworks of next-generation sequencing (NGS) and predictive models to support cancer diagnosis and patient stratification. Rapid developments in machine learning are enabling opportunities for improved clinical decision making in the healthcare industry, however, several key challenges hinder its utility by clinicians and researchers. The application of deep learning for medical predictions often results in a hindered ability to interpret the decision made by the classifier. Healthcare professionals require informative tools that can explain their predictions. Domain experts need to ensure a level of trust in predictive models by evaluating the usefulness, reliability, and internal logic of the system.

The ability to interpret the behaviour of a machine learning model can provide valuable insight into the internal logic of the classifier and the structural importance of the features. As the application of predictive systems is integrated deeper into the industrial and scientific domains, it is becoming increasingly important to be able to explain the basis of their decisions. Certain models benefit from inherent transparency in interpretation. These ~~techniques~~ provide a direct link to the features used to make a prediction. Unfortunately, transparent models such as decision trees, sparse linear models and rule-based systems have inferior predictive performance compared to more complex model abstractions such as random forest classifiers, support vector machines and deep neural networks. The increased complexity of the more advanced models, however, makes interpreting the underlying logic of the system a difficult task. Despite the prevalence of diverse forms of machine learning models, not many systems provide explanations of their decisions in biomedical applications. In other domains, model-agnostic explanations have been used to address this problem. Model-agnostic explanations are of interest due to their wide applicability. These interpretation methods are used to explain the behaviour of models where the internal logic of the system is not directly available for inspection. As well, model-agnostic methods are flexible in that they can derive explanations from any underlying model, making them desirable for use in medical predictive analytics.

ACM-BCB '19, September 07–10, 2019, Niagara, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

In this paper, we focus on extending a method for model-agnostic explanations to a clinically relevant learning task. We assess methods that leverage interpretable dimensionality reduction and model-agnostic explanations to help understand the underlying behaviour of a multimodal machine learning model. We use differential expression (DE) and clustered gene filtering (CGF) to extract a meaningful subset of genes from RNA-seq and SNV, respectively. We then utilize these features to predict the class of eight cancer cell types with a gated multimodal unit (GMU) based deep neural network. Finally, we used model-agnostic explanations to analyze the behaviour of our deep learning model. The performance of these methods affirms that model-agnostic explanations are useful for interpreting how a complex underlying model uses genetic information to make a prediction.

2 BACKGROUND

Clustering cancer cells with NGS data has been a well-studied area of computational biology. Several approaches have been developed to stratify cancer cells and benign cells using supervised learning techniques. In addition, various clustering methods have been proposed to identify cancer sub-class from multi-platform data, but very few of them attempt to interpret the influence of input features [13, 20, 29].

A key focus for researchers has involved producing expression and sequence-based features with a manageable dimensionality. Multi-platform biomedical datasets present numerous challenges for machine learning and statistical approaches. Biological data is often high-dimensional, noisy and sparse. High-throughput transcriptome sequencing and genome-wide genotyping arrays can produce tens of thousands to millions of features, making the identification of biomarkers a central issue in cancer research [32]. Representation learning for regularized and data-driven feature identification has thus emerged as a critical component of both the dimensionality reduction and model interpretation paradigms.

Various unsupervised methods have been used for dimensionality reduction and classification of sequencing data. Techniques such as stacked denoising autoencoders (SDAEs) have been used to acquire low dimension non-linear feature sets from breast cancer RNA expression data [30]. Transformative autoencoders have achieved some success, but these techniques result in encodings that lack direct interpretability. Furthermore, LASSO regression has been applied as a simple method for selecting gene expression features in a principled way [2, 27]. However, LASSO will only provide a useful set of selected features, not necessarily the most important features for a given application. Accordingly, conventional DE analysis has remained an integral component in cancer research [25]. Quantitative changes in expression levels between cancer cell types allow the identification of specifically relevant genes through statistical analysis [13]. For SNV data, recent work has shown superior performance with mutual information based feature selection. Due to the sparse nature of discrete point mutation SNV data, a clustered gene filtering approach was developed to identify subsets of the most informative gene regions to effectively provide a form of interpretable dimensionality reduction [36].

In recent work, common strategies for evaluating feature importance in neural networks has been through gradient-based methods,

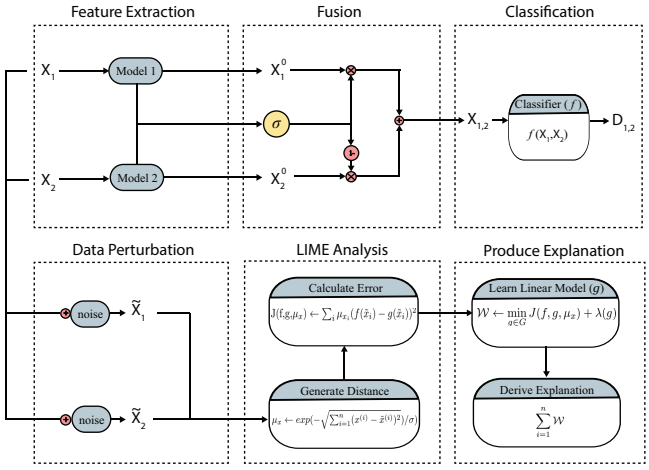

Figure 1: Schematic representation of LIME workflow.

perturbation techniques or utilizing surrogate models [6, 24, 26]. Gradient-based methods have especially dominated in biomedical imaging. This is due to the wide-ranging applicability of gradient-based methods, such as guided Grad-Cam, with convolutional neural networks (CNN) [6, 21, 26]. Guided backpropagation and Grad-Cam has been applied with a CNN to trace the most significant genes in gene expression profiles converted to 2D images [21]. However, guided Grad-Cam is specifically designed to leverage spatial information in CNNs, and thus are not particularly transferable to the dense layers of other neural networks. Alternatively, surrogates models approximate the behaviour of a complex model by using an interpretable model [9]. Interpretable models, such as generalized regression models or decision trees, are trained to approximate the predictions of an underlying model, and global explanations are derived from analyzing the surrogate. More recently, perturbation analysis has been used to model the impact of local perturbations to explain the sensitivity of machine learning models [5, 14]. These methods employ permuting the input and observing the variation to the model output. Local perturbation analysis allows the determination of the specific output variance caused by permuting the elements of the input for a training example. Recently, a combination of these approaches was developed to extend the utility of surrogate models with permutation analysis, producing an algorithm referred to as local interpretable model-agnostic explanations (LIME) [24]. LIME generates interpretable explanations by approximating the prediction of any classifier locally for a given training example. Local explanations of the underlying model are captured by training an interpretable model on perturbations of the input data. LIME generates a sample set of perturbed examples in the neighborhood of the local instance and uses an interpretable model to draw a decision boundary. Explanations are derived from analyzing the parameters of the decision boundary. Formally, a local surrogate model g is defined through the following expression:

$$\min_{g \in G} J(f, g, \mu_{x_i}) + \lambda(g), \tag{1}$$

where $J(f, g, \mu_{x_i})$ is a cost function that measures how closely the surrogate model g approximates the underlying model f while

keeping the model complexity $\lambda(g)$ low. The proximity measure μ_{x_i} determines the size of the neighborhood around an example x_i. For a training example, the probability that a classifier maps the input to a class label k is denoted by $y_k = f(x_i)$. Accordingly, LIME works to optimize the expression in (1) to interpret why f maps feature vector x_i to a class label k.

In order to produce an explanation, LIME first builds a dataset of perturbed instances $\tilde{x}$ by adding noise Z_i to the mass center of the training data. The noise, $Z_i \sim N(0, \sigma^2)$, is drawn from a zero-mean normal distribution with variance σ^2. The underlying model can then be used to generate a sample set that is weighted by their proximity to the selected instance. The surrogate model can then be trained using a cost function of the mean squared error:

$$J(f, g, \mu_{x_i}) = \sum_i \mu_{x_i}(f(\tilde{x_i}) - g(\tilde{x_i}'))^2, \quad (2)$$

where $\tilde{x}'$ is the interpretable representation of the perturbed data point. The learned weights of the trained model g form an n dimensional vector where each weight corresponds to a feature in training vector x_i. The magnitude of the n-th weight, $|w_n|$, defines the importance of that features on the prediction. The feature effect is defined by the polarity of the weight, where $w_n > 0$ or $w_n < 0$ suggests that the feature has a positive or negative influence on the prediction of the given class, respectively.

3 MATERIALS AND METHODS

In this study, the LIME procedure was extended to explain the use of RNA-seq and SNV in a multimodal machine learning model. The predictive behaviour of the underlying model was examined using a linear model to approximate the decision boundary in the neighborhood of each correctly labeled instance. The workflow for our method is shown in Fig. 1. This process involved (1) individual preprocessing of the RNA-seq and SNV data with interpretable dimensionality reduction, (2) generating cancer cell class predictions with the multimodal machine learning model, (3) using gene-wise model-agnostic explanations to interpret the underlying model, and (4) a clinical analysis of the explanatory genes.

3.1 Transcriptome Expression

In this study, we utilized RNA sequence (RNA-seq) transcriptome expression profiling as the first input modality. The RNA-seq data was derived from the HTSeq-Counts expression quantification framework [3]. The expression profile contained the feature dimensionality of all assayed genes in the HTSeq-Counts pipeline. As a result, the RNA-seq expression data contained the normalized expression counts of 60484 genes for each cell mass sample.

For the transcriptome expression data, interpretable features were produced by computing the significantly differentially expressed genes. The $\log_2$(fold change) was computed between the median tumour cell mass expression and healthy cell mass expression. The most statistically significant features were identified by fitting the differential expression to a Gaussian distribution and computing a two-tailed p-value. Features that match the preselected experimental dimensions were acquired by selecting the top most significant deferentially expressed genes using the two-tailed p-values.

3.2 Single Nucleotide Variation

The second input modality was SNV data. This data was obtained in the form of masked somatic mutations, derived from a MuTect2 Variant Aggregation and Masking workflow [7]. The analysis of the raw SNV data was based on the variant occurrence frequency of the genetic data. Variation occurrence was mapped to every listed gene for all available cell samples. This was performed by mapping mutated genes to cell samples in the raw SNV data, and accumulating the number of mutations for each respective cell sample. After preprocessing the SNV data, the variant occurrence frequency was obtained for 20516 human genes for each cell mass sample.

Clustered gene filtering (CGF) was used to select an interpretable subset of the most discriminatory genes based on the variant occurrence frequency of the SNV data [36]. The genes are selected based on high mutation frequency because genes with more mutations are likely of more interest. The procedure involves filtering the genes into groups based on a distance threshold, d_{cgf}, and then selecting the top, n_{cgf}, genes from each group. Interpretable dimensionality reduction was controlled algorithmically through the modulation of distance threshold d_{cgf}, and the group element threshold n_{cgf}. The distance threshold dictates how similar the mutation profiles of two genes need to be grouped together, and the group element threshold is the number of genes kept from each group.

3.3 Deep Gated Multimodal Unit

The multimodal biomedical classification was conducted with a deep gated multimodal unit (dGMU) [4]. This model is defined by the function of a representation network and a decision network as shown in Fig. 2.

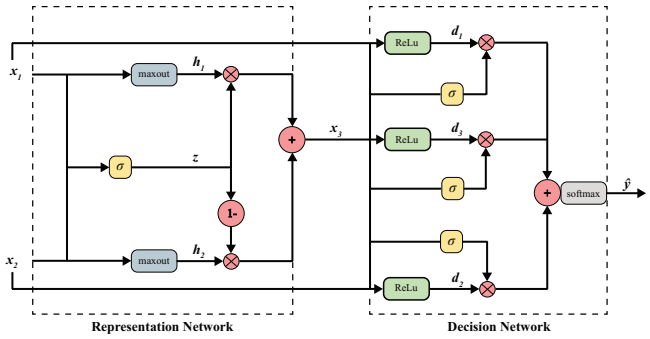

Figure 2: Deep Gated Multimodal Unit.

In the representation network, the input modalities learn a latent representation of the combined input data. Each modality becomes the input for a multilayer perceptron (MLP) with a maxout activation function, $maxout(\cdot)$ [12]. In Fig. 2, this produces $h_1 = maxout(\theta_{h1} \cdot x_1)$ and $h_2 = maxout(\theta_{h2} \cdot x_2)$, for modalities x_1 and x_2, respectively. Activated by the sigmoid activation function, $\sigma(\cdot)$, the gating neuron, $z = \sigma(\theta_z \cdot [x_1, x_2])$, ties both modalities and controls their contribution to the output of the unit. The output of the representation network is governed by the following equation:

$$x_3(x_1, x_2; \Theta_R) = z * h_1 + (1 - z) * h_2, \quad (3)$$

where latent space, $x_3(x_1, x_2; \Theta_R)$, depends on inputs x_1 and x_2, and $\Theta_R = \{\theta_{h1}, \theta_{h2}, \theta_{x3}\}$ is the set of parameters used for encoding the latent space.

In the representation network, each modality becomes the input to an MLP with a rectified linear unit (ReLU) activation function. Here, gating neuron $\sigma(\cdot)$ controls the untied contributions of decision gates d_1, d_2, and d_3. The decision network is governed by the following equation:

$$\hat{y}(x_1, x_2, x_3; \Theta_D) = \sum_{i=1}^{3} ReLu(\theta_{di} \cdot x_i)\sigma(\theta_{di} \cdot [x_1, x_2, x_3]), \quad (4)$$

where the network output, $\hat{y}(x_1, x_2, x_3; \Theta_D)$, depends on inputs x_1, x_2, and x_3, and $\Theta_D = \{\theta_{d1}, \theta_{d2}, \theta_{d3}, \theta_{g1}, \theta_{g2}, \theta_{g3}\}$ is the set of network parameters used across the untied gates in the decision network.

The dGMU model was implemented with original code in Tensorflow version 1.11.0 on an Nvidia Tesla K80 GPU [1]. The global loss was computed using the softmax cross entropy loss with L_2 regularization, and model parameters were learned using batch stochastic gradient descent with ADAM optimization [18].

3.4 Gene-Wise Interpretable Explanations

To find a gene-wise explanation, the LIME procedure is used to approximate the dGMU model with a linear model of class G, such that $g(\tilde{x}) = {w_g}^T \tilde{x}$. Perturbed instance $\tilde{x}$ is generated by individually noising each feature by drawing from a normal distribution. The mean and standard deviation is taken from each feature in the original dataset X. The perturbed instance is weighted using an exponential kernel learned over a Euclidean distance by letting $\mu_{x_i}(\tilde{x}) = exp(-\sqrt{\sum_{i=1}^{n}(x - \tilde{x})^2})/\sigma)$. The kernel width σ is defined as 0.75 times the square root of the number of training instances (default value for σ is used as established in [24]). With a locally weighted squared error J, as defined in Eq. (1), we learn the weights w_g of the sparse linear model via least squares. Each trained model provides interpretable explanations through the learned weights. The magnitude of a coefficient relates to the importance of the respective gene in sample x_i. Furthermore, genes with a positive weight coefficient are positively correlated with the prediction of the dGMU model and genes with a negative weight coefficient are negatively correlated. Accordingly, the explanation of a single prediction provides an interpretable framework by indicating the genes that are most influential. Specifically, a single LIME explanation can explain how the RNA-seq expression or SNV of the gene correlates with the model prediction.

The gene-wise explanations for a single prediction provide locally faithful insight into the logic of the classifier. In order to assess the global fidelity of the model, gene-wise explanations are pooled to evaluate the reliability of the predictions as a whole. A procedure for generating gene-wise LIME explanations is summarized in Algorithm 1. Gene-wise explanations are extended to understand the set of individual instances associated with correctly labelled predictions. Explanations for a set of correctly labelled instances are relevant in understanding the reliability of the classifier and assessing how the model behaves globally. For a given cancer class k, we can denote the dataset of correctly labelled instances as $\mathcal{X}_k$.

Algorithm 1 Gene-Wise Global Importance with LIME

Require: Data matrix X, Perturbed data $\tilde{X}$
Require: Decision function f, True labels y
Require: Number of samples M, Class k, Kernel width σ

1: **procedure** GeneLIME($X, \tilde{X}, f, \sigma, N, k$)
2: $\mathcal{X}_k \leftarrow \{\}$
3: **for** $i \in \{1 \dots M\}$ **do**
4: **if** $f(X^{(i)}) = k$ and $y^{(i)} = k$ **then**
5: $\mathcal{X}_k \leftarrow \mathcal{X}_k \cup X^{(i)}$
6: **for all** $x^{(i)} \in \mathcal{X}_k$ **do**
7: Initialize $w_g^{(i)}$
8: $g^{(i)} \leftarrow (w_g^{(i)})^T \tilde{x}^{(i)}$
9: $\mu_x^{(i)} \leftarrow exp\left(-\sqrt{\sum_{i=1}^{n}(x^{(i)} - \tilde{x}^{(i)})^2}\right)/\sigma\Big)$
10: $J(f, g, \mu_x) = \sum_i \mu_x^{(i)}(f(\tilde{x}^{(i)}) - g(\tilde{x}^{(i)}))^2$
11: $\mathcal{W} \leftarrow \min_{g \in G} J(f, g, \mu_x) + \lambda(g)$
12: $\mathcal{G} \leftarrow \sum_{i=1}^{n} \mathcal{W}$
13: **return** $\mathcal{G}$

The process of producing the matrix $\mathcal{X}_k$ is shown in lines 2 to 5 in Algorithm 1. Furthermore, we can denote the process of deriving an explanation from a subset of samples with a function $\xi(\cdot)$. Applying the function $\xi(\cdot)$ is equivalent to performing lines 6 to 11 in Algorithm 1. We now construct an $n \times p$ dimensional explanation matrix by setting $\mathcal{W} = \xi(\mathcal{X}_k)$. The matrix $\mathcal{W}$ represents the local importance of all n genes for each of the p correctly labelled instances for a given class. The gene-wise global weights can then be pooled in an n dimensional vector $\mathcal{G} = \sum_{i=1}^{n} \mathcal{W}_{ij}$. Accordingly, genes that explain more instances will be ranked with higher importance.

4 RESULTS AND DISCUSSION

4.1 dGMU Model Interpretation

We examined the functional enrichment of the top 400 interpretable gene components through a GO term and KEGG pathway analysis. The top 400 genes that promote positive explanations for the eight cancer types were identified as having significantly enriched GO terms and related pathways. The biological process related GO terms with a p-value smaller than 10^{-10} and the related KEGG pathways with p-value smaller than 10^{-3} are presented on Table 1. Many of the statistically significant pathways and terms are related to DNA replication, DNA repair, and cell cycle processes. This suggests that the genes ~~most attributed~~ to explaining the cancer classifications are related to cell proliferation and tumor growth. Furthermore, an additional review of literature was used to identify relationships between the significantly enriched pathways and the cancer types. The enrichment analysis of LIHC identified the carbon metabolism (hsa01200) KEGG pathway, and the response to insulin (GO:0032868), response to activity (GO:0014823), and fatty acid metabolic process (GO:0006631) GO terms. The identification of these biological processes supports significant research describing the pathophysiological link between the human bodies response to insulin and the incidence of LIHC [19, 28]. Insulin stimulates the liver to store glucose, and the liver is the primary site for converting excess carbohydrates into fatty acids. Dysregulated cellular

metabolism, where aberrant oncogenic signals alter the expression of metabolic enzymes, is a reoccurring theme in cancer cells. Currently, there is substantial evidence supporting dysregulated fatty acid metabolism and lipid metabolic reprogramming in LIHC [23, 33]. Through the application of LIME, we identified that the dGMU model is using biologically relevant information to stratify cancer classifications. These results suggest that a domain expert can use the interpretable gene components to understand why the dGMU model correctly classifies true positive cancer instances.

The cell division (GO:0051301) GO term was found to be significantly enriched in four cancer types. For HNSC, the pathway related genes were BUB1, LIG1, BIM, CIB1, SAC3D1, SPC24, BORA, BIRC, ECT2, KIF14, BUB3, and NCAPG. For KIRP, the genes were MAD2L2, ZWINT, CDCA3, CDK5, CDK7, KIRF2C, PARD3B, PRKCE, CDT1, BUB1, and TACC1. For LUAD, the genes were ATAD3B, BUB1B, BUB1, DSN1, NEK2, BIRC5, CDC25C, CDC6, CHEK2, KIF18B, NCAPG, RCC1, SGO1, UBE2C. Lastly, for LUSC the genes were ATAD3B, BUB1, LIG1, DSN1, MAD2L2, SPC25, ZWINT, MCM5, PRKCE, RCC2, TACC1, UBE2C. The greatest overlapping similarity was shared between the two lung cancers LUAD and LUSC, where four genes were shown to be shared. Despite representing the same biological process related to cell division, between the four cancer types, the gene sets were observed to be quite heterogeneous. This suggests that the genes identified as interpretable components have potential applications as biomarkers.

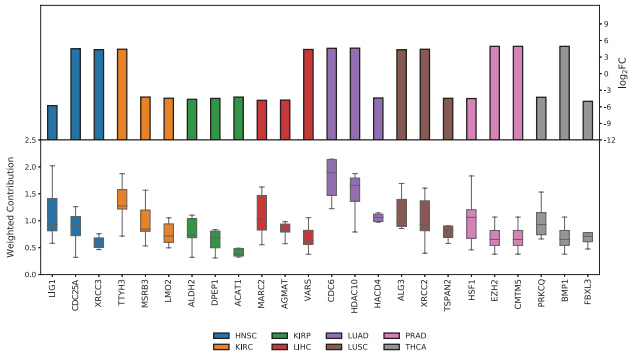

Figure 3: Top 3 explanatory genes for each cancer class.

The intuition of the dGMU model was further investigated by examining the top three explanatory genes derived from LIME analysis for each cancer class. The weighted contribution of these genes along with their respective $\log_2$FC from the differential expression analysis are shown in Fig. 3. For HNSC, CDC25A was identified as one of the explanatory genes. CDC25A is a protein-coding gene that performs an integral role in cell cycle progression. In literature, CDC25A is a known oncogene that is overexpressed in head and neck cancers [11]. This validates the explanation derived from the model that found CDC25A as a key explanatory gene with an overexpressed $\log_2$FC of 5.9. For THCA, LIME identified PRKCQ and BMP1 as the top two explanatory genes for the dGMU model. PRKCQ has been identified as having a potential role in the progression of thyroid cancer, and BMP1 is a known oncogene with potential gene interactions that are influential in the carcinogenesis of thyroid cancer [10, 35]. For both KIRC and KIRP the top explanatory genes, TTYH3 and ALDH2, were identified as prognostic markers for kidney cancer [31]. Accordingly, through the application of LIME explanations, the dGMU model has shown a substantial utility of biologically relevant information for predicting cancer type class.

The location and variability of explanations were visualized using 2D embeddings of the RNA-seq input data. The RNA-seq data were embedded into 2D images by ordering the genes based on gene function and then reshaping the 3025×1 arrays into 55×55 images. An example is shown for five correctly classified HNSC RNA-seq profiles in Fig. 4. The first row shows the 2D embedding of the RNA-seq instances and the second row shows the respective top five positive and negative gene explanations. On the second row, the positive gene explanations that encouraged the prediction of the correct class were labeled in red, and the negative gene explanations were labeled in blue. The circled regions indicate a cluster with a high density of explanatory genes between examples. Although the explanatory genes were determined locally for a given instance, a general consistency in positive explanations remained between RNA-seq data input.

During the dGMU model training scheme, the cumulative weighted contributions for the RNA-seq and SNV features were examined. We found that as the model increased in efficacy, the influence of the RNA-seq modality increasingly dominated in weighted contribution as shown in the top part of Fig. 5. The red kernel density function for the RNA-seq modality progressively separates and settles at a larger average value than the blue kernel density function for the SNV modality. This suggests that the RNA-seq modality contributes stronger explanatory information on average than the SNV modality. This makes sense as the baseline linear model obtained a higher accuracy with the RNA-seq modality than the SNV modality as shown by the associated line chart of error rates illustrated alongside the labelled training epochs on the bottom half of Fig. 5.

The genetic variations of the top 3 explanatory SNV gene regions were visualized across the cell samples for each cancer class on Fig. 6. Each column represents a sample and each row a different gene. The fields are labeled to indicate the category of SNV that is present in the region of the respective gene. The left barplot shows the frequency of variations for each gene, and the associated cancer class is labeled on the right. The mutation frequency of the top 3 SNV gene regions varied widely across the cancer types. The lung cancers LUAD and LUSC had the highest presence of SNVs and the remaining cancers had very sparse variations across cell samples. An exception to this appeared for the BRAF gene in the THCA cancer cell samples. The genetic aberrations of all top three genes have been implicated in the development of various cancers, but genetic variations in HMCN1 and BRCA2 were only found in one case each [16, 17, 22]. Genetic variations in BRAF were found in more than half of the cell samples, the majority of which were missense mutations. BRAF is a protein-coding gene involved in the regulation of signalling pathways that influence cell division, differentiation, and secretion. Mutations in this gene are recurrent in THCA and the missense point mutation in which a single nucleotide change results in valine 600 to glutamic acid

Table 1: Summary of Enriched Gene Ontology Terms and Related Pathways.

Cancer Name	Enriched GO term and Related Pathway			
	ID	Name	Enrichment	P-value
HNSC	hsa04110	Cell cycle	5.2	4.1E-5
	hsa03030	DNA replication	9.8	3.1E-4
	hsa04914	Progesterone-mediated oocyte maturation	5.4	6.2E-4
	GO:0051301	Cell division	4.7	3.3E-11
	GO:0007062	Sister chromatid cohesion	9.2	8.3E-11
	GO:0008283	Cell population proliferation	4.4	4.3E-15
KIRC	GO:0007162	Negative regulation of cell adhesion	16.5	1.8E-13
	GO:0001666	Response to hypoxia	4.4	5.3E-13
KIRP	hsa04210	Apoptotic process	18.2	1.3E-5
	hsa04914	Progesterone-mediated oocyte maturation	5.4	6.2E-4
	GO:0051301	Cell division	4.7	3.3E-11
	GO:0007062	Sister chromatid cohesion	9.2	8.3E-11
LIHC	hsa01200	Carbon metabolism	6.4	7.04E-4
	GO:0032868	Response to insulin	8.6	2.6E-13
	GO:0014823	Response to activity	10.8	5.9E-13
	GO:0006631	Fatty acid metabolic process	8.9	1.0E-12
LUAD	hsa00630	Glyoxylate and dicarboxylate metabolism	11	9.7E-4
	GO:0001525	Angiogenesis	4.1	1.9E-14
	GO:0031568	G1/S transition of mitotic cell cycle	5.9	4.1E-14
	GO:0007062	Sister chromatid cohesion	6.4	2.5E-15
	GO:0051301	Cell division	6.0	8.2E-15
LUSC	hsa03440	Homologous recombination	28.2	8.6E-6
	hsa03030	DNA replication	13.6	1.9E-4
	GO:0051301	Cell division	5.4	2.8E-17
	GO:0000724	Double-strand break repair via homologous recombination	16.4	2.3E-14
PRAD	hsa04530	Tight junction	6.5	2.2E-4
THCA	GO:0006260	DNA replication	10.2	3.3E-12
	GO:0006974	Cellular response to DNA damage stimulus	4.4	5.03E-13
	GO:0006915	Apoptotic process	2.5	1.1E-12

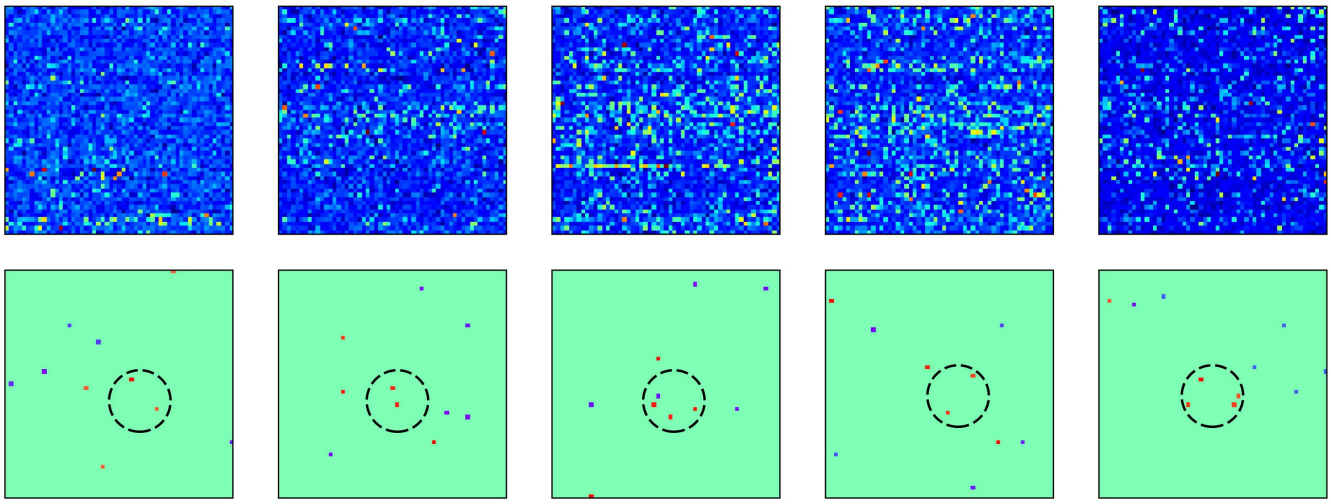

Figure 4: 2D embedding of RNA-seq and explanation heatmap with a localization of persistent explanations.

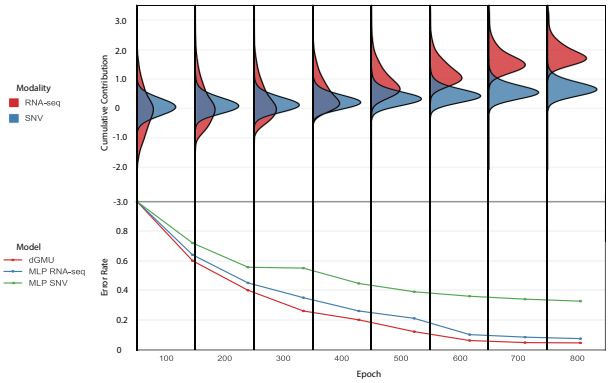

Figure 5: Distribution of cumulative contribution for positive explanations over a range of training epochs.

(V600E) is the most prevalent [34]. Although THCA has a low mortality rate, the presence of the V600E mutation is associated with faster cancer growth and a higher death rate [34]. Accordingly, the interpretable local explanations derived from LIME indicate that the dGMU model draws from clinically relevant information. This trend is found across the different cancer types. The LIME algorithm indicated SNVs in cancer-related genes in all cancer types which provides reasonable explanations that a domain expert can use to understand the prediction of the dGMU model.

5 CONCLUSION

The LIME algorithm was extended to facilitate the interpretation of multi-platform genomic data. We demonstrated the use of this algorithm on a multimodal neural network to generate gene-wise RNA-seq and SNV explanations for the classification of correctly labelled instances. We found that gene-wise explanations are useful for revealing clinically relevant genes used by the machine learning model to make accurate predictions. We also demonstrated that the explanations derived from multi-platform genomic data are helpful for identifying potential biomarkers and validating the predictive influence of known oncogenes. The additional insight gained by examining the explanations is helpful to gain trust in the predictions of the dGMU model. For a given classification, a domain expert can obtain the relative contributions of the modalities and the top explanatory RNA-seq expression and SNV gene regions. In the future, we would like to evaluate enhanced interpretable representations that incorporate the interaction between modalities. This involves incorporating known pathways and gene-gene relationships as a part of the model. We believe that correlating deeper biological relationships will help facilitate greater insight into the underlying machine learning model.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning.. In *OSDI*, Vol. 16. 265–283.

[2] Zakariya Yahya Algamal and Muhammad Hisyam Lee. 2015. Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications* 42, 23 (2015), 9326–9332.

[3] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. 2015. HTSeqâĂŤa Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 2 (2015), 166–169.

[4] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated Multimodal Units for Information Fusion. *arXiv preprint arXiv:1702.01992* (2017).

[5] Emanuele Borgonovo and Elmar Plischke. 2016. Sensitivity analysis: a review of recent advances. *European Journal of Operational Research* 248, 3 (2016), 869–887.

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 839–847.

[7] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* 31, 3 (2013), 213.

[8] ENCODE Project Consortium et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 7414 (2012), 57.

[9] Mark Craven and Jude W Shavlik. 1996. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*. 24–30.

[10] Anthony A Firek, Mia C Perez, Amber Gonda, Li Lei, Iqbal Munir, Alfred A Simental, Frances E Carr, Benjamin J Becerra, Marino De Leon, and Salma Khan. 2017. Pathologic significance of a novel oncoprotein in thyroid cancer progression. *Head & neck* 39, 12 (2017), 2459–2469.

[11] Daniela Gasparotto, Roberta Maestro, Sara Piccinin, Tamara Vukosavljevic, Luigi Barzan, Sandro Sulfaro, and Mauro Boiocchi. 1997. Overexpression of CDC25A and CDC25B in head and neck cancers. *Cancer Research* 57, 12 (1997), 2366–2368.

[12] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. 2013. Maxout networks. *arXiv preprint arXiv:1302.4389* (2013).

[13] Yang Guo, Shuhui Liu, Zhanhuai Li, and Xuequn Shang. 2017. Towards the classification of cancer subtypes by using cascade deep forest model in gene expression data. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*. IEEE, 1664–1669.

[14] Bertrand Iooss and Paul Lemaître. 2015. A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*. Springer, 101–122.

[15] Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya. 2015. Big data analytics in bioinformatics: A machine learning perspective. *arXiv preprint arXiv:1506.05101* (2015).

[16] Electron Kebebew, Julie Weng, Juergen Bauer, Gustavo Ranvier, Orlo H Clark, Quan-Yang Duh, Daniel Shibru, Boris Bastian, and Ann Griffin. 2007. The prevalence and prognostic value of BRAF mutation in thyroid cancer. *Annals of surgery* 246, 3 (2007), 466.

[17] Chie Kikutake, Minako Yoshihara, Tetsuya Sato, Daisuke Saito, and Mikita Suyama. 2018. Intratumor heterogeneity of HMCN1 mutant alleles associated with poor prognosis in patients with breast cancer. *Oncotarget* 9, 70 (2018), 33337.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Xu Li, Xiaocong Wang, and Pujun Gao. 2017. Diabetes mellitus and risk of hepatocellular carcinoma. *BioMed research international* 2017 (2017).

[20] Muxuan Liang, Zhizhong Li, Ting Chen, and Jianyang Zeng. 2015. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 12, 4 (2015), 928–937.

[21] Boyu Lyu and Anamul Haque. 2018. Deep learning based tumor type classification using gene expression data. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 89–96.

[22] Jacqueline Mersch, Michelle A Jackson, Minjeong Park, Denise Nebgen, Susan K Peterson, Claire Singletary, Banu K Arun, and Jennifer K Litton. 2015. Cancers associated with BRCA 1 and BRCA 2 mutations other than breast and ovarian. *Cancer* 121, 2 (2015), 269–275.

[23] Hayato Nakagawa, Yuki Hayata, Satoshi Kawamura, Tomoharu Yamada, Naoto Fujiwara, and Kazuhiko Koike. 2018. Lipid metabolic reprogramming in hepatocellular carcinoma. *Cancers* 10, 11 (2018), 447.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[25] Raheleh Roudi, Alireza Korourian, Ahmad Shariftabrizi, and Zahra Madjd. 2015. Differential expression of cancer stem cell markers ALDH1 and CD133 in various lung cancer subtypes. *Cancer investigation* 33, 7 (2015), 294–302.

[26] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.

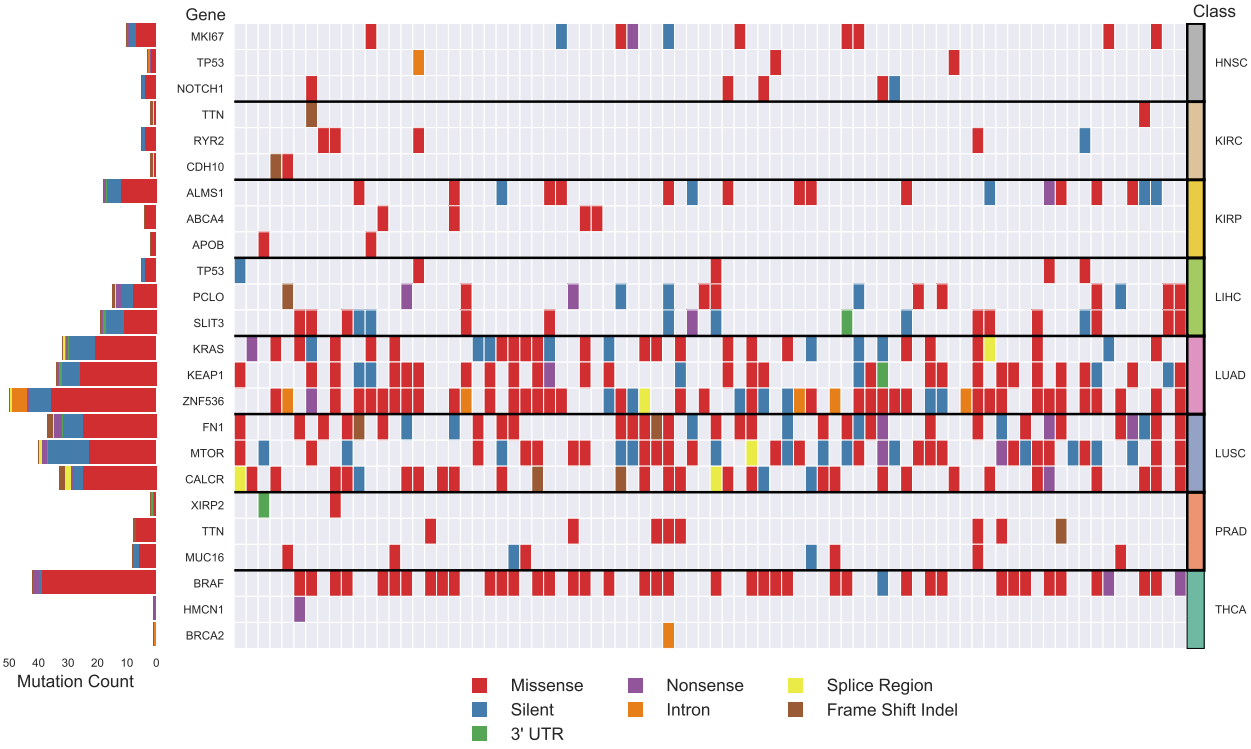

Figure 6: Top 3 explanatory single nucleotide variations for each cancer class.

[27] Shirish Krishnaj Shevade and S Sathiya Keerthi. 2003. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* 19, 17 (2003), 2246–2253.

[28] Mandeep Kumar Singh, Bhrigu Kumar Das, Sandeep Choudhary, Deepak Gupta, and Umesh K Patil. 2018. Diabetes and hepatocellular carcinoma: A pathophysiological link and pharmacological management. *Biomedicine & Pharmacotherapy* 106 (2018), 991–1002.

[29] Arida Ferti Syafiandini, Ito Wasito, Setiadi Yazid, Aries Fitriawan, and Mukhlis Amien. 2016. Cancer subtype identification using deep learning approach. In *Computer, Control, Informatics and its Applications (IC3INA), 2016 International Conference on*. IEEE, 108–112.

[30] Vítor Teixeira, Rui Camacho, and Pedro G Ferreira. 2017. Learning influential genes on cancer gene expression data with stacked denoising autoencoders. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*. IEEE, 1201–1205.

[31] Mathias Uhlen, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, Muhammad Arif, Zhengtao Liu, Fredrik Edfors, et al. 2017. A pathology atlas of the human cancer transcriptome. *Science* 357, 6352 (2017), eaan2507.

[32] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26, 12 (2010), i237–i245.

[33] Mingda Wang, Jun Han, Hao Xing, Han Zhang, Zhenli Li, Lei Liang, Chao Li, Shuyang Dai, Mengchao Wu, Feng Shen, et al. 2016. Dysregulated fatty acid metabolism in hepatocellular carcinoma. *Hepatic oncology* 3, 4 (2016), 241–251.

[34] Mingzhao Xing, Ali S Alzahrani, Kathryn A Carson, David Viola, Rossella Elisei, Bela Bendlova, Linwah Yip, Caterina Mian, Federica Vianello, R Michael Tuttle, et al. 2013. Association between BRAF V600E mutation and mortality in patients with papillary thyroid cancer. *Jama* 309, 14 (2013), 1493–1501.

[35] Yaping Xu, Yue Deng, Zhenhua Ji, Haibin Liu, Yueyang Liu, Hu Peng, Jian Wu, and Jingping Fan. 2014. Identification of thyroid carcinoma related genes with mRMR and shortest path approaches. *PLoS one* 9, 4 (2014), e94022.

[36] Yuchen Yuan, Yi Shi, Changyang Li, Jinman Kim, Weidong Cai, Zeguang Han, and David Dagan Feng. 2016. DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. *BMC bioinformatics* 17, 17 (2016), 476.