2020

MACHINE LEARNING APPROACHES FOR BIOMARKER IDENTIFICATION AND SUBGROUP DISCOVERY FOR POST-TRAUMATIC STRESS DISORDER

Liangqun Lu

MACHINE LEARNING APPROACHES FOR BIOMARKER IDENTIFICATION AND SUBGROUP DISCOVERY FOR POST-TRAUMATIC STRESS DISORDER

by

Liangqun Lu

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Biological Sciences

The University of Memphis

May 2020

ABSTRACT

Post-traumatic stress disorder (PTSD) is a psychiatric disorder caused by environmental and genetic factors resulting from alterations in genetic variation, epigenetic changes and neuroimaging characteristics. There is a pressing need to identify reliable molecular and physiological biomarkers for accurate diagnosis, prognosis, and treatment, as well to deepen the understanding of PTSD pathophysiology. Machine learning methods are widely used to infer patterns from biological data, identify biomarkers, and make predictions. The objective of this research is to apply machine learning methods for the accurate classification of human diseases from genome-scale datasets, focusing primarily on PTSD.

The DoD-funded Systems Biology of PTSD Consortium has recruited combat veterans with and without PTSD for measurement of molecular and physiological data from blood or urine samples with the goal of identifying accurate and specific PTSD biomarkers. As a member of the Consortium with access to these PTSD multiple omics datasets, we first completed a project titled "Clinical Subgroup-Specific PTSD Classification and Biomarker Discovery". We applied machine learning approaches to these data to build classification models consisting of molecular and clinical features to predict PTSD status. We also identified candidate biomarkers for diagnosis, which improves our understanding of PTSD pathogenesis. In a second project, entitled "Multi-Omic PTSD Subgroup Identification and Clinical Characterization", we applied methods for integrating multiple omics datasets to investigate the complex, multivariate nature of the biological systems underlying PTSD. We identified an optimal 2 PTSD subgroups using two different machine learning approaches from 82 PTSD positive samples, and we found that the subgroups exhibited different remitting behavior as inferred from subjects recalled at a later time point. The results from our association, differential expression, and classification analyses demonstrated the distinct clinical and molecular features characterizing these subgroups.

Taken together, our work has advanced our understanding of PTSD

biomarkers and subgroups through the use of machine learning approaches. Results from our work should strongly contribute to the precise diagnosis and eventual treatment of PTSD, as well as other diseases. Future work will involve continuing to leverage these results to enable precision medicine for PTSD.

TABLE OF CONTENTS

LIST OF TABLES

Tables **Pages**

LIST OF FIGURES

Chapter 1

Introduction

1.1 Post-Traumatic Stress Disorder (PTSD)

1.1.1 PTSD Epidemiology

Post-Traumatic stress disorder (PTSD) is a mental disorder that can develop after exposure to serious traumatic events, such as combat, violence, warfare, traffic collisions, or other life-threatening threats. PTSD symptoms may appear within a month or longer after the traumatic events. From the mayo clinic website ("Post-Traumatic Stress Disorder (PTSD) - Symptoms and Causes" 2018), the symptoms can be generally grouped into four types: intrusive memories, avoidance, negative changes in thinking and mood, and changes in physical and emotional reactions. Intrusive memories include recurrent, unwanted distressing memories of the traumatic events. Avoidance means to avoid thinking or talking about the traumatic event. Negative changes in thinking and mood result in problems such as negative thoughts, hopelessness, detachment from family and friends and emotional numbness. Changes in physical and emotional reactions may include irritability, hypervigilance, self-destructive behaviors and trouble sleeping and concentrating. These symptoms cause serious problems in work and life, including significant problems in social or work situations and in relationships. Their frequency interferes with being able to perform normal daily tasks. PTSD severity is often associated with co-occurring conditions such as anxiety disorders.

The current diagnosis of PTSD is from the Clinician Administered PTSD Scale (CAPS - 5) from the Diagnostic and Statistical Manual of Mental Disorders (DSM-V), published in 2013 by the American Psychiatric Association. CAPS - 5 queries PTSD symptoms using 7 criteria and scores are added up to the final assessment. The detailed descriptions and the differences compared with the previous version DSM-IV are shown in the Figure 1.1, as summarized [1].

One of the first large epidemiological analyses of PTSD was performed on

Criterion*	Description	Specific examples	Requirements	Compared with DSM-IV
Criterion A	Exposure to stressor	• Direct exposure • Witnessing trauma • Learning of a trauma • Repeat or extreme indirect exposure to aversive details	DSM-5 recognizes that exposure to trauma can occur either by direct or indirect confrontation with extreme trauma	Specific definition of details of the stressor needed, including repeated experience or extreme exposure to details of events
Criterion B	Intrusion symptoms	• Recurrent memories • Traumatic nightmares • Dissociative reactions (flashbacks) • Psychological distress at traumatic reminders • Marked physiological reactivity to reminders	At least one of these five examples is required	No change, but further clarification of the dissociative quality of flashbacks needed
Criterion C	Persistent avoidance	• Trauma-related thoughts or feelings • Trauma-related external reminders such as people, places or activities	At least one of these two examples is required	DSM-IV did not separate the avoidance criterion
Criterion D	Negative alterations in cognitions and mood	• Dissociative amnesia • Persistent negative beliefs and expectations • Persistent distorted blame of self or others for causing trauma • Negative trauma-related emotions: fear, horror, guilt, shame and anger • Diminished interest in activities • Detachment or estrangement from others • Inability to experience positive emotions	At least two of these seven examples are required	DSM-IV noted social estrangement and restricted the range of affect; numbing redefined to positive rather than all affects
Criterion E	Alterations in arousal and reactivity	• Irritable and aggressive behaviour • Self-destructive and reckless behaviour • Hypervigilance • Exaggerated startle • Problems concentrating • Sleep disturbance	At least two of these six examples are required	Self-destructive and risk-taking behaviours were not defined in DSM-IV
Criterion F	Duration	Must experience criteria B, C, D and F for >1 month	Acute stress disorder is diagnosed for symptoms occurring for <1 month post trauma	No change
Criterion G	Functional significance	Impairment in social, occupational or other domains	Disability in at least one of these domains is required	No change
Criterion H	Exclusion	Not attributable to medication, substance use or other illness	Symptoms must not be secondary to other causes	Not stated in DSM -IV
Subtypes	• Dissociative subtype: used when depersonalization and derealization occur in tandem with other symptoms described above. • Delayed subtype: used to describe the emergence of symptoms following a period post trauma in which symptoms were not present or were present at a subthreshold level.			

DSM, Diagnostic and Statistical Manual of Mental Disorders; PTSD, post-traumatic stress disorder. *Criteria according to DSM-5 (REF. 1).

Fig. 1.1: DSM-5 criteria for PTSD, referred from [1] and [2]

Vietnam War veterans in the United States. Results from this study show that 28% of veterans who had experienced combat developed PTSD, though 11% are still experiencing PTSD 40 years after the combat [3]. Another longitudinal epidemiological study on war fighters serving in Iraq or Afghanistan concluded a 13% PTSD occurrence rate in combat-exposed infantry units while 6% in the general population [4]. According to NIMH statistics ("NIMH » Post-Traumatic Stress Disorder (PTSD)" n.d.) for the whole prevalence, about 6.8% adults in the US experience PTSD at some point in their lives while 3.5% have 12-month prevalence as seen in Figure 1.2.

Across different populations and cultures, the prevalence of PTSD varies significantly.

1.1.2 PTSD Pathophysiology

It has been widely recognized that PTSD is the result of genetic and environmental interaction [5] [6]. In a study of PTSD etiology, Keane et al [7]

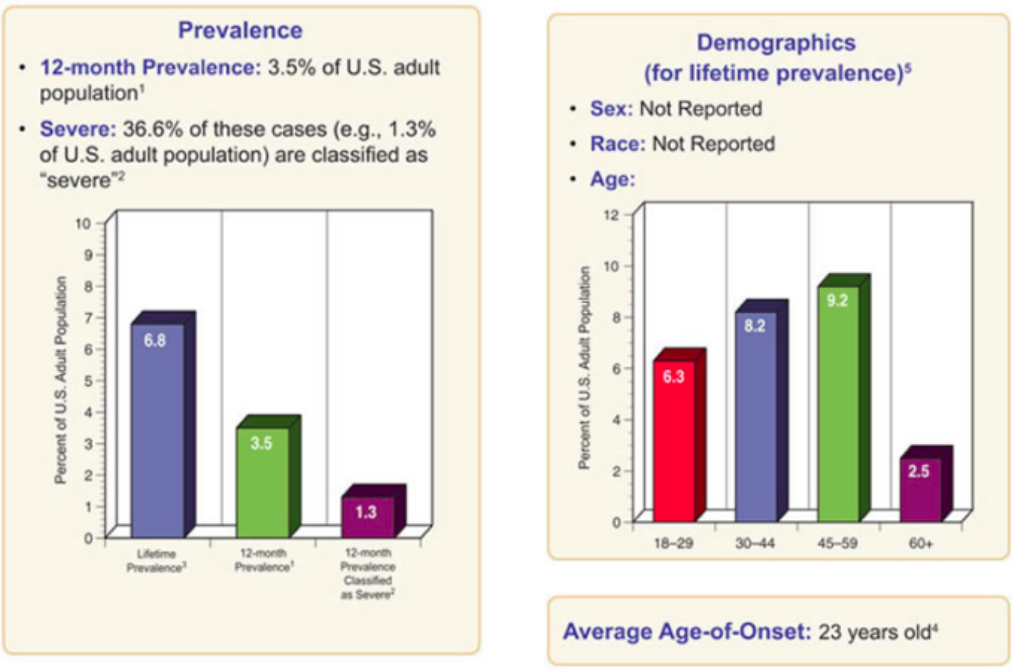


Fig. 1.2: PTSD Prevalence and Demographics across US From NIMH

proposed the triggering model of PTSD and analyzed potential PTSD risk factors divided into three major categories: (a) pre-existing factors specific to the individual, (b) factors related to the traumatic event, including one's immediate response during the trauma, and (c) events that occur following the trauma. Pre-existing Factors include familial psychology, demographic factors (gender, age, race, marital status), prior trauma and life adversity and psychopathology prior to the trauma. A conditional model of PTSD etiology illustrated the processes of PTSD development in the consideration of the three category factors (Figure 1.3).

Systems biology is an approach in biomedical research to integrate information from different scales and gain the benefit of data integration to understand complicated biological systems [8] [9]. These scales consist of measurements in the "omics" level including molecular data–Genomic, Transcriptomic, Epigenetic, Proteomic and Metabolic–and non-molecular data–biological Images and Clinical and Physiological measurements. Given the known pathogenesis spanning neural biology to genomics and genetics discovered

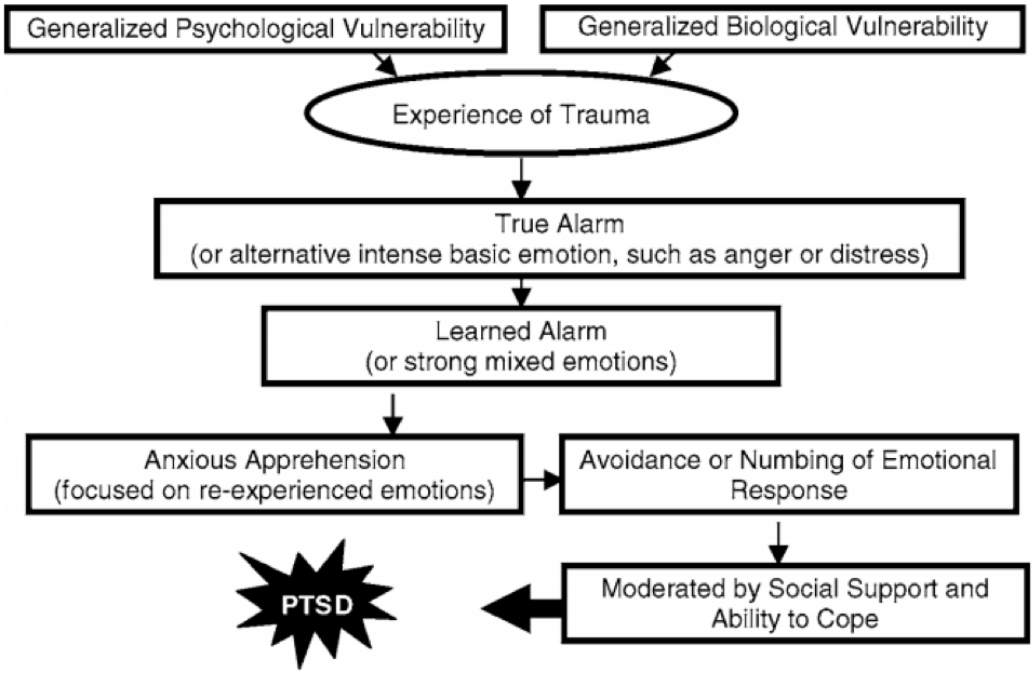


Fig. 1.3: A conditioning model of the etiology of Post-Traumatic stress disorder (PTSD), referred from [7]

in previous PTSD studies, systems biology has been applied to evaluate these discoveries and obtain new biological insights.

PTSD is a psychiatric disorder caused by genetic and environmental interactions, involving a series of biological changes from stress and fear to PTSD [1] [10] [9]. Biological understanding of PTSD has progressed in deciphering the interplay between environmental stimulation, stress responses/reactions, and pathology in light of alterations in brain circuitry and neurochemistry and cellular, immune, endocrine, metabolic, and genetic factors. Characteristic changes in brain regions including hippocampus, amygdala and prefrontal cortex (PFC) have been identified in patients with PTSD. These areas relate to abnormal responses of fear, stress and cognitive deficits, which help to explain the development of PTSD [11]. The hypothalamic-pituitary-adrenal (HPA) axis constitutes the central coordinator of the mammalian neuroendocrine stress response systems and exhibits low cortisol levels in PTSD cases [12]. Molecular-level studies on glucocorticoid signalling confirmed alterations of the HPA axis that reflect exaggerated responses. The increased

secretion of corticotropin-releasing hormone from the hypothalamus results in activated glucocorticoid receptors (GRs). The complex of GRs and cortisols, bound by chaperone proteins including FK506-binding protein 5 (FKBP5), translocates to the nucleus and binds to glucocorticoid response elements (GRE) to ultimately affect transcription of a number of genes [13][1]. The neurochemical features of PTSD found in brain circuits that regulate/integrate stress and fear responses include catecholamine, serotonin, amino acids, peptides, and opioid neurotransmitters. Neuropeptide Y (NPY), which encodes a neuropeptide that is widely expressed in the central nervous system and influences many physiological processes, has been shown to be protective against the development of PTSD. Decreased NPY levels from combat veterans suggests resilience to PTSD by contributing to noradrenergic hyperactivity [14] [15].

In PTSD candidate gene research, previously reported genes confirmed the alterations in PTSD neurobiology, as well as expanded to other biological systems [16] [17] [18]. FKBP5, an important regulator of the stress system by altering GR sensitivity, was reported to have polymorphisms associated with PTSD through interactions with child abuse severity [19] [20] and gene expression modulation by DNA methylation [21]. Catechol-O-methyltransferase (COMT), a critical enzyme involved in the breakdown of the catecholamine neurotransmitters, was reported to play an important role in fear processing, and a genotype change (SNP rs4680) led to impaired fear inhibition in PTSD [22]. Brain-derived neurotrophic factor (BDNF), involved in the neural plasticity underlying the extinction of fear and stress, was identified in relation to anxiety and PTSD [23].Using a hypothesis-free approach, genome-wide association studies (GWAS) have discovered that genes such as retinoid-related orphan receptor alpha (RORA), Cordon-Bleu WH2 Repeat Protein (COBL), Phosphoribosyl Transferase Domain Containing 1 (PRTFDC1) and lincRNA AC068718.1 are associated with PTSD ([24] [25] [26] [27]).

Beyond standard genetics analyses, DNA methylation studies have

identified methylation levels changed due to PTSD in genes Solute Carrier Family 6 Member 4 (SLC6A4), Solute Carrier Family 6 Member 3 (SLC6A3), FKBP Prolyl Isomerase 5 (FKBP5), Nuclear Receptor Subfamily 3 Group C Member 1 (NR3C1) [28]. Neuroimaging genetics studies have also helped identify intermediate phenotypes of PTSD that clarify the functional link between genes and disease phenotype by characterizing gene-specific neurobiological traits associated with PTSD [29]. However, variations in sample size and research methods in these studies have made conclusive identification of PTSD candidate genes difficult. In particular, most of the identified candidate genes lack validation in other independent studies.

Although survey-based PTSD diagnosis is well established using DSM-5 based on criteria from seven specific aspects of the disorder, the accuracy of such diagnoses is influenced by many factors. For instance, some PTSD symptoms are not easily uncovered via survey, which leads to potential cases going undetected. Also, some cases may not wish to be identified and thus do not volunteer information, making it difficult for clinicians to make a correct assessment. For a more objective assessment of PTSD, researchers have worked to identify molecular diagnostic biomarkers for clinical evaluation and treatment as well as to elucidate pathophysiology from these biomarkers ([30] [31]). Candidate biomarkers can be derived from neurobiological, molecular, behavioral, and clinical data and phenotypes associated with PTSD, including neuroendocrine data, brain region changes, genetic and epigenetic molecules and psychophysiological measurements. Comorbidities between PTSD, physical illness and inflammation also directs biomarker investigation to inflammatory systems, such as elevated expression of pro-inflammatory cytokines and C-reactive protein (CRP) [31] [32]. A previous review provided schematic overview of potential biomarkers of PTSD as seen in Figure 1.4.

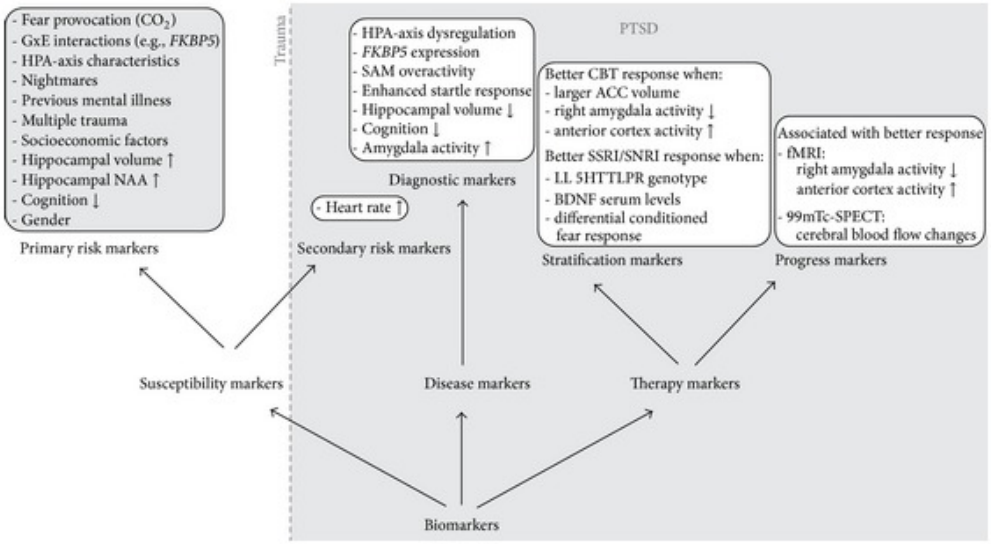


Fig. 1.4: Schematic Overview of PTSD Biomarkers from [30]

1.1.3 PTSD Subgroups

The heterogeneity in the triggering and development of PTSD poses challenges in the diagnosis and treatment of the disorder, and patient diversity makes it difficult for treatment to be equally beneficial to all individuals. Therefore, identification of patient subgroups by inferring patterns underlying groups is a step toward precision medicine and the development of responsive treatment [33]. Multiple studies have worked to discover PTSD subgroups and identify unique biological signatures for the groups. The discovery of subgroups requires techniques from data mining, statistics, and machine learning to learn interesting variables to represent and classify groups [34]. In a study uncovering heterogeneities in the progression of early PTSD symptoms, three trajectories–rapid remitting, slow remitting and non-remitting were proposed based on longitudinal data from the Jerusalem Trauma Outreach and Prevention Study [35]. Using the same cohort of 957 trauma survivors, Galatzer-Levy et al. [36] later applied a support vector machine (SVM) approach to predict the non-remitting PTSD group from information collected within 10 days of a traumatic event. Using a number of psychophysiological features, the authors obtained a mean Area Under Receiver Operating Characteristics Curve (AUC)

of 0.78. They also found that the non-remitting phenotype was attributed to features including nightmares, age, and pulse. In a subsequent study of the same cohort, Galatzer-Levy et al. applied SVMs with feature engineering to build prediction models for trajectories, reaching an AUC of 0.82 using combined demographic, psychophysiological, neuroendocrine and clinical information ([37]). In a recent study, Maron-Katz et al. identified two PTSD subgroups using resting-state functional MRI data from 87 veterans, and these subgroups showed differences between visual and sensorimotor network connection ([16]). Despite these findings, there is a lack of studies discovering PTSD subgroups using multiple scales of biological systems and represented by unique biological signatures.

1.1.4 PTSD Consortium and Data

In 2012, the Department of Defense initiated a multi-site "PTSD Systems Biology Consortium" encompassing researchers from more than 10 institutions, including New York University (NYU), Icahn School of Medicine at Mount Sinai (ISMMS), University of California at San Francisco (UCSF) and at Santa Barbara (UCSB), U.S. Army Center for Environmental Health Research (USACEHR), Emory University, Institute for Systems Biology (ISB), and Harvard University. The primary goals of the PTSD Systems Biology Consortium included identifying a panel of sensitive and specific biomarkers from molecular, physiological, and/or demographic data for PTSD diagnosis, especially in warzone-related cases.

For subject recruitment, PTSD-positive and PTSD-negative participants were selected using the following criteria: 1) male veteran between 20 and 60 years old, 2) deployment in Operation Enduring Freedom and/or Operation Iraqi Freedom, 3) PTSD positive participants with at least 40 CAPS score, 4) PTSD negative participants with less than 20 CAPS score. For consistency, all study participants were evaluated using the DSM-IV PTSD assessment upon recruitment.

For research purposes, there are three sequential cohorts, a Training cohort also called the discovery cohort (82 positive and 82 negative samples), a Test cohort also called the validation cohort (28 positive and 39 negative samples) and a Recall cohort (14 positive, 10 subthreshold positive and 29 negative samples). In the recall cohort, the samples were recalled from the samples existing in the training cohort.

For each of the recruited subjects, blood and urine samples were taken and used to isolate DNA, RNA, protein, metabolites, and endocrine markers for downstream analyses. Physiological measures (e.g., pulse, blood pressure, body mass index) were also collected. Molecular and physiological data from the Training cohort were initially designated for hypothesis generation regarding potential diagnostic biomarkers and underlying biological mechanisms of PTSD, while data from the Test cohort were designated for hypothesis testing and attempted replication of the training sample findings. These subject groups were designed to be age- and ethnicity-matched to minimize biases within and confounding covariates between case and control groups. In total, the genome-wide "omics" measurements include DNA-level methylation of CpG sites, single nucleotide polymorphisms, expression of miRNA-Plasma, miRNA-Deplete and miRNA-Exosome fractions, metabolites, proteins and selected endocrine markers. The miRNA-Plasma fraction measures total miRNA, while miRNA-Exosome and miRNA-Deplete measure miRNA enriched and depleted in exosomes, respectively. These rich omics data sets from the PTSD Systems Biology Consortium were utilized for the following analyses.

1.2 Machine Learning Approaches

High-throughput technologies of molecular biology have advanced to allow examination of associations between omics data—DNA, RNA, metabolites, miRNAs, proteins—and biological conditions, particularly human diseases. Here, omics data refers to the plethora of molecules associated with each "ome" (e.g., genome, transcriptome, proteome) such as DNA or RNA. Next-generation

sequencing technologies have revolutionized the biological sciences by providing an ultra-high throughput way to profile DNA and RNA and thus allow omics data to be generated quickly and economically. Mass spectrometry [38] [39] allows us to efficiently identify and quantify proteins, metabolites and lipids in cells, capturing underlying cellular variations in response to physiological and pathological changes. Although these powerful advances have enabled examination of hundreds of thousands of molecules at the same time, the large-scale nature of data from genomes, transcriptomes, and proteomes have created challenges for data analysis. Machine learning approaches have been developed and applied to elucidate complex biological systems, identify molecular signatures and predict clinical outcomes from such large biomedical datasets ([40] [41]).

Machine learning encompasses algorithms and statistical models applied by computer systems to perform specific tasks based on patterns in data and without using explicit instructions. With more computational resources and elegant algorithms, machine learning has enhanced many real-life experiences including human visual perception. Machine learning has also become an integral part of an ever-growing number of healthcare systems and industries. It has been applied to predict pharmaceutical properties of molecular compounds and targets for drug discovery, perform pattern recognition and segmentation on medical images to enable faster diagnoses and tracking of disease progression, design generative algorithms for computational augmentation of existing clinical and imaging data sets, and develop deep learning techniques for multimodal data sources such as genomic and clinical data that can be combined to make new predictive models [42]. Machine learning mainly uses two types of learning algorithms, supervised and unsupervised learning (Figure 1.5).

In biomedical applications, high-dimensional sets of molecules (variables or features) and small sample sizes are particularly challenging in integrative analysis where multiple omics data are combined to identify unique signatures.

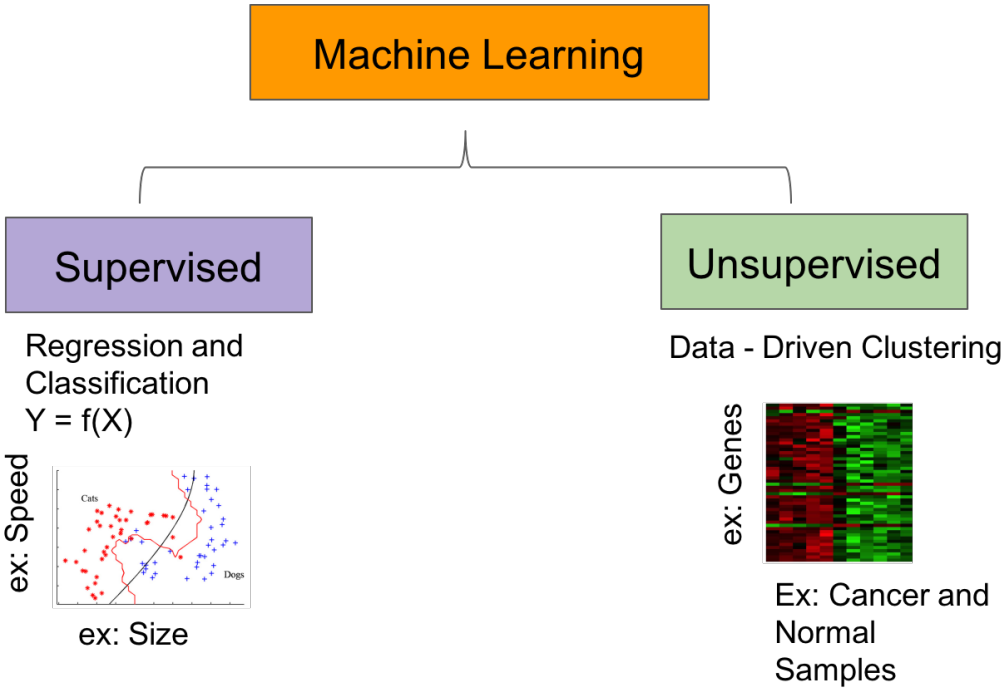


Fig. 1.5: Machine Learning Introduction

This is called the ”curse of dimensionality,” where the number of features is substantially higher than the number of samples, making most machine learning algorithms vulnerable to overfitting [43]. Dimension reduction is the process of reducing the number of variables under consideration by obtaining a set of principal variables. Therefore, dimension reduction is quite useful in machine learning applications for biomedical big data research.

Feature engineering is the process of transforming raw data into fewer features that better represent the underlying problem for predictive models, avoiding overfitting as well as resulting in improved model accuracy on unseen data. Feature engineering generally consists of feature extraction and feature selection approaches to reduce dimensions. Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. Principal component analysis (PCA) is a widely-used feature extraction approach where principal components are extracted to represent the raw data. Feature selection is the process where features which contribute most to predictive performance or output of interest

are automatically or manually selected. Feature selection techniques are widely used to simplify models for interpretation, provide shorter training time, help avoid the curse of dimensionality and enable enhanced generalization by reducing overfitting. The removal of irrelevant features is also beneficial to reduce noise and increase model accuracy. Unlike feature extraction methods such as PCA, which creates new features from functions of the original features, feature selection returns a subset of the features. Three main categories of feature selection algorithms include filters, wrappers and embedded methods. Both feature engineering and feature selection have been successfully used in medical applications, where they can not only reduce dimensionality but also help us understand the causes of a disease ([44]).

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Principal components summarize a large set of correlated variables with a smaller number of variables explaining most of the variability in the data, providing a low-dimensional representation that can be used to produce derived variables for use in supervised learning problems and visualization of observations or variables.

1.2.1 Supervised Learning

Supervised learning algorithms build a mathematical model from a set of data that contains both the inputs and the desired outputs. The data consists of a set of training examples, and each example is a pair consisting of an input object (typically a vector) and a desired output value. Through iterative optimization, an algorithm learns patterns relating the input to the output and improves the performance of predicting output from input. An optimal scenario will allow for the algorithm to correctly determine the outputs for unseen instances. Once the model is trained successfully, it can be applied for prediction in new data (Figure 1.6). One example of supervised learning in biomedicine is

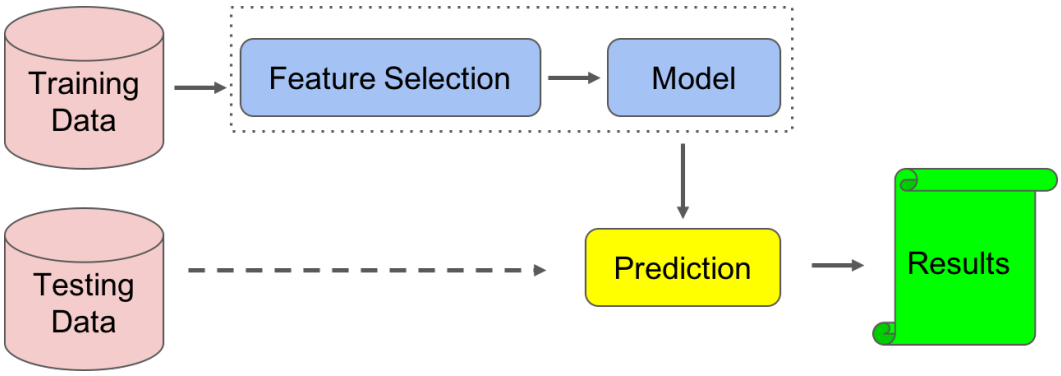


Fig. 1.6: Supervised learning workflow

to train a model using demographic and anthropometric features such as gender, height, education and career to predict a patient's clinical outcome.

Based on the nature of outputs, supervised learning algorithms are further categorized into tasks of regression, where outputs are continuous, and classification, where outputs are binary or categorical. Classification is the task of identifying to which class a new observation belongs, while regression estimates the quantitative relationship between input features and outputs. Examples of supervised classification include building a model to predict whether or not a patient has liver cancer, or building a model to predict whether a patient with liver cancer will get better or worse. Regression examples include building a model to estimate gene expression using gene mutation and copy number variation data. When applied to high-dimensional data such as gene expression for diagnosis prediction, feature engineering is recommended to reduce the problem dimension as well to avoid overfitting.

The procedure of efficient supervised learning consists of the following steps: 1) Determine the type of training examples, 2) Gather a training set, 3) Determine the input feature representation of the learning function, 4) Determine the structure of the learned function and corresponding learning algorithm, 5) Complete the design, 6) Evaluate the accuracy of the learned function.

Logistic Regression is a frequently used supervised classification approach which models the probability of samples falling into a certain group or class (Generalized Linear Models 1989). A binary logistic regression model

estimates the probability of belonging to one or the other group for each sample. Typically, a group probability above 0.5 is treated as a prediction for membership in that group, while a probability below 0.5 predicts membership in the other group. The regression coefficients are estimated for either univariate or multivariate input features, and training can be optimized using gradient descent which maximizes the log-likelihood. Evaluation metrics for logistic regression include Accuracy, Receiver Operating Characteristic (ROC) curve, and area under the ROC curve (AUC). Accuracy refers to the rate of correct group predictions over all samples. ROC curves are created by plotting the true positive rate (TPR, sensitivity) against the false positive rate (FPR, 1 - specificity) at various group probability thresholds. AUC is interpreted as the probability that the classifier will assign a higher group probability to a randomly chosen member of that group than to a randomly chosen member of the other group. Logistic regression has the advantages that it uses a probabilistic framework and can predict binary outcome variables. However, this technique generally assumes independence between input features and is suitable for predicting either discrete (group membership) or continuous (group probability) outcomes.

Two regularization approaches–Lasso (least absolute shrinkage and selection operator) and Ridge regression–can be applied to logistic regression to avoid model overfitting. Lasso involves adding an L1 penalty which refers to an absolute value of magnitude of coefficients to the loss function which represents methods of evaluating model learning for the given data. and in the end can perform both variable selection and regularization to enhance prediction accuracy and model interpretability. Ridge adds an L2-form penalty which refers to the squared magnitude of coefficients and can shrink the regression coefficients toward zero to reduce model complexity and multi-collinearity of input features.

Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms used for classification and regression analyses. SVM constructs a hyperplane or set of hyperplanes in a high- or

infinite-dimensional space. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class. SVMs use hinge loss for optimization, and as mentioned above, Recursive Feature Elimination is often applied to reduce the number of features. Besides binary classification, SVM can also be applied to classification problems with more than two classes.

An important feature of SVMs is the use of kernel functions that enable operation in a high-dimensional, implicit feature space without requiring computation of the coordinates of data in that space. Kernels provide a mapping of the problem from the input space to this higher-dimensional space (called the feature space) by performing a nonlinear transformation. SVMs then use a linear model in this new high-dimensional feature space, which corresponds to a nonlinear model in the input space.

SVMs have the following advantages: 1) SVM hyperplane is robust to outliers, 2) use of a regularization parameter helps to prevent overfitting, 3) a variety of kernel functions are supported, 4) classifier training is equivalent to solving a convex optimization problem, which is algorithmically efficient. Disadvantages include: 1) optimization of the regularisation and kernel parameters and choice of kernel must be conducted separately from training, 2) some kernel methods can be quite sensitive to overfitting, 4) the hinge loss used in SVM optimization results in sparse sets of important features.

SVM has proven to be successful in classifying high-dimensional cancer samples such as ovarian cancer tissues, normal ovarian tissues and other normal tissues ([45]).

Cross Validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. In the case of supervised classification, all samples are randomly divided into training and validation sets. The classification model is then fit using the training set, and its performance is quantified by assessing errors made in predicting

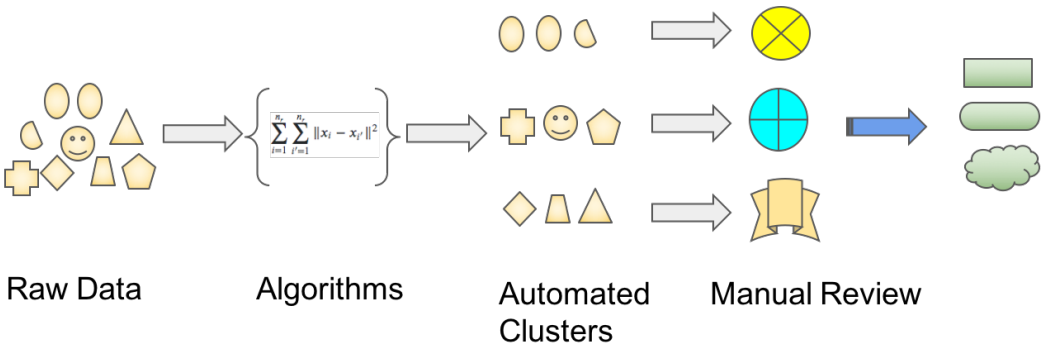


Fig. 1.7: Unsupervised learning workflow

responses for the validation set. In Leave-One-Out (LOO) Cross-Validation, all data is split into n-1 sample training and 1 sample validation sets, followed by model training and validation as described above. The process repeats n times and the average performance of the model is computed over all splits. A more general form of LOO, K-fold Cross-Validation, randomly splits all samples into k folds with k-1 folds used for training and 1 fold for validation. The model is trained and evaluated k times, and average performance is quantified over the k splits. In supervised learning, cross-validation is commonly used to perform model evaluation while avoiding overfitting.

1.2.2 Unsupervised Learning and Data Clustering

In contrast to supervised learning, unsupervised learning is the task of inferring a function to describe a hidden structure in data that is missing output values (e.g., classes or labels). Goals of unsupervised learning include understanding relationships between observations, visualizing data in an informative way and discovering subgroups among the variables or observations (Figure 1.7). One downside to this approach is that, without label information, evaluation of predicted structure in the data is not straightforward. Specific applications of unsupervised learning include clustering, PCA, anomaly detection, and latent variable identification. A number of studies have applied unsupervised clustering in prostate cancer subtype identification [46], novel breast cancer subgroups [47] [48], glioma subtype discovery [49].

Clustering is the task of grouping a set of objects in such a way that

objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). Since the notion of a cluster is not precisely defined, many different clustering algorithms have been developed to discover clusters. Examples of clustering methods include: Connectivity models (e.g., hierarchical clustering), which builds clusters based on distance connectivity where samples are more related to nearby samples than samples far away; Centroid models (e.g., K-means), which represents each cluster by a single mean vector; Graph models (e.g., clique-based), which represent a cluster by a subset of nodes connected by edges; Biologically-inspired models (e.g., unsupervised neural networks). Clustering can be categorized as either hard(each object belongs to a single cluster) or soft (each object belongs to every cluster to a certain degree). As mentioned above, one disadvantage of clustering is the difficulty in evaluating predicted clusters, particularly in independent test data sets. Thus, these methods are commonly performed as part of exploratory data analysis.

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. For a given value of K (number of clusters), the goal of K-means is to partition all training data samples into K distinct, non-overlapping clusters by identifying clusters for which within-cluster variation is as small as possible.

The K-means algorithm involves the steps below. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments. 1. Iterate until the cluster assignments stop changing: 2. For each of the clusters, compute the cluster centroid. This is a vector of the feature means for the observations in the cluster. 3. Assign each observation to the cluster whose centroid is closest in terms of Euclidean distance.

This algorithm is guaranteed to decrease the sum of within-cluster variations at each step. However, it is not guaranteed to reach the global minimum, because results depend on the initial, random cluster assignments of each observation in Step 1 and individual runs may terminate at suboptimal

local minima. In practice, the K-means algorithm is run multiple times from different random initial configurations, and the best clustering result (smallest sum of within-cluster variations) is selected.

The K-means algorithm involves repeatedly assigning points to the closest cluster centroid. To do so, K-Means requires computing of pairwise Euclidean distances between data points, because the sum of squared deviations from a centroid is equal to the sum of pairwise squared Euclidean distances between points divided by the number of points. The term "centroid", which derives from Euclidean geometry, refers to a multivariate mean calculated in euclidean space.

A second clustering approach, spectral clustering ([50]), uses the eigenvectors (spectrum) and eigenvalues of a matrix to define cluster membership. These eigenvectors function as indicators of cluster membership. Importantly, although small perturbations such as adding a few edges linking clusters or removing edges from inside the clusters will increase eigenvalues and change the corresponding eigenvectors, this does not generally cause the underlying cluster structure to be lost. Like K-means, spectral clustering technique requires the number of desired clusters to be specified ([51]).

As previously mentioned, it is particularly challenging to evaluate unsupervised clustering results. Popular approaches consider both internal and external evaluation. Internal evaluation is typically summarized by a quality score, although it is not always clear which metric should be used to compute this score. Examples of popular metrics include the following:

Silhouette coefficient: average distance within the cluster against average distance outside the cluster. It is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample with the formula (b - a) / max(a, b) of each sample. The score ranges from -1 to 1, with a negative value indicating that the average distance within the cluster is greater than the distance outside the cluster.

1.2.3 Deep Learning

Deep learning is a class of machine learning algorithms based on artificial neural networks (ANNs) and representation learning. It utilizes a cascade of multi-layered deep neural networks (DNNs) for feature extraction and data transformation. Deep learning attempts to generate abstractions from large-scale data including images and texts. Deep learning can be applied for supervised or unsupervised tasks. An advantage of deep learning architectures is that they automatically perform multiple levels of nonlinear data transformation and supervised or unsupervised learning of feature representations. Deep learning architectures such as autoencoders, recurrent neural networks and convolutional neural networks have been applied to fields in computer vision, natural language processing and biomedical data science ([52] [53]). A summary of deep learning architectures is shown in the neural network zoo [54].

ANNs were originally inspired by the complex neurobiological systems in the brain which learn to perform tasks by generating representations without task-specific programming [55]. An ANN is based on a collection of connected units called artificial neurons, analogous to biological neurons in a biological brain. Each connection between neurons can transmit a signal in a manner similar to a synapse. The receiving neuron can process incoming signals and then signal downstream neurons connected to it. Neurons in ANNs may have a state, generally represented by real numbers, typically between 0 and 1. Neurons and synapses may also have weights that vary as learning proceeds, which can increase or decrease the strength of signals that are sent downstream. Typically, neurons are organized in layers, with each layer potentially performing different transformations to their inputs. Signals travel from the first (input) to the last (output) layer, possibly after traversing inner layers multiple times. The original goal of the neural network approach was to solve problems in the same way that a human brain would. However, attention has shifted over time to matching specific cognitive tasks which led to deviations from biology such as the

backpropagation optimization technique, in which information is passed from output to input layers while adjusting the network weights to reflect that information.

DNNs are ANNs with multiple internal layers between the input and output layers. Like other ANNs, a DNN learns the correct mathematical operations to transform the network input into the output. Although there exists a variety of DNN architectures, most DNNs rely on similar techniques for feature extraction and training, such as feedforward and backpropagation passes. In the feedforward pass, the network is activated by an input to the first layer, which then spreads the activation to the final layer along the weighted connections and generates the prediction or reconstruction results. In the backpropagation pass, the weights of connections are tuned by minimizing the difference between the predicted output and the real output. By combining these techniques with activation functions, optimization objectives and optimization methods, deep learning models can be implemented and applied to specific tasks.

Activation functions make up the nonlinear layers in all deep learning models, and their combination with other layers enable nonlinear transformations from the input to the output. An optimization objective is typically composed of a loss function and a regularization term, with the former measuring the discrepancy between predicted and actual network output and the latter used to reduce test set error and avoid overfitting. Optimization methods are strategies used to achieve minima of the objective function by selecting appropriate hyperparameters which means a number of parameters before the training processes and their combination set can be optimized based on the performance in the evaluation data. Stochastic gradient descent (SGD) [56] and its variants are commonly-used optimization methods which update network weights (parameters) by a step corresponding to the Jacobian matrix which as a matrix calculates partial derivatives of a vector function for weight updates. For example, the Adaptive Gradient Algorithm (AdaGrad) [57] technique updates

weights according to the accumulation of squared gradients, which can converge rapidly when applied to convex functions. RMSProp, an AdaGrad algorithm, has been an effective and popular method for parameter optimization.

Tensorflow (TF) is a widely used end-to-end open source deep learning platform implemented in the Python language. Keras is a highly wrapped deep learning framework running on top of TF or other deep learning platforms. These two deep learning frameworks have implementations on various deep learning architectures.

An autoencoder (AE) is one class of deep learning architectures used to learn input data representations in an unsupervised manner [58]. The purposes of this representation include dimensionality reduction and reconstruction of the input with the removal of noise. An AE is constituted by two main parts: an encoder that maps the input into a code, and a decoder that maps the code to a reconstruction of the original input. In terms of architecture, AEs are feedforward, non-recurrent neural networks very similar to the multilayer perceptron (MLP)–they have an input layer, an output layer and one or more hidden layers connecting them, with the output layer having the same number of nodes as the input layer. Additional applications of AEs include feature learning and learning generative models which indicate the architectures can estimate training data distribution and generate new samples from the same distribution. Various AE variants exist to prevent autoencoders from simply learning the identity function and to improve their ability to capture important information and learn richer representations. Examples include denoising AE, sparse AE and variational AE, as shown in Figure 1.8.

Variational autoencoder (VAE) models inherit the general autoencoder architecture of both an encoder and a decoder and are trained to reduce the reconstruction error between input and output. However, VAEs make strong assumptions of the input distribution over the latent space which refers to the middle hidden layer with the fewest neurons in AEs. In practice, encoded

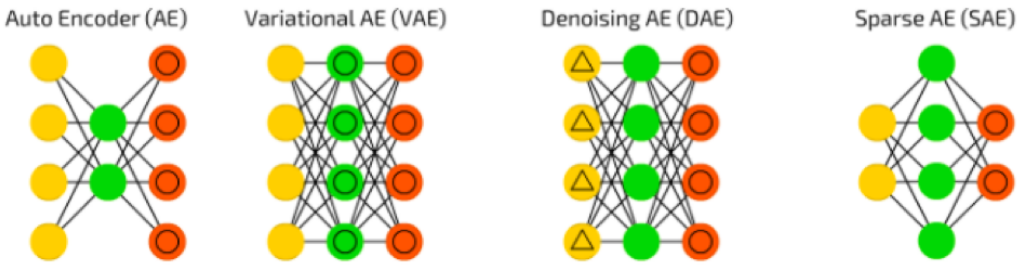


Fig. 1.8: Architectures for AE, VAE, DAE and SAE

distributions are chosen to be normal so the encoder can be trained to return the mean and covariance matrix of a multivariate gaussian distribution. The distributions returned by the encoder are then enforced to be close to a standard normal distribution. The loss function of VAEs is composed of a reconstruction loss and Kullback-Leibler divergence. The reconstruction loss measures the distance between input and output, while the Kullback-Leibler divergence measures the similarity between the distribution of the encoder and a standard normal distribution. This loss function has a closed form that can be directly expressed in terms of the means and covariance matrices of the encoded distributions. The VAE model is trained in the following steps: 1) the input is encoded as a distribution over the latent space, 2) a point from the latent space is sampled from that distribution, 3) the sampled point is decoded and the reconstruction error computed, 4) the reconstruction error is backpropagated through the network. An important application of VAEs includes learning deep generative models to generate new data using the learned distribution. One such study in biomedical research trained a VAE on cancer gene expression data and identified biological patterns in the encoded features [59].

1.3 Research Objectives

Given the great promise of machine learning approaches for biomarker discovery and precision medicine, our research aims to apply these techniques for subgroup discovery and classification of human diseases such as PTSD or cancer using genome scale datasets. With multiple omics data sets from the PTSD Systems Biology Consortium, the first of our projects involved implementing

clinical subgroup-specific PTSD classification and biomarker discovery. Specifically, we applied machine learning approaches to classify PTSD positive patients in three molecular datasets–miRNA-Exosome, miRNA-Deplete and Metabolomics–and select discriminative features as potential biomarkers. Our second project aimed to identify PTSD subgroups through multiple omics data integration via statistical and deep learning approaches. For biological interpretation, we performed clinical characterization and differential expression analysis based on the identified subgroups. We also utilized subgroup labels to improve diagnostic classification. Finally, we worked on two supplementary projects involving the development of a novel feature selection technique and histopathological image feature extraction for classification using pre-trained convolutional neural networks. Taken together, our research aims to explore applications of machine learning in biomedicine and contribute to the precise diagnosis and eventual treatment of PTSD, as well as other diseases.

Chapter 2

Clinical Subgroup-Specific PTSD Classification and Biomarker Discovery

2.1 Abstract

Post-Traumatic Stress Disorder (PTSD) is a psychiatric disorder caused by environmental and genetic factors resulting from alterations in genetic variation, epigenetic changes and neuroimaging characteristics. There is a pressing need to identify reliable molecular and physiological biomarkers for accurate diagnosis, prognosis, and treatment, as well to deep the underpinning of pathophysiology.

Using a cohort of 234 samples with 166 in training and 68 in validation, applied machine learning approaches to classify PTSD patients in three molecular datasets miRNA-Exosome, miRNA-Deplete and Metabolomics. We first divided patients into multiple sets of two subgroups based on the values of 112 clinical and endocrine measurements. We then performed supervised classification across all samples and within each subgroup using two feature selection strategies (Recursive Feature Elimination (RFE) and ANOVA), four classifiers (logistic regression (LR), support vector machines (SVM), random forests (RF), and extra trees (ET)), and 10-fold nested cross validation. We evaluated each subgroup for significantly improved classification performance by computing empirical false discovery rates (FDRs) based on accuracy and AUC values. Finally, we combined those significant clinical features with molecular measurements and constructed an overall PTSD classifier. We fit all data in the best classification model in training and selected features as biomarkers.

In total, 85 clinical subgroups from 72/112 clinical and endocrine features lead improved classification performance compared to the baseline from all samples in training, with 38 yielding improved performance in more than one method. Tree-based models yielded the greatest number of improved subgroups in the metabolomics and miRNA from exosomes datasets, while Logistic

Regression showed the greatest improvement in the miRNA depleted of exosomes dataset. Using our overall PTSD classifier with molecular and clinical features, we observed that the majority of classification models show improved accuracy in both training and testing. In metabolomics, the overall model achieved the best AUC 0.79 ± 0.13 and ACC 0.722 ± 0.078 at ANOVA-SVM, significantly better than the baseline models at ACC with only molecular or clinical features. In miRNA-Exosome, ANOVA-LR model reached the best AUC 0.758 ± 0.097 and ACC 0.701 ± 0.116 with ACC significance compared to the clinical model baseline. In miRNA-Deplete, RFE-SVM model reached the best AUC 0.677 ± 0.134 and ACC 0.605 ± 0.128. These best models have fair performance in validation as well. We also selected the features in these models and listed as potential biomarkers consisting of molecular and clinical features.

We applied machine learning approaches in multiple types of PTSD data and built classification models consisting of molecular and clinical features to predict PTSD patients. We also provide candidate biomarkers for the diagnostics, which improves the pathogenesis understanding of PTSD. Hopefully, our work contribute to the precise diagnostics and treatment at PTSD.

2.2 Materials and Methods

2.2.1 Study Samples

The Department of Defense-funded Systems Biology of PTSD Consortium has recruited over 200 male combat veterans with and without PTSD for the purposes of identifying diagnostics biomarkers. Whole blood samples taken from each subject have been used to isolate DNA, RNA, protein, metabolites, and endocrine markers for subsequent study. In the first stage, two cohorts were collected–a discovery cohort of 166 samples: 83 PTSD positive and 83 PTSD negative, followed by a validation cohort of 68 samples: 29 PTSD positive and 39 PTSD negative. For this study, we analyzed three molecular datasets from these two cohorts–two measuring miRNA derived from plasma either enriched (miRNA-Exosome) or depleted (miRNA-Deplete) in exosomes, and one

measuring metabolomics. We also included clinical characteristics from four categories including PTSD scales, demographic information, biochemical and anthropometric body measurements, and endocrine measurements from blood and urine. After filtering these clinical characteristics for numeric measurements with less than 10% missing values in the discovery cohort, we were left with 112 features, among which 24 are binary and 88 are continuous. The datasets used in this study were collected and preprocessed in or before January 2017.

2.2.2 Clinical Feature Association Analysis

For the 112 clinical features, we first performed statistical association analysis between those measurements and PTSD labels. In both training and validation cohorts, we used Fisher's Exact test for binary features and t-test for continuous features. We obtained p-values and used the threshold 0.05 to determine significant associations.

2.2.3 Clinical Subgroups

For each clinical feature, we split samples from the training cohort into two subgroups based on their feature values. For binary features, subgroups were created based on which of the two values each sample held. For continuous features, subgroups were based on the median measurement: a higher expression subgroup contains samples whose measurements were greater than or equal to the median, with the remaining samples belonging to a lower expression subgroup. We used the same criteria to split samples and obtain subgroups in the validation cohort. After clinical subgroup creation, we split the corresponding measurements from each of the three omics datasets–miRNA-Exosome, miRNA-Depleted and metabolomics–accordingly. In the following analyses, we only considered clinical subgroups with at least 10 samples each from the training cohort and at least two samples each from the validation cohort.

2.2.4 Missing Value Imputation

Although we kept clinical features with less than 10% missing measurements, any missing values still prohibited the direct use of many

supervised machine learning methods. Thus, we first used the method Imputer to generate missing values based on mean values of nearest neighbors [60]. We then fed the input data to machine learning algorithms for classification.

2.2.5 Supervised Classification

Feature selection is a common preprocessing step in machine learning, especially when applied to high-dimensional biological data, as it is effective in reducing dimensionality, removing irrelevant information, decreasing model variability and increasing learning accuracy [61]. We applied two types of feature selection in supervised classification. The first is a filtering approach, which selects the most relevant features manually or automatically before classification. We chose Analysis of variance (ANOVA) as a commonly-used filtering feature selection method. The second is an embedding approach, which considers feature selection and classification simultaneously as part of an analysis pipeline. For this approach, we chose recursive feature elimination (RFE), which has the potential to yield better performance [62].

For performing supervised classification, we included four classifiers: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Extra Tree (ET). LR builds a linear boundary between classes and predicts the maximum probability for being in one class. SVM builds a hyperplane between classes and uses kernel tricks to perform linear or nonlinear classification. Both RF and ET are tree-based classifiers differing in the setup of randomized trees. For each classifier, we evaluated a series of hyperparameters including the following: l1 and l2 regularization (LR), linear, polynomial, and radial basis function kernels (SVM), cost penalty ($10^{-4,4}$ and $2^{-4,4}$) and gamma ($10^{-4,-1}$) (SVM), tree numbers and max features (RF and ET).

When using ANOVA feature selection, we selected the best model for each classifier using a grid search across a range of 10% to 100% top-scoring features. When performing RFE-based feature selection and classification, the lowest-scoring 10% of features were eliminated at each iteration and the best

model selected across all iterations. To evaluate classification performance, we computed metrics including AUC (Area under ROC curve) and accuracy (ACC). To avoid overfitting, we applied 10-fold nested cross validation (NCV) to obtain the average performance of each classifier. Specifically, we divided all samples into 10 folds, in which nine folds were used for training with one left for external testing. In those nine folds, 9/10 were used for tuning classifier hyperparameters, followed by internal testing in the remaining fold. For both ANOVA- and RFE-based classification, we computed average model performance based on 10-fold NCV.

2.2.6 Classification Performance Comparison

For the clinical subgroup comparison, we used a paired t-test to detect mean differences in AUC and ACC between each subgroup and the baseline where all samples were used for classification. We used a p-value threshold of 0.05 to select significant differences. When performing comparisons between models composed of molecular data, clinical data and mixed data, we also applied paired t-tests to detect significant differences in performance.

2.2.7 Biomarker Discovery

We first selected the best overall models from the three omics datasets, considering both ANOVA- and RFE-based feature selection methods. We then fit models of all molecular and clinical data together using the same hyperparameter tuning strategy as above. All selected features were considered to be candidate PTSD biomarkers. For the tree-based classifiers, the rankings of these biomarkers were calculated using variable importance, while for LR or SVM classifiers the rankings were computed from the feature coefficients.

2.3 Results

In this study, we applied machine learning approaches to multiple PTSD omics datasets and built classification models to predict PTSD status as well as identify candidate biomarkers. Using the clinical and molecular data described

above, we built classification models composed of integrated clinical and molecular features and identified diagnostic biomarkers from both data types.

2.3.1 Association Analysis of Clinical and Endocrine Features With PTSD

We first applied statistical tests (t-tests for continuous features and Fisher's Exact test for binary features) to detect associations between PTSD status and clinical and endocrine characteristics. Given the 112 clinical and endocrine features (24 binary and 88 continuous) passing the filters described above, we identified 29 (26%) that are significantly associated with PTSD (Figure 2.1). Of these 29, three consist of physiological measurements (Figure **??**-A), two of demographic information (Figure 2.1-B), 17 of biochemical measurements (Figure 2.1-C), three of blood endocrine measurements (Figure 2.1-D), and four of PTSD scales. The majority of these features show significant associations in the training cohort but not in the validation cohort. Only three–pulse, glucose (fasting glucose) and rbc (red blood cell count)–are significantly associated with PTSD in both cohorts, where PTSD patients have higher levels of these features.

2.3.2 Clinical Subgroup Classification Performance

Despite the heterogeneity of PTSD pathology, there is currently no staging approach for suggesting distinct therapeutic treatments depending on the degree of biological progression of the disorder [63]. Inspired by findings that subtypes of PTSD vary in time to recovery and disorder severity, we worked to determine whether clinical subgroups within each omic dataset could enable improved classification performance compared to a baseline considering all samples. The analysis flow of such clinical subgroup classification is sketched in Figure 2.2. For each of three omics datasets–miRNA-exosome, miRNA-deplete and metabolomics–we split all samples to obtain two clinical subgroups for each of 112 clinical measurements. We then built classification models for all subgroups individually as well as for a baseline consisting of all samples for

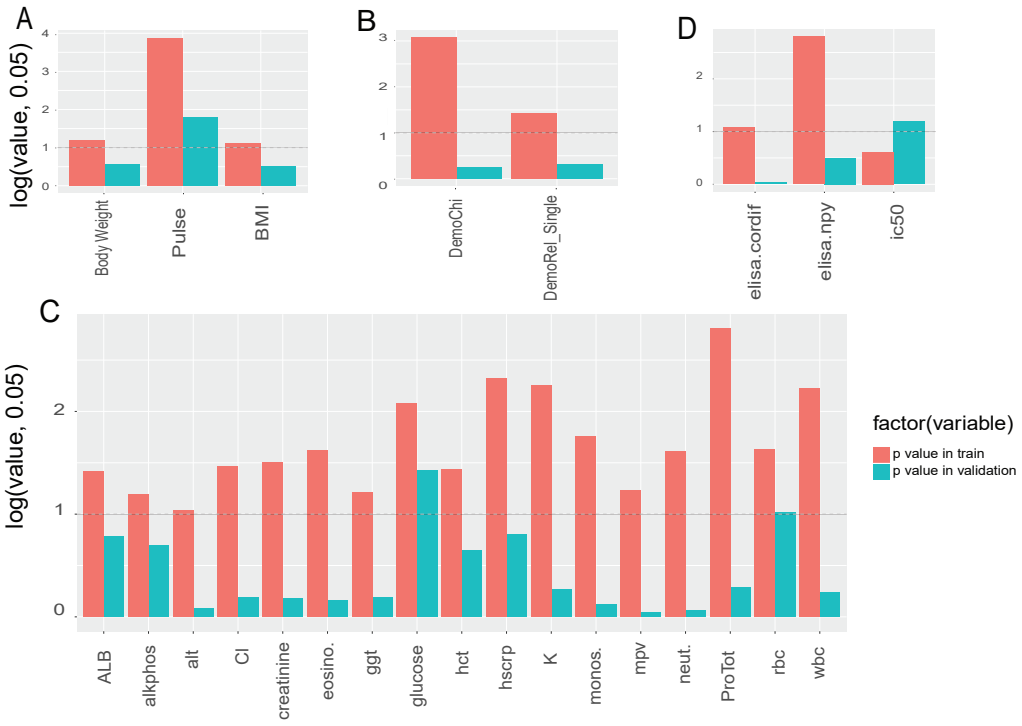


Fig. 2.1: Bar Plots of Clinical Characteristics Significantly Associated With PTSD, x axis indicates clinical or endocrine features, y axis indicates the log transform p values, red color indicates the significance in training dataset while blue color indicates in validation dataset, A) in physiological measurements, B) in clinical biological background, C) in biochemical measurements, D) in endocrine from blood.

performance comparison. We evaluated two feature selection approaches–ANOVA and Recursive Feature Elimination (RFE)–and four classifiers–Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Extra Tree (ET). We utilized 10-fold cross validation to evaluate model performance using the criteria Area Under ROC Curve (AUC) and Accuracy (ACC). Additional details about clinical subgroups and classification models can be seen in the Materials and Methods section above.

Overall, 85 clinical subgroups from 72 clinical and endocrine features show improved classification performance with significantly improved training AUC or ACC compared to the baseline. We also listed a list of top clinical subgroups which showed AUC improvement in the validation (Vad) data in Table 2.1. Across the eight classification methods tested (two feature selection approaches times four classifiers), we found that use of the tree-based classifiers

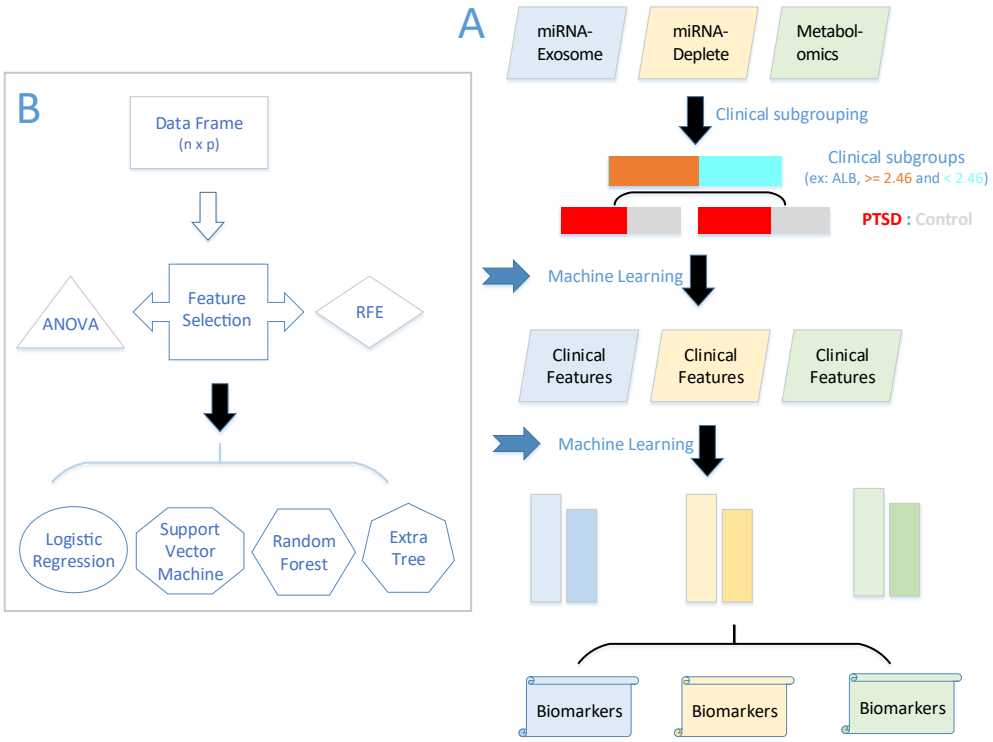


Fig. 2.2: PTSD Clinical Subgroup Classification Workflow, A) In the analysis flow from subgrouping to the biomarker discovery, each omic dataset is split based on each clinical or endocrine measurement and ends up with two clinical subgroups consisting of PTSD and control samples individually. For each subgroup, a machine learning approach is used to build classification models and compared to the baseline from all samples. A set of subgroups with improved classification performance are selected for each omic dataset. With the combined molecular and clinical features, machine learning is used to build classification models. Then best models are selected for three datasets and biomarkers are discovered from the selected features. B) A machine learning approach is used to build classification models for datasets in A part. For each dataset with samples and features, two types of feature selection methods ANOVA and Recursive Feature Elimination (RFE) and 4 classifiers Logistic Regression, Support Vector Machine, Random Forest and Extra tree are pipelined and implemented.

RF and ET led to the greatest number of subgroups with improved performance in Metabolomics and miRNA-Exosome data, while use of LR resulted in the most improvements in miRNA-Deplete data. In Metabolomics, use of the top two models–ANOVA-RF and RFE-RF–led to 19 and 16 improved subgroups, respectively, in training, and 17 and 13 improved subgroups, respectively, in validation. In the miRNA-Exosome dataset, the top three methods–ANOVA-ET, ANOVA-RF and RFE-ET–led to 13, 9 and 9 improved subgroups, respectively,

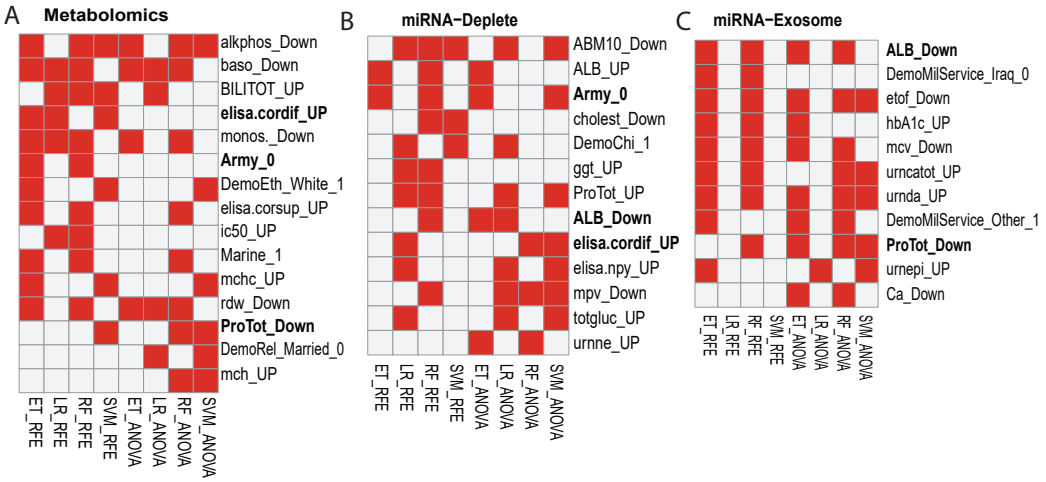


Fig. 2.3: Heatmaps Showing Improved Performance in Three Datasets miRNA-Exosome, miRNA-Deplete and Metabolomics, rows indicate different methods from feature selections and classifiers while columns indicate the clinical or endocrine subgroups, A) in Metabolomics, b) in miRNA-Deplete, C) in miRNA-Exosome, red color indicates detected improvement while grey indicates no improvement.

in training and 4, 9 and 4 improved subgroups, respectively, in validation. Finally, in the miRNA-Deplete dataset, the top two methods–ANOVA-LR and RFE-LR–led to 26 and 22 improved subgroups, respectively, in training and 9 and 17 improved subgroups, respectively, in validation. Overall, 38 subgroups showed significant improvement for more than one classification method using the same feature selection approach. Of these, 11 are from miRNA-Exosome data, 12 from miRNA-Deplete data and 15 from Metabolomics data as seen in Figure 2.3(A-C). We also found that 4 subgroups showed improvements in more than one dataset: `ALBDown (albumin)`, `Army0 (Military Service)`, `elisa.cordifUP (cortisol1 - cortisol2)` and `ProTotDown (total serum protein)` as seen in Figure 2.4. Taken together, the improved classification performance for particular subgroups suggests that inclusion of the corresponding clinical measurements in classification would improve overall PTSD status prediction.

2.3.3 Classification From Clinical and Molecular Features

The complex network of clinical and molecular interactions shows promise for enabling our understanding of the biological dysregulation responsive to

Table 2.1: Performance Improved Clinical Subgroups

Data Sets	Methods	Clinical Subgroups	Classifier	AUC	VadAUC	AUC p value
Metabolomics	ANOVA	monos.-Down	ET	0.883 ± 0.172	0.762	0.0253
Metabolomics	RFE	monos.-Down	RF	0.821 ± 0.148	0.721	0.034
Metabolomics	RFE	baso-Down	LR	0.831 ± 0.145	0.685	0.0116
miRNA-Exosome	ANOVA	etof-Down	SVM	0.907 ± 0.161	0.679	0.0278
Metabolomics	RFE	baso-Down	RF	0.822 ± 0.114	0.658	0.0111
Metabolomics	RFE	BILITOT-UP	RF	0.782 ± 0.087	0.653	0.0392
Metabolomics	RFE	monos.-Down	LR	0.842 ± 0.217	0.646	0.04
Metabolomics	ANOVA	rdw-Down	ET	0.858 ± 0.123	0.644	0.0157
Metabolomics	ANOVA	ldl-Down	SVM	0.839 ± 0.08	0.634	0.0244
Metabolomics	ANOVA	rdw-Down	RF	0.869 ± 0.14	0.633	0.0131
Metabolomics	RFE	rdw-Down	RF	0.855 ± 0.14	0.631	0.007
miRNA-Exosome	RFE	ABM3B-UP	LR	0.818 ± 0.129	0.631	0.0229
Metabolomics	ANOVA	ABM6-UP	SVM	0.854 ± 0.114	0.63	0.0241
Metabolomics	ANOVA	BILITOT-UP	LR	0.848 ± 0.094	0.627	0.0413
Metabolomics	RFE	alkphos-Down	RF	0.795 ± 0.097	0.609	0.0254
Metabolomics	RFE	elisa.cordif-UP	SVM	0.849 ± 0.156	0.604	0.0427
miRNA-Deplete	RFE	Army-0	RF	0.708 ± 0.186	0.604	0.359
Metabolomics	RFE	ic50-UP	LR	0.772 ± 0.109	0.602	0.0483
Metabolomics	RFE	rdw-Down	ET	0.84 ± 0.14	0.593	0.0203
Metabolomics	RFE	baso-Down	ET	0.864 ± 0.115	0.59	0.0031
Metabolomics	RFE	bthftof-UP	RF	0.842 ± 0.123	0.59	0.0061
Metabolomics	ANOVA	baso-Down	RF	0.849 ± 0.145	0.59	0.0339

PTSD triggers [64] [?]. After obtaining the clinical and endocrine features contributing to improved subgroup classification, we combined the top 15 clinical features with each molecular omics dataset to build overall classification models. Our working hypothesis is that the complementary information provided by molecular data and clinical features can contribute to better classification performance. For comparison, we also built baseline classifiers using only molecular or clinical data (with the same 15 clinical features). We applied a one-sided paired t-test to compare training performance between the mixed-feature model and two baselines. In total, we evaluated eight combined classification models for each dataset and compared performance with eight molecular baseline and eight clinical baseline models. From our results, the majority of combined models outperformed the baselines with a higher AUC in the validation data set; specifically, we saw improvement in 23/24 combined model-dataset pairs compared to molecular data and 18/24 pairs compared to clinical data. Of these improvements, 5 were statistically significantly different from the molecular baseline and 10 were significantly different from the clinical baseline.

In the metabolomics data, all of the combined models showed improved performance relative to the two sets of baseline models (16/16 comparisons total) in the training data, and 14/16 comparisons showed improvement in the validation data. In particular, the ANOVA-based linear SVM model with clinical and molecular features achieved AUC 0.79 $\pm$ 0.13 and ACC 0.722 $\pm$ 0.078, which is significantly better than the performance of the clinical data-only model (p-values: 0.04 and 0.002) or the molecular data-only model (p-values: 0.091 and 0.005). This suggests that the complementation between metabolites and clinical features yields a better distinction between PTSD case and control. Using RFE-based models, the best performance observed is from ET with AUC 0.769 $\pm$ 0.081 and ACC 0.681 $\pm$ 0.088 (training data), along with improved performance in the validation data.

In the miRNA-Deplete dataset, the best ANOVA- and RFE-based models both use ET, with AUC 0.725 ± 0.139 and ACC 0.72 ± 0.097 (ANOVA) and AUC 0.724 ± 0.134 and ACC 0.64 ± 0.154 (RFE). In the miRNA-Exosome dataset, the best models are ANOVA-LR and RFE-ET, with AUC 0.758 ± 0.097 and ACC 0.701 ± 0.116 for the former and AUC 0.766 ± 0.154 and ACC 0.705 ± 0.139 for the latter. The results in the miRNA-Deplete and miRNA-Exosome datasets suggest more consistent performance using ANOVA- rather than RFE-based feature selection, which may be due to overfitting when using RFE with a small sample size. Given these best-performing classification models, we next selected subsets of features as candidate biomarkers for each dataset.

2.3.4 Biomarker Discovery

Given the best-performing combined clinical and molecular models for each dataset (Figure B.5), we selected the most relevant features as candidate biomarkers. To do so, we ranked features by coefficients (LR and SVM) or variable importance (RF and ET). Our result obtained all of the candidate biomarkers, while Table 2.2 shows the top 10 candidate biomarkers from each dataset using ANOVA. We selected the best models from each dataset using each of the two feature selection approaches in an attempt to identify general relevant features rather than model-specific features. Interestingly, the top 10 features from the ANOVA-SVM and RFE-ET combined clinical-metabolomics models have 8 features in common, including the 2 clinical features ProTot and monos.. In the miRNA-Deplete dataset, the top 10 features (6 clinical and 4 molecular) were selected from the ANOVA-ET and RFE-ET models, with 5 of the clinical features (DemoChi, Pulse, ALB, elisa.npy and ProTot) in common between the models. In the miRNA-Exosome data, the top 10 features for the ANOVA-LR and RFE-ET models were molecular, with only one molecule (hsa-miR-200b-3p) in common between the models. These results confirmed our hypothesis about the complementary nature of the information contained in the clinical and molecular data. Our results also suggest that the contribution from clinical

features may be redundant in the miRNA-Exosome dataset (no clinical features were in the top 10), while clinical features appear to play an important role in accurate classification using the miRNA-Deplete dataset.

Table 2.2: Top10 Candidate Biomarkers of the ANOVA Approach

miRNA-Exosome ANOVA	miRNA-Deplete ANOVA	Metabolomics ANOVA
hsa-miR-7-1-5p (M)	DemoChi (C)	5-oxoproline (M)
hsa-miR-200b-3p (M)	ProTot (C)	ProTot (C)
hsa-miR-3613-5p (M)	ALB (C)	lactate (M)
hsa-miR-376c-3p (M)	ggt (C)	monos. (C)
hsa-miR-29c-5p (M)	hsa-miR-185-5p (M)	hypoxanthine (M)
hsa-miR-486-2-5p (M)	ABM10 (C)	docosapentaenoate (n3-DPA; 22:5n3) (M)
hsa-miR-941-1-3p (M)	hsa-miR-3940-3p (M)	tyrosine (M)
hsa-miR-590-5p (M)	elisa.npy (C)	2-hydroxypalmitate (M)
hsa-miR-4454-5p (M)	hsa-miR-1304-5p (M)	glutamine (M)
hsa-miR-502-3p (M)	hsa-miR-574-3p (M)	docosahexaenoate (DHA; 22:6n3) (M)

*C: Clinical features, M: Molecular features

2.4 Discussion

In this study, we built classification models for clinical subgroups, then combined the clinical features leading to significant improvement with molecular features to build predictive models of PTSD status. We then selected important features as candidate diagnostic biomarkers. Specifically, 72 clinical measurements demonstrated improved classification performance through subgroup splitting, among which 21 are significantly associated with PTSD. In the top 10 clinical biomarkers, 8 are both significantly associated with PTSD status in the training data and contribute significantly to subgroup-specific classification improvements. The combination of association with PTSD as well as improvement in PTSD prediction leads to more easily interpretable biomarkers. Knowledge of these biomarkers may also suggest PTSD candidate genes or biological pathways related to PTSD pathogenesis. We note that inconsistencies in performance observed between training and validation datasets may be due to population differences between the cohorts as well as the smaller number of validation samples.

PTSD has the characteristics of hyperarousal and exaggerated startle responses. It was previously reported that PTSD patients have a high risk of developing cardiovascular issues, such as increased heart rate and blood pressure [65] and increased BMI and weight [66] as a result of traumas. Associations of relationship status (being single) and having children with PTSD may suggest that the symptoms lead to (or result from) issues with trust, closeness and communication in relationships and/or with family. Three blood endocrine biomarker–elisa.npy (plasma neuropeptide Y), elisa.cordif (plasma cortisol dexamethasone suppression test 1 - plasma cortisol dexamethasone suppression test 2), and ic50 (Peripheral Blood Mononuclear Cell (PBMC) Lysozyme IC50-DEX)–showed association with PTSD, which demonstrates a consistency with known neuroendocrine alterations resulting from PTSD. PTSD is also associated with enhanced pro-inflammatory cytokines in PBMC [67]. Counts of monocytes, which secrete chemokines during inflammation, are useful biomarkers along with ProTot (total serum protein) in metabolomics data, although they do not appear to be as informative when combined with miRNA data. The lactate, one of our top metabolite biomarkers, has also been reported as a candidate biomarker for PTSD diagnosis [68]. Although our discoveries may be limited by the sample size, to our knowledge this study is the first to use clinical and endocrine features to define PTSD subgroups for improved diagnostic performance.

Data integration enables understanding of complex biological systems in multiple dimensions [69][70]. In our work, we evaluated the classification performance of models integrating clinical and molecular data in both training and validation cohorts. For machine learning approaches, we applied two feature selection techniques and four classifiers in order to find the most suitable model for each PTSD dataset. In clinical subgroup classification, we found that tree-based models yielded the largest number of improved groups in miRNA-Exosome and metabolomics data, while LR yields the most in

miRNA-Deplete data. In an overall classification with clinical and molecular features, we observed good performance using ET, LR or SVM. Despite our success, our best classification performance observed (0.79 AUC) highlights the ongoing challenges in accurately predicting PTSD status.

As survey-based PTSD diagnoses used to determine PTSD status have several important limitations, use of molecular biomarkers have the potential to more objectively predict PTSD status. In this study, we identified candidate biomarkers for each omics dataset comprising both molecular and clinical features. Importantly, the molecular candidate biomarkers enabled accurate PTSD prediction and also provided novel insights for understanding the pathophysiology of PTSD. Since the dataset used in this study is from combat veterans, it is likely that our findings are most relevant for patients with a similar background.

Several other limitations exist in the current study. First, the PTSD Systems Biology Consortium continually improves sample collection and experimental assays used to generate data. The work described here is based on data collected in or before January 2017. A study of the most recent data from the Consortium would require repeating all steps of our analysis on the latest data. Second, the relatively small sample sizes of study cohorts pose challenges in the application of machine learning approaches. The working sample sizes become even smaller when splitting samples into clinical subgroups. We presented our statistical test results using raw p-values. However, given the number of clinical variables used for subgroup generation, a more robust analysis would require controlling for multiple testing by adjusting the original p-values. Future work using datasets with larger sample sizes will include this additional step.

2.5 Conclusions

We applied machine learning approaches to multiple types of PTSD data to explore the diagnostic potential of clinical subgroups and build classification

models consisting of molecular and clinical features to predict PTSD status. We also provided candidate diagnostic biomarkers, knowledge of which improves our understanding of PTSD pathogenesis. We expect that our work will contribute to more precise diagnosis and treatment of PTSD.

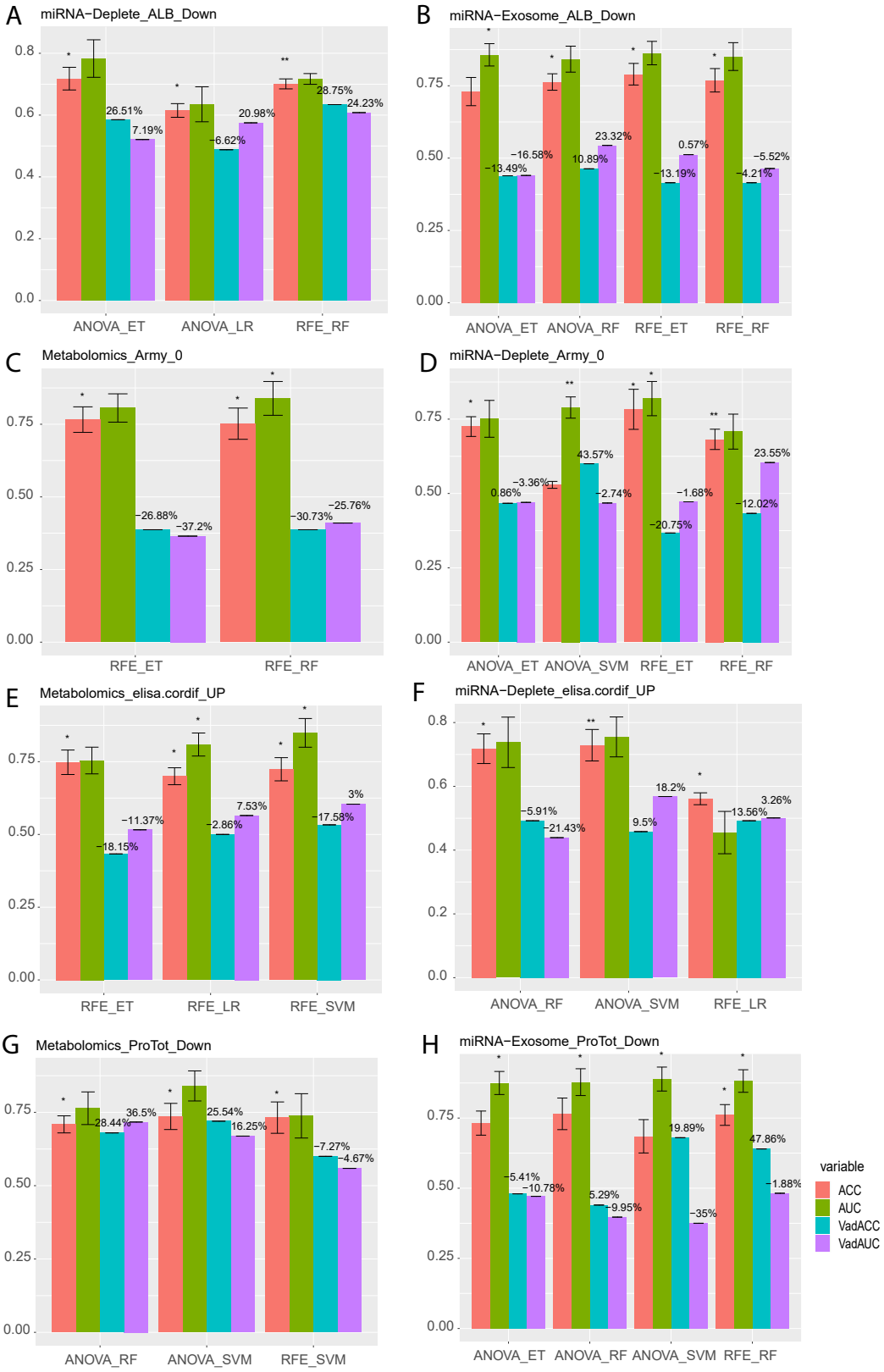


Fig. 2.4: Bar Plots Showing the Performance of Clinical Subgroups, x axis indicates the combined feature selection and classifiers, y axis indicates the evaluation metrics of AUC and ACC, red and green bars presenting AUC and ACC in training dataset while blue and pink presenting AUC and ACC in validation dataset, asteroids used to show statistical the significance and the percentages to show the increase compared to the baseline. A-B indicate `ALBDown` group, C-D indicate `Army0`, E-F indicate `elisa.cordifUP` and G-H indicate `ProTotDown`.

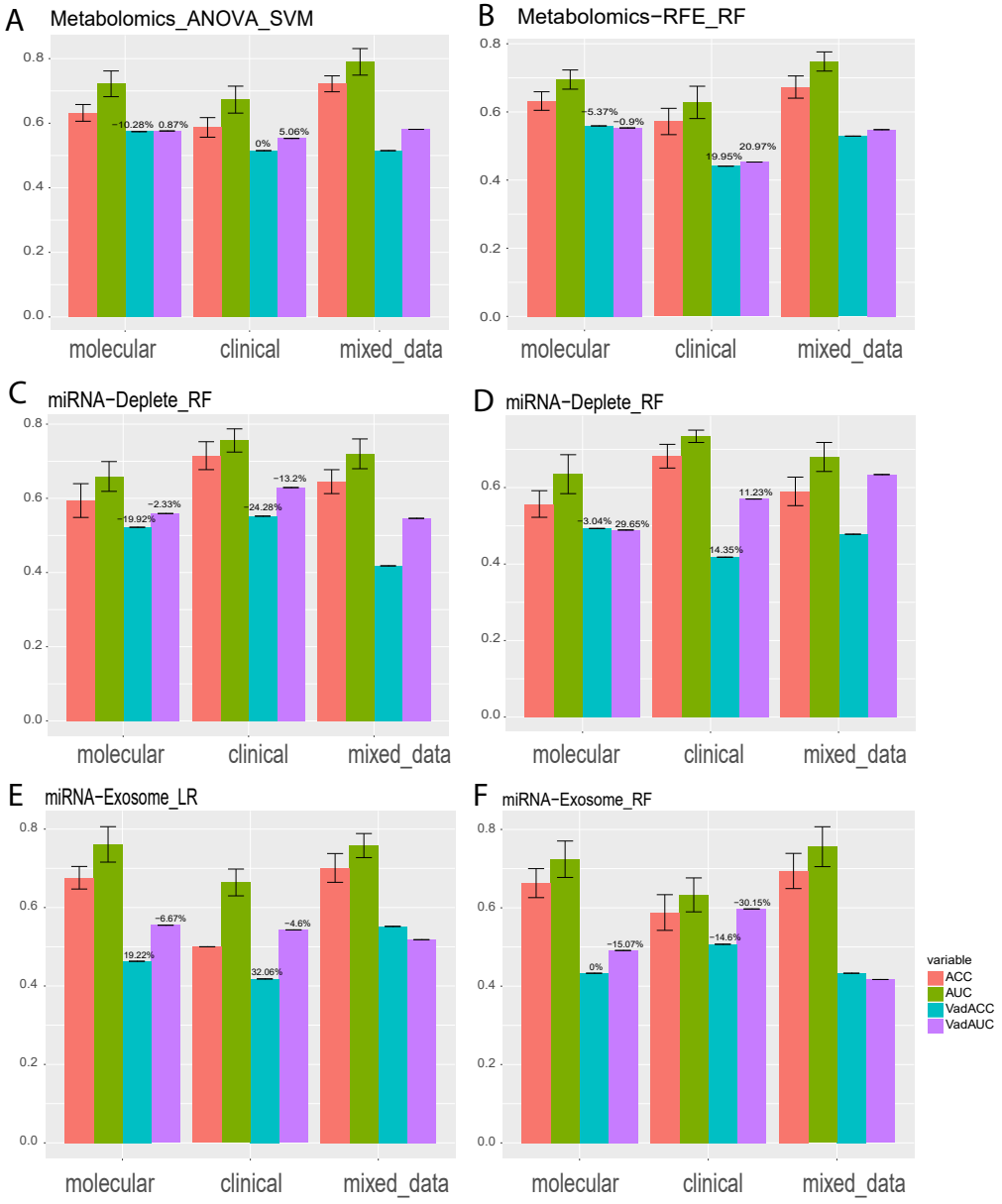


Fig. 2.5: Bar Plots of Overall Classification Models Including Molecular and Clinical Features, x axis indicates data source molecular data, clinical data and combined data while y axis indicates the evaluation metrics of AUC and ACC, the percentages indicate the performance in the combined data compared to the baseline in molecular or clinical data, negative means poor improvement while positive means better improvement in combined model.

Chapter 3

Multi-omic data integration to discover subgroups of PTSD

3.1 Abstract

Post-Traumatic Stress Disorder (PTSD) is a psychiatric disorder caused by environmental and genetic factors resulting from alterations to gene expression, DNA methylation, and neuroimaging characteristics. There is a pressing need to identify reliable molecular and physiological biomarkers for accurate diagnosis, prognosis, and treatment of PTSD, as well as to deepen our understanding of its pathophysiology. Multiple omic data integration enables investigation of the complex, multivariate nature of the biological systems underlying PTSD and is essential for identifying molecular subgroups of the disorder.

Given 284 total samples (124 PTSD positives) from four omics data sets (miRNA enriched in exosomes, miRNA depleted for exosomes, total miRNA, and Metabolomics) in cohorts of Training, Validation and Recall, we used two methods–Similarity Network Fusion (SNF) and Variational Autoencoder (VAE)–to integrate the data sets and identify subgroups. SNF performs integration by efficiently fusing sample similarity matrices from each data set into one network representing the full spectrum of underlying data. Spectral clustering can then be used to identify subgroups from this network. The VAE method uses a symmetric deep neural network to reconstruct multiple omics input data sets by estimating data distributions and identifying representative hidden variables. K-means clustering is then used to identify subgroups from the lower dimensional hidden variables. In order to interpret the subgroups, we tested the associations between identified subgroups and clinical characteristics. We also calculated differentially expressed molecules between subgroups in each omics dataset. We then built supervised classification models for PTSD diagnosis with/without subgroups and compared the accuracy of predicting PTSD status in the context of subgroups to the accuracy of predicting PTSD

status without any knowledge of subgroups. Finally, we built a classification model to predict subgroups in PTSD positive samples.

Our results suggest the presence of two PTSD subgroups in 82 training PTSD positive samples, both SNF- and VAE-based methods. These subgroups show significant differences in recalled sample PTSD status (p-value 0.0213). We also found that a majority of samples associated with the same subgroups when comparing results from the two methods. Upon statistical testing for association of the subgroups with over 600 clinical features, we found a significant association with features including heart rate and insulin. The two identified subgroups also exhibit a number of differentially expressed molecules from each omics data set. For diagnostic classification, we observed improved performance for subgroup-aware PTSD status prediction in total miRNA and Metabolomics data sets using SNF-based subgroups and in all four omics data sets using VAE-based subgroups. Finally, using our classification model for subgroup prediction, we found that identified subgroups in the validation cohort were significantly associated with many of the same clinical features associated with subgroups from the training cohort.

We integrated four omics datasets and discovered two clinically-plausible PTSD subgroups. These subgroups showed significant association with clinical features and a collection of differentially expressed molecules between them. Supervised classification using a subgroup-aware classifier suggested improved PTSD diagnostic potential. Future work will involve leveraging knowledge of these subgroups to enable precision medicine for PTSD.

3.2 Introduction

Single "omics" data sets have helped explain diagnosis and progression for complex disorders, but the information contained is limited to one modality. As different layers of biological systems are often relevant and interdependent, multiple omics data integration using mRNA expression, miRNA expression, protein, DNA methylation and metabolomics can utilize complementary

information and hidden coherent biological signatures to discover biomarkers for diagnosis, progression and treatment in human diseases [71] [72]. Several studies have summarized the variety of approaches, challenges and the potential benefits of using data integration to understand biological systems [70] [73] [74] [75]. Of these approaches, use of deep neural networks and deep learning is particularly promising.

Deep learning, a class of machine learning techniques, utilizes a cascade of multi-layered artificial neural networks for automatic feature extraction and representation learning. Deep learning architectures such as autoencoders, recurrent neural networks and convolutional neural networks have been applied to fields in computer vision, natural language processing and biomedical data science [52] [53]. Autoencoders (AEs), a type of deep learning model consisting of an encoder and decoder, learns data reconstruction and efficient representative features in an unsupervised manner. Specific variants of AEs include Denoising Autoencoders [76], Adversarial Autoencoders [77], and Variational Autoencoders. AEs have been successfully applied in biomedical research to solve tasks such as subgroup discovery in liver cancer [78], neuroblastoma cancer subtype discovery [79], unsupervised cancer detection using Adversarial AEs [80] and feature construction and knowledge extraction using a Denoising AE [81]. A number of studies have applied AEs for multiple omics data integration to enable neuroblastoma clinical endpoint prediction [82], high-risk neuroblastoma subtype prediction [79], risk stratification of bladder cancer [83], liver cancer survival prediction [84] and evaluation of colorectal cancer subtypes and cell lines [85]. However, few studies have made use of multiple omics data integration to discover subgroups of PTSD.

In this study, we integrated four omics data sets from 82 PTSD positive samples using two different approaches–Similarity Network Fusion (SNF) and Variational Autoencoder (VAE). Using these approaches, we identified subgroups of PTSD which showed significant differences in recalled sample PTSD status. In

order to interpret the subgroups biologically, we performed differential expression and clinical characteristic association analyses. We also built a subgroup-aware PTSD diagnostic model and a PTSD subgroup prediction model which showed good performance in multiple sample cohorts. These models and our associated findings regarding PTSD subgroups should contribute to a better understanding of PTSD pathogenesis and improved clinical applications for PTSD patient stratification.

3.3 Materials & Methods

3.3.1 Study Samples

The DoD-funded Systems Biology of PTSD Consortium has recruited over 200 male combat veterans with and without PTSD for the purposes of identifying diagnostic biomarkers. For subject recruitment, PTSD-positive and PTSD-negative participants were selected using the following criteria: 1) male veteran between 20 and 60 years old, 2) deployment in Operation Enduring Freedom and/or Operation Iraqi Freedom, 3) PTSD positive participants with at least 40 CAPS score, 4) PTSD negative participants with less than 20 CAPS score. For consistency, all study participants were evaluated using the DSM-IV PTSD assessment upon recruitment. For research purposes, there are three cohorts: a Training cohort also called the discovery cohort (82 PTSD positive cases and 82 PTSD negative controls), a Test cohort also called the validation cohort (28 positive and 39 negative samples) and a Recall cohort (14 positive, 10 subthreshold positive and 29 negative samples). In the recall cohort, a subset of samples from the training cohort were recalled and reassessed an average of three years later.

For each of the recruited subjects, blood and urine samples were taken and used to isolate DNA, RNA, protein, metabolites, and endocrine markers for downstream analyses. Physiological measures (e.g., pulse, blood pressure, body mass index) were also collected. Molecular and physiological data from the Training cohort were initially designated for hypothesis generation regarding

potential diagnostic biomarkers and underlying biological mechanisms of PTSD, while data from the Test cohort were designated for hypothesis testing and attempted replication of the training sample findings. For this study, we analyzed four molecular datasets from these cohorts–three measuring miRNA derived from either total plasma (miRNA-Plasma), plasma enriched in exosomes (miRNA-Exosome), or plasma depleted in exosomes (miRNA-Deplete), and one measuring metabolomics–with the goal of identifying and clinically characterizing multi-omic PTSD subgroups.

In addition to molecular features, 613 clinical and physiological features (hereafter referred to as clinical features) belonging to four categories including demographic information, PTSD diagnostic assessment scores, biochemical and anthropometric body measurements, and endocrine measurements from blood and urine, were collected for each subject in the three cohorts. In our analysis, we first removed clinical features with missing values in the training or validation cohorts, which left 602 features remaining. We then assigned each clinical feature to one of three categories based on which of the following criteria were satisfied for both training and validation cohorts: 1) "Binary" if two unique values exist, 2) "Categorical" if at most five unique values exist, 3) "Continuous" if more than five unique values exist. In total, we identified 101 binary, 427 categorical and 74 continuous features. To perform statistical association analyses, we used Fisher's Exact Test for binary and categorical features and t-test for continuous features. We adjusted p-values for multiple testing using the Benjamini-Hochberg (BH) [86] method and considered any feature with adjusted $p \leq 0.05$ as significantly associated.

3.3.2 Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. We applied PCA on the concatenation of all four omics

data sets from the 82 PTSD positive samples of the Training cohort and used the maximum number (82) of extracted components to represent the input data. We then used K-means clustering as described below to discover the optimal PTSD subgroups.

3.3.3 Similarity Network Fusion Data Integration

Similarity Network Fusion (SNF) is a computational approach for performing data integration of multiple omics data sets [87]. SNF first calculates a sample similarity matrix individually for each data set, which may include mRNA expression, DNA methylation and miRNA expression, among others. Next, SNF iteratively integrates these similarity matrices into an overall sample similarity matrix using graph fusion. This approach helps reduce data set-specific noise and bias to capture complementary information from omics datasets of different modalities. We used the R library "snftools" to integrate our four study datasets–miRNA-Exosome, miRNA-Deplete, miRNA-Plasma and metabolomics–from the PTSD positive samples in the Training cohort. We then discovered optimal subgroups from the overall sample similarity matrix using the spectral clustering approach implemented in the snftools library and described below.

3.3.4 Variational Autoencoder Model

As described above, an Autoencoder (AE) is one class of deep learning architectures used to learn input data representations in an unsupervised manner. The purposes of this representation include dimensionality reduction and reconstruction of the input with the removal of noise. An AE is constituted by two main parts: an encoder that maps the input into a code, and a decoder that maps the code to a reconstruction of the original input. The encoder and decoder are symmetric in terms of layer structure. The simplest AE model is one hidden layer which refers to the input layer, one hidden layer in the encoder, the symmetric hidden layer in the decoder and the output layer. The output of the last encoder hidden layer is also called latent space, which usually has reduced

dimensions of features compared to the input and the features are also called latent variables or latent vectors. The output of the decoder is the reconstruction which is trained to be close to the input. A Variational Autoencoder (VAE) is an AE variant which inherits the general autoencoder architecture of both an encoder and a decoder and is trained to reduce the reconstruction error between input and output. Moreover, VAEs learn the distribution of samples by estimating a mean and standard deviation vectors which are used to sample a latent space to be fed to the decoder. VAEs also add a Kullback-Leibler divergence term to the reconstruction loss which measures the difference between a standard Gaussian and the estimated distribution.

In our work, we implemented a one-hidden layer VAE to reconstruct the input data, which in this case is the concatenation of four omics data sets from the 82 PTSD positive samples of the Training cohort. For model training, we tuned hyperparameters including learning rate (ranging from 0.0001 to 0.1) and dropout rate (ranging from 0.1 to 0.5). For each hyperparameter combination, we trained the corresponding VAE for 500 epochs, at which point the training process has converged. We then selected nodes from the hidden layer as a reduced dimensional representation of the input data. Using this representation, we applied K-means clustering as described below to select the optimal PTSD subgroups.

3.3.5 Unsupervised Clustering

Spectral clustering [50] [88] is a technique that uses the eigenvectors (spectrum) and eigenvalues of a matrix to define cluster membership.

This approach is based on the fact that if a graph (network defined by sample similarity matrix) is formed by k disjoint cliques (clusters), then the samples are projected into a lower dimensional space to have more obvious clusters where Graph Laplacian Matrix, a matrix representation of a graph, is used to calculate eigenvalues and eigenvectors. The eigenvectors of the sample similarity matrix function as indicators of cluster membership. The eigengap

refers to the difference between consecutive eigenvalues. Importantly, although small perturbations such as adding a few edges linking clusters or removing edges from inside the clusters will increase eigenvalues and change the corresponding eigenvectors, this does not generally cause the underlying structure to be lost. This clustering technique requires the number of desired clusters to be specified [51]. We applied spectral clustering to the integrated sample similarity matrix from SNF and calculated the eigengaps for numbers of clusters ranging from 2 to 10. We then selected the optimal number of subgroups based on the maximum eigengap value.

K-means is another method of unsupervised clustering, originating from signal processing, that is popular for cluster analysis in data mining. For a given value of K (number of clusters), the goal of K-means is to partition all training data samples into K distinct, non-overlapping clusters by identifying clusters for which within-cluster variation is as small as possible. For our work, we used the K-means implementation from the Python Scikit-learn library. As with spectral clustering, we evaluated numbers of clusters ranging from 2 to 10 and calculated the Silhouette score to determine the optimal number. The Silhouette score indicates the separation distance between a given set of clusters, with a large score (close to 1) corresponding to good separation and a small score (close to 0) corresponding to poor separation. We thus chose the optimal number of subgroups based on the maximum Silhouette score.

3.3.6 Subgroup Recall Status Test

Among 59 total samples in the recall cohort, the 29 PTSD cases were recalled from PTSD positive samples in the training cohort. Upon reassessment on average three years later, these 29 samples were diagnosed as either "Positive", "Negative" or "Positive Subthreshold." For the identified PTSD subgroups, we compared recalled PTSD status between the groups and used Fisher's Exact test to test for differences. We used a p-value threshold of 0.05 to determine if the subgroups showed significantly different recall statuses.

3.3.7 Differential Expression Analysis

For the PTSD subgroups, we also identified differentially expressed (DE) molecules between groups in each of the four omics data sets. In addition, we identified DE molecules between PTSD positive and PTSD negative samples in the Training and Validation cohorts. We used the R package "Limma" for DE analysis and applied a Benjamini-Hochberg adjusted p-value threshold of 0.05 to determine significantly DE molecules. For the three group (Control vs Subgroup 1 vs Subgroup 2) test, we compared each pair, then selected intersected significantly DE molecules.

3.3.8 Supervised Diagnosis Classification

Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms used for classification and regression analyses. SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space. An important feature of SVMs is the use of kernel functions that enable operation in a high-dimensional, implicit feature space without requiring computation of the coordinates of data in that space. Kernels provide a mapping of the problem from the input space to this higher-dimensional space (called the feature space) by performing a nonlinear transformation. We first used SVM with a linear kernel to train diagnostic classifiers based on PTSD positive and negative labels for each of the four omics data sets in the Training cohort. We then validated these classifiers using data from the Validation cohort. We then used SVM to train a PTSD subgroup-aware classifier based on labels of the identified subgroups as well as PTSD negative labels. To assess accuracy, we considered any sample classified into one of the predicted subgroups as a PTSD positive sample and PTSD negative otherwise. During classifier training, we applied recursive feature elimination and cross-validation for feature selection and accuracy evaluation, respectively. Specifically for the latter, we used 10-fold nested cross-validation to perform training and testing as well as optimize classifier hyperparameters. For the linear

SVM, we evaluated values for the cost hyperparameter (C) in a series of $10^{(}-4,3)$ and $2^{(}-4,4)$. We selected the optimal hyperparameter based on the highest average accuracy achieved in the training data and used the corresponding trained classifier for subsequent evaluation.

3.3.9 Subgroup Prediction

We also used a linear SVM to construct a PTSD subgroup prediction model. Here, the training labels are the predicted subgroups from the Training cohort, and we use the resulting classifier to predict subgroups for the Validation cohort. We used AUC as the evaluation metric for predictions based on the training data, and we applied the same hyperparameter tuning and cross-validation schemes as above. The best-performing hyperparameter was selected based on the average AUC from cross-validation, and the corresponding trained classifier was used to predict subgroup labels in the validation data. If all subgroup predictions for the testing data set happen to be the same, we instead chose the trained classifier using the next-best hyperparameter value as determined by average AUC.

3.4 Results

In this work, we aimed to integrate multiple omics data sets to discover PTSD subgroups and their associated clinical characteristics. Specifically, we integrated three miRNA data sets and 1 metabolomics data set, and we applied three different approaches–SNF, VAE and PCA–to perform data integration and unsupervised subgroup discovery. We used the criteria of cluster separation and differences in recall sample PTSD status to guide hyperparameter tuning and select the optimal PTSD subgroups. For the subgroups identified using SNF and VAE, we performed the following downstream analyses: 1) Differential expression analysis between subgroups, 2) Subgroup clinical association analysis, 3) Subgroup-aware diagnostic model creation and evaluation, 4) Subgroup prediction model construction and evaluation. Figure 3.1 shows our overall workflow.

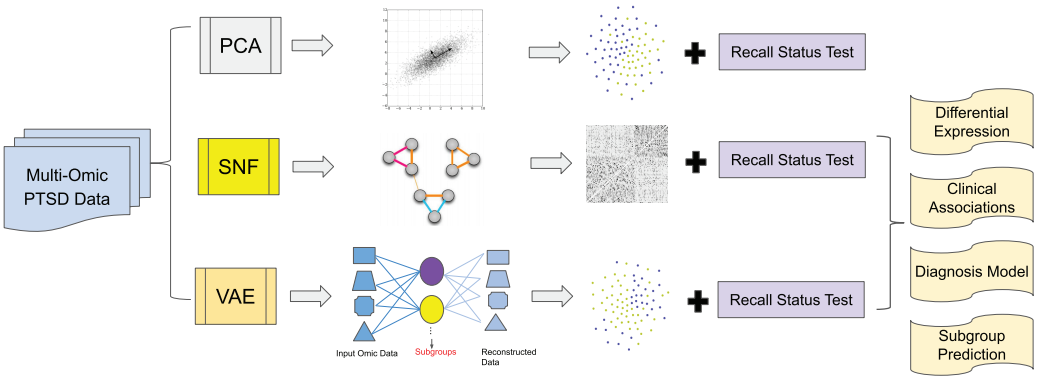


Fig. 3.1: Workflow for PTSD subgroup identification and associated clinical characterization. From the left side, we utilized PTSD multiple omic data sets and applied methods PCA, SNF and VAE to discover subgroups in parallel as seen in the middle. For the identified PSTD subgroups, we analyzed the status change for the recalled samples, followed by differential expression, clinical association and diagnosis model. We also built classification models to predict PTSD subgroups.

3.4.1 PTSD Subgroup Identification

As described above, we utilized four omics data sets for our analyses–miRNA-Exosome, miRNA-Deplete, miRNA-Plasma and metabolomics–from three patient cohorts–Training (discovery), Test (validation) and Recall. Across all cohorts, miRNA-Exosome contains 209 molecular features, miRNA-Deplete contains 310 features, miRNA-Plasma contains 284 features and metabolomics contains 168 features. The counts of samples and features can be seen in Table 3.1.

Table 3.1: PTSD Multiple Omic Data Sets and Cohorts

Data Sets	Feature Counts	Cohorts	Samples P:(SP):N
miRNA-Exosome	209	Training	82:82
miRNA-Deplete	310	Validation	28:39
miRNA-Plasma	284	Recall	14:10:29
Metabolomics	168	All	124:10:150
All	971		

*P: PTSD positive; SP: Subthreshold PTSD positive; N: PTSD negative

We first removed a known batch effect for location of sample recruitment from the four omics datasets using the function "removeBatchEffect" from the R Limma package. We performed this batch correction on data from all cohorts. We also removed one feature (EDTA) from the metabolomics data set, as

changes in its levels were known to reflect this batch effect. The above procedures left us with a concatenated training data set of 82 PTSD positive samples measured for 971 total molecular features. We applied data integration and unsupervised clustering to this dataset to discover PTSD subgroups.

As described above, Similarity Network Fusion (SNF) performs omics data integration through the construction and fusion of sample similarity matrices from each data set [87]. To begin, we computed normalized pairwise squared euclidean distances for each data set and then applied SNF to integrate the data sets and obtain an overall 82 x 82 sample similarity matrix. We next used spectral clustering to determine the optimal number of clusters within the matrix. We selected the optimal number of subgroups based on the maximum value of the eigengap, which reflects the degree of separation between clusters. We then tuned SNF hyperparameters and tested for significant associations between candidate subgroups and recall PTSD status. We identified an optimal 2 subgroups, consisting of 48 and 34 samples, respectively, with an eigengap of 0.166 and recall status p-value of 0.0213. We visualized a lower-dimensional representation of the subgroup assignments using the t-SNE approach. Figure 3.2 A-C displays the numbers of subgroups versus their eigengaps and recall status p-values, the overall sample similarity matrix and the t-SNE visualization.

Variational autoencoders (VAEs) provide an alternative means for data integration by performing reconstruction and dimensional reduction of input data. We applied a VAE model to the four PTSD omics datasets as an 82 sample x 970 feature matrix, and we used the 28 sample x 970 matrix from the validation cohort for model evaluation. We tuned a series of hyperparameters to optimize the VAE model for our data set. We then applied K-means clustering to the reduced-dimensional hidden variables of the VAE to discover subgroups. Similarly as with SNF, we tested for significant associations of recall status with candidate subgroups. We used four criteria to choose the optimal number of clusters: 1) model convergence with a low validation loss, 2) $>$0.1 Silhouette

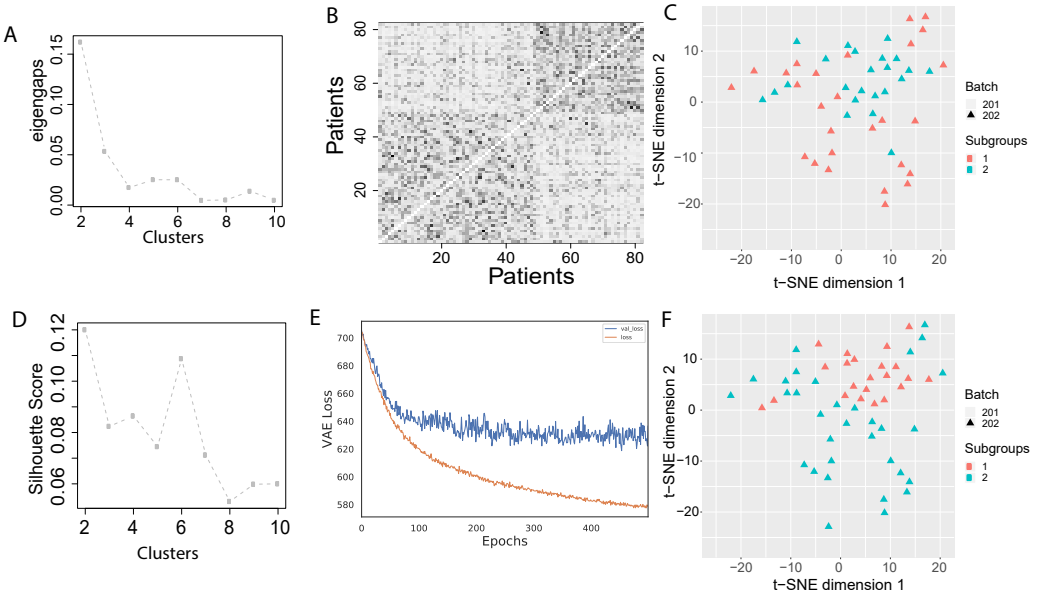


Fig. 3.2: SNF and VAE Subgroup Identification, A) Eigen gaps in SNF and spectral clustering, B) Patient-patient similarity matrix with subgroup blocks, C) SNF-derived subgroup visualization in training PTSD+ samples, D) VAE training process in 500 epochs, E) K-means clustering in VAE hidden variables, F) VAE-derived subgroup visualization in training PTSD+ samples.

score from K-means clusters, 3) fewer than 5 subgroups, 4) an association test between candidate subgroups and recall PTSD status. We then visualized the subgroups as before using t-SNE. We again identified an optimal 2 subgroups, consisting of 50 and 32 samples, respectively, with a Silhouette score of 0.12 and recall status p-value of 0.0213. Figure 3.2 D-F) show the numbers of subgroups versus their Silhouette scores and recall status p-values, the training and validation losses as training progresses, and the t-SNE subgroup visualization.

We also evaluated PCA as a simpler data integration method for comparison with SNF and VAE. We selected all 82 principal components for data integration, which represent 100% of the variance in the input data. Similar to the VAE approach above, we applied K-means clustering on the reduced-dimensional set of principal components and selected optimal clusters based on Silhouette score. We identified an optimal 2 subgroups, consisting of 38 and 44 samples, respectively, with a Silhouette score of 0.142 and recall status p-value of 0.142. Unlike with the SNF and VAE subgroups, the PCA subgroups

did not demonstrate a significant association with recall status, so we did not perform downstream analyses of these groups. Table 3.2 summarizes results of the three approaches used for subgroup identification.

Table 3.2: Subgroup Identification and Fisher's Exact Test on Recall Status

Methods	Subgroups	Separation	Batch Effect Test p value	Recall Status Test p value
SNF	48 – 34	0.166	0.2456	0.0213
PCA	38 – 44	0.142	0.8157	0.1423
VAE	50 – 32	0.12	0.8119	0.0213

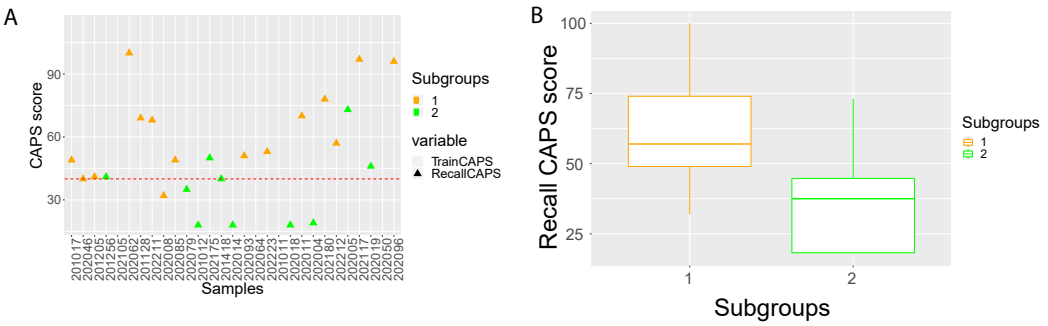
Given the SNF and VAE subgroups, we next compared group membership between the two methods to detect subgroup overlap. A majority of the 82 samples were placed in the same subgroups by the two methods, with the exception of 14 samples that had switched membership (Table 3.3). When considering only the subset of recalled samples, we found that all 29 samples were placed in the same subgroups by both methods. Furthermore, recalled samples in subgroup 2 have lower CAPS scores upon recall than those in subgroup 1(t-test p-value 0.02). Figure 3.3-A indicates the change in CAPS for the recalled samples between Training and Recall cohorts, and Figure 3.3-B summarizes the difference in recall CAPS scores for the two subgroups. We used t test and found p value 0.028 for the CAPS score difference between the two subgroups in the Recall status.

Table 3.3: Overlap of Subgroups Identified Using SNF and VAE

SNF	VAE	Counts
1	1	42
1	2	6
2	1	8
2	2	26

3.4.2 Clinical Characterization Of Subgroups

To enable biological interpretation of the PTSD subgroups, we performed association analysis of 602 clinical characteristics, among which 101 are binary, 427 are categorical and 74 are continuous. We applied Fisher's Exact tests on binary or categorical features and t-tests on continuous features. We then



C

Subgroups\ Recall Status	Positive	Negative	Subthreshold
1	12	0	5
2	4	4	4

Fig. 3.3: Subgroup Recall Status Change, A) CAPS score change in SNF-derived training PTSD+ subgroups, B) CAPS score change in Recall samples using SNF-derived subgroups, C) Confusion matrix between SNF-derived subgroups and Recall status changes.

performed multiple test adjustment of the original p-values using the Benjamini-Hochberg (BH) method. We performed this analysis separately for the subgroups identified by the SNF and VAE methods. For comparison, we also tested for clinical associations with case and control status in the Training and Validation cohorts. In the latter comparison, 332 clinical features were significantly associated with case-control in both cohorts. Among those features, 62 were significantly associated with SNF-based subgroups while 60 were significantly associated with VAE-based subgroups. Taken together, 59 clinical features from six categories were significantly associated with subgroups detected by both methods. These 59 clinical features were from 5 categories and the members from the same category look quite similar. For the further analyses, we chose one member from each category, but two from the ABM category in the total of six clinical features. Figure 4 summarizes the values of six clinical features–ABM10 (Body Mass Index), ABM2 (Pulse), `FASA_C`, insulin, psai1a, SAS23–from five categories between the subgroups and control samples. Figure 3.4A displays values for SNF-based subgroups, while Figure 3.4B illustrates VAE-based subgroups. We note that ABM10 and ABM2 show significantly

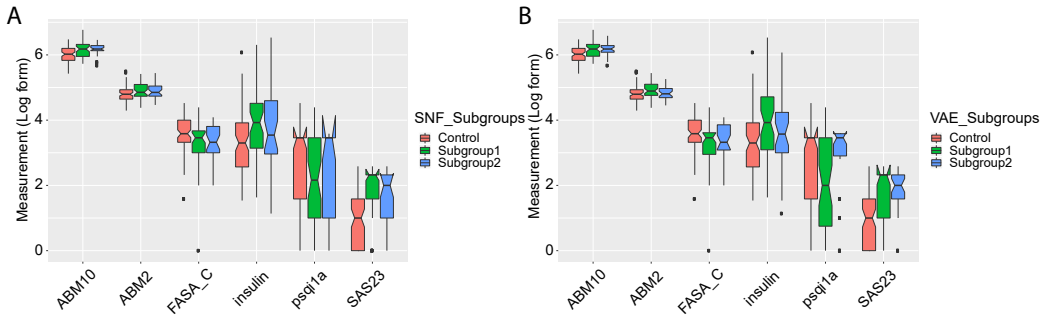


Fig. 3.4: Clinical Association with Predicted PTSD Subgroups in Training Data Set. A) six selected clinical features expression in PTSD- and SNF-based identified PTSD subgroups, B) six selected clinical features expression in PTSD- and VAE-based identified PTSD subgroups.

higher values in cases than controls, while `FASA_C` exhibits the opposite trend. Table 3.4 lists a detailed result of the association analysis for the top 30 clinical features. The full names for clinical categories are provided in the Appendix C. The FDR indicates the False Discovery Rate on the original p values.

3.4.3 Differential Expression Between Subgroups

To better understand differences in molecular expression between subgroups, we applied differential expression (DE) analysis. We used the R package "Limma" to test for expression differences across the four omics data sets between (1) case and control samples, (2) SNF-based subgroups, (3) VAE-based subgroups, and (4,5) three-way comparisons for both SNF and VAE (control vs subgroup 1 vs subgroup 2). We considered significantly differentially expressed molecules as those with BH-adjusted p-values ≤ 0.05. We detected the largest numbers of DE molecules in the miRNA-Deplete and miRNA-Exosome data sets from the SNF and VAE subgroup comparisons. In the miRNA-Plasma data set, the highest overall number (247) of DE molecules were detected between cases and controls, although DE molecules from the subgroup comparisons (206 from SNF and 211 from VAE) were a close second. Figure 3.5A and Figure 3.5C illustrate these findings. Figure 3.5B and Figure 3.5D summarize the expression of the top five most significantly DE molecules (all miRNAs) from the

Table 3.4: Clinical Association Test With the Identified PTSD Subgroups

Features	Categories	SNF-Subgroups FDR	VAE-Subgroups FDR	Train Case-Control FDR	Validation Case-Control FDR
ABM9-Syst	NYU-ABM	8.25E-83	8.20E-83	1.96E-169	2.28E-50
ABM9-Dias	NYU-ABM	4.22E-76	4.17E-76	8.63E-154	2.78E-47
WVisWorkMemConf2	NYU-WAIS-WMS	8.39E-69	8.28E-69	8.83E-144	6.56E-45
ABM10	NYU-ABM	1.65E-67	1.62E-67	9.90E-127	2.30E-41
WVisWorkMemConf1	NYU-WAIS-WMS	5.30E-64	5.22E-64	4.48E-134	1.32E-41
WVisWorkIndex	NYU-WAIS-WMS	1.13E-63	1.11E-63	8.12E-134	9.46E-42
ABM2	NYU-ABM	2.20E-61	2.12E-61	4.01E-126	3.94E-32
WMemLetterRS	NYU-WAIS-WMS	2.60E-59	2.49E-59	8.30E-123	3.48E-43
ABM1	NYU-ABM	6.92E-59	6.86E-59	3.19E-119	9.36E-36
ABM6	NYU-ABM	1.28E-57	1.24E-57	1.41E-124	8.85E-47
WMemDigRS	NYU-WAIS-WMS	2.59E-57	2.49E-57	6.10E-115	8.66E-32
WLongestForwRS	NYU-WAIS-WMS	3.23E-54	2.55E-54	1.77E-128	7.88E-34
WLongDSSRS	NYU-WAIS-WMS	3.48E-54	2.34E-54	3.67E-112	9.01E-34
ABM4	NYU-ABM	2.21E-53	2.12E-53	6.79E-116	6.46E-40
WDigSpanForwardRS	NYU-WAIS-WMS	5.21E-51	4.60E-51	8.88E-109	3.82E-29
DemoAge	NYU-Background	9.60E-50	9.13E-50	1.42E-101	2.15E-36
WProcSpeedCodRS	NYU-WAIS-WMS	2.80E-48	2.72E-48	1.07E-104	8.49E-35
WMemLetterAgeSS	NYU-WAIS-WMS	3.75E-46	3.17E-46	9.74E-56	2.58E-30
WVisWorkMemSum	NYU-WAIS-WMS	1.75E-44	1.60E-44	2.22E-98	2.70E-30
WVocRS	NYU-WAIS-WMS	9.39E-44	8.94E-44	3.31E-96	4.96E-43
WDigSpanBackRS	NYU-WAIS-WMS	2.77E-42	2.22E-42	1.50E-94	7.20E-25
WMemDigAgeSS	NYU-WAIS-WMS	5.08E-42	4.22E-42	3.11E-90	1.24E-25
WSymbolSpan-RS	NYU-WAIS-WMS	3.82E-38	3.54E-38	4.36E-90	1.45E-26
FASF-C	NYU-FAS	1.64E-37	1.42E-37	9.56E-85	7.67E-24
WProcSpeedCodAgeSS	NYU-WAIS-WMS	2.89E-37	2.33E-37	4.18E-88	1.31E-28
WSymbolSpan-SS	NYU-WAIS-WMS	3.44E-37	2.82E-37	1.62E-91	3.47E-28
FASS-C	NYU-FAS	4.62E-36	4.08E-36	1.67E-85	3.09E-23
WSpatialAdd-SS	NYU-WAIS-WMS	5.13E-36	4.21E-36	1.07E-84	1.07E-27
WLongestBackRS	NYU-WAIS-WMS	1.17E-35	6.93E-36	1.05E-92	1.89E-25
WSpatialAdd-RS	NYU-WAIS-WMS	2.01E-35	1.76E-35	5.87E-81	1.30E-26

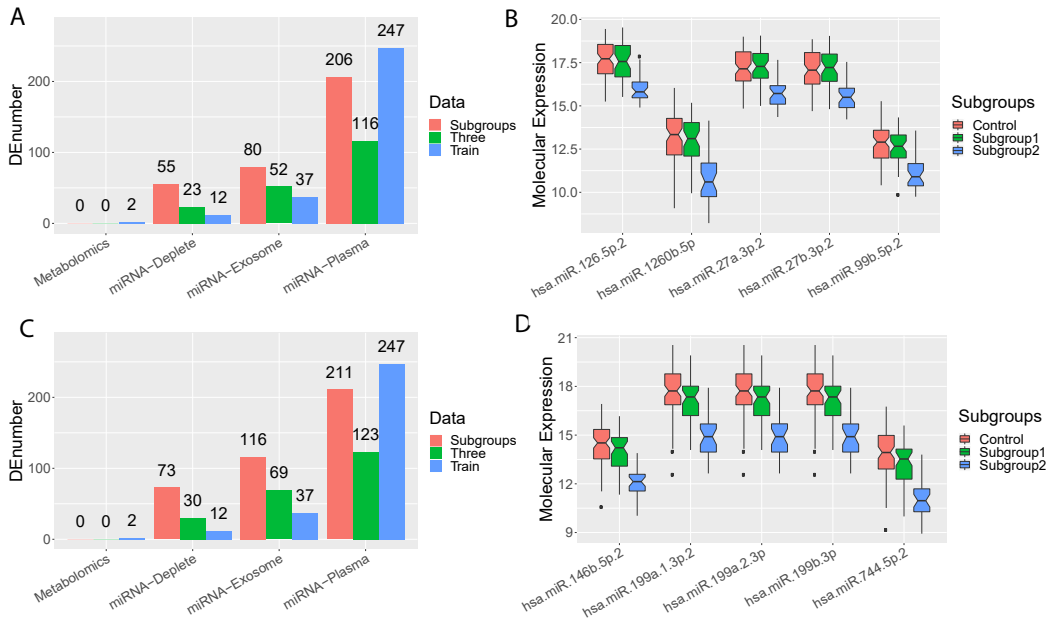


Fig. 3.5: Molecular Differential Expression With Identified Subgroups, A) Barplot of differentially expressed molecule numbers in SNF-based subgroups, B) Top DE molecules in SNF-based subgroups, C) Barplot of differentially expressed molecule numbers in VAE-based subgroups, D) Top DE molecules in VAE-based subgroups.

three-way comparison for SNF and VAE subgroups, respectively. More detailed lists of DE molecules show in Table 3.5 of SNF and Table 3.6 of VAE.

3.4.4 Supervised Classification Using Subgroup Labels

In order to detect whether knowledge of the discovered subgroups was relevant to PTSD diagnosis, we next constructed a supervised classification model to discriminate between case and control samples within each PTSD omic data set. Our working hypothesis is that a subgroup-aware classifier can glean additional information from data that will be useful for diagnosis. We trained an SVM classifier to distinguish between three classes (control, subgroup 1 and subgroup 2) in the Training cohort, and we evaluated classification performance using prediction accuracy in the Validation cohort. Given that subgroups were not yet defined in the latter cohort, we considered predictions for subgroups 1 or 2 equivalent to PTSD positive predictions when calculating Validation accuracy. For comparison, we also trained a baseline two-class classifier (case versus

Table 3.5: Top DE Molecules Between SNF-based Subgroups

Molecules	DataSets	Train FDR	Subgroup FDR	Three FDR
hsa.miR.1260b.5p	miRNA-Plasma	0.00023	2.81E-09	8.19E -14
hsa.miR.27b.3p.2	miRNA-Plasma	0.002	2.81E-09	1.58E -12
hsa.miR.27a.3p.2	miRNA-Plasma	0.003	2.81E-09	2.69E -12
hsa.miR.99b.5p.2	miRNA-Plasma	1.83E-05	4.09E-09	2.44E -13
hsa.miR.126.5p.2	miRNA-Plasma	0.0003	4.09E-09	1.25E -12
hsa.miR.126.3p.2	miRNA-Plasma	4.36E-06	7.46E-09	6.14E -15
hsa.miR.125a.5p.2	miRNA-Plasma	1.82E-05	7.46E-09	8.87E -13
hsa.miR.223.3p.2	miRNA-Plasma	2.75E-05	7.46E-09	1.64E -13
hsa.miR.335.5p.2	miRNA-Plasma	0.00048	8.13E-09	4.45E -12
hsa.miR.223.3p.1	miRNA-Exosome	0.0026	1.10E-08	9.01E -13
hsa.miR.199a.1.3p.2	miRNA-Plasma	1.15E-05	1.10E-08	2.55E -14
hsa.miR.199a.2.3p	miRNA-Plasma	1.15E-05	1.10E-08	2.54E -14
hsa.miR.199b.3p	miRNA-Plasma	1.15E-05	1.10E-08	2.55E -14
hsa.miR.146b.5p.2	miRNA-Plasma	1.43E-05	1.10E-08	3.65E -13
hsa.miR.494.3p.2	miRNA-Plasma	1.07E-05	1.65E-08	6.69E -14
hsa.miR.361.5p.2	miRNA-Plasma	0.00039	1.65E-08	2.87E -11
hsa.miR.376c.3p.2	miRNA-Plasma	2.30E-05	1.68E-08	5.92E -14
hsa.let.7d.3p.2	miRNA-Plasma	0.0016	1.82E-08	3.42E -10
hsa.miR.24.1.3p.2	miRNA-Plasma	1.39E-05	1.96E-08	4.61E -13
hsa.miR.24.2.3p	miRNA-Plasma	1.39E-05	1.96E-08	4.61E -13
hsa.miR.425.3p	miRNA-Plasma	6.02E-05	2.34E-08	6.86E -12
hsa.miR.10a.5p.2	miRNA-Plasma	0.059	2.34E-08	2.28E-08
hsa.miR.199a.1.3p.1	miRNA-Exosome	0.007	2.37E-08	2.43E -11
hsa.miR.376a.1.3p.2	miRNA-Plasma	4.07E-05	3.18E-08	1.41E -12
hsa.miR.376a.2.3p	miRNA-Plasma	4.07E-05	3.18E-08	1.41E -12
hsa.miR.340.5p.2	miRNA-Plasma	0.00021	3.18E-08	5.27E -12
hsa.miR.130a.3p.2	miRNA-Plasma	0.000299	3.18E-08	2.96E -11
hsa.miR.744.5p.2	miRNA-Plasma	3.52E-06	3.32E-08	1.01E -13
hsa.miR.628.3p.2	miRNA-Plasma	7.92E-05	3.51E-08	6.36E -12
hsa.miR.146a.5p.2	miRNA-Plasma	1.39E-05	3.56E-08	1.93E -12
hsa.miR.382.5p.2	miRNA-Plasma	4.86E-05	3.65E-08	8.15E -12
hsa.miR.409.3p.2	miRNA-Plasma	9.28E-06	4.41E-08	1.09E -12
hsa.miR.23a.3p.2	miRNA-Plasma	2.65E-05	4.41E-08	3.29E -12
hsa.miR.23b.3p.2	miRNA-Plasma	2.77E-05	4.41E-08	4.21E -12
hsa.miR.433.3p.2	miRNA-Plasma	0.00048	4.41E-08	1.63E -10
hsa.miR.130b.5p.2	miRNA-Plasma	0.0005	4.41E-08	2.29E -11
hsa.miR.197.3p.2	miRNA-Plasma	3.17E-05	4.79E-08	1.30E -11
hsa.miR.326.3p	miRNA-Plasma	7.20E-05	4.79E-08	5.15E -12
hsa.miR.28.3p.2	miRNA-Plasma	7.97E-06	5.76E-08	1.15E -12
hsa.miR.584.5p.2	miRNA-Plasma	0.00078	5.76E-08	6.31E -11
hsa.miR.151a.3p.2	miRNA-Plasma	1.15E-05	6.54E-08	9.58E -13

Table 3.6: Top DE Molecules Between VAE-based Subgroups

Molecules	DataSets	Train FDR	Subgroup FDR	Three FDR
hsa.miR.326.3p	miRNA-Plasma	7.20E-05	3.08E-09	1.27E-13
hsa.miR.146a.5p.2	miRNA-Plasma	1.39E-05	3.45E-09	1.61E-13
hsa.miR.126.3p.2	miRNA-Plasma	4.36E-06	3.70E-09	7.77E-15
hsa.miR.199a.1.5p.2	miRNA-Plasma	4.36E-06	3.70E-09	2.55E-14
hsa.miR.199a.2.5p	miRNA-Plasma	4.36E-06	3.70E-09	2.55E-14
hsa.miR.23a.3p.2	miRNA-Plasma	2.65E-05	3.70E-09	2.61E-13
hsa.miR.23b.3p.2	miRNA-Plasma	2.77E-05	3.70E-09	3.47E-13
hsa.miR.1260b.5p	miRNA-Plasma	0.0002	3.70E-09	7.87E-13
hsa.miR.27a.3p.2	miRNA-Plasma	0.003	3.70E-09	1.85E-11
hsa.miR.330.3p.2	miRNA-Plasma	3.28E-06	3.76E-09	5.10E-15
hsa.miR.423.3p.2	miRNA-Plasma	4.36E-06	3.76E-09	3.39E-13
hsa.miR.28.3p.2	miRNA-Plasma	7.97E-06	3.76E-09	9.21E-14
hsa.miR.221.3p.2	miRNA-Plasma	2.99E-05	3.76E-09	9.72E-13
hsa.miR.130b.3p.2	miRNA-Plasma	4.07E-05	3.76E-09	3.00E-12
hsa.miR.27b.3p.2	miRNA-Plasma	0.002	3.80E-09	1.27E-11
hsa.miR.584.5p.2	miRNA-Plasma	0.00078	4.04E-09	4.09E-12
hsa.miR.128.2.3p	miRNA-Plasma	9.46E-05	4.77E-09	4.90E-12
hsa.miR.128.1.3p.2	miRNA-Plasma	7.00E-05	4.80E-09	3.55E-12
hsa.miR.181d.5p.1	miRNA-Plasma	1.39E-05	5.40E-09	1.78E-12
hsa.miR.151a.3p.2	miRNA-Plasma	1.15E-05	5.69E-09	9.45E-14
hsa.miR.99b.5p.2	miRNA-Plasma	1.83E-05	5.69E-09	1.25E-12
hsa.miR.425.3p	miRNA-Plasma	6.02E-05	5.69E-09	3.02E-12
hsa.miR.197.3p.2	miRNA-Plasma	3.17E-05	5.77E-09	2.38E-12
hsa.miR.339.5p.2	miRNA-Plasma	7.25E-05	5.77E-09	2.60E-12
hsa.miR.21.3p.2	miRNA-Plasma	9.28E-06	6.69E-09	6.42E-14
hsa.miR.181b.2.5p	miRNA-Plasma	1.39E-05	6.69E-09	4.69E-12
hsa.miR.130a.3p.2	miRNA-Plasma	0.0003	7.58E-09	1.32E-11
hsa.miR.361.5p.2	miRNA-Plasma	0.00039	8.15E-09	3.52E-11
hsa.miR.151a.5p.2	miRNA-Plasma	1.33E-05	8.45E-09	4.74E-13
hsa.miR.125a.5p.2	miRNA-Plasma	1.82E-05	8.45E-09	3.24E-12
hsa.miR.181b.1.5p.2	miRNA-Plasma	1.59E-05	8.54E-09	8.02E-12
hsa.miR.1307.5p.2	miRNA-Plasma	2.65E-05	9.52E-09	1.26E-12
hsa.miR.652.3p.2	miRNA-Plasma	1.68E-05	1.06E-08	8.77E-12
hsa.miR.4286.5p.2	miRNA-Plasma	0.0004	1.13E-08	3.54E-11
hsa.miR.130b.5p.2	miRNA-Plasma	0.0005	1.33E-08	9.84E-12
hsa.miR.191.5p.2	miRNA-Plasma	5.47E-07	1.39E-08	3.37E-14
hsa.miR.22.5p.2	miRNA-Plasma	0.00017	1.45E-08	3.04E-12
hsa.miR.1307.3p.2	miRNA-Plasma	2.22E-06	1.95E-08	8.74E-13
hsa.miR.331.3p.1	miRNA-Plasma	2.22E-06	1.95E-08	7.51E-14
hsa.miR.142.3p.2	miRNA-Plasma	1.07E-05	1.95E-08	4.76E-13

control) using the Training cohort and evaluated using the Validation cohort. We built and evaluated classifiers using SNF-based and VAE-based subgroups separately. In the evaluation, we mainly compared the prediction accuracy in the validation data after the model finished training. To note, the training accuracy refers to accuracy in three groups in three-class models but it refers to accuracy in two groups in two-class models.

Our results showed that classification using the miRNA-Plasma and metabolomics data sets was improved by SNF-based subgroup-aware classifiers, with Validation accuracies increasing over baseline from 0.567 to 0.582 and 0.552 to 0.582, respectively (Table 3.7). For the VAE-based subgroups, all four omics data sets showed Validation accuracy improvement with subgroup aware classifiers (Table 3.8). The classifier for the miRNA-Deplete data set showed the best diagnostic performance overall, with Validation accuracy of 0.612 compared to a 0.582 baseline. The improvements observed as a result of subgroup-aware classification suggest the value of including multi-omic subgroup identification in a diagnostic for PTSD.

Table 3.7: Supervised Classification Performance Using SNF-based Subgroups

Data	Two Class Model		Three Class Model	
	Discovery ACC	Validation ACC	Discovery ACC	Validation ACC
miRNA-Plasma	0.594 ± 0.113	0.567	0.573 ± 0.112	0.582
miRNA-Exosome	0.547 ± 0.103	0.552	0.5 ± 0.092	0.448
miRNA-Deplete	0.623 ± 0.07	0.582	0.5 ± 0.019	0.582
Metabolomics	0.629 ± 0.108	0.552	0.531 ± 0.046	0.582

Table 3.8: Supervised Classification Performance Using VAE-based Subgroups

Data	Two Class Model		Three Class Model	
	Discovery ACC	Validation ACC	Discovery ACC	Validation ACC
miRNA-Plasma	0.594 ± 0.113	0.567	0.545 ± 0.087	0.582
miRNA-Exosome	0.547 ± 0.103	0.552	0.5 ± 0.034	0.582
miRNA-Deplete	0.623 ± 0.07	0.582	0.554 ± 0.08	0.612
Metabolomics	0.629 ± 0.108	0.552	0.507 ± 0.084	0.582

3.4.5 PTSD Subgroup Prediction

Stratification of PTSD samples is an important first step in the practice of precision medicine for PTSD. To facilitate this step, we also constructed a

PTSD subgroup binary classifier. Given the two subgroups identified in the 82 PTSD positive Training samples, we trained a multi-omic binary SVM classifier using all four data sets. Since subgroup assignments do not exist a priori in the Validation cohort, we used 10-fold cross validation to tune the SVM hyperparameter C based on AUC, as described in the Materials and Methods section. We trained a classifier using the optimal hyperparameter, determined by the largest mean AUC, on Training cohort data, followed by application of this classifier to predict subgroups in PTSD cases from the Validation cohort. In cases where all Validation predictions were for a single subgroup, we instead selected the second-best performing hyperparameter for model training and validation. Figure 3.6A and Figure 3.6C show Receiver Operating Characteristic (ROC) curves for SNF and VAE subgroup classifiers, respectively. The training AUC for SNF-based subgroups was 0.93, while the AUC for VAE-based subgroups reached higher to 0.98. Validation predictions based on the two subgroup definitions were similar, with the difference being that five additional samples were predicted to be from subgroup 2 using the SNF-based classifier. Figures 6B and 6D show t-SNE visualizations of the predicted Validation subgroups derived from SNF- and VAE-based classifiers, respectively.

In addition, we performed clinical feature association analysis given our predicted subgroups for the Validation cohort. We used t-tests for continuous features and Fisher's Exact Tests for binary or categorical features. As before, we adjusted raw p-values for multiple testing using the BH method. We found that the same six clinical features that were significantly DE between subgroups from the Training cohort were also significantly DE in the Validation cohort. Figure 3.7A and Figure 3.7B summarize the values of these features between the SNF- and VAE-based Validation subgroups, respectively. The consistency observed in the clinical characterization of Training and Validation subgroups supports the biological plausibility and robustness of the identified PTSD subgroups.

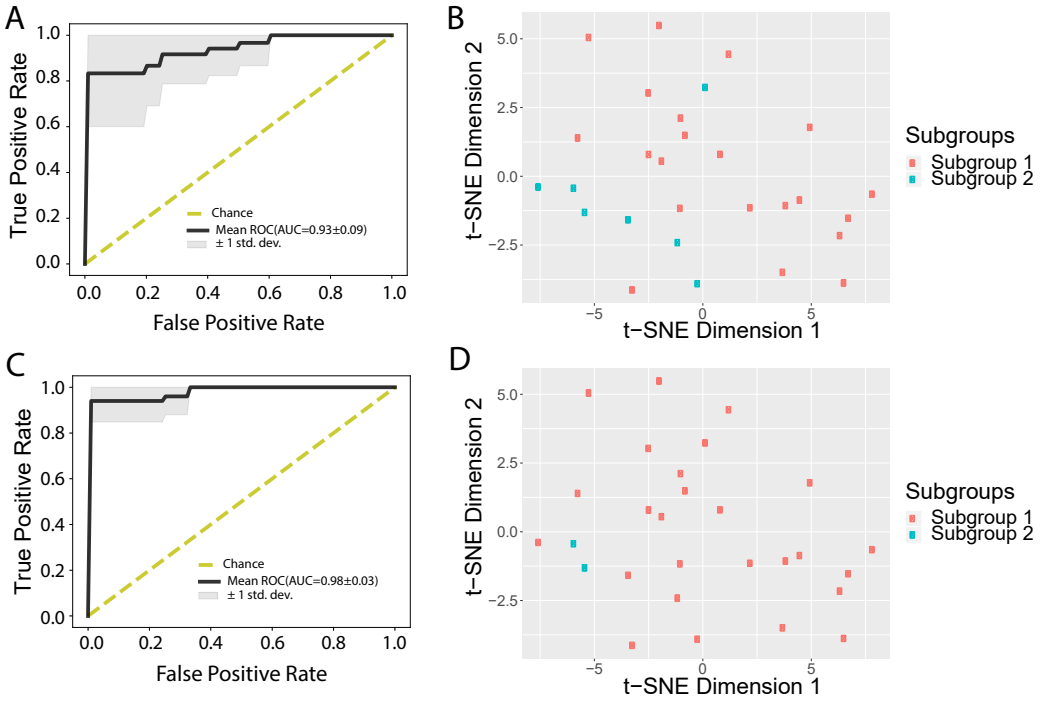


Fig. 3.6: PTSD Subgroup Classification AUC Plot and Subgroup Visualization, A) SNF identified subgroups AUC plot using all omic data sets, B) Visualization of SNF-based subgroups in validation data set, C) VAE identified subgroups AUC plot using all omic data sets, D) Visualization of VAE-based subgroups in validation data set.

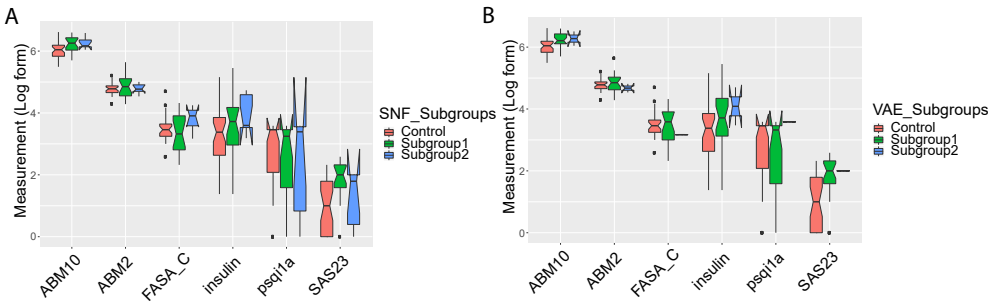


Fig. 3.7: Clinical Association with Predicted PTSD Subgroups in Validation Data Set. A) six selected clinical features expression in PTSD- and SNF-based predicted PTSD subgroups, B) six selected clinical features expression in PTSD- and VAE-based predicted PTSD subgroups.

3.5 Discussion

In this study, we applied three methods–SNF, VAE and PCA–to integrate and discover subgroups from four PTSD omics datasets. The subgroups identified by SNF and VAE show good separation and significant association with recall sample PTSD status. Moreover, the SNF and VAE subgroups largely overlapped, and the subgroup assignment of the recalled sample subset is the same between the two methods. These subgroups also exhibited considerable numbers of DE molecules in each omic dataset. To clinically characterize the subgroups, we identified 59 significantly associated clinical features associated with the subgroups identified by both methods as well as with case-control status. In a diagnostic model, knowledge of these subgroups improved the performance of classifying PTSD positive and negative samples. We also observed excellent classification performance for subgroup prediction models constructed based on SNF and VAE subgroups.

The complexity and heterogeneity of biological systems have made data integration crucial in the advancement of understanding in support of precision medicine [89]. SNF and VAE have been applied to subgroup discovery in cancers [78] [79] [80]; however, our work represents the first time these two approaches have been applied to PTSD. VAEs have been successfully applied in image generation [90]. In our study, VAEs learn the distribution of latent variables and reconstruct input data that are likely to be generative models for PTSD subgroup/biomarker analysis.

PTSD symptoms often exhibit the characteristics of hyperarousal and exaggerated startle responses. It was previously reported that PTSD patients have a high risk of developing cardiovascular issues, such as increased heart rate/pulse rate and blood pressure [65] and increased BMI and weight [66] as a result of traumas. PTSD re-experiencing symptoms were inversely associated with high-frequency pulse rate variability [91]. It was reported that the cortisol and Glucocorticoid Receptor (GR) alterations in the dysregulation of the

hypothalamic-pituitary-adrenal axis (HPA) -axis are associated with insulin resistance [92] [93]. A recent study discovered both heart rate and insulin levels as potential biomarkers for PTSD [68]. MicroRNAs (miRNAs) are involved in critical aspects of PTSD pathophysiology and are also potential biomarkers [94]. The expression changes of miRNAs could decrease vulnerability to stress or promote resilience, as evidenced by a decreased expression of miR-99b-5p and miR-27a-3p in rats leading to an increased vulnerability to stress [95]. The miRNAs miR-199b and miR-24 were also reported to be correlatively implicated in fear- and trauma-related disorders [96]. MicroRNAs miR-99b-5p and miR-27a-3p were two of the most significant molecules we found to be differentially expressed between subgroups.

Several limitations exist in the current study. First, the subjects analyzed in our study were restricted to a pool of men deployed to Iraq and/or Afghanistan with moderate to severe cases of combat-related PTSD compared to a pool of similarly combat-exposed asymptomatic controls. Moreover, the subjects were diagnosed using DSM-IV, which is an older version of the now current DSM-5. This may complicate efforts to reconcile findings from our study with newer datasets collected using DSM-5. Second, the relatively small sample sizes of our study cohorts pose challenges to the application of machine learning approaches, especially deep learning methods such as VAE. We used 82 PTSD positive samples measured for 971 features from the Training cohort for data integration and subgroup discovery, while we validated the discovered subgroups using 28 samples from the Validation cohort. Although our training of the VAE model showed converged minimum losses for both cohorts, this small sample size may cause model overfitting and decreased generalization. Moreover, the assumption of a Gaussian distribution in VAE formulation may exhibit a potential limitation for our study. Finally, there are no other independent datasets that can be used to validate our results. Future work will involve using datasets with larger sample sizes and including additional test datasets to

further improve our understanding of PTSD stratification the unique signatures underlying PTSD subgroups.

3.6 Conclusions

We integrated four omics datasets and discovered two clinically-plausible PTSD subgroups. These subgroups showed significant association with clinical characterization and molecular differential expression. Supervised classification with a subgroup-aware classifier showed improved accuracy for PTSD diagnosis. Future work will involve leveraging knowledge of these subgroups to enable precision medicine for PTSD.

Chapter 4

Conclusions

The objective for this research was to apply machine learning approaches for disease subgroup discovery and classification, applied in particular to PTSD and, secondarily, cancer. Overall, we completed two applications on PTSD and two supplementary applications on various forms of cancer.

In chapter 2, we presented the project "Clinical Subgroup-Specific PTSD Classification and Biomarker Discovery." Using a cohort of 234 samples with 166 for training and 68 for validation, we applied machine learning approaches to classify PTSD patients based on three molecular datasets (miRNA-Exosome: miRNAs enriched in exosomes, miRNA-Deplete: miRNAs in plasma depleted for exosomes, and Metabolomics). We first divided patients into multiple sets of two subgroups based on values of 112 clinical and endocrine measurements. We then performed supervised classification across all samples and within each subgroup using two feature selection strategies (Recursive Feature Elimination [RFE] and ANOVA), four classifiers (logistic regression [LR], support vector machine [SVM], random forest, and extra trees), and 10-fold nested cross validation. We evaluated each subgroup for significantly improved classification performance by statistical tests based on accuracy and AUC values. Finally, we combined those significant clinical features with molecular measurements and constructed an overall PTSD classifier. We fit all data using the best classification model from training and selected features as biomarkers. In total, 85 clinical subgroups from 72 clinical and endocrine features led to improved classification performance compared to the baseline, among which 38 yielded improved performance using more than one method. Tree-based models yielded the greatest number of improved subgroups in the Metabolomics and miRNA-Exosome datasets, while Logistic Regression showed the greatest improvement in the miRNA-Deplete dataset. Using an overall PTSD classifier including both molecular and clinical features, we observed that the majority of classification models show improved

accuracy in both training and testing. Applied to Metabolomics data, the overall classifier achieved the best AUC = 0.79 ± 0.13 and accuracy (ACC) = 0.722 ± 0.078 using ANOVA-SVM, which was a significantly better ACC than the baseline models composed of only molecular or clinical features. Using miRNA-Exosome data, the ANOVA-LR classifier reached the best AUC = 0.758 ± 0.097 and ACC = 0.701 ± 0.116, with a significantly higher ACC than the baseline model with clinical features. In the miRNA-Deplete data set, the RFE-SVM classifier reached the best AUC = 0.677 ± 0.134 and ACC = 0.605 ± 0.128. These best-performing models showed fair performance in the validation samples as well. Finally, we selected the resulting molecular and clinical features from these models and listed them as potential biomarkers for PTSD.

In chapter 3, we described the project "Multi-Omic PTSD Subgroup Identification and Clinical Characterization." Given 284 total samples (124 PTSD positives) from four omics data sets (miRNA-Exosome: miRNAs enriched in exosomes, miRNA-Deplete: miRNAs in plasma depleted for exosomes, miRNA-Plasma: total miRNAs in plasma, and Metabolomics) in cohorts of Training, Validation and Recall, we applied two methods–Similarity Network Fusion (SNF) and Variational Autoencoder (VAE)–to integrate the data sets. SNF performs integration by efficiently fusing sample similarities matrices from each data set into one network representing the full spectrum of underlying data. Next, spectral clustering is used to identify subgroups from this network. The VAE method uses a symmetric deep neural network to reconstruct multiple omics input data sets by estimating data distributions and identifying representative hidden variables. K-means clustering is then used to identify subgroups from the lower-dimensional hidden variables. In order to interpret the subgroups, we tested the associations between identified subgroups and clinical characteristics. We also calculated differentially expressed molecules between subgroups in each omics dataset. We then built supervised classification models for PTSD diagnosis with/without subgroups and compared the accuracy of

predicting PTSD status in the context of subgroups to the accuracy of predicting PTSD status without any knowledge of subgroups. Finally, we built a classification model to predict subgroups in PTSD positive samples. Our results suggest the presence of 2 PTSD subgroups in 82 training PTSD positive samples, using both SNF- and VAE-based methods. These subgroups show a significant association with recalled sample PTSD status change (p-value 0.0213). We also found that a majority of samples associated with the same subgroups when comparing results from the two methods. Upon statistical testing for association of the subgroups with over 600 clinical features, we found a significant association with features including heart rate and insulin. The two identified subgroups also exhibit a number of differentially expressed molecules from each omics data set. For diagnostic classification, we observed improved performance for subgroup-aware PTSD status prediction in miRNA-Plasma and Metabolomics data sets using SNF-based subgroups and in all four omics data sets using VAE-based subgroups. Finally, using our classification model for subgroup prediction, we found that identified subgroups in the validation cohort were significantly associated with many of the same clinical features associated with subgroups from the training cohort.

In two supplementary chapters, we discussed the projects "GEOlimma: Differential Expression Analysis and Feature Selection Using Pre-Existing Microarray Data" and "Prognostic Analysis of Histopathological Images Using Pre-Trained Convolutional Neural Networks: Application to Hepatocellular Carcinoma (HCC)". In the former, we first quantified differential gene expression across 2481 pairwise comparisons from 602 curated Gene Expression Omnibus (GEO) Datasets, and we converted differential expression frequencies to DE prior probabilities. Genes with high DE prior probabilities show enrichment in cell growth and death, signal transduction, and cancer-related biological pathways, while genes with low prior probabilities were enriched in sensory system pathways. We then applied GEOlimma to four differential expression

comparisons within two human disease datasets and performed differential expression, feature selection, and supervised classification analyses. Our results suggest that use of GEOlimma provides greater experimental power to detect DE genes compared to the popular Limma technique, due to its increased effective sample size. Furthermore, in a supervised classification analysis using GEOlimma as a feature selection method, we observed similar or better classification performance than Limma given small, noisy subsets of an asthma dataset. Due to its focus on gene-level differential expression, GEOlimma also has the potential to be applied to other high-throughput biological datasets.

In the latter supplementary project, we applied three pre-trained CNN models–VGG 16, Inception V3, and ResNet 50–to extract features from HCC histopathological images. Sample visualization and classification analyses based on these features showed a very clear separation between cancer and normal samples. In a univariate Cox regression analysis, 21.4% and 16% of image features on average were significantly associated with overall survival and disease-free survival, respectively. We also observed significant correlations between these features and integrated biological pathways derived from gene expression and copy number variation. Using an elastic net regularized CoxPH model of overall survival constructed from Inception image features, we obtained a concordance index (C-index) of 0.789 and a significant log-rank test (p = 7.6E18). We also performed unsupervised classification to identify HCC subgroups from image features. The optimal two subgroups discovered using Inception model image features showed significant differences in both overall (C-index = 0.628 and p = 7.39E-07) and disease-free survival (C-index = 0.558 and p = 0.012). Our work demonstrates the utility of extracting image features using pre-trained models by using them to build accurate prognostic models of HCC as well as highlight significant correlations between these features, clinical survival, and relevant biological pathways. Image features extracted from HCC histopathological images using the pre-trained CNN models VGG 16, Inception

V3 and ResNet 50 can accurately distinguish normal and cancer samples. Furthermore, these image features are significantly correlated with survival and relevant biological pathways.

Despite the progress we have made applying machine learning to disease subgroup discovery and classification, we note several potential limitations in our studies. First, we note that the initial study cohorts for PTSD were restricted to a pool of men deployed to Iraq and/or Afghanistan with moderate to severe cases of combat-related PTSD compared to a pool of similarly combat-exposed asymptomatic controls. During data collection, PTSD diagnosis was formalized using DSM-IV criteria for consistency across all cohorts. However, the more recently released DSM-5 details three additional categories of symptoms, which complicates efforts to reconcile newer datasets with those collected using DSM-IV. As a result, our discoveries of biomarkers and subgroups may require further validation in samples using updated collection criteria. Second, the sample sizes of our PTSD cohorts are relatively small for machine learning given the high dimensionality of the data. Taking all study cohorts together, there are a total of 234 samples. To protect against overfitting, we have applied cross-validation to train classification models, which enables internal evaluation of these models before validating in a test data set. However, a more robust strategy will involve validating and expanding our proposed biomarkers using additional independently collected data sets. Along these lines, the PTSD Systems Biology Consortium has recently collected a new cohort of nearly 1,800 active duty soldiers from Fort Campbell, KY assayed at some combination of time points before and after deployment. As before, blood samples and clinical measurements were collected from these individuals, and the same molecular markers were isolated and extracted from the blood samples. These data offer the potential to independently verify our previously discovered biomarkers as well as improve the identification of both diagnostic and prognostic biomarkers. Furthermore, such longitudinal data also allow the calculation of "delta

measurements"–differences in clinical and molecular features pre- and post-deployment, which enables assessment of effects from the combat events. Delta measurements redefine features in terms of their differences between two time points, allowing patient-specific, PTSD-nonspecific variability to be subtracted away. The data from these "Fort Campbell cohort" samples thus provide a unique opportunity to identify robust diagnostic and prognostic biomarkers for PTSD. At last, we found limitations in visualizing features of pre-trained CNNs for the HCC project.

In order to further improve understanding of our discoveries related to PTSD, several future directions of research are needed. First, we plan to include larger collections of PTSD data sets and better remove noise from input data. Specifically, methods for missing value imputation and batch effect removal could potentially improve downstream data analysis. In addition, public cancer data bases and recent applications in cancer studies may provide opportunities for transfer learning to PTSD. Second, we plan to develop and apply customized machine learning methods for particular data sources. The success of deep learning applications in computer vision and natural language processing may suggest possible transferable applications to biomedical research. However, more work is required to build customized approaches and benchmarks that are particularly beneficial for biomedical applications such as precision medicine. Furthermore, validation from additional clinical studies could help confirm the biological relevance of our discovered biomarkers. At last, for the GEOlimma project, future work will involve the application of GEOlimma to RNA-seq data and develop disease such as asthma specialized GEOlimma approaches.

In conclusion, my dissertation research has involved working collaboratively to improve our knowledge and understanding of PTSD biomarkers and subgroups. We have also explored a range of applications of machine learning to cancer research using public data sets. Future work will

continue leveraging knowledge of machine learning and deep learning to ultimately enable precision medicine for human diseases.

REFERENCES

[1] R. Yehuda, C. W. Hoge, A. C. McFarlane, E. Vermetten, R. A. Lanius, C. M. Nievergelt, S. E. Hobfoll, K. C. Koenen, T. C. Neylan, and S. E. Hyman, "Post-traumatic stress disorder," *Nature Reviews Disease Primers*, vol. 1, p. 15057, Oct. 2015.

[2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub, May 2013.

[3] C. R. Marmar, W. Schlenger, C. Henn-Haase, M. Qian, E. Purchia, M. Li, N. Corry, C. S. Williams, C.-L. Ho, D. Horesh, K.-I. Karstoft, A. Shalev, and R. A. Kulka, "Course of posttraumatic stress disorder 40 years after the vietnam war: Findings from the national vietnam veterans longitudinal study," *JAMA Psychiatry*, vol. 72, no. 9, pp. 875–881, Sept. 2015.

[4] B. C. Kok, R. K. Herrell, J. L. Thomas, and C. W. Hoge, "Posttraumatic stress disorder associated with combat service in iraq or afghanistan: reconciling prevalence differences between studies," *J. Nerv. Ment. Dis.*, vol. 200, no. 5, pp. 444–450, May 2012.

[5] C. M. Sheerin, M. J. Lind, K. Bountress, N. R. Nugent, and A. B. Amstadter, "The genetics and epigenetics of PTSD: Overview, recent advances, and future directions," *Curr Opin Psychol*, vol. 14, pp. 5–11, Apr. 2017.

[6] L. E. Duncan, B. N. Cooper, and H. Shen, "Robust findings from 25 years of PTSD genetics research," *Curr. Psychiatry Rep.*, vol. 20, no. 12, p. 115, Oct. 2018.

[7] T. M. Keane, A. D. Marshall, and C. T. Taft, "Posttraumatic stress disorder: etiology, epidemiology, and treatment outcome," *Annu. Rev. Clin. Psychol.*, vol. 2, pp. 161–197, 2006.

[8] M. W. Kirschner, "The meaning of systems biology," *Cell*, vol. 121, no. 4, pp. 503–504, May 2005.

[9] G. S. Thakur, B. J. Daigle, Jr, K. R. Dean, Y. Zhang, M. Rodriguez-Fernandez, R. Hammamieh, R. Yang, M. Jett, J. Palma, L. R. Petzold, and F. J. Doyle, 3rd, "Systems biology approach to understanding post-traumatic stress disorder," *Mol. Biosyst.*, vol. 11, no. 4, pp. 980–993, Apr. 2015.

[10] A. Shalev, I. Liberzon, and C. Marmar, "Post-Traumatic stress disorder," *N. Engl. J. Med.*, vol. 376, no. 25, pp. 2459–2469, June 2017.

[11] J. E. Sherin and C. B. Nemeroff, "Post-traumatic stress disorder: the neurobiological impact of psychological trauma," *Dialogues Clin. Neurosci.*, vol. 13, no. 3, pp. 263–278, 2011.

[12] R. Yehuda and J. Seckl, "Minireview: Stress-related psychiatric disorders with low cortisol levels: a metabolic hypothesis," *Endocrinology*, vol. 152, no. 12, pp. 4496–4503, Dec. 2011.

[13] R. Yehuda, "Post-Traumatic stress disorder," *N. Engl. J. Med.*, vol. 346, no. 2, pp. 108–114, Jan. 2002.

[14] ——, "Advances in understanding neuroendocrine alterations in PTSD and their therapeutic implications," *Ann. N. Y. Acad. Sci.*, vol. 1071, pp. 137–166, July 2006.

[15] A. M. Rasmusson, R. L. Hauger, C. A. Morgan, J. D. Bremner, D. S. Charney, and S. M. Southwick, "Low baseline and yohimbine-stimulated plasma neuropeptide Y (NPY) levels in combat-related PTSD," *Biol. Psychiatry*, vol. 47, no. 6, pp. 526–539, Mar. 2000.

[16] A. Maron-Katz, Y. Zhang, M. Narayan, W. Wu, R. T. Toll, S. Naparstek, C. De Los Angeles, P. Longwell, E. Shpigel, J. Newman, D. Abu-Amara, C. Marmar, and A. Etkin, "Individual patterns of abnormality in Resting-State functional connectivity reveal two Data-Driven PTSD subgroups," *Am. J. Psychiatry*, vol. 177, no. 3, pp. 244–253, Mar. 2020.

[17] M. W. Logue, A. B. Amstadter, D. G. Baker, L. Duncan, K. C. Koenen, I. Liberzon, M. W. Miller, R. A. Morey, C. M. Nievergelt, K. J. Ressler, A. K. Smith, J. W. Smoller, M. B. Stein, J. A. Sumner, and M. Uddin, "The psychiatric genomics consortium posttraumatic stress disorder workgroup: Posttraumatic stress disorder enters the age of Large-Scale genomic collaboration," *Neuropsychopharmacology*, vol. 40, no. 10, pp. 2287–2297, Sept. 2015.

[18] L. M. Almli, N. Fani, A. K. Smith, and K. J. Ressler, "Genetic approaches to understanding post-traumatic stress disorder," *Int. J. Neuropsychopharmacol.*, vol. 17, no. 2, pp. 355–370, Feb. 2014.

[19] E. B. Binder, R. G. Bradley, W. Liu, M. P. Epstein, T. C. Deveau, K. B. Mercer, Y. Tang, C. F. Gillespie, C. M. Heim, C. B. Nemeroff, A. C. Schwartz, J. F. Cubells, and K. J. Ressler, "Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults," *JAMA*, vol. 299, no. 11, pp. 1291–1305, Mar. 2008.

[20] P. Xie, H. R. Kranzler, J. Poling, M. B. Stein, R. F. Anton, L. A. Farrer, and J. Gelernter, "Interaction of FKBP5 with childhood adversity on risk for post-traumatic stress disorder," *Neuropsychopharmacology*, vol. 35, no. 8, pp. 1684–1692, July 2010.

[21] T. Klengel, D. Mehta, C. Anacker, M. Rex-Haffner, J. C. Pruessner, C. M. Pariante, T. W. W. Pace, K. B. Mercer, H. S. Mayberg, B. Bradley, C. B. Nemeroff, F. Holsboer, C. M. Heim, K. J. Ressler, T. Rein, and E. B. Binder, "Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions," *Nat. Neurosci.*, vol. 16, no. 1, pp. 33–41, Jan. 2013.

[22] S. D. Norrholm, T. Jovanovic, A. K. Smith, E. Binder, T. Klengel, K. Conneely, K. B. Mercer, J. S. Davis, K. Kerley, J. Winkler, C. F. Gillespie, B. Bradley, and K. J. Ressler, "Differential genetic and epigenetic regulation of catechol-o-methyltransferase is associated with impaired fear inhibition in posttraumatic stress disorder," *Front. Behav. Neurosci.*, vol. 7, p. 30, Apr. 2013.

[23] R. Andero and K. J. Ressler, "Fear extinction and BDNF: translating animal models of PTSD to the clinic," *Genes Brain Behav.*, vol. 11, no. 5, pp. 503–512, July 2012.

[24] M. W. Logue, C. Baldwin, G. Guffanti, E. Melista, E. J. Wolf, A. F. Reardon, M. Uddin, D. Wildman, S. Galea, K. C. Koenen, and M. W. Miller, "A

genome-wide association study of post-traumatic stress disorder identifies the retinoid-related orphan receptor alpha (RORA) gene as a significant risk locus," *Mol. Psychiatry*, vol. 18, no. 8, pp. 937–942, Aug. 2013.

[25] P. Xie, H. R. Kranzler, C. Yang, H. Zhao, L. A. Farrer, and J. Gelernter, "Genome-wide association study identifies new susceptibility loci for posttraumatic stress disorder," *Biol. Psychiatry*, vol. 74, no. 9, pp. 656–663, Nov. 2013.

[26] G. Guffanti, S. Galea, L. Yan, A. L. Roberts, N. Solovieff, A. E. Aiello, J. W. Smoller, I. De Vivo, H. Ranu, M. Uddin, D. E. Wildman, S. Purcell, and K. C. Koenen, "Genome-wide association study implicates a novel RNA gene, the lincRNA AC068718.1, as a risk factor for post-traumatic stress disorder in women," *Psychoneuroendocrinology*, vol. 38, no. 12, pp. 3029–3038, Dec. 2013.

[27] C. M. Nievergelt, A. X. Maihofer, M. Mustapic, K. A. Yurgil, N. J. Schork, M. W. Miller, M. W. Logue, M. A. Geyer, V. B. Risbrough, D. T. O'Connor, and D. G. Baker, "Genomic predictors of combat stress vulnerability and resilience in U.S. marines: A genome-wide association study across multiple ancestries implicates PRTFDC1 as a potential PTSD gene," *Psychoneuroendocrinology*, vol. 51, pp. 459–471, Jan. 2015.

[28] J. Voisey, R. M. Young, B. R. Lawford, and C. P. Morris, "Progress towards understanding the genetics of posttraumatic stress disorder," *J. Anxiety Disord.*, vol. 28, no. 8, pp. 873–883, Dec. 2014.

[29] A. M. Hull, "Neuroimaging findings in post-traumatic stress disorder. systematic review," *Br. J. Psychiatry*, vol. 181, pp. 102–110, Aug. 2002.

[30] U. Schmidt, S. F. Kaltwasser, and C. T. Wotjak, "Biomarkers in posttraumatic stress disorder: overview and implications for future research," *Dis. Markers*, vol. 35, no. 1, pp. 43–54, July 2013.

[31] V. Michopoulos, S. D. Norrholm, and T. Jovanovic, "Diagnostic biomarkers for posttraumatic stress disorder: Promising horizons from translational neuroscience research," *Biol. Psychiatry*, vol. 78, no. 5, pp. 344–353, Sept. 2015.

[32] H. J. Kang, S. Yoon, and I. K. Lyoo, "Peripheral biomarker candidates of posttraumatic stress disorder," *Exp. Neurobiol.*, vol. 24, no. 3, pp. 186–196, Sept. 2015.

[33] W. Loh, L. Cao, and P. Zhou, "Subgroup identification for precision medicine: A comparative review of 13 methods," *WIREs Data Mining Knowl Discov*, vol. 9, no. 5, p. 604, Sept. 2019.

[34] S. Helal, "Subgroup discovery algorithms: A survey and empirical evaluation," *J. Comput. Sci. Technol.*, vol. 31, no. 3, pp. 561–576, May 2016.

[35] I. R. Galatzer-Levy, Y. Ankri, S. Freedman, Y. Israeli-Shalev, P. Roitman, M. Gilad, and A. Y. Shalev, "Early PTSD symptom trajectories: persistence, recovery, and response to treatment: results from the jerusalem trauma outreach and prevention study (J-TOPS)," *PLoS One*, vol. 8, no. 8, p. e70084, Aug. 2013.

[36] I. R. Galatzer-Levy, K.-I. Karstoft, A. Statnikov, and A. Y. Shalev, "Quantitative forecasting of PTSD from early trauma responses: a machine learning application," *J. Psychiatr. Res.*, vol. 59, pp. 68–76, Dec. 2014.

[37] I. R. Galatzer-Levy, S. Ma, A. Statnikov, R. Yehuda, and A. Y. Shalev, "Utilization of machine learning for prediction of post-traumatic stress: a re-examination of cortisol in the prediction and pathways to non-remitting PTSD," *Transl. Psychiatry*, vol. 7, no. 3, p. e0, Mar. 2017.

[38] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, Mar. 2003.

[39] G. A. N. Gowda and D. Djukovic, "Overview of mass spectrometry-based metabolomics: opportunities and challenges," *Methods Mol. Biol.*, vol. 1198, pp. 3–12, 2014.

[40] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping, "Machine learning and integrative analysis of biomedical big data," *Genes*, vol. 10, no. 2, Jan. 2019.

[41] G. Rong, A. Mendez, E. Bou Assi, B. Zhao, and M. Sawan, "Artificial intelligence in healthcare: Review and prediction case studies," *Proc. Est. Acad. Sci. Eng.*, Jan. 2020.

[42] P. Shah, F. Kendall, S. Khozin, R. Goosen, J. Hu, J. Laramie, M. Ringel, and N. Schork, "Artificial intelligence and machine learning in clinical development: a translational perspective," *NPJ Digit Med*, vol. 2, p. 69, July 2019.

[43] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nat. Rev. Cancer*, vol. 8, no. 1, pp. 37–49, Jan. 2008.

[44] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Comput. Biol. Med.*, vol. 112, p. 103375, Sept. 2019.

[45] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, Oct. 2000.

[46] S. Gao, Z. Qiu, Y. Song, C. Mo, W. Tan, Q. Chen, D. Liu, M. Chen, and H. Zhou, "Unsupervised clustering reveals new prostate cancer subtypes," *Transl. Cancer Res.*, vol. 6, no. 3, pp. 561–572, 2017.

[47] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, METABRIC Group, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, June 2012.

[48] J. Wu, Y. Cui, X. Sun, G. Cao, B. Li, D. M. Ikeda, A. W. Kurian, and R. Li, "Unsupervised clustering of quantitative image phenotypes reveals breast cancer subtypes with distinct prognoses and molecular pathways," *Clin. Cancer Res.*, vol. 23, no. 13, pp. 3334–3342, July 2017.

[49] A. Li, J. Walling, S. Ahn, Y. Kotliarov, Q. Su, M. Quezado, J. C. Oberholtzer, J. Park, J. C. Zenklusen, and H. A. Fine, "Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes," *Cancer Res.*, vol. 69, no. 5, pp. 2091–2099, Mar. 2009.

[50] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," *J. ACM*, vol. 51, no. 3, pp. 497–515, May 2004.

[51] H. Almeida, D. Guedes, W. Meira, and M. J. Zaki, "Is there a best quality metric for graph clusters?" in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Sept. 2011, pp. 44–59.

[52] P. Baldi, "Deep learning in biomedical data science," *Annu. Rev. Biomed. Data Sci.*, vol. 1, no. 1, pp. 181–205, July 2018.

[53] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of deep learning in biomedicine," *Mol. Pharm.*, vol. 13, no. 5, pp. 1445–1454, May 2016.

[54] F. van Veen, "The neural network zoo - the asimov institute," https://www.asimovinstitute.org/neural-network-zoo/, Sept. 2016, accessed: 2020-4-7.

[55] T. M. Mitchell, *Machine Learning*, 1st ed. USA: McGraw-Hill, Inc., 1997.

[56] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, no. 3, pp. 400–407, 1951.

[57] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 61, pp. 2121–2159, 2011.

[58] J. Schmidhuber, "Deep learning in neural networks: An overview," Apr. 2014.

[59] G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," *Pac. Symp. Biocomput.*, vol. 23, pp. 80–91, 2018.

[60] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, June 2001.

[61] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and Semi-Supervised feature selection: A review on gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 13, no. 5, pp. 971–989, Sept. 2016.

[62] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, Jan. 2002.

[63] A. C. McFarlane, E. Lawrence-Wood, M. Van Hooff, G. S. Malhi, and R. Yehuda, "The need to take a staging approach to the biological mechanisms of PTSD and its treatment," *Curr. Psychiatry Rep.*, vol. 19, no. 2, p. 10, Feb. 2017.

[64] T. C. Neylan, E. E. Schadt, and R. Yehuda, "Biomarkers for combat-related PTSD: focus on molecular networks from high-dimensional data," *Eur. J. Psychotraumatol.*, vol. 5, Aug. 2014.

[65] E. A. Dedert, P. S. Calhoun, L. L. Watkins, A. Sherwood, and J. C. Beckham, "Posttraumatic stress disorder, cardiovascular, and metabolic disease: a review of the evidence," *Ann. Behav. Med.*, vol. 39, no. 1, pp. 61–78, Feb. 2010.

[66] S. C. McLeay, W. M. Harvey, M. N. Romaniuk, D. H. Crawford, D. M. Colquhoun, R. M. Young, M. Dwyer, J. M. Gibson, R. A. O'Sullivan, G. Cooksley, C. R. Strakosch, R. M. Thomson, J. Voisey, and B. R. Lawford, "Physical comorbidities of post-traumatic stress disorder in australian vietnam war veterans," *Med. J. Aust.*, vol. 206, no. 6, pp. 251–257, Apr. 2017.

[67] H. Gola, H. Engler, A. Sommershof, H. Adenauer, S. Kolassa, M. Schedlowski, M. Groettrup, T. Elbert, and I.-T. Kolassa, "Posttraumatic stress disorder is associated with an enhanced spontaneous production of pro-inflammatory cytokines by peripheral blood mononuclear cells," *BMC Psychiatry*, vol. 13, p. 40, Jan. 2013.

[68] K. R. Dean, R. Hammamieh, S. H. Mellon, D. Abu-Amara, J. D. Flory, G. Guffanti, K. Wang, B. J. Daigle, Jr, A. Gautam, I. Lee, R. Yang, L. M. Almli, F. S. Bersani, N. Chakraborty, D. Donohue, K. Kerley, T.-K. Kim, E. Laska, M. Young Lee, D. Lindqvist, A. Lori, L. Lu, B. Misganaw, S. Muhie, J. Newman, N. D. Price, S. Qin, V. I. Reus, C. Siegel, P. R. Somvanshi, G. S. Thakur, Y. Zhou, PTSD Systems Biology Consortium, L. Hood, K. J. Ressler, O. M. Wolkowitz, R. Yehuda, M. Jett, F. J. Doyle, 3rd, and C. Marmar, "Multi-omic biomarker identification and validation for diagnosing warzone-related post-traumatic stress disorder," *Mol. Psychiatry*, Sept. 2019.

[69] B. Berger, J. Peng, and M. Singh, "Computational solutions for omics data," *Nat. Rev. Genet.*, vol. 14, no. 5, pp. 333–346, May 2013.

[70] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: Recent progress in Multi-Omics data integration methods," *Front. Genet.*, vol. 8, p. 84, June 2017.

[71] L. Zhao, V. H. F. Lee, M. K. Ng, H. Yan, and M. F. Bijlsma, "Molecular subtyping of cancer: current status and moving toward clinical applications," *Brief. Bioinform.*, vol. 20, no. 2, pp. 572–584, Mar. 2019.

[72] A. Elefsinioti, T. Bellaire, A. Wang, K. Quast, H. Seidel, M. Braxenthaler, G. Goeller, A. Christianson, D. Henderson, and J. Reischl, "Key factors for successful data integration in biomarker research," *Nat. Rev. Drug Discov.*, vol. 15, no. 6, pp. 369–370, June 2016.

[73] F. R. Pinu, D. J. Beale, A. M. Paten, K. Kouremenos, S. Swarup, H. J. Schirra, and D. Wishart, "Systems biology and Multi-Omics integration: Viewpoints from the metabolomics research community," *Metabolites*, vol. 9, no. 4, Apr. 2019.

[74] C. Wu, F. Zhou, J. Ren, X. Li, Y. Jiang, and S. Ma, "A selective review of Multi-Level omics data integration using variable selection," *High Throughput*, vol. 8, no. 1, Jan. 2019.

[75] E. López de Maturana, L. Alonso, P. Alarcón, I. A. Martín-Antoniano, S. Pineda, L. Piorno, M. L. Calle, and N. Malats, "Challenges in the integration of omics and Non-Omics data," *Genes*, vol. 10, no. 3, Mar. 2019.

[76] A. Creswell and A. A. Bharath, "Denoising adversarial autoencoders," *IEEE Trans Neural Netw Learn Syst*, vol. 30, no. 4, pp. 968–984, Apr. 2019.

[77] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," Nov. 2015.

[78] L. Beggel, M. Pfeiffer, and B. Bischl, "Robust anomaly detection in images using adversarial autoencoders," Jan. 2019.

[79] L. Zhang, C. Lv, Y. Jin, G. Cheng, Y. Fu, D. Yuan, Y. Tao, Y. Guo, X. Ni, and T. Shi, "Deep Learning-Based Multi-Omics data integration reveals two prognostic subtypes in High-Risk neuroblastoma," *Front. Genet.*, vol. 9, p. 477, Oct. 2018.

[80] W. Bulten, "Unsupervised cancer detection using deep learning and adversarial autoencoders," https://www.wouterbulten.nl/blog/tech/unsupervised-cancer-detection-using-deep-learning-adversarial-autoencoders/, Oct. 2018, accessed: 2019-5-9.

[81] J. Tan, M. Ung, C. Cheng, and C. S. Greene, "Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders," *Pac. Symp. Biocomput.*, pp. 132–143, 2015.

[82] M. Francescatto, M. Chierici, S. Rezvan Dezfooli, A. Zandonà, G. Jurman, and C. Furlanello, "Multi-omics integration for neuroblastoma clinical endpoint prediction," *Biol. Direct*, vol. 13, no. 1, p. 5, Apr. 2018.

[83] O. B. Poirion, K. Chaudhary, and L. X. Garmire, "Deep learning data integration for better risk stratification models of bladder cancer," *AMIA Jt Summits Transl Sci Proc*, vol. 2017, pp. 197–206, May 2018.

[84] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep Learning–Based Multi-Omics integration robustly predicts survival in liver cancer," *Clin. Cancer Res.*, vol. 24, no. 6, pp. 1248–1259, Mar. 2018.

[85] J. Ronen, S. Hayat, and A. Akalin, "Evaluation of colorectal cancer subtypes and cell lines using deep learning," Nov. 2018.

[86] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," pp. 289–300, 1995.

[87] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nat. Methods*, vol. 11, no. 3, pp. 333–337, Mar. 2014.

[88] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, Aug. 2007.

[89] T. Hulsen, S. S. Jamuar, A. R. Moody, J. H. Karnes, O. Varga, S. Hedensted, R. Spreafico, D. A. Hafler, and E. F. McKinney, "From big data to precision medicine," *Front. Med.*, vol. 6, p. 34, Mar. 2019.

[90] D. P. Kingma and M. Welling, "Auto-Encoding variational bayes," Dec. 2013.

[91] D. W. Grupe, T. Imhoff-Smith, J. Wielgosz, J. B. Nitschke, and R. J. Davidson, "A common neural substrate for elevated PTSD symptoms and reduced pulse rate variability in combat-exposed veterans," *Psychophysiology*, vol. 57, no. 1, p. e13352, Jan. 2020.

[92] P. Anagnostis, V. G. Athyros, K. Tziomalos, A. Karagiannis, and D. P. Mikhailidis, "Clinical review: The pathogenetic role of cortisol in the metabolic syndrome: a hypothesis," *J. Clin. Endocrinol. Metab.*, vol. 94, no. 8, pp. 2692–2701, Aug. 2009.

[93] R. M. Reynolds, K. E. Chapman, J. R. Seckl, B. R. Walker, P. M. McKeigue, and H. O. Lithell, "Skeletal muscle glucocorticoid receptor density and insulin resistance," *JAMA*, vol. 287, no. 19, pp. 2505–2506, May 2002.

[94] N. P. Daskalakis, A. C. Provost, R. G. Hunter, and G. Guffanti, "Noncoding RNAs: Stress, glucocorticoids, and posttraumatic stress disorder," *Biol. Psychiatry*, vol. 83, no. 10, pp. 849–865, May 2018.

[95] C. Snijders, L. de Nijs, D. G. Baker, R. L. Hauger, D. van den Hove, G. Kenis, C. M. Nievergelt, M. P. Boks, E. Vermetten, F. H. Gage, and B. P. F. Rutten, "MicroRNAs in post-traumatic stress disorder," *Curr. Top. Behav. Neurosci.*, Oct. 2017.

[96] C. P. Murphy and N. Singewald, "Potential of microRNAs as novel targets in the alleviation of pathological fear," *Genes Brain Behav.*, vol. 17, no. 3, p. e12427, Mar. 2018.

[97] C. A. Harrington, C. Rosenow, and J. Retief, "Monitoring gene expression using DNA microarrays," *Curr. Opin. Microbiol.*, vol. 3, no. 3, pp. 285–291, June 2000.

[98] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat. Rev. Genet.*, vol. 10, p. 57, Jan. 2009.

[99] R. Govindarajan, J. Duraiyan, K. Kaliyappan, and M. Palanisamy, "Microarray and its applications," *J. Pharm. Bioallied Sci.*, vol. 4, no. Suppl 2, pp. S310–2, Aug. 2012.

[100] R. B. Stoughton, "Applications of DNA microarrays in biology," *Annu. Rev. Biochem.*, vol. 74, pp. 53–82, 2005.

[101] K. Van Den Berge, K. M. Hembach, C. Soneson, S. Tiberi, L. Clement, M. I. Love, R. Patro, and M. D. Robinson, "RNA sequencing data: hitchhiker's guide to expression analysis," *Annual Review of Biomedical Data Science*, vol. 2, 2018.

[102] Y. Hou, B. Gao, G. Li, and Z. Su, "MaxMIF: A new method for identifying cancer driver genes through effective data integration," *Adv. Sci.*, vol. 5, no. 9, p. 1800640, Sept. 2018.

[103] A. Alkhateeb, I. Rezaeian, S. Singireddy, D. Cavallo-Medved, L. A. Porter, and L. Rueda, "Transcriptomics signature from Next-Generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer," *Cancer Inform.*, vol. 18, p. 1176935119835522, Mar. 2019.

[104] J. Han, M. Chen, Y. Wang, B. Gong, T. Zhuang, L. Liang, and H. Qiao, "Identification of biomarkers based on differentially expressed genes in papillary thyroid carcinoma," *Sci. Rep.*, vol. 8, no. 1, p. 9912, July 2018.

[105] H. D. Gliddon, J. A. Herberg, M. Levin, and M. Kaforou, "Genome-wide host RNA signatures of infectious diseases: discovery and clinical translation," *Immunology*, vol. 153, no. 2, pp. 171–178, 2018.

[106] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinformatics*, vol. 2015, p. 198363, June 2015.

[107] P. V. Nazarov, A. Muller, T. Kaoma, N. Nicot, C. Maximo, P. Birembaut, N. L. Tran, G. Dittmar, and L. Vallar, "RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples," *BMC Genomics*, vol. 18, no. 1, p. 443, June 2017.

[108] Y. Wang, C. Barbacioru, F. Hyland, W. Xiao, K. L. Hunkapiller, J. Blake, F. Chan, C. Gonzalez, L. Zhang, and R. R. Samaha, "Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays," *BMC Genomics*, vol. 7, p. 59, Mar. 2006.

[109] J. J. Chen, H.-M. Hsueh, R. R. Delongchamp, C.-J. Lin, and C.-A. Tsai, "Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data," *BMC Bioinformatics*, vol. 8, p. 412, Oct. 2007.

[110] C. Wei, J. Li, and R. E. Bumgarner, "Sample size for detecting differentially expressed genes in microarray experiments," *BMC Genomics*, vol. 5, p. 87, Nov. 2004.

[111] S. Boluki, M. S. Esfahani, X. Qian, and E. R. Dougherty, "Incorporating biological prior knowledge for bayesian learning via maximal knowledge-driven information priors," *BMC Bioinformatics*, vol. 18, no. Suppl 14, p. 552, Dec. 2017.

[112] D. McNeish, "On using bayesian methods to address small sample problems," *Struct. Equ. Modeling*, vol. 23, no. 5, pp. 750–773, Sept. 2016.

[113] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000.

[114] The Gene Ontology Consortium, "The gene ontology resource: 20 years and still GOing strong," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D330–D338, Jan. 2019.

[115] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[116] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, Jan. 2017.

[117] B. J. Daigle, Jr and R. B. Altman, "M-BISON: microarray-based integration of data sources using networks," *BMC Bioinformatics*, vol. 9, p. 214, Apr. 2008.

[118] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, "GeneRank: using search engine technology for the analysis of microarray experiments," *BMC Bioinformatics*, vol. 6, p. 233, Sept. 2005.

[119] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, Jan. 2002.

[120] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets–update," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D991–5, Jan. 2013.

[121] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, and A. Brazma, "ArrayExpress update–simplifying data submissions," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D1113–6, Jan. 2015.

[122] B. J. Daigle, Jr, A. Deng, T. McLaughlin, S. W. Cushman, M. C. Cam, G. Reaven, P. S. Tsao, and R. B. Altman, "Using pre-existing microarray datasets to increase experimental power: application to insulin resistance," *PLoS Comput. Biol.*, vol. 6, no. 3, p. e1000718, Mar. 2010.

[123] J. M. Engreitz, B. J. Daigle, Jr, J. J. Marshall, and R. B. Altman, "Independent component analysis: mining microarray data for fundamental human gene expression modules," *J. Biomed. Inform.*, vol. 43, no. 6, pp. 932–944, Dec. 2010.

[124] R. D. Kim and P. J. Park, "Improving identification of differentially expressed genes in microarray studies using information from public databases," *Genome Biol.*, vol. 5, no. 9, p. R70, Aug. 2004.

[125] R. Chen, A. A. Morgan, J. Dudley, T. Deshpande, L. Li, K. Kodama, A. P. Chiang, and A. J. Butte, "FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease," *Genome Biol.*, vol. 9, no. 12, p. R170, Dec. 2008.

[126] M. Crow, N. Lim, S. Ballouz, P. Pavlidis, and J. Gillis, "Predictability of human differential gene expression," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 13, pp. 6491–6500, Mar. 2019.

[127] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.

[128] Z. He and W. Yu, "Stable feature selection for biomarker discovery," *Comput. Biol. Chem.*, vol. 34, no. 4, pp. 215–225, Aug. 2010.

[129] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inf. Sci.*, vol. 282, pp. 111–135, Oct. 2014.

[130] H. Abusamra, "A comparative study of feature selection and classification methods for gene expression data of glioma," *Procedia Comput. Sci.*, vol. 23, pp. 5–14, Jan. 2013.

[131] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Stat. Appl. Genet. Mol. Biol.*, vol. 3, p. Article3, Feb. 2004.

[132] ——, "limma: Linear models for microarray data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, Eds. New York, NY: Springer New York, 2005, pp. 397–420.

[133] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, Apr. 2015.

[134] M. F. Moffatt, M. Kabesch, L. Liang, A. L. Dixon, D. Strachan, S. Heath, M. Depner, A. von Berg, A. Bufe, E. Rietschel, A. Heinzmann, B. Simma, T. Frischer, S. A. G. Willis-Owen, K. C. C. Wong, T. Illig, C. Vogelberg, S. K. Weiland, E. von Mutius, G. R. Abecasis, M. Farrall, I. G. Gut, G. M. Lathrop, and W. O. C. Cookson, "Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma," *Nature*, vol. 448, no. 7152, pp. 470–473, July 2007.

[135] K. I. Mills, A. Kohlmann, P. M. Williams, L. Wieczorek, W.-M. Liu, R. Li, W. Wei, D. T. Bowen, H. Loeffler, J. M. Hernandez, W.-K. Hofmann, and T. Haferlach, "Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome," *Blood*, vol. 114, no. 5, pp. 1063–1072, July 2009.

[136] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterprofiler: an R package for comparing biological themes among gene clusters," *OMICS*, vol. 16, no. 5, pp. 284–287, May 2012.

[137] W. Luo and C. Brouwer, "Pathview: an R/Bioconductor package for pathway-based data integration and visualization," *Bioinformatics*, vol. 29, no. 14, pp. 1830–1831, July 2013.

[138] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[139] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, Oct. 2005.

[140] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[141] H. Trevor, T. Robert, and F. Jh, "The elements of statistical learning: data mining, inference, and prediction," 2009.

[142] A. K. Pandurangan, T. Divya, K. Kumar, V. Dineshbabu, B. Velavan, and G. Sudhandiran, "Colorectal carcinogenesis: Insights into the cell death and signal transduction pathways: A review," *World J. Gastrointest. Oncol.*, vol. 10, no. 9, pp. 244–259, Sept. 2018.

[143] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011.

[144] C. Huttenhower, E. M. Haley, M. A. Hibbs, V. Dumeaux, D. R. Barrett, H. A. Coller, and O. G. Troyanskaya, "Exploring the human genome with functional maps," *Genome Res.*, vol. 19, no. 6, pp. 1093–1106, June 2009.

[145] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts," *Genome Biol.*, vol. 15, no. 2, p. R29, Feb. 2014.

[146] C. B. Giles, C. A. Brown, M. Ripperger, Z. Dennis, X. Roopnarinesingh, H. Porter, A. Perz, and J. D. Wren, "ALE: automated label extraction from GEO metadata," *BMC Bioinformatics*, vol. 18, no. Suppl 14, p. 509, Dec. 2017.

[147] M. N. Gurcan, L. Boucheron, A. Can, and others, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, 2009.

[148] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 34–42, Feb. 2018.

[149] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[150] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[151] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights Imaging*, June 2018.

[152] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, June 2017.

[153] J. Xu, C. Zhou, B. Lang, and Q. Liu, "Deep learning for histopathological image analysis: Towards computerized diagnosis on cancers," in *Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Large-Scale Datasets*, L. Lu, Y. Zheng, G. Carneiro, and L. Yang, Eds. Cham: Springer International Publishing, 2017, pp. 73–95.

[154] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," p. 2402, 2016.

[155] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[156] T. K. Yoo, J. Y. Choi, J. G. Seo, B. Ramasubramanian, S. Selvaperumal, and D. W. Kim, "The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment," pp. 677–687, 2019.

[157] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, the CAMELYON16 Consortium, M. Hermsen, Q. F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. C. van Dijk, P. Bult, F. Beca, A. H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H.-J. Lin, P.-A. Heng, C. Haß, E. Bruni, Q. Wong, U. Halici, M. Ü. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y.-W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori,

K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. Ahmady Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, and R. Venâncio, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, Dec. 2017.

[158] N. Hegde, J. D. Hipp, Y. Liu, M. Emmert-Buck, E. Reif, D. Smilkov, M. Terry, C. J. Cai, M. B. Amin, C. H. Mermel, P. Q. Nelson, L. H. Peng, G. S. Corrado, and M. C. Stumpe, "Similar image search for histopathology: SMILY," *NPJ Digit Med*, vol. 2, p. 56, June 2019.

[159] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[160] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for Computer-Aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[161] S. Dabeer, M. M. Khan, and S. Islam, "Cancer diagnosis in histopathological image: CNN based approach," *Informatics in Medicine Unlocked*, vol. 16, p. 100231, Jan. 2019.

[162] J. M. Llovet, J. Zucman-Rossi, E. Pikarsky, B. Sangro, M. Schwartz, M. Sherman, and G. Gores, "Hepatocellular carcinoma," *Nat Rev Dis Primers*, vol. 2, p. 16018, Apr. 2016.

[163] C. Guichard, G. Amaddeo, S. Imbeaud, Y. Ladeiro, L. Pelletier, I. B. Maad, J. Calderaro, P. Bioulac-Sage, M. Letexier, F. Degos, B. Clément, C. Balabaud, E. Chevet, A. Laurent, G. Couchy, E. Letouzé, F. Calvo, and J. Zucman-Rossi, "Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma," *Nat. Genet.*, vol. 44, no. 6, pp. 694–698, June 2012.

[164] Y. Totoki, K. Tatsuno, S. Yamamoto, Y. Arai, F. Hosoda, S. Ishikawa, S. Tsutsumi, K. Sonoda, H. Totsuka, T. Shirakihara, H. Sakamoto, L. Wang, H. Ojima, K. Shimada, T. Kosuge, T. Okusaka, K. Kato, J. Kusuda, T. Yoshida, H. Aburatani, and T. Shibata, "High-resolution characterization of a hepatocellular carcinoma genome," *Nat. Genet.*, vol. 43, no. 5, pp. 464–469, May 2011.

[165] K. Schulze, S. Imbeaud, E. Letouzé, L. B. Alexandrov, J. Calderaro, S. Rebouissou, G. Couchy, C. Meiller, J. Shinde, F. Soysouvanh, A.-L. Calatayud, R. Pinyol, L. Pelletier, C. Balabaud, A. Laurent, J.-F. Blanc, V. Mazzaferro, F. Calvo, A. Villanueva, J.-C. Nault, P. Bioulac-Sage, M. R. Stratton, J. M. Llovet, and J. Zucman-Rossi, "Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets," *Nat. Genet.*, vol. 47, no. 5, pp. 505–511, May 2015.

[166] A. Ally, M. Balasundaram, R. Carlsen, E. Chuah, A. Clarke, N. Dhalla, R. A. Holt, S. J. Jones, D. Lee, Y. Ma, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, D. Cheung,

T. Wong, D. Brooks, A. G. Robertson, R. Bowlby, K. Mungall, S. Sadeghi, L. Xi, K. Covington, E. Shinbrot, D. A. Wheeler, R. A. Gibbs, L. A. Donehower, L. Wang, J. Bowen, J. M. Gastier-Foster, M. Gerken, C. Helsel, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, S. B. Gabriel, M. Meyerson, C. Cibulskis, B. A. Murray, J. Shih, R. Beroukhim, A. D. Cherniack, S. E. Schumacher, G. Saksena, C. S. Pedamallu, L. Chin, G. Getz, M. Noble, H. Zhang, D. Heiman, J. Cho, N. Gehlenborg, G. Saksena, D. Voet, P. Lin, S. Frazer, T. Defreitas, S. Meier, M. Lawrence, J. Kim, C. J. Creighton, D. Muzny, H. Doddapaneni, J. Hu, M. Wang, D. Morton, V. Korchina, Y. Han, H. Dinh, L. Lewis, M. Bellair, X. Liu, J. Santibanez, R. Glenn, S. Lee, W. Hale, J. S. Parker, M. D. Wilkerson, D. N. Hayes, S. M. Reynolds, I. Shmulevich, W. Zhang, Y. Liu, L. Iype, H. Makhlouf, M. S. Torbenson, S. Kakar, M. M. Yeh, D. Jain, D. E. Kleiner, D. Jain, R. Dhanasekaran, H. B. El-Serag, S. Y. Yim, J. N. Weinstein, L. Mishra, J. Zhang, R. Akbani, S. Ling, Z. Ju, X. Su, A. M. Hegde, G. B. Mills, Y. Lu, J. Chen, J.-S. Lee, B. H. Sohn, J. J. Shim, P. Tong, H. Aburatani, S. Yamamoto, K. Tatsuno, W. Li, Z. Xia, N. Stransky, E. Seiser, F. Innocenti, J. Gao, R. Kundra, H. Zhang, Z. Heins, A. Ochoa, C. Sander, M. Ladanyi, R. Shen, A. Arora, F. Sanchez-Vega, N. Schultz, K. Kasaian, A. Radenbaugh, K.-D. Bissig, D. D. Moore, Y. Totoki, H. Nakamura, T. Shibata, C. Yau, K. Graim, J. Stuart, D. Haussler, B. L. Slagle, A. I. Ojesina, P. Katsonis, A. Koire, O. Lichtarge, T.-K. Hsu, M. L. Ferguson, J. A. Demchok, I. Felau, M. Sheth, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. Zhang, C. M. Hutter, H. J. Sofia, R. G. Verhaak, S. Zheng, F. Lang, S. Chudamani, J. Liu, L. Lolla, Y. Wu, R. Naresh, T. Pihl, C. Sun, Y. Wan, C. Benz, A. H. Perou, L. B. Thorne, L. Boice, M. Huang, W. K. Rathmell, H. Noushmehr, F. P. Saggioro, D. P. da Cunha Tirapelli, C. G. C. Junior, E. D. Mente, O. de Castro Silva, F. A. Trevisan, K. J. Kang, K. S. Ahn, N. H. Giama, C. D. Moser, T. J. Giordano, M. Vinco, T. H. Welling, D. Crain, E. Curley, J. Gardner, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, R. Kelley, J.-W. Park, V. S. Chandan, L. R. Roberts, O. F. Bathe, C. H. Hagedorn, J. T. Auman, D. R. O'Brien, J.-P. A. Kocher, C. D. Jones, P. A. Mieczkowski, C. M. Perou, T. Skelly, D. Tan, U. Veluvolu, S. Balu, T. Bodenheimer, A. P. Hoyle, S. R. Jefferys, S. Meng, L. E. Mose, Y. Shi, J. V. Simons, M. G. Soloway, J. Roach, K. A. Hoadley, S. B. Baylin, H. Shen, T. Hinoue, M. S. Bootwalla, D. J. V. D. Berg, D. J. Weisenberger, P. H. Lai, A. Holbrook, M. Berrios, and P. W. Laird, "Comprehensive and integrative genomic characterization of hepatocellular carcinoma," *Cell*, vol. 169, no. 7, pp. 1327–1341, 2017.

[167] J. Calderaro, G. Couchy, S. Imbeaud, G. Amaddeo, E. Letouzé, J.-F. Blanc, C. Laurent, Y. Hajji, D. Azoulay, P. Bioulac-Sage, J.-C. Nault, and J. Zucman-Rossi, "Histological subtypes of hepatocellular carcinoma are related to gene mutations and molecular tumour classification," *J. Hepatol.*, vol. 67, no. 4, pp. 727–738, Oct. 2017.

[168] K. Chaudhary, O. B. Poirion, L. Lu, S. Huang, T. Ching, and L. X. Garmire, "Multi-modal meta-analysis of 1494 hepatocellular carcinoma samples reveals significant impact of consensus driver genes on phenotypes," *Clin. Cancer Res.*, Sept. 2018.

[169] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep learning based multi-omics integration robustly predicts survival in liver cancer," *Clin. Cancer Res.*, p. clincanres.0853.2017, Jan. 2017.

[170] M. Kojiro, "Histopathology of liver cancers," *Best Pract. Res. Clin. Gastroenterol.*, vol. 19, no. 1, pp. 39–62, Feb. 2005.

[171] M. Schlageter, L. M. Terracciano, S. D'Angelo, and P. Sorrentino, "Histopathology of hepatocellular carcinoma," *World J. Gastroenterol.*, vol. 20, no. 43, pp. 15 955–15 964, Nov. 2014.

[172] J. Cheng, J. Zhang, Y. Han, X. Wang, X. Ye, Y. Meng, A. Parwani, Z. Han, Q. Feng, and K. Huang, "Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis," *Cancer Res.*, vol. 77, no. 21, pp. e91–e100, Nov. 2017.

[173] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Velázquez Vega, D. J. Brat, and L. A. D. Cooper, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Natl. Acad. Sci. U. S. A.*, p. 201717139, Mar. 2018.

[174] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM," *Bioinformatics*, vol. 26, no. 12, pp. i237–45, June 2010.

[175] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, Xiaojun Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, June 2009, pp. 1107–1110.

[176] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PLoS One*, vol. 12, no. 6, p. e0177544, June 2017.

[177] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *Image Analysis and Recognition*. Springer International Publishing, 2018, pp. 737–744.

[178] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.*, vol. 23, no. 4, pp. 291–299, Aug. 2001.

[179] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-fei, "Imagenet: A large-scale hierarchical image database," in *In CVPR*, 2009.

[180] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *J. Mach. Learn. Res.*, vol. 2, no. Dec, pp. 125–137, 2001.

[181] M. J. Pencina and R. B. D'Agostino, "Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation," *Stat. Med.*, vol. 23, no. 13, pp. 2109–2123, July 2004.

[182] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, and V. C. Raykar, "On ranking in survival analysis: Bounds on the concordance index," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 1209–1216.

[183] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[184] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[185] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, Feb. 1979.

[186] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, Apr. 2015.

[187] A. Rastogi, "Changing role of histopathology in the diagnosis and management of hepatocellular carcinoma," *World J. Gastroenterol.*, vol. 24, no. 35, pp. 4000–4013, Sept. 2018.

[188] L.-X. Qin, "The prognostic molecular markers in hepatocellular carcinoma," p. 385, 2002.

[189] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sept. 2017.

[190] I. Park and H.-S. Lee, "EphB/ephrinB signaling in cell adhesion and migration," *Mol. Cells*, vol. 38, no. 1, pp. 14–19, Jan. 2015.

[191] H.-Q. Xi, X.-S. Wu, B. Wei, and L. Chen, "Eph receptors and ephrins as targets for cancer therapy," *J. Cell. Mol. Med.*, vol. 16, no. 12, pp. 2894–2909, Dec. 2012.

[192] R.-X. Li, Z.-H. Chen, and Z.-K. Chen, "The role of EPH receptors in cancer-related epithelial-mesenchymal transition," *Chin. J. Cancer*, vol. 33, no. 5, pp. 231–240, May 2014.

[193] P. P. Le, J. R. Friedman, J. Schug, J. E. Brestelli, J. Brandon Parker, I. M. Bochkis, and K. H. Kaestner, "Glucocorticoid Receptor-Dependent gene regulatory networks," *PLoS Genet.*, vol. 1, no. 2, p. e16, Aug. 2005.

[194] K. M. Mueller, J.-W. Kornfeld, K. Friedbichler, L. Blaas, G. Egger, H. Esterbauer, P. Hasselblatt, M. Schlederer, S. Haindl, K.-U. Wagner, D. Engblom, G. Haemmerle, D. Kratky, V. Sexl, L. Kenner, A. V. Kozlov, L. Terracciano, R. Zechner, G. Schuetz, E. Casanova, J. A. Pospisilik, M. H. Heim, and R. Moriggl, "Impairment of hepatic growth hormone and glucocorticoid receptor signaling causes steatosis and hepatocellular carcinoma in mice," *Hepatology*, vol. 54, no. 4, pp. 1398–1409, Oct. 2011.

[195] D. J. J. Waugh and C. Wilson, "The interleukin-8 pathway in cancer," *Clin. Cancer Res.*, vol. 14, no. 21, pp. 6735–6741, Nov. 2008.

[196] A. Benedicto, I. Romayor, and B. Arteta, "Role of liver ICAM-1 in metastasis," *Oncol. Lett.*, vol. 14, no. 4, pp. 3883–3892, Oct. 2017.

[197] N. Goossens, X. Sun, and Y. Hoshida, "Molecular classification of hepatocellular carcinoma: potential therapeutic implications," *Hepat Oncol*, vol. 2, no. 4, pp. 371–379, 2015.

[198] Y. Hoshida, S. M. B. Nijman, M. Kobayashi, J. A. Chan, J.-P. Brunet, D. Y. Chiang, A. Villanueva, P. Newell, K. Ikeda, M. Hashimoto, G. Watanabe, S. Gabriel, S. L. Friedman, H. Kumada, J. M. Llovet, and T. R. Golub, "Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma," *Cancer Res.*, vol. 69, no. 18, pp. 7385–7392, Sept. 2009.

[199] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, A. Madabhushi, and F. González, "High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection," *PLoS One*, vol. 13, no. 5, p. e0196828, May 2018.

[200] D. B. Nagarkar, E. Mercan, D. L. Weaver, T. T. Brunyé, P. A. Carney, M. H. Rendi, A. H. Beck, P. D. Frederick, L. G. Shapiro, and J. G. Elmore, "Region of interest identification and diagnostic agreement in breast pathology," *Mod. Pathol.*, vol. 29, no. 9, pp. 1004–1011, Sept. 2016.

[201] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. W. M. van der Laak, "Stain specific standardization of Whole-Slide histopathological images," *IEEE Trans. Med. Imaging*, vol. 35, no. 2, pp. 404–415, Feb. 2016.

[202] M. W. Lafarge, J. P. W. Pluim, K. A. J. Eppenhof, P. Moeskops, and M. Veta, "Domain-Adversarial neural networks to address the appearance variability of histopathology images," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer International Publishing, 2017, pp. 83–91.

[203] T. J. Alhindi, S. Kalra, K. H. Ng, A. Afrin, and H. R. Tizhoosh, "Comparing LBP, HOG and deep features for classification of histopathology images," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–7.

[204] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," Feb. 2018.

[205] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[206] A. C. Quiros, R. Murray-Smith, and K. Yuan, "Pathology GAN: Learning deep representations of cancer tissue," July 2019.

[207] H. Uzunova, S. Schultz, H. Handels, and J. Ehrhardt, "Unsupervised pathology detection in medical images using conditional variational autoencoders," pp. 451–461, 2019.

Appendix A

GEOlimma: Differential Expression Analysis and Feature Selection Using Pre-Existing Microarray Data

A.1 Abstract

Differential expression and feature selection analyses are essential steps for the development of accurate diagnostic/prognostic classifiers of complicated human diseases using transcriptomics data. These steps are particularly challenging due to the curse of dimensionality and the presence of technical and biological noise. A promising strategy for overcoming these challenges is the incorporation of pre-existing transcriptomics data in the identification of differentially expressed (DE) genes. This approach has the potential to improve the quality of selected genes, increase classification performance, and enhance biological interpretability. While a number of methods have been developed that use pre-existing data for differential expression analysis, existing methods do not leverage the identities of experimental conditions to create a robust metric for identifying DE genes.

In this study, we propose a novel differential expression and feature selection method—GEOlimma—which combines pre-existing microarray data from the Gene Expression Omnibus (GEO) with the widely-applied Limma method for differential expression analysis. We first quantify differential gene expression across 2481 pairwise comparisons from 602 curated GEO Datasets, and we convert differential expression frequencies to DE prior probabilities. Genes with high DE prior probabilities show enrichment in cell growth and death, signal transduction, and cancer-related biological pathways, while genes with low prior probabilities were enriched in sensory system pathways. We then applied GEOlimma to four differential expression comparisons within two human disease datasets and performed differential expression, feature selection, and supervised classification analyses. Our results suggest that use of GEOlimma provides greater experimental power to detect DE genes compared to Limma,

due to its increased effective sample size. Furthermore, in a supervised classification analysis using GEOlimma as a feature selection method, we observed similar or better classification performance than Limma given small, noisy subsets of an asthma dataset.

Our results demonstrate that GEOlimma is a more effective method for differential gene expression and feature selection analyses compared to the standard Limma method. Due to its focus on gene-level differential expression, GEOlimma also has the potential to be applied to other high-throughput biological datasets.

A.2 Introduction

DNA microarrays and RNA sequencing (RNA-Seq) have become indispensable experimental tools for characterizing the effects of biological interventions on genome-wide gene expression ("transcriptomics") [97] [98]. Applications of these tools have been transformative in many areas of biological research, including cancer biology, biomarker discovery, and drug target identification [99] [100] [101]. These applications often involve differential expression analysis: the isolation of differentially expressed (DE) genes between healthy and disease conditions. Knowledge of DE genes facilitates the discovery of causative genes and gene pathways for a disease of interest. For example, many studies of carcinogenesis focus on identifying the genes directly responsible for promoting cancer occurrence ("driver genes") out of all DE genes [102] . Furthermore, DE gene identification is an important first step for disease biomarker discovery. The discovery of biomarkers from transcriptomics data typically involves selecting the most discriminative genes between a healthy and diseased state or between different disease states [103] . A comprehensive list of DE genes provides a biologically plausible set of candidates for these discriminative genes and can greatly streamline the search [104]. Common applications of transcriptomics-derived biomarkers include predicting diagnosis, prognosis, and therapeutic response for a disease of interest through a process

known as supervised classification [105]. In this context, DE gene identification can be viewed as a means of performing feature selection for classification. In general, feature selection is a process for dimensionality reduction that removes redundant or irrelevant features (genes), reduces classification model complexity, and improves classification performance [106].

Despite their widespread use for DE gene identification, transcriptomics data are notorious for their inclusion of technical and biological noise [107]. This noise complicates differential expression analysis by reducing the accuracy of DE gene identification relative to other assays (e.g., real-time or quantitative PCR [108]), lowering the reproducibility of experiments conducted on different platforms [109], and reducing the statistical power associated with the detection of DE genes at a particular fold change [110]. A straightforward strategy for mitigating the effects of noise is to increase the number of replicates assayed ("sample size") for each condition of interest. However, this practice can be cost prohibitive or even impossible for conditions with limited sample availability. Furthermore, even with larger sample sizes, transcriptomics data pose a considerable challenge to feature selection methods due to the curse of dimensionality. Specifically, it is well known that optimal fitting of classification models (including the selection of features) breaks down when the feature dimensionality is substantially larger than the sample size [43].

One promising solution for the above challenges is to incorporate prior biological knowledge into differential expression and feature selection analyses [111]. This Bayesian approach can mitigate problems associated with a small sample size [112], while also improving biological interpretability of the resulting DE genes/features [106]. Prior biological knowledge for transcriptomics data can take several forms, including pre-existing transcriptomics data from other studies, data from complementary high-throughput assays (e.g., chromatin immunoprecipitation or protein-protein interactions), and gene functional annotation (e.g., Gene Ontology [113] [114] or KEGG [115] [116]). For the

purposes of this study, we will focus on the first type of knowledge, although we note that analytical methods are available to incorporate the other types as well [117] [118]. Thanks to functional genomics repositories like the Gene Expression Omnibus (GEO)[119] [120] and ArrayExpress [121], transcriptomics data from over 2.5 million samples are publicly available. Furthermore, the size of this resource is growing exponentially, with numbers of samples in GEO doubling every 3-4 years.

Over the last 15 years, a number of methods have been developed that use prior knowledge in the form of transcriptomics data to inform differential expression analyses [122] [123] [124] [125] [126]. However, these methods typically either ignore the identities of the many experimental conditions in the pre-existing data, or they do not leverage these identities to create a rigorous statistical metric for identifying DE genes. For example, the SVD Augmented Gene expression Analysis Tool (SAGAT) uses singular value decomposition (SVD) to extract transcriptional modules from pre-existing DNA microarray data [122]. These modules, which contain no information regarding assayed conditions, are then incorporated into a statistical analog of the two-sample t-test to improve the accuracy of DE gene identification. In contrast, a very recent study made direct use of the experimental conditions in pre-existing data to characterize empirical prior probabilities of differential expression [126]. However, although these prior probabilities were predictive of differential expression patterns, they were not explicitly utilized in a Bayesian statistical framework for identifying DE genes. Relatedly, although there have been many studies contributing novel or adapted feature selection methodologies for classification of biomedical data [127] [128] [129] [61] [130], to our knowledge no method combines an experimental condition-aware analysis of pre-existing data with a statistically principled means of feature selection.

To address these shortcomings, we propose a novel differential expression and feature selection approach—GEOlimma—that leverages pre-existing

GEO-derived transcriptomics data. As described below, our proposed method modifies the popular Linear Models for Microarray and RNA-Seq Data ("Limma") method [131][132] . Specifically, GEOlimma incorporates empirical prior probabilities of differential expression (DE prior probabilities) in a Bayesian statistical test for DE genes. We first describe the computation and biological characterization of DE prior probabilities from a large collection of pre-existing DNA microarray experiments from GEO. Next, we apply GEOlimma and Limma to four benchmark differential expression comparisons from two validation datasets. Our results demonstrate a substantial increase in experimental power for identifying DE genes due to use of GEOlimma. Finally, we explore GEOlimma's ability to improve feature selection for classification across the four benchmark comparisons.

A.3 Materials & Methods

A.3.1 GEOlimma Method Formulation

We developed the GEOlimma method by combining the widely-used differential expression (DE) analysis method Limma, which is typically used to analyze gene expression microarray and RNA-seq data and assess differential expression between biological conditions. Limma uses empirical Bayesian methods to provide stable DE predictions, which is particularly useful when the number of sample replicates is small. However, one simplifying assumption made by Limma is that the DE prior probabilities for each gene are identical (set 0.01 by default). GEOlimma combines the Bayesian nature of Limma with gene-level DE prior probabilities calculated from large-scale microarray datasets to better select genes that are biologically relevant to a comparison of interest.

The Gene Expression Omnibus (GEO) is a public data repository for high-throughput gene expression data including microarray and RNA-seq data [120]. GEO DataSets (GDS) are a subset of the repository that store curated gene expression datasets, along with the original data (GEO Series) and experimental platform information. GPL570, also known as the `HG-U133_Plus_2`

Affymetrix Human Genome U133 Plus 2.0 Array, is one of the best-represented human genome microarray platforms in GEO, with 149,049 samples available (as of June 7, 2019). GPL570 measures over 47,000 human transcripts, which consist of the Human Genome U133 Set plus 6,500 additional genes. In this study, we downloaded all 602 GPL570 GEO DataSets (GDS) (current as of June 7, 2019). Specifically, for each dataset we obtained normalized, log-transformed expression values at the probeset level. We then mapped these probesets to the non-redundant Entrez Gene IDs (provided by the Bioconductor R package hgu133plus2.db) and obtained gene-level expression values by computing medians across any probe sets mapping to the same gene. With the minimum requirement of 5 samples in each group, we performed pairwise DE analysis among the largest possible collection of non-overlapping sample groups from each GDS experiment. Specifically, for each DE comparison, we applied the Limma moderated t-test [133] (using the "lmFit" and "eBayes" functions) to calculate differential expression p-values for each gene. Given a list of p-values for a particular comparison, we adjusted for multiple hypothesis testing using the Benjamini-Hochberg (BH) procedure [86]. Genes with adjusted p-values (false discovery rates or FDRs) $\leq$0.05 for a given pairwise comparison were considered DE for that comparison. We calculated the DE frequencies across all comparisons for each gene and converted these frequencies to DE prior probabilities (P(DE)) as follows:

$$P(DE_i) = \frac{\sum_j I(Adj.P_{ij} \leq 0.05)}{M} \tag{A.1}$$

where $i \in \{1, \ldots, N\}$ indexes each gene, $j \in \{1, \ldots, M\}$ indexes each comparison, $Adj.P_{ij}$ represents the FDR for the i-th gene in the j-th comparison, and I($\cdot$) is the indicator function.

We chose human asthma and cancer validation datasets present as GEO Series (GSE) but not as GEO DataSets (GDS), in order to avoid double counting data. The asthma dataset [134] consists of 404 total samples

(transformed lymphoblastoid cell lines) taken from 268 children afflicted with asthma and 136 healthy children. The cancer dataset [135] consists of 870 total bone marrow samples, of which 202, 164, and 69 are from individuals with acute myeloid leukemia (AML), myelodysplastic syndrome (MDS), and neither AML nor MDS, respectively. We considered the three possible comparisons between these three groups. In total, we evaluated four comparisons: Asthma vs Non-asthma, Nonleukemia vs AML, Nonleukemia vs MDS, and AML vs MDS.

For a given comparison, we compute GEOlimma DE posterior probabilities using Bayes' theorem:

$$P(DE_i \mid Data) = \frac{P(Data \mid DE_i)\, P(DE_i)}{P(Data)} \tag{A.2}$$

where Data represents the samples making up the given comparison, $P(Data \mid DE_i)$ denotes the likelihood of the Data, as calculated by limma [131], $P(DE_i)$ is the previously calculated DE prior probability, and $P(Data)$ is a normalization constant [131]. Given these posterior probabilities, we then calculate B scores (log odds of DE) for each gene as follows:

$$B_i = \log\left[\frac{P(DE_i \mid Data)}{1 - P(DE_i \mid Data)}\right] \tag{A.3}$$

We implemented GEOlimma as modified R functions based on code from the Limma package.

A.3.2 Enrichment Analysis for Gene Sets

To explore the DE prior probabilities biologically, we conducted KEGG Enrichment Analysis using the R package ClusterProfiler [136]. Specifically, we identified enriched KEGG pathways using the hypergeometric test in both the top and bottom 500 most/least frequently DE genes, separately. Pathways with BH-adjusted p-values less than 0.05 were considered significantly enriched. We used the Pathview R package [137] to visualize the location of DE genes in particular KEGG pathways.

A.3.3 Differential Expression Analysis

Evaluation datasets As described above, we downloaded the GSEs for two evaluation datasets from GEO. As with the GDS data, we mapped normalized, log-transformed expression values at the probeset level to non-redundant Entrez Gene IDs and consolidated expression values by computing medians across probe sets mapping to the same gene. We included all genes with unique probe mappings (20283 total) for subsequent analyses. For each of the four evaluation comparisons, we performed DE analysis on all samples using both GEOlimma and Limma. Genes were considered DE if their BH-adjusted p-value $\leq$0.05 (Limma) or their B score exceeded the smallest Limma B score for genes with adjusted p-value $\leq$0.05 (GEOlimma).

Sample Visualization To visualize samples, we first used Principal Component Analysis (PCA) to reduce the dimensionality of genes as features. We visualized the first two components of PCA. We further applied the t-Distributed Stochastic Neighbor Embedding (t-SNE) method to visualize the first 10 PCA components in 2 dimensions. t-SNE can reduce the dimensionality of data based on conditional probabilities that preserve local similarity. We used a t-SNE implementation that makes Barnes-Hut approximations, allowing it to be applied on large real-world datasets [138]. We set the perplexity to 15, and sample points were colored using the group information.

Experimental power To quantify the performance improvement achieved by GEOlimma vs Limma, we performed DE analysis on small sample size subsets for each comparison. As detailed below, we started with the minimum subset size at which the group proportions for a given comparison could be maintained and generated all non-overlapping sample subsets of this size. We then increased this subset size by the smallest possible sample increment and repeated the generation of subsets. For each sample subset, we first applied both GEOlimma and Limma and ranked genes by their corresponding B scores. Next, using the Limma DE genes previously identified from all samples as the

ground truth (see Results section for specific numbers), we applied the R package ROCR [139] to calculate Area under the ROC curves (AUCs) for the B score-ranked genes of each subset. We calculated the performance improvement of GEOlimma over Limma for each subset as the difference in AUC between the two methods. In addition, we converted these AUC improvements into gains in effective sample size by constructing and interpolating from a "standard curve" of mean Limma AUC values calculated across the full range of possible sample sizes. As an example, if GEOlimma delivered an AUC improvement of 0.1 over Limma for a subset of size 10, the GEOlimma effective sample size is simply the sample size of the standard curve corresponding to an AUC value 0.1 higher than the mean Limma AUC value for 10-sample subsets.

A.3.4 Supervised Classification

We performed supervised classification for each comparison in the evaluation datasets using both GEOlimma and Limma as feature selection methods. Scikit-learn (sklearn) [140] is a Python module implementing machine learning algorithms. It enables various tasks such as dimensionality reduction, classification, regression and model selection. The sklearn classification pipeline involves sequentially applying feature selection, classification, parameter optimization and model selection to yield final classification results. We first used the Python rpy2 module to build a connection between sklearn and the R language, followed by creating customized feature selection methods for Limma and GEOlimma which we compiled into the sklearn pipeline function. For classification training, we first sampled 10 subsets of 40 samples (20 from each of the two groups) at random and selected the 1000 genes with largest variance across these samples. Next, we fed data from each subset to the sklearn pipeline function and performed either Limma or GEOlimma-based feature selection by selecting subsets of 100-1000 genes (in increments of 100) with the highest B scores. We selected the Logistic Regression [141] classifier for classification. We also included L1 and L2 penalties as hyperparameters and applied 10-fold cross

validation to train the model and optimize the hyperparameters. We used classification AUC as the criterion to evaluate classification performance. A high AUC represents both high recall and high precision, which translate to low false positive and false negative rates. For classification testing, we sampled an additional 40 samples to evaluate the training models. We used a Wilcoxon signed-rank test to identify significant AUC differences between performing feature selection using Limma or GEOlimma.

A.4 Results

In this study, we developed a gene expression feature selection method, GEOlimma, in which gene-level differential expression (DE) prior probabilities were derived from large-scale microarray data freely available from the Gene Expression Omnibus (GEO). We first explored enriched biological pathways in genes with either high or low DE prior probabilities. We then applied GEOlimma to DE analysis and supervised classification tasks on a collection of four validation datasets.

A.4.1 Biological Analysis of DE Prior Probabilities

The goal of differential expression analysis is to identify differences in gene expression across biological conditions in order to discover functional genes and pathways involved in a biological process of interest. The Limma method [133] is an empirical Bayesian approach for identifying DE genes that has been widely applied. However, an important limitation of this method is that the prior probabilities for differential expression are set to be constant for all genes. This implies that all genes have the same chance of being expressed differently, which is not biologically realistic [126]. Therefore, we developed and applied GEOlimma, which uses a large collection of GEO datasets to compute gene level DE prior probabilities (see Methods section). We first downloaded the 602 GEO DataSets (GDS) currently available from the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array), followed by performing pairwise DE analysis among the largest possible collection of non-overlapping sample groups

(number of samples 5) from each GDS experiment. We identified DE genes using a Benjamini-Hochberg false discovery rate (FDR) threshold of 0.05. By repeating this procedure for every GDS, we calculated DE frequencies for 21025 distinct Entrez genes (20283 genes with unique gene mappings) across all experiments (2481 pairwise comparisons total) and converted these to prior probabilities of DE. Given gene-level DE prior probabilities, we can then compute posterior probabilities of DE for a given biological experiment using Bayes' theorem. Figure B.1 shows the distribution of DE prior probabilities, which ranged between 0.0048 and 0.1769 and appeared to have two modes. The median probability is 0.069, which we note is roughly seven times higher than the default constant prior probability used by Limma (0.01). Figure SFigure1 lists the top most frequently DE genes, including TUBA1A (tubulin alpha 1a), CD24, and SERPINB1 (serpin family B member 1), with DE prior probabilities of 0.1769, 0.1761, and 0.1693, respectively. The three least frequently DE genes were LOC102725116, TMCO5A (transmembrane and coiled-coil domains 5A), and LINC01492 (long intergenic non-protein coding RNA 1492), with DE prior probabilities of 0.0048, 0.0056, and 0.0060, respectively. Generally speaking, we hypothesize that genes with high prior probabilities of DE are more likely to be implicated in human disease and thus could function as biomarkers, while those with low DE prior probabilities represent constitutively expressed genes that are required for the maintenance of basic cellular functions (i.e., housekeeping genes).

In order to improve our biological understanding of the calculated DE prior probabilities, we performed gene set enrichment analysis (GSEA) based on KEGG pathways with the top 500 most and least frequently DE genes, respectively. Table B.1 lists significantly enriched pathways (BH-adjusted p-value $\leq$0.05), which include 19 pathways from the most frequently DE genes and 4 from the least frequently DE genes. The most significant pathway in the former category is hsa04110: Cell cycle (adjusted p = 7.83E-08); Figure B.2 illustrates the frequently DE genes mapped in this pathway. Two additional

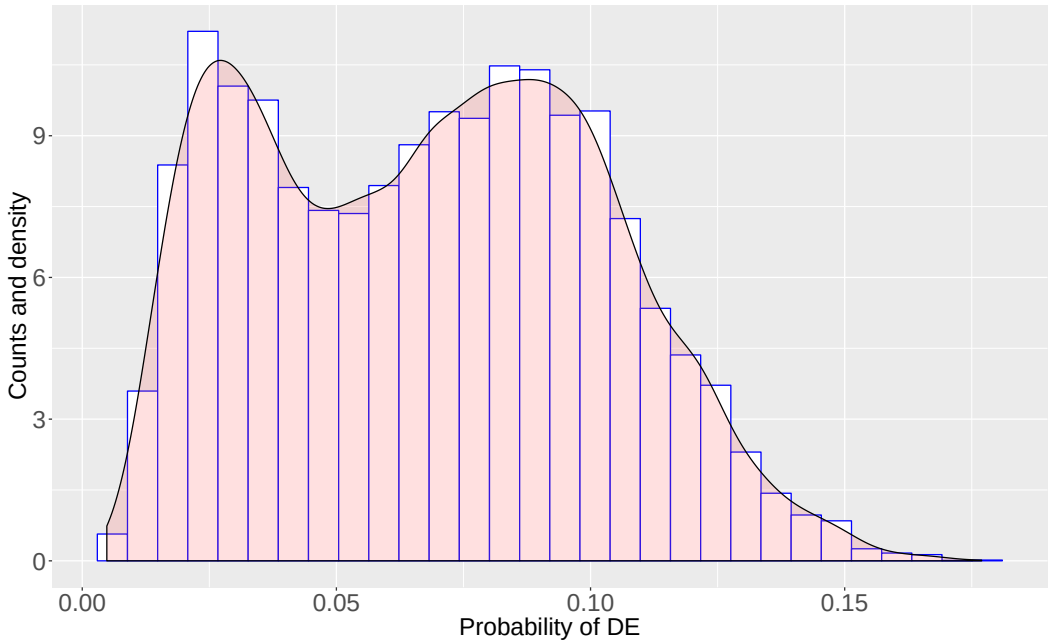


Fig. A.1: Distribution of DE prior probabilities for 20283 genes, calculated from 2481 pairwise comparisons made within 602 curated GEO Datasets.

pathways in this category directly related to cell growth and death include hsa04115: p53 signaling pathway and hsa04210: Apoptosis. We also identified six cancer-specific frequently DE pathways: hsa05222: Small cell lung cancer, hsa05206: MicroRNAs in cancer, hsa05218: Melanoma, hsa05202: Transcriptional misregulation in cancer, hsa05205: Proteoglycans in cancer, and hsa05220: Chronic myeloid leukemia. Finally, the two frequently DE pathways hsa04068: FoxO signaling pathway and hsa04668: TNF signaling pathway function in Signal transduction. We note that signal transduction pathways are involved in cell death mechanisms that function in colorectal carcinogenesis progression [142].

The 4 least frequently DE pathways include two sensory system pathways: hsa04740: Olfactory transduction and hsa04742: Taste transduction, Signaling molecules and interaction pathway. The other two significant pathways in this category were hsa04080: Neuroactive ligand-receptor interaction and hsa05320: Autoimmune thyroid disease. Our results suggest that genes belonging to these pathways show relatively stable expression across different biological conditions.

Table A.1: KEGG Enrichment Analysis of top 500 genes with high and low DE prior probabilities.

Pathway IDs	Description	GeneRatio	BgRatio	Pvalue	P value Adjustment	Source
hsa04110	Cell cycle	21/242	124/7528	2.94E-10	7.83E-08	HighPrior
hsa05222	Small cell lung cancer	13/242	93/7528	7.47E-06	9.94E-04	HighPrior
hsa04115	p53 signaling pathway	11/242	72/7528	1.61E-05	1.43E-03	HighPrior
hsa05169	Epstein-Barr virus infection	18/242	201/7528	7.63E-05	3.57E-03	HighPrior
hsa05206	MicroRNAs in cancer	23/242	299/7528	8.53E-05	3.57E-03	HighPrior
hsa05218	Melanoma	10/242	72/7528	9.09E-05	3.57E-03	10
HighPrior						
hsa05202	Transcriptional misregulation in cancer	17/242	186/7528	9.40E-05	3.57E-03	HighPrior
hsa04210	Apoptosis	14/242	136/7528	1.12E-04	3.71E-03	HighPrior
hsa05205	Proteoglycans in cancer	17/242	201/7528	2.42E-04	7.15E-03	HighPrior
hsa04068	FoxO signaling pathway	13/242	132/7528	3.05E-04	8.11E-03	HighPrior
hsa05418	Fluid shear stress and atherosclerosis	13/242	139/7528	5.05E-04	1.22E-02	HighPrior
hsa05220	Chronic myeloid leukemia	9/242	76/7528	6.87E-04	1.52E-02	HighPrior
hsa03030	DNA replication	6/242	36/7528	8.98E-04	1.84E-02	HighPrior
hsa05130	Pathogenic Escherichia coli infection	7/242	55/7528	1.77E-03	3.36E-02	HighPrior
hsa04540	Gap junction	9/242	88/7528	1.97E-03	3.50E-02	HighPrior
hsa01524	Platinum drug resistance	8/242	73/7528	2.25E-03	3.74E-02	HighPrior
hsa05167	Kaposi sarcoma-associated herpesvirus infection	14/242	186/7528	2.60E-03	3.83E-02	HighPrior
hsa04380	Osteoclast differentiation	11/242	128/7528	2.67E-03	3.83E-02	HighPrior
hsa04668	TNF signaling pathway	10/242	110/7528	2.74E-03	3.83E-02	HighPrior
hsa04740	Olfactory transduction	17/68	448/7528	2.91E-07	3.08E-05	LowPrior
hsa04742	Taste transduction	8/68	83/7528	6.70E-07	3.55E-05	LowPrior
hsa04080	Neuroactive ligand-receptor interaction	14/68	338/7528	1.40E-06	4.96E-05	LowPrior
hsa05320	Autoimmune thyroid disease	4/68	53/7528	1.28E-03	3.39E-02	LowPrior

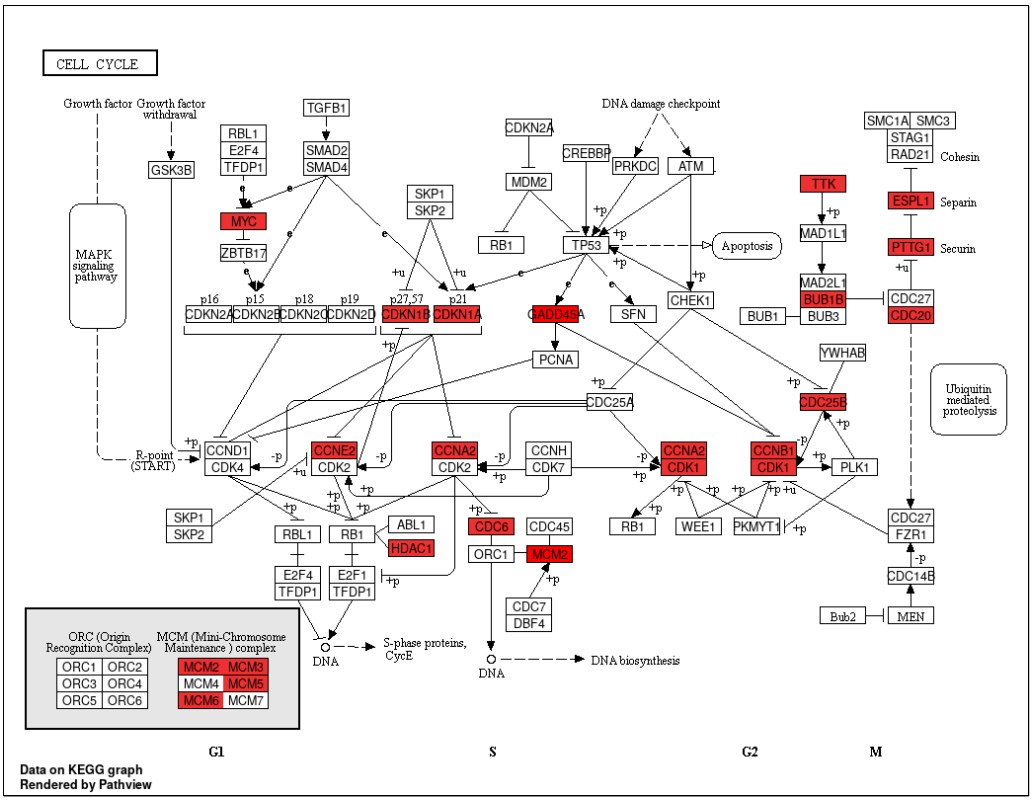


Fig. A.2: Significantly enriched Cell Cycle pathway from genes with high DE prior probabilities. The red shaded blocks indicate genes with high prior probabilities.

A.4.2 GEOLimma method application on four validation datasets

We investigated the utility of gene-specific DE prior probabilities by performing DE analysis with GEOlimma in four evaluation datasets. Specifically, we selected two GEO series—GSE8052 and GSE15061— from platform GPL570 that enabled four DE comparisons to be made. Importantly, neither of these datasets was represented by a GEO GDS, meaning that none of the resulting comparisons were involved in DE prior probability computation. The four comparisons include Asthma vs Non-asthma (GSE8052) and three comparisons from GSE15061: Nonleukemia (Nonleuk) vs Myelodysplastic syndrome (MDS), Nonleuk vs acute myeloid leukemia (AML), and AML vs MDS. The probes for each of the datasets are represented by 20283 genes with unique mappings. Any genes without available DE prior probabilities were assigned the median value of all prior probabilities. We first identified DE genes

using GEOlimma as well as the standard Limma method. This allowed us to compare the two methods, as well as characterize the extent of differential expression present in each comparison. For Limma, we considered genes to be DE if their BH-adjusted p-value ≤ 0.05. In contrast, as GEOlimma enables the calculation of a modified B score only (see Methods), we selected a B score threshold for GEOlimma significance based on the smallest Limma B score for which the Limma adjusted p-value ≤ 0.05. Using these criteria, we identified DE genes based on all relevant samples for each of the four comparisons described above. To assess the effect of small sample sizes on GEOlimma/Limma performance, we also randomly sampled 10 subsets of 40 samples (20 in each class) for each comparison and calculated the mean and standard deviation of the number of DE genes across these subsets using both methods. Table B.2 lists details of each DE comparison along with summaries of our analysis results using both Limma and GEOlimma. We note that for the Asthma comparison, there are no significant DE genes based on all samples (as well as in subsets) using the Limma method. Therefore, we were not able to quantify the number of DE genes for this comparison using GEOlimma. In the remaining three comparisons, our results demonstrate that GEOlimma identifies more DE genes than Limma when applied to either all samples or 40-sample subsets. Figure B.3 A helps illustrate why this is, by examining the distributions of Limma and GEOlimma B scores for the Asthma comparison. Despite the lack of significant DE genes in this comparison, use of GEOlimma results in a wider B score distribution with a marked shift to higher values compared to Limma. This difference is due to the diverse set of gene-specific DE prior probabilities used by GEOlimma, the median value of which is substantially higher than the constant value used by Limma. The potential increase in numbers of DE genes identified by GEOlimma also suggests that use of a small constant DE prior probability may result in overly conservative DE gene identification. In our PCA and t-SNE visualizations of all samples (Figures B.3 B and C), we note the lack of clear separation

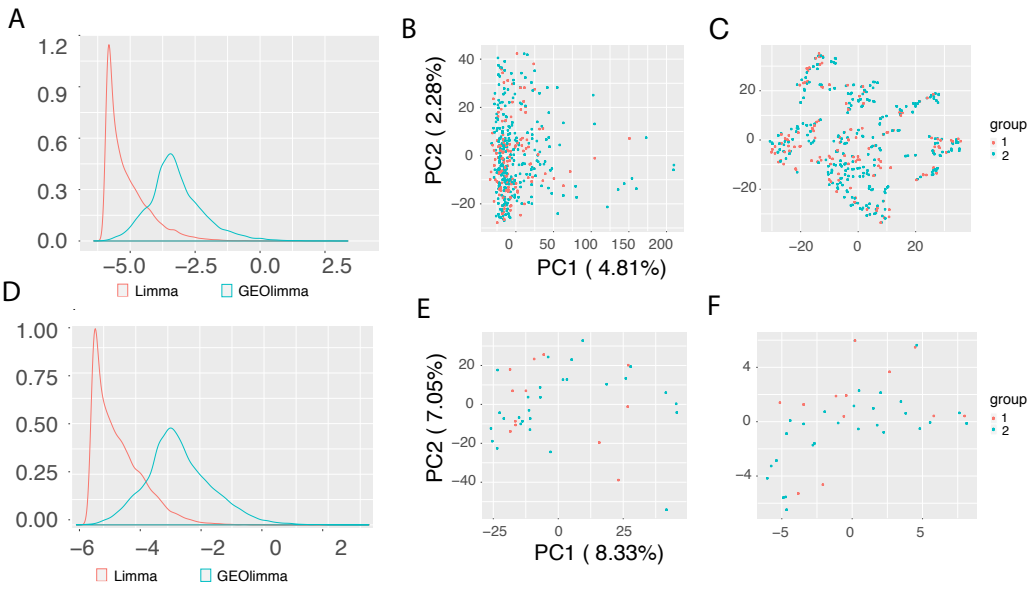


Fig. A.3: B score change and sample visualizations of asthma dataset. The top figures are generated from all samples; the bottom figures are drawn using a random subset with 40 samples. (A) and (D) depict Limma and GEOlimma B score distributions of all genes, (B) and (E) show PCA visualizations of samples, and (C) and (F) show t-SNE sample visualizations.

between the Asthma and Non-asthma groups, which helps explain why no significant DE genes were detected. Figures B.3 D, E, and F show the same information for a randomly selected subset of 40 samples. We note that the B scores have a similar distribution as that of all samples.

Table A.2: Differential expression comparison details Limma and GEOlimma DE gene counts for all samples and 10 subsets of 40 samples of each comparison.

Datasets	Samples	Limma DEGs	GEOlimma DEGs	Limma DEGs*	GEOlimma DEGs*
Asthma vs non-asthma [134]	268:136	0	–	0	–
Nonleuk vs MDS [135]	164:69	2619	5823	98.5±161.4	404.3±600.9
Nonleuk vs AML [135]	202:69	8610	13379	2788.9±901.5	5879.3±1415.2
AML vs MDS [135]	164:202	10975	15337	2881.5±1068.7	6017±1666.7

* indicates DE tests of subset samples

When looking at the top 20 most significantly DE genes for each comparison, we noted that use of GEOlimma changes the order of these genes compared to Limma, with an overall higher average B score (Figures SFigure2).

To further explore this phenomenon, we counted the genes in common for the top 100 to 1000 most significantly DE genes between GEOlimma and Limma across 10 randomly selected 40-sample subsets for each comparison. The average overlap percentages were 67.3% for the Asthma comparison, 87% for Nonleuk vs MDS, while over 95% for both AML vs MDS (95.2%) and Nonleuk vs AML (95.5%)(Figure SFigure3). These results suggest that GEOlimma DE prior probabilities have a larger effect on the resulting DE gene list for datasets showing a more modest overall degree of differential expression (e.g., Asthma and Nonleuk vs MDS comparisons).

In order to explore the practical benefits of using GEOlimma, we compared the accuracy of DE gene identification between GEOlimma and Limma for each of the four DE comparisons. For each comparison, we first performed DE analysis on all samples using Limma, with the resulting significant DE genes (n = 1241 [FDR $\leq$ 0.4], 2619 [FDR $\leq$ 0.05], 8610 [FDR $\leq$ 0.05], and 10975 [FDR $\leq$ 0.05] for the Asthma, Nonleuk vs MDS, Nonleuk vs AML, and AML vs MDS comparisons) being treated as the ground truth. We note that we relaxed the significance thresholds for the Asthma comparison in order to include a sufficient number of DE genes for subsequent evaluation. Next, we randomly generated non-overlapping sample subsets for each comparison based on the minimum sample size at which the group proportions of the dataset could be maintained. For example, as GSE8052 contains 66% Asthma and 34% Non-asthma samples, the smallest sample size considered was 6 (4 Asthma, 2 Non-asthma) in order to ensure 2 samples per group. We then increased this sample size in increments of 3 to also consider subsets of 9, 12, and 15 samples. We then applied both GEOlimma and Limma on each of the sample subsets to determine which method best recovered the ground truth. Specifically, we used the R package ROCR [139] to compute areas under the receiver operating characteristic curve (AUCs) given the GEOlimma/Limma B scores and the ground truth. Figure B.4 depicts the AUC improvement of GEOlimma over

Limma for all four comparisons. Notably, GEOlimma consistently increases the average AUC for each of the subset sizes, with an overall average AUC improvement of 0.04. Furthermore, in the three comparisons made within GSE15061, GEOlimma increases AUC for every subset tested. Interestingly, the AUC improvement is largest for the smallest sample sizes evaluated and decreases slightly as sample size increases. This further supports the assertion that GEOlimma has a bigger impact on datasets with more modest expression differences (as would result from a small sample size). To confirm that these improvements result specifically from the DE prior probabilities learned using publicly available GPL570 data, we randomly shuffled the prior probabilities and repeated the above analysis. As seen in Figure SFfigure4, GEOlimma using randomized prior probabilities consistently decreases AUC compared to Limma.

To quantify the experimental power gained by using GEOlimma, we converted AUC values into effective sample size. Specifically, for each of the evaluation datasets, we first calculated AUCs resulting from applying Limma to all non-overlapping sample subsets ranging in size from the minimum number needed to maintain group proportions (described above) to the total number of replicates. For example, in the Asthma comparison we considered all subsets of size 6 to 402 in increments of 3. These AUCs enabled us to fit a "standard curve" for each comparison, from which we could interpolate the mean number of samples gained by using GEOlimma given initial numbers of 6, 9, 12, and 15 (Asthma) samples. Figure SFigure5 presents the AUC standard curves and Table B.3 summarizes the distribution of GEOlimma effective sample sizes for each comparison. Overall, GEOlimma leads to a substantial increase in mean effective sample sizes, particularly when applied to smaller subsets, where we observed gains of 157-288% for the smallest sample sizes evaluated for each comparison. The Asthma comparison shows the largest relative increases across all subsets, with the mean GEOlimma effective sample size more than doubling that of Limma even for the largest subset tested (m = 15). These results

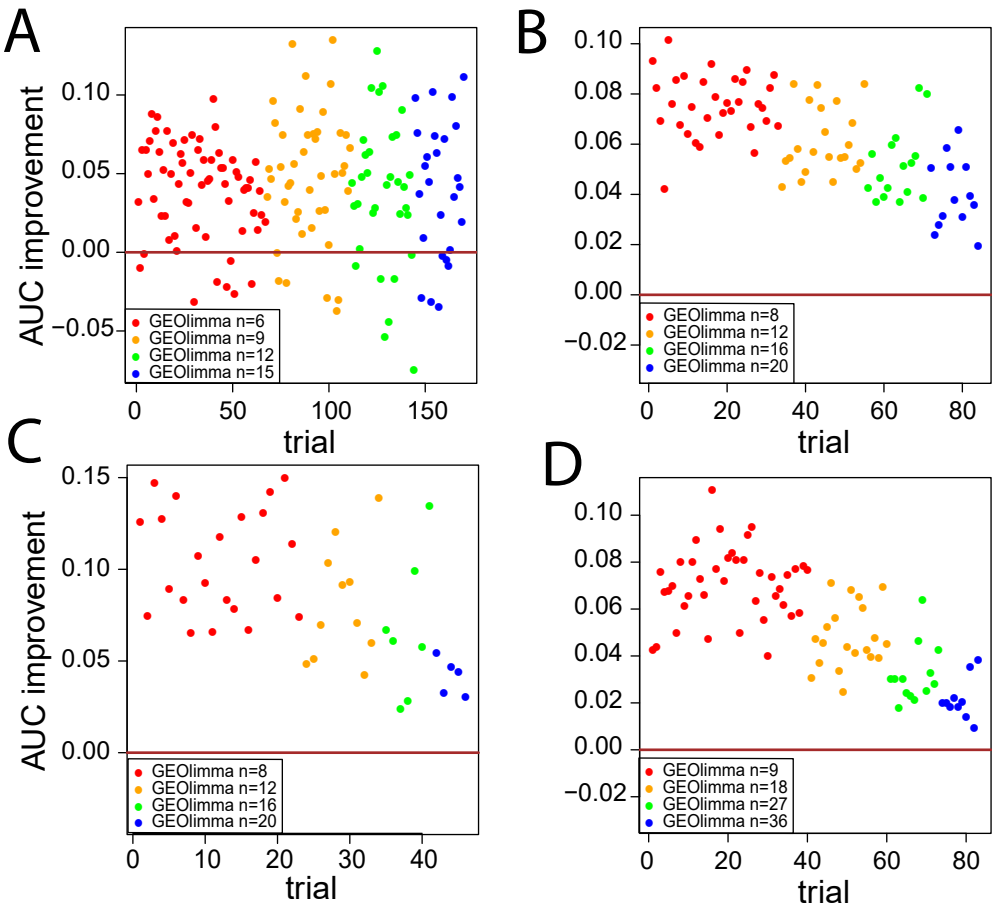


Fig. A.4: Area under the ROC curve (AUC) improvement of GEOlimma over Limma for identifying DE genes from a range of data subset sizes: A) Asthma vs Non-asthma comparison, B) Nonleukemia vs AML comparison, C) Nonleukemia vs MDS comparison, D) AML vs MDS comparison.

demonstrate the gains in experimental power for DE gene discovery that are possible with the use of GEOlimma.

A.4.3 Classification performance using GEOLimma feature selection method

Feature selection is a critical step in supervised classification for diagnosis, prognosis and treatment. Here we compare the abilities of GEOlimma and Limma as feature selection methods to perform accurate classification on the four evaluation datasets. To focus on the most challenging classification tasks for each comparison, we randomly sampled subsets of size 20 from each of the two groups. Specifically, we generated 10 pairs of subsets for training, with each pair containing 40 total samples (20 per group). In the same manner, we also

generated an additional 10 pairs of samples for testing. During training, we performed 10-fold cross-validation to estimate model performance. Given the large numbers of genes present in these datasets, we focused on the 1000 genes with the highest variance across all samples within each comparison. Within these 1000 genes, we selected the top 100-1000, in increments of 100, using either Limma or GEOlimma and performed classification using a logistic regression (LR) classifier. For each sampled subset, we applied a one-sided (hypothesis: GEOlimma AUC > Limma AUC) paired Wilcoxon test to compare the AUC differences between GEOlimma and Limma at each feature size (10 total). Because of the near perfect AUC observed for subsets of the AML vs MDS and Nonleuk vs AML comparisons, we only evaluated AUC differences for the Asthma and Nonleuk vs MDS comparisons using the Wilcoxon test. Table A.3 shows the mean AUC differences of Asthma for each of the 10 pairs of subsets.Although many of the subsets do not show a significantly higher GEOlimma AUC, we note that the average GEOlimma - Limma AUC difference for both training and testing subsets is positive. Furthermore, subset pairs 7 and 9 show a significant GEOlimma AUC improvement in both training and testing subsets, while none of the negative AUC differences observed were significantly less than 0 (hypothesis: Limma AUC > GEOlimma AUC) in training sets. Figure B.5 shows the GEOlimma and Limma AUC values at each number of features for subset pairs 7 (A) and 9 (B). For the Nonleuk vs MDS comparison, we find no significant differences between GEOlimma and Limma AUCs in training or testing subset pairs. Figure B.5(C) shows one example of a training pair for this comparison. Overall, our results suggest that use of GEOlimma for feature selection can provide moderate improvements in classification performance for datasets with a modest overall degree of differential expression (e.g., Asthma comparison). For datasets with more pronounced degrees of differential expression, use of GEOlimma resulted in very similar classification performance compared to Limma.

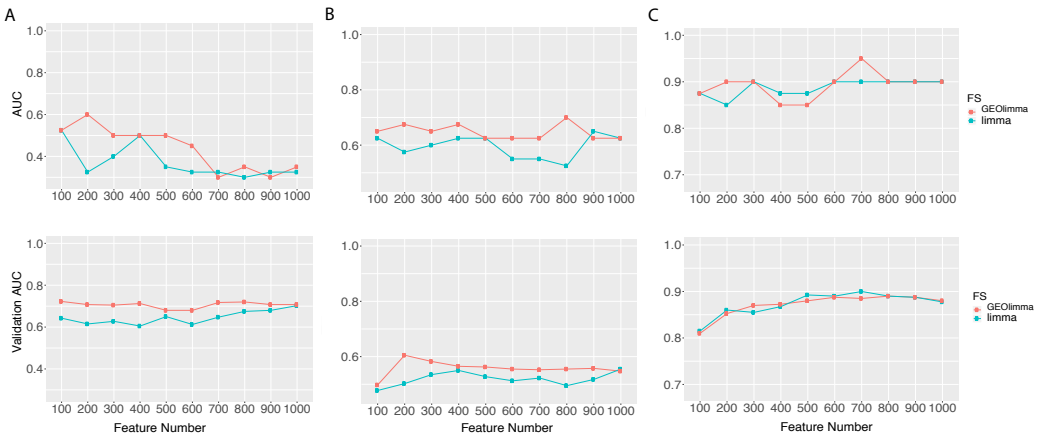


Fig. A.5: Classification performance of data subsets using a logistic regression classifier with GEOlimma and Limma feature selection methods. The x-axis indicates the number of selected features; y-axis indicates classification AUC. The top three plots display training AUC values; the bottom three plots depict validation AUCs. A) Asthma vs Non-asthma subset 7 AUCs, B) Asthma vs Non-asthma subset 9 AUCs, C) Nonleukemia vs MDS subset 9 AUCs.

Table A.3: Differences in classification performance (GEOlimma AUC - Limma AUC) for 10 data subsets of the Asthma comparison. Bold p-values (Wilcoxon signed-rank test) denote statistically significant AUC improvements of GEOlimma over Limma.

Sample Order	AUCdiff	Wilcox p Value	VadAUCdiff	Wilcox p Value
1	0.0075	2.36E-01	-0.01425	9.78E-01
2	-0.0275	8.53E-01	0.01375	1.43E-01
3	-0.0075	6.07E-01	-0.025	9.97E-01
4	-0.0075	7.79E-01	0.00525	2.39E-01
5	-0.0175	9.31E-01	0.0105	3.12E-01
6	-0.03	8.97E-01	-0.0085	7.23E-01
7	0.0675	**3.98E-02**	**0.06025**	9.77E-04
8	0.025	1.04E-01	-0.007	7.54E-01
9	0.0525	**1.23E-02**	**0.0385**	1.95E-03
10	-0.0075	7.36E-01	-0.001	7.93E-01

*asthma dataset on LR classification

A.5 Discussion

In this study, we developed a differential expression feature selection method, GEOlimma, in which we calculated gene-level differential expression (DE) prior probabilities from large-scale GEO transcriptomics data and incorporated them into a Bayesian framework. In a DE analysis, GEOlimma detected a larger number of DE genes in four comparisons within two evaluation datasets, compared to Limma. By analyzing small sample subsets of each dataset, we showed that knowledge-driven GEOlimma substantially improved

experimental power in terms of effective sample size. Furthermore, in a supervised classification analysis, GEOlimma used as a feature selection technique led to similar or better classification performance than standard Limma given noisy, small sample subsets from the Asthma comparison.

We also biologically characterized genes with especially high or low DE prior probabilities using KEGG pathway enrichment analysis. The strongest signal came from genes with high DE prior probabilities, where we detected enrichment in cell growth and death, signal transduction and cancer-related pathways. Cell growth and death are fundamental biological processes; however, deregulation of these processes is often involved in carcinosis. Specifically, resisting cell death and sustaining proliferative signaling were reported to be hallmarks of cancer [143]. This prevalence of enriched cancer-specific pathways may be indicative of an over-representation of cancer-related studies in data repositories such as GEO, which has been previously reported [144] [123]. However, while we saw excellent improvements in experimental power in differential expression analysis of three cancer-related comparisons, we note that the largest relative increases in effective sample size were observed in the Asthma comparison. This suggests that GEOlimma can also provide a substantial benefit to datasets that are unrelated to cancer.

We closely modeled GEOlimma after the widely-used differential expression analysis method Limma. Since its first publication nearly 15 years ago, papers describing the Limma method [145] [133] [132] have been cited over 10,000 times for applications in differential expression analysis of DNA microarray or RNA-Seq transcriptomics data. For the latter application, the more recently-developed voom method [145] adapts the Limma empirical Bayesian framework to read count data, which enables computation of posterior DE probabilities for RNA-Seq experiments. Although we only applied GEOlimma to DNA microarray data in this study, our approach is readily transferable to RNA-Seq data through the use of the voom methodology.

In this study, we made use of all available GPL570 GEO datasets (GDS), which we acknowledge represent a relatively small subset of all available GPL570 data at GEO. We made this selection in large part due to the high-quality curation of GDS datasets compared to the more abundant GSEs, which allowed us to easily perform multiple differential expression comparisons within each dataset. Given recent advances in natural language processing and the extraction of experimental metadata (e.g., [146]), an exciting future direction is the automatic annotation and inclusion of the larger number of GSEs (5154 for GPL570 as of June 2019) in the DE prior probability calculations. Such an expansion of a pre-existing data collection would enable subdivision and calculation of condition-specific DE prior probabilities (e.g., stem cell-related or viral infection-related), which could further improve GEOlimma performance when applied to the analysis of related datasets. One final future direction is the generalization of GEOlimma DE prior probabilities from individual values to probability distributions. In this case, DE hyperprior parameters could be calculated from pre-existing data rather than explicit prior probabilities. This modification would enable a more nuanced adjustment of DE posterior probabilities by GEOlimma given the biological characteristics of the dataset of interest.

A.6 Conclusions

Overall, our results demonstrate that GEOlimma effectively utilized pre-existing transcriptomics data for improved differential expression and feature selection analyses. Due to its focus on gene-level differential expression, GEOlimma also has the potential to be applied to other high-throughput biological datasets.

Appendix B

Prognostic Analysis of Histopathological Images Using Pre-Trained Convolutional Neural Networks: Application to Hepatocellular Carcinoma

B.1 Abstract

Histopathological images contain rich phenotypic descriptions of the molecular processes underlying disease progression. Convolutional neural networks (CNNs), state-of-the-art image analysis techniques in computer vision, automatically learn representative features from such images which can be useful for disease diagnosis, prognosis, and subtyping. Hepatocellular carcinoma (HCC) is the sixth most common type of primary liver malignancy. Despite the high mortality rate of HCC, little previous work has made use of CNN models to explore the use of histopathological images for prognosis and clinical survival prediction of HCC.

We applied three pre-trained CNN models – VGG 16, Inception V3, and ResNet 50 – to extract features from HCC histopathological images. Sample visualization and classification analyses based on these features showed a very clear separation between cancer and normal samples. In a univariate Cox regression analysis, 21.4% and 16% of image features on average were significantly associated with overall survival and disease-free survival, respectively. We also observed significant correlations between these features and integrated biological pathways derived from gene expression and copy number variation. Using an elastic net regularized CoxPH model of overall survival constructed from Inception image features, we obtained a concordance index (C-index) of 0.789 and a significant log-rank test (p = 7.6E18). We also performed unsupervised classification to identify HCC subgroups from image features. The optimal two subgroups discovered using Inception model image features showed significant differences in both overall (C-index = 0.628 and p = 7.39E-07) and disease-free survival (C-index = 0.558 and p = 0.012). Our work

demonstrates the utility of extracting image features using pre-trained models by using them to build accurate prognostic models of HCC as well as highlight significant correlations between these features, clinical survival, and relevant biological pathways.

Image features extracted from HCC histopathological images using the pre-trained CNN models VGG 16, Inception V3 and ResNet 50 can accurately distinguish normal and cancer samples. Furthermore, these image features are significantly correlated with survival and relevant biological pathways.

B.2 Introduction

Histopathological images contain rich phenotypic descriptions of the molecular processes underlying disease progression and have been used for diagnosis, prognosis, and subtype discovery [147]. These images contain visual features such as nuclear atypia, mitotic activity, cellular density, tissue architecture and higher-order patterns, which are typically examined by pathologists to diagnose and grade lesions. The recent accumulation of scanned and digitized whole slide images (WSI) has enabled wide application of machine learning algorithms to extract useful information and assist in lesion detection, classification, segmentation, and image reconstruction [148].

Deep learning is a machine learning method based on deep neural networks that has been widely applied in recent computer vision and natural language processing tasks [149]. A convolutional neural network (CNN), a class of deep learning architecture commonly used in computer vision, automatically learns representative features from images. CNNs have been dominant since their astonishing results at the ImageNet Large Scale Visual Recognition Competition (ILSVRC) [150]. In various studies, CNNs have shown good performance when applied to medical images, including those from radiology [151] [152] [153]. Additional applications of CNNs in the areas of diabetic retinopathy screening [154], skin lesion classification [155], age-related macular degeneration diagnosis [156] and lymph node metastasis detection [157] have

demonstrated expert-level performance in these tasks. In addition, a recent study applied CNN models to develop a content-based histopathological image retrieval tool for improving search efficiency of large histopathological image has archived [158]. Compared with traditional machine learning techniques, CNNs have achieved significantly improved performance in the areas of image registration for localization, detection of anatomical and cellular structures, tissue segmentation, and computer-aided disease prognosis and diagnosis [159].

One disadvantage of CNNs is their need for massive amounts of data, which can be a challenge for biomedical image analysis studies. Furthermore, deep feature learning depends on the size and degree of annotation of images, which are often not standardized across different datasets. One possible solution for analyzing image datasets with a small sample size is transfer learning, in which pre-trained CNN models from large-scale natural image datasets are applied to solve biomedical image tasks. In a previous study of CNN models applied to both thoraco-abdominal lymph node detection and interstitial lung disease classification, transfer-learning from large scale annotated image datasets (ImageNet) was consistently beneficial in both tasks [160]. Furthermore, in a breast cancer study [161], CNNs used for feature extraction followed by supervised classification achieved 99.86% accuracy for the positive class.

The overarching goal of this work is to evaluate the potential of transfer learning for histopathological image analysis of hepatocellular carcinoma (HCC). Primary liver cancer is the sixth most common liver malignancy, with a high mortality and morbidity rate. HCC is the representative type, resulting from the malignant transformation of hepatocytes in a cirrhotic, non-fibrotic, or minimally fibrotic liver [162]. With the development of high-throughput technologies, a number of "omics" research studies have helped elucidate the mechanisms of HCC molecular pathogenesis, which in turn have significantly contributed to our understanding of cancer genomics, diagnostics, prognostics, and therapeutics [163] [164] [165] [166]. In particular, the most frequent mutations and

chromosome alterations leading to HCC were identified in the TERT promoter as well as the CTNNB1, TP53, AXIN1, ARID1A, NFE2L2, ARID2 and RPS6KA3 genes [166]. The biological pathways Wnt/β-catenin signaling, oxidative stress metabolism, and Ras/mitogen-activated protein kinase (MAPK) were reported to be involved in liver carcinogenesis [163]. Frequent TP53-inactivating mutations, higher expression of stemness markers (KRT19, EPCAM) and the tumor marker BIRC5, and activated Wnt and Akt signaling pathways were also reported to associate with stratification of HCC samples ([166]). The histological subtypes of HCC have been shown relate to particular gene mutations and molecular tumour classification [167]. Two recent studies have demonstrated strong connections between molecular changes and disease phenotypes. In a meta-analysis of 1494 HCC samples, consensus driver genes were identified that showed strong impacts on cancer phenotypes [168]. In addition, a deep learning-based multi-omics data integration study produced a model capable of robust survival prediction [169]. These and other recent findings may help to translate our knowledge of HCC biology into clinical practice [167].

At the pathological level, HCC exhibits as a morphologically heterogeneous tumour. Although HCC neoplastic cells often grow in cords of variable thickness lined by endothelial cells mimicking the trabeculae and sinusoids of normal liver, other architectural patterns are frequently observed and numerous cytological variants recognized. Though histopathologic criteria for diagnosing classical, progressed HCC are well established and known, it is challenging to detect increasingly small lesions in core needle biopsies during routine screenings. Such lesions can be far more difficult to distinguish from one another than progressed HCC, which is usually diagnosed in a straightforward manner using hematoxylin and eosin staining [170] [171]. Although prognostication increasingly relies on genomic biomarkers that measure genetic alterations, gene expression changes, and epigenetic modifications, histology remains an important tool in predicting the future course of a patient's disease.

Previous studies [172] [173] indicated the complementary nature of information provided by histopathological and genomic data. Quantitative analysis of histopathological images and their integration with genomics data require innovations in integrative genomics and bioimage informatics.

In this study, we applied pre-trained CNN models on HCC histopathological images to extract image features and characterize the relationships between images, clinical survival and biological pathways. We first downloaded Hematoxylin and Eosin (H&E) stained whole slide images from HCC subjects (421 tumor samples and 105 normal tissue adjacent to tumor samples) from the National Cancer Institute Genomic Data Commons Data Portal. After image normalization, we then applied three pre-trained CNN models–VGG 16, Inception V3, and ResNet 50–to extract representative image features. Using these features, we (1) performed classification between cancer and normal samples, (2) constructed models associating image features with clinical survival, (3) discovered potential HCC subgroups and characterized subgroup survival differences, and (4) calculated correlations between image features and integrated biological pathways. To the best of our knowledge, this is the first study to extract HCC image features using pre-trained CNN models and assess correlations between image features and integrated pathways. Our results indicate the feasibility of applying CNN models to histopathological images to better understand disease diagnosis, prognosis, and pathophysiology.

B.3 Materials & Methods

B.3.1 HCC Datasets

We downloaded HCC histopathological images of diagnostic slides (access by TCGA-LIHC Diagnostic Slide Images) from the National Cancer Institute Genomic Data Commons Data Portal on January 23, 2019. In addition to images, this Portal also provides multiple molecular datasets (e.g., Transcriptomics, DNA Methylation, Copy Number Variation) and clinical information for the same cohort. In total, we obtained 966 H&E stained whole

slide images from 421 scanned HCC subjects (421 tumor samples and 105 normal tissue samples adjacent to tumors). The images were digitized and stored in .svs files, which contain pyramids of tiled images with differing levels of magnification and resolution. We used the Python modules OpenSlide and DeepZoomGenerator to read those image files. Most of the files contained three or four levels of sizes and resolutions, with level 4 corresponding to the highest resolution (median pixels: 89640 x 35870) and level 3 comprising 1/16th the size of level 4 (median pixels: 5601 x 2249.5). To reduce memory usage and processing time, we extracted either level 3 images or downsampled level 4 images (if available) by a factor of 16 to the level 3 equivalent. We removed two files which were either corrupted or did not contain level 3 or 4 information. In total, we used 964 files in our analysis.

We downloaded clinical files containing overall survival (OS) and disease free survival (DFS) information on January 23, 2019 from the cBioPortal for Cancer Genomics website (https://www.cbioportal.org/). The cBioPortal provides visualization, analysis and downloading of large-scale cancer genomics data sets. Importantly, cBioPortal includes data for the same patient cohort from which the HCC images were taken. When performing OS analysis, the event of interest is death (event = 1), while the censored event is being alive (event = 0). Thus, the number of days for event 1 and event 0 are the number of days until death and number of days until last contact, respectively. In DFS analysis, the event of interest is new tumor occurrence (event = 1), while the censored event is the lack of detection of a new tumor (event = 0). In this case, the number of days for event 1 and event 0 are the number of days until detection of a new tumor and number of days until last contact, respectively.

We downloaded molecular pathway information, including integrated gene expression and copy number variation data, on January 28, 2019 from the Broad GDAC Firehose (https://gdac.broadinstitute.org/). This resource provides an open access web portal for exploring analysis-ready, standardized TCGA data

including the cohort from which the TCGA-Liver Hepatocellular Carcinoma image files were collected. Using this pathway information, we applied the PAthway Representation and Analysis by Direct Inference on Graphical Models (PARADIGM) algorithm [174] to infer Integrated Pathway Levels (IPLs). Briefly, PARADIGM predicts the activities of molecular concepts including genes, complexes, and processes and measures using a belief propagation strategy within the pathway context. Given the copy numbers and gene expression measurements of all genes, this belief propagation iteratively updates hidden states reflecting the activities of all of the genes in a pathway so as to maximize the likelihood of the observed data given the interactions within the pathway. In the end, the IPLs reflect both the data observed for that pathway as well as the neighborhood of activity surrounding the pathway. We used the analysis-ready file of IPLs generated by PARADIGM for correlation analyses between image features and biological pathways.

B.3.2 Image Pre-Processing and Feature Extraction

For each of the 964 image files from 421 tumor and 105 normal samples, we performed stain-color normalization as described in previous image studies [175] [176] [177]. After color normalization, we performed 50 random color augmentations. We followed a previous study [178] and first deconvolved the original RGB color into H&E color density space. We then estimated a specific stain matrix for a given input image and multiplied the pixels with a random value from the range [0.7, 1.3] to obtain the color augmented image. We repeated the process to generate 50 augmentations. Next, we randomly selected 20 crops of size 256 x 256 and 512 x 512 pixels from each augmented image. We separately input each crop to the three pre-trained CNN models (VGG 16, Inception V3, and ResNet 50), each of which generated a total of 20 sets of features. Within each model, we combined all sets of features associated with an image into a single set by computing median values of features across all crops of all augmented images.

Deep CNN models such as VGG 16, Inception V3 and ResNet 50 contain millions of parameters that require extensive training on large datasets. When properly trained, these models have reached state-of-the-art performance in tasks such as image recognition and classification. To avoid the challenges of training an entire CNN from scratch, we used pre-trained versions of these models to extract histopathological image features in an unsupervised manner. This transfer learning approach was essential given the relatively small sample size of the HCC cohort. For the Inception and ResNet models, we used nodes in the second-to-last convolutional layer as image features. For the VGG model, we concatenated nodes from the last 4 convolutional layers (`block2_conv2`,`block3_conv3`, `block4_conv3`, `block5_conv3`) as image features. In each case, the CNN network weights had been pre-trained using ImageNet data [179]. We implemented the above steps using Keras, a popular Python framework for deep learning.

B.3.3 Sample Visualization

To visualize samples, we first used Principal Component Analysis (PCA) to reduce the dimensionality of image features. We then applied the t-Distributed Stochastic Neighbor Embedding (t-SNE) method to visualize the first 10 components in 2 dimensions. The t-SNE method reduces data dimensionality based on conditional probabilities that preserve local similarity. We applied a t-SNE implementation that uses Barnes-Hut approximations, allowing it to be applied on large real-world datasets [138]. We set the perplexity to 15, and colored the sample points using the group information.

B.3.4 Supervised Classification from Image Features

We applied a linear Support Vector Machine (SVM) classifier [180] to discriminate between cancer and normal samples using the extracted image features (derived as described above). We used stratified 6-fold cross validation to train the model. To evaluate classifier performance, we visualized the Receiver Operating Characteristic (ROC) curve generated using cross-validation, with

false positive rate on the X axis and true positive rate on the Y axis. We calculated the Area under the ROC curve (AUC) for each cross-validation fold, as well as the overall mean value. We also plotted the 2-class precision-recall curve to visualize the tradeoff between precision and recall for different prediction thresholds. A high AUC represents both high recall and high precision, which translates to low false positive and negative rates. Using average precision (AP), we summarized the mean precision achieved at each prediction threshold. We used the Python module Scikit-learn to perform classification with a linear SVM, setting the parameter C to its default value of 1.0.

B.3.5 Survival Analysis

To perform univariate survival analysis for each image feature individually, we applied Cox Proportional Hazards (CoxPH) regression models using the R package 'survival' for both overall (OS) and disease-free survival (DFS). We used a log-rank test to select significant image features with p-value ≤ 0.05.

For multivariate survival analysis, we used the R package 'glmnet' to build separate CoxPH OS models based on image features from each of the three pre-trained CNN models. We applied elastic net regularization with fixed alpha = 0.5, which corresponds to equal parts lasso and ridge regularization. In order to learn the optimal penalty coefficient lambda, we applied 10-fold cross validation. We evaluated models with the Concordance index (C-index) and a log-rank test. The C-index quantifies the quality of rankings and can be interpreted as the fraction of all pairs of individuals whose predicted survival times are correctly ordered [181] [182]. A C-index of 0.5 indicates that predictions are no better than random.

B.3.6 Subgroup Discovery

Using the Scikit-learn Python module, we applied K-means clustering across all cancer samples to discover HCC subgroups. Specifically, we clustered all image features which were significantly associated with both overall and

disease-free survival. The K-means algorithm [183] clusters samples by minimizing within-cluster sum-of-squares distances for a given number of groups (K), which we varied between 2-12. To identify the optimal number of subgroups, we applied two metrics: the mean Silhouette coefficient and the Davies-Bouldin index. The Silhouette coefficient [184] takes values between -1 and 1, and it is calculated based on the mean intra-cluster distance and the mean nearest-cluster distance for each sample. Higher positive Silhouette values correspond to good cluster separation, values near 0 indicate overlapping clusters, and negative values indicate assignment of samples to the wrong cluster. The Davies-Bouldin index [185] is calculated based on the average similarity between each cluster and its most similar one, where an index close to 0 indicates a good partition. Given the optimal number of subgroups, we constructed CoxPH models to detect survival differences between the subgroups, again using C-index and log-rank test for evaluation. We fit Kaplan–Meier curves to visualize the survival probabilities for each subgroup.

B.3.7 Correlation Between Image Features and Pathways

We calculated the Pearson correlation between image features and Integrated Pathway Levels (IPLs) using the scipy Python module. Pearson correlation coefficients range between -1 and 1, with 0 implying no correlation. Each correlation coefficient is accompanied by a p-value, which indicates the significance of the coefficient in either the positive or negative direction. To correct for multiple hypothesis testing, we adjusted p-values using the Benjamini & Hochberg (BH) method [86]. We selected significant correlations between image features and IPLs as those whose adjusted p-values were ≤ 0.05.

B.3.8 Differential Expression Analysis

To identify differentially expressed (DE) pathways between two HCC subgroups, we applied the widely-used "Limma" R package [186]. We selected significantly DE pathways as those whose Benjamini & Hochberg (BH)-adjusted p-values were ≤ 0.1.

B.4 Results

In this study, we made use of pre-trained CNN models VGG 16, Inception V3 and ResNet 50 to extract features from HCC histopathological whole slide images. We first downsampled the whole slide images, normalized the color, and generated augmented images. We then aggregated the features extracted from randomly selected crops using pre-trained CNN models. Using these image features, we performed survival analysis and subgroup discovery. We also performed correlation analysis between image features and integrated biological pathways. The workflow of these analysis steps can be seen in Figure B.1.

B.4.1 Image Feature Extraction and Survival Analysis

Histopathology assessment is mandatory in HCC diagnosis [187], and the characteristics such as tumor number, size, cell differentiation and grade, and presence of satellite nodules were reported to be prognostic biomarkers [188]. Given a histopathological image, CNNs enable efficient feature extraction and representation using convolutional, pooling, and fully connected network layers. To examine image features relevant to HCC, we first downloaded HCC histopathological images from the National Cancer Institute Genomic Data Commons Data Portal. In addition to images, this Portal also provides multiple molecular datasets and clinical information for the same cohort of samples. We downloaded a total of 966 .svs image files from 421 cancer tissues and 105 tumor-adjacent normal tissues, of which 964 had sufficient information for the following analysis. For all image files, we used the equivalent of level 3 magnification (median 5601 x 2249.5 pixels) as described in the Materials & Methods section. We performed staining color normalization, followed by image augmentation to improve sample variety. We randomly selected 20 crops of sizes 512 x 512 pixels or 256 x 256 pixels from each augmented image. The 20 512 x 512 crops represent 41.6% of the input image pixels on average, while the 20 256 x 256 crops represent 10.4% on average.

The deep CNN models VGG 16, Inception V3 and ResNet 50 contain

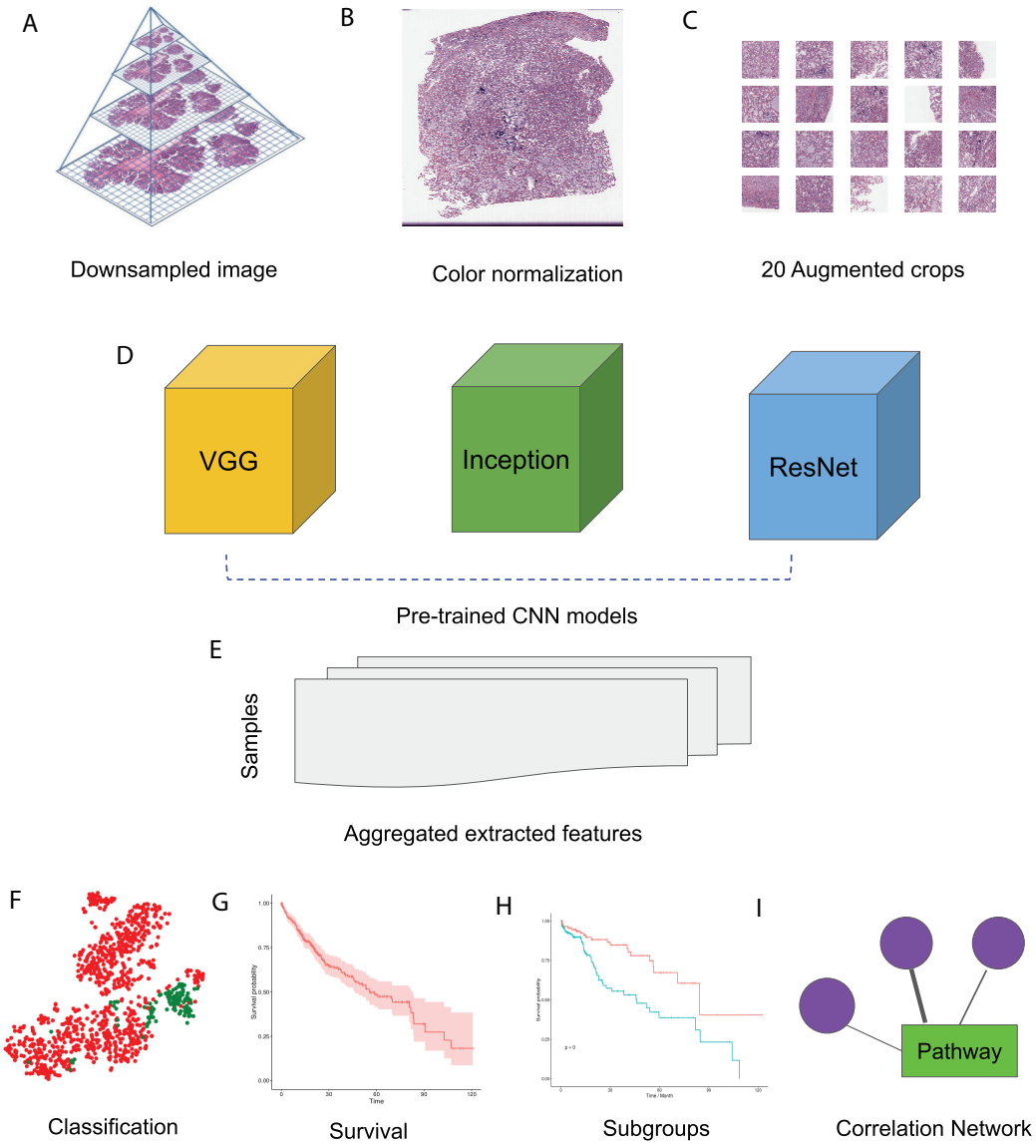


Fig. B.1: HCC image analysis flow. A-1) For whole slide .svs files, downsampled images were generated, B-2) color normalization was performed, C-3) 50 augmented images were made for each original image and 20 crops were selected at random from each augmented image, D-4) three CNN models, VGG 16, Inception 3 and ResNet 50 were applied to extract features from each crop, E-5) features from all crops were aggregated and 50 sets of image features were obtained from each CNN model, F-6) image features were used for classification, G-7) image features were fit for survival analysis, H-8) image features were used for subgroup discovery, I-9) correlation between image features and biological pathways.

millions of parameters, and extensive training of these models has led to state-of-the-art performance in image recognition and detection [189]. Given the small sample size in our cohort, we extracted features from each image crop by applying pre-trained versions of these models. This approach, which is a form of

transfer learning, allows us to avoid the challenges of CNN model training from scratch. For the Inception and ResNet models, we chose all nodes in the second-to-last network layer as features after excluding the final fully-connected layers. For the VGG model, we chose all nodes from the last four convolutional layers as features. For each full image, we combined features from the 20 random crops into a single set of features representing that image.

In total, we obtained 1408, 2408, and 2408 features for each image using the VGG 16, Inception V3, and ResNet 50 models, respectively. To aggregate these features across all augmented images, we computed median values for each feature. We then visualized cancer and normal samples in the context of these features by using PCA to reduce the feature dimensionality followed by applying t-SNE to the first 10 principal components. We also performed supervised classification of the samples using a linear Support Vector Machine applied to each set of image features. Figure B.2 shows these results using features derived from 256 x 256 crop sizes, with classification performance displayed as receiver operating characteristic (ROC) and two-class precision-recall curves. The average AUC achieved by all three models is between 0.99 and 1, illustrating the clear separation achieved between tumor and normal samples using the extracted image features. Similarly, the AUCs achieved for features derived from 512 x 512 crop sizes were very close to 1. To compare this performance with that of an alternate method, we also applied PCA (randomized SVD) and SVD (full SVD) on the downsampled images without augmentation. Specifically, we extracted the first 100 principal components (PCA) or singular vectors (SVD) as features and performed supervised classification. Figure S1 shows that performance using PCA- and SVD-derived features is very poor. Finally, we performed classification on features derived without using image augmentation. Here, performance is only slightly worse, with AUCs ranging between 0.98 and 0.99 (Figure S2).

We next compared the performance of a simpler network to that of the three CNNs evaluated in this study. Specifically, we applied a MobileNet v1

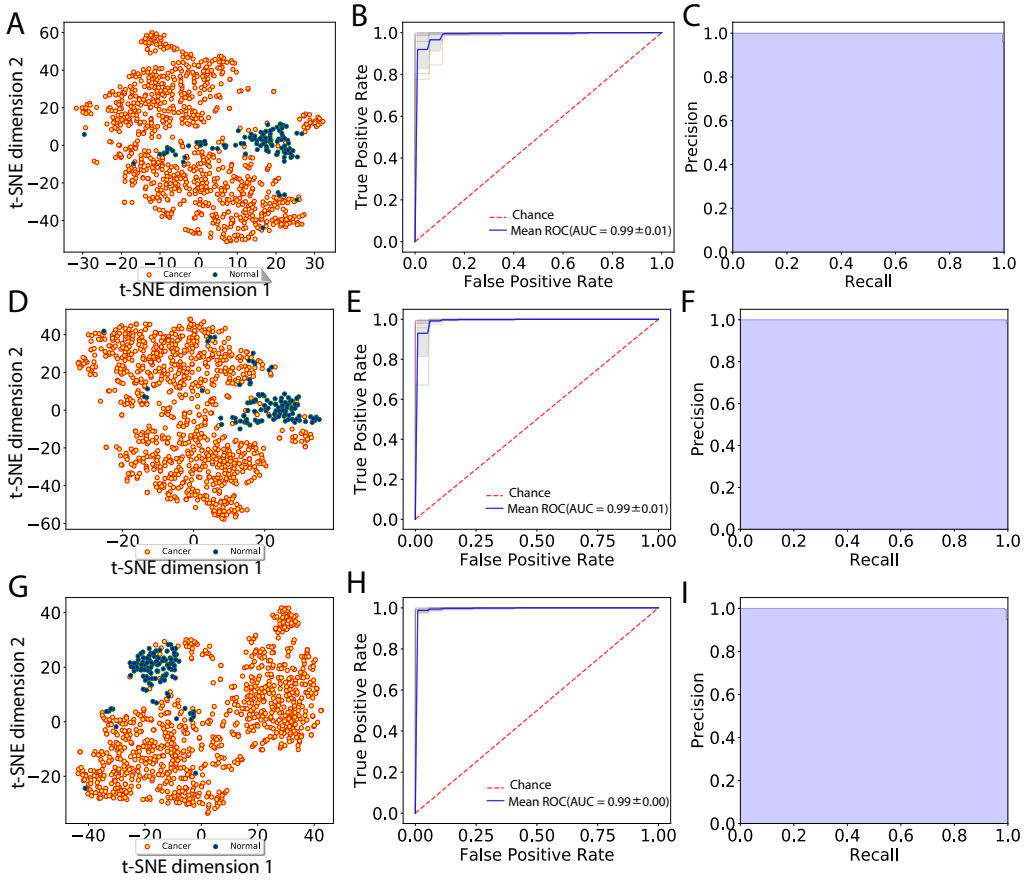


Fig. B.2: Visualization of extracted image features and classification between cancer and normal samples. A), D) and G) indicate t-SNE visualization, B), E) and H) indicate ROC curves from linear SVM and C), F) and I) indicate Recall and Precision curves measured using VGG image features, Inception features and ResNet features, respectively.

pre-trained network, which has many fewer tunable parameters ($4.2x10^6$) than VGG 16 ($1.4x10^8$), Inception v3 ($2.4x10^7$), and ResNet 50 ($2.3x10^7$). As with the other networks, we removed the final layer of MobileNet v1 and used the network to extract features for each image. We aggregated these features as before, followed by performing SVM classification. We found that the classification performance using MobileNet v1 was indistinguishable from those achieved by the larger networks. This result suggests that the pre-trained networks used in our study contain many more tunable parameters than are strictly necessary to yield very good classification performance.

We also explored reduction of model complexity by selecting smaller and smaller subsets of pre-trained CNN image features for classification. Figure S3 displays performance using randomly-selected image feature subsets of size 10, 25, 50 and 100 in each of the three pre-trained CNNs using 256 x 256 pixel crops. Our results show that when using smaller and smaller sets of features, classification AUC reached as low as 0.84, which was substantially worse than our original results. However, using random sets of 100 features led to performance that was nearly as good as that achieved using all features. Overall, our results show that use of CNN-derived image features is extremely effective for distinguishing HCC tumor from normal samples, which suggests that pre-trained CNN models capture the most relevant characteristics from HCC histopathological images.

To aid in interpreting CNN-derived image features of HCC, we visualized feature mappings of VGG model convolutional layer blocks when applied to 256 x 256 pixel crops of histopathological images (Figure B.3 and Figure S4). We note that the first convolutional layers tend to resemble the original image, but subsequent layers seem to intensify partial objects. In order to study whether the CNN-derived image features are associated with clinical survival, we next performed univariate CoxPH regression survival analysis on each feature. We obtained clinical information for each sample from the cBioPortal for Cancer

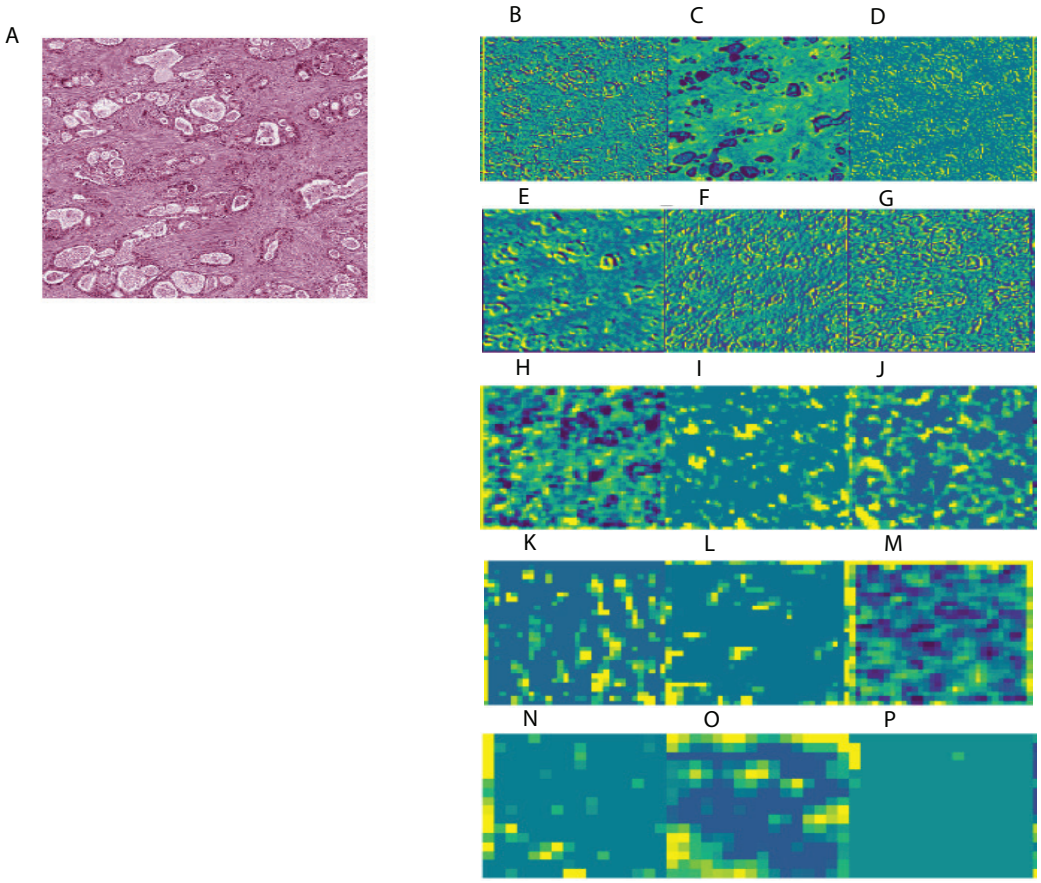


Fig. B.3: Example of feature mapping visualization in VGG 16 model in one cancer sample. A) shows an image patch with 256 x 256 pixels. B-P) indicates the corresponding feature mapping from convolutional block 1 (B-D) to convolutional block 5 (N-P)

Genomics, as described in the Materials & Methods section. For the subjects with multiple histopathological images, we computed median feature values across the images for the following survival analysis. For each image feature, we applied CoxPH regression models for both overall survival (OS) and disease-free survival (DFS) and selected significantly associated features (p-value ≤ 0.05) based on a Score (log-rank) test. We also validated the predictive ability of the survival models using Concordance index (C-index). Table B.1 shows the number of significant features for each model and survival type. 21.4% and 16% of image features on average were significantly associated with OS and DFS, respectively. Each model had a slightly different number of significant features, with more features associated with OS than DFS.

Table B.1: Significant Image Feature Number from Univariate CoxPH Regression Models

Model	Feature Number	Crop Size	Significant Features of OS	Significant Features of DFS
VGG	1408	256	272 (19.3%)	219 (15.6%)
Inception	2048	256	574 (28.0%)	294 (14.4%)
ResNet	2048	256	522 (25.5%)	385 (18.8%)
VGG	1408	512	300 (21.3%)	201 (14.3%)
Inception	2048	512	356 (17.4%)	290 (14.2%)
ResNet	2048	512	347 (17.0%)	390 (19.0%)

Finally, we performed multivariate CoxPH regression analyses for each survival type on all image features from each model. We employed elastic net regularization using equally weighted lasso and ridge regularization during model training. Optimal hyperparameters were selected using 10-fold cross-validation and subsequently used for model prediction. Overall, we identified three multivariate OS models with the following log-rank p-values and C-indices: 1.2E-23 and 0.788 (VGG), 7.6E-18 and 0.789 (Inception), and 1.2E-12 and 0.739 (ResNet) from the 256 x 256 crop sizes. Table B.2 displays the C-indices and p-values achieved for each pre-trained network, image crop size and survival type. The Inception-derived model achieved the highest indices of 0.789 at OS and 0.744 at DFS. Overall, our results show that CNN-derived image features are significantly associated with clinical survival and can be used to build accurate survival models.

Table B.2: Multivariate CoxPH Regression Model in Three Models

Model	Crop	Survival	C-index	P value
VGG	256	OS	0.788 ± 0.022	1.2E-23
Inception	256	OS	**0.789 ± 0.021**	7.6E-18
ResNet	256	OS	0.739 ± 0.025	1.2E-12
VGG	256	DFS	0.655 ± 0.019	1.5E-08
Inception	256	DFS	**0.744 ± 0.018**	3.2E-13
ResNet	256	DFS	0.7 ± 0.019	4.1E-11

B.4.2 Subgroup Discovery from Image Features

To investigate whether our CNN-derived image features relate to HCC prognosis, we next used these features to discover subgroups within tumor samples. We considered all image features which were significantly associated with both OS and DFS. Using these features, we clustered the tumor samples using K-means (K = 2-12) and used both Silhouette coefficients and Davies-Bouldin values to choose the optimal number of subgroups. As shown in Figure B.4, two subgroups were determined to be optimal for all three models. We visualized these subgroups using t-SNE to reduce dimensionality.

We then examined survival differences between the subgroups. For each model and survival type, we generated Kaplan-Meier survival curves stratified by subgroup. Our results (Figure B.5) note that the subgroups discovered using the Inception and ResNet models show a significant difference in both OS and DFS using a log-rank test. The two subgroups from Inception have the most significant OS difference, with p-value 7.39E-07 and C-index 0.628, followed by the two subgroups from ResNet with p-value 0.001 and C-index 0.582. We also observed significant differences in DFS between subgroups in both models, with p-values and C-indices of 0.012 and 0.558 (Inception) and 0.014 and 0.56 (ResNet), respectively. For the VGG model, we only detected a significant difference for DFS (p-value 0.007 and C-index 0.536). In all models, we note that the second subgroup ("group 2") has consistently better OS and DFS survival than the first subgroup ("group 1"). Table B.3 shows the subgroup overlap between the three models. Overall, 176 samples from the Inception group 1 were also labeled group 1 in VGG and ResNet models. In contrast, 109 samples from the Inception group 2 were identified as group 2 in ResNet but group 1 in VGG. Taken together, the significant survival differences detected between sample subgroups demonstrate the feasibility of discovering clinically-relevant HCC subgroups using CNN-derived image features.

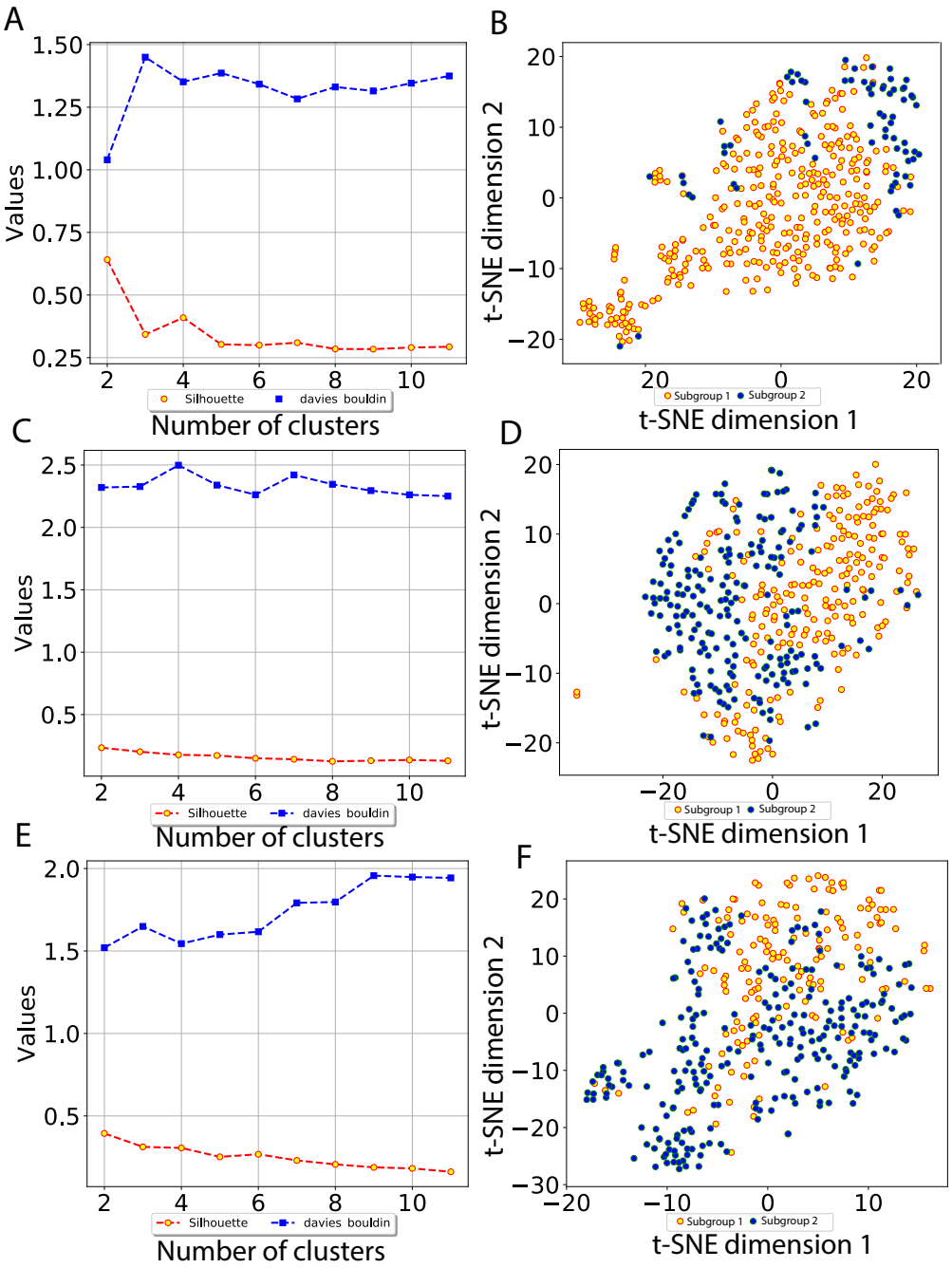


Fig. B.4: Subgroup discovery from image features using 256 x 256 pixel crop size. A), C) and E) display two different metrics for selecting the optimal number of clusters, and B), D) and F) indicate the t-SNE visualization of best clusters using VGG image features, Inception image features and ResNet image features, respectively.

B.4.3 Correlation Between Image Features and Biological Pathways

Previous studies examined the molecular mechanisms underlying HCC [163] [164] [165] [166]. To relate our CNN-derived image features to such

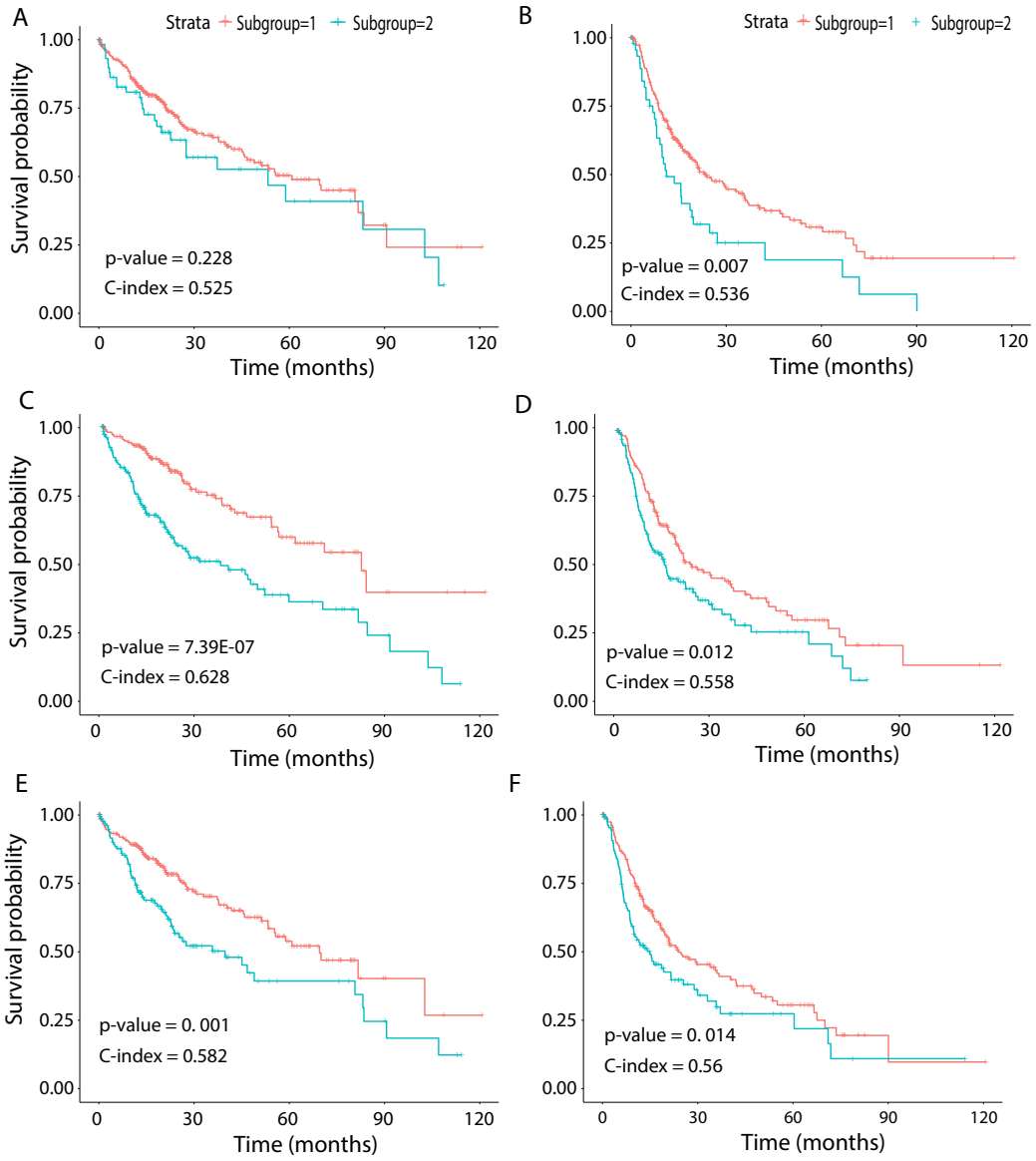


Fig. B.5: Survival analysis from discovered subgroups. A), C) and E) correspond to the CoxPH model applied to OS, B), D) and F) correspond to DFS. The two groups are indicated in red and green, using VGG image features, Inception image features and ResNet image features, respectively.

mechanisms, we identified correlations between features and a collection of molecular pathways. Specifically, we first obtained integrated pathway levels (IPLs) using the Firehose Genome Browser, which provides analysis-ready files inferred from both gene expression and DNA copy number variation using the PARADIGM algorithm [174]. IPLs indicate the predicted activities of biological concepts using both copy number and gene expression data (described in Materials & Methods). The IPL matrix contains a total of 7202 entities derived

Table B.3: Overlaps of Subgroup (1/2) Frequency Counts Between Three Pre-trained CNNs

Inception	VGG 16	ResNet	Sample Count
1	**1**	**1**	**176**
1	1	2	18
1	2	1	20
1	2	2	4
2	1	1	48
2	**1**	**2**	**109**
2	2	1	16
2	2	2	30

from 3656 concepts in 135 merged pathways. Each entity is annotated with the concept (gene) and pathway index as shown by the example `19_EPHB3`. Here, the EPHB3 gene participates in EPHB forward signaling whose pathway index is 19. We first computed Pearson correlation coefficients between these IPLs and each feature significantly associated with both OS and DFS. We then selected significantly correlated IPL-feature pairs based on Benjamini & Hochberg (BH) [86] -adjusted p-values ≤ 0.05. With 256 x 256 crop sizes, 90 (out of 97), 199 (out of 203) and 192 (out of 203) survival-associated image features from the VGG, Inception and ResNet models, respectively, were significantly correlated with IPLs. On average, 90.2% of the image features showed a significant correlation, with Pearson correlation coefficients ranging between -0.536 and 0.385.

Finally, we performed differential expression analysis to identify IPL differences between each pair of sample subgroups. For each model, we selected pathways with BH-adjusted p-values 0.05. Surprisingly, we found no significant pathways at this threshold for all three models and both crop sizes. After relaxing the p-value threshold to 0.1, we detected five significant entities from two pathways: 19: EPHB forward signaling (EPHB3, ROCK1, Ephrin B1/EPHB3) and 66: Glucocorticoid receptor regulatory network (IL8, ICAM1). The two entities at pathway 66 were calculated between two subgroups from Inception model with 256 x 256 crops while the three entities at pathway 19 were from VGG model with 512 x 512 crops. Figure B.6 shows a network

visualization of these pathways with significantly-correlated image features. The nodes represent image features and pathways, while the thickness of the edges denote the observed Pearson correlation coefficients. The numbers on the image feature nodes were assigned according to the order from the initial feature extraction. We note that some image features showed correlation with more than one entity from the same pathway, while others seemed to be related to only one entity. Overall, 31 out of 49 image features with significant correlations were found using the Inception model, of which three features (324, 1859, and 1292) were correlated with pathway 19: EPHB forward signaling. The VGG model identified a total of four significantly-associated features (two each of 870 and 871) from 256 x 256 and 512 x 512 crops. Feature 870 showed correlation with only 19: EPHB forward signaling, while feature 871 was correlated with both 19: EPHB forward signaling and 66: Glucocorticoid receptor regulatory network. The observation that consecutive features from the VGG model were correlated with similar pathways suggests that these features represent related attributes of the original images. In addition, it is noteworthy that the model with the largest proportion of significantly-associated features (Inception) also showed the most significant survival analysis results.

B.5 Discussion

In this study, we applied the pre-trained CNN models VGG 16, Inception V3, and ResNet 50 to extract features from HCC histopathological whole slide images. Using these image features, we observed clear separation between cancer and normal samples both visually (t-SNE) and through supervised classification. By performing univariate CoxPH regression, we identified averages of 21.4% and 16.0% of image features significantly associated with overall (OS) and disease-free survival (DFS), respectively. Many of these image features were also significantly associated with OS in a multivariate CoxPH regression model. We utilized the CNN-derived image features to discover HCC subgroups, with the

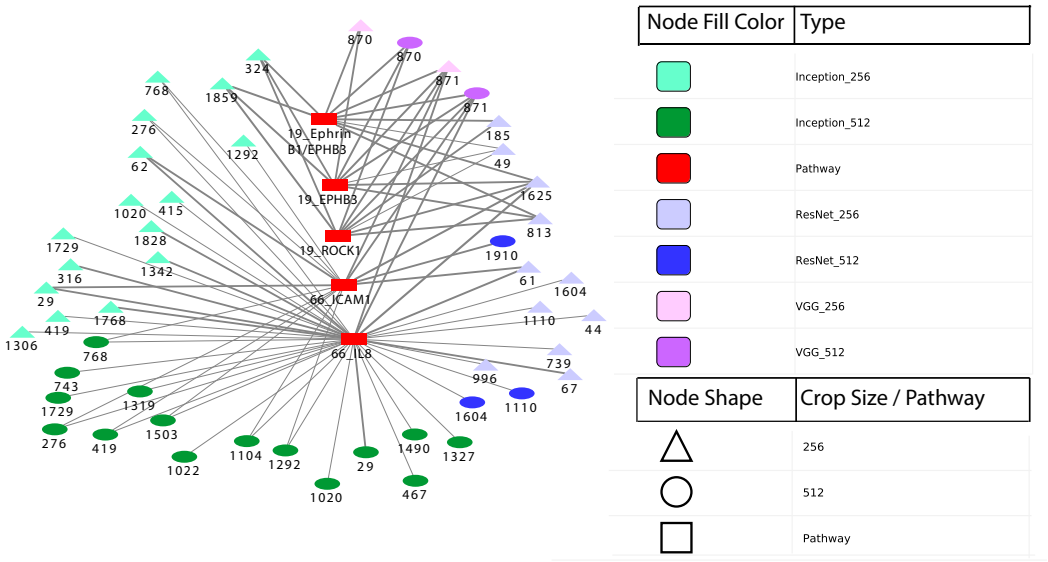


Fig. B.6: Correlation network between image features and example pathways. Colors of nodes indicate CNN models VGG, Inception and ResNet, as well as pathways. The labeled names of image features consist of the model name, crop size and feature order number. The thickness of each edge corresponds to the magnitude of correlation coefficients ranging between -0.536 and 0.385 that were statistically significant with the range.

optimal two subgroups showing a significant difference in both OS and DFS in the Inception model.

Notably, we found that 90.2% of the image features significantly associated with both OS and DFS were also significantly correlated with measures of integrated pathway levels (IPLs). The five most significant IPL entities were found in two pathways–EPHB forward signaling and Glucocorticoid receptor (GR) regulatory network–implying a potential role for these pathways in the prognosis of HCC. Previous studies of EPHB forward signaling have shown that it induces cell repulsion and controls actin cell adhesion and migration [190]. It has also been reported that EPHB receptors and ephrin ligands are involved in carcinogenesis and cancer progression [191]. Finally, the EPHB3 receptor also inhibits the Wnt signaling pathway [192], which was reported to be useful for HCC stratification [166]. In addition, previous studies have reported that the glucocorticoid receptor binds promoters, interacts with other transcription factors [193], and causes hepatocellular carcinoma [194] in

mice when receptor signaling is impaired. GR regulatory network member Interleukin-8 (IL8), a proinflammatory CXC chemokine, was reported to promote malignant cancer progression [195], while member Intercellular cell adhesion molecule-1 (ICAM-1) has functions in immune and inflammatory responses and was reported to play a role in liver metastasis [196]. We note that a previous study performed integration of genomic data and cellular morphological features of histopathological images for clear cell renal cell carcinoma, finding that an integrated risk index from genomics and histopathological images correlated well with survival [172]. In addition, a second study [173] developed a CNN model using both histopathological and genomic data from brain tumors, which surpassed the current state of the art in predicting overall survival.

Stratification of patients is an important step to better understand disease mechanisms and ultimately enable personalized medicine. Previous studies of HCC have suggested molecular-level subgroups [197] [198] [169]. In the recent study, the authors applied deep learning to integrate three omic datasets from 360 HCC patients (the same cohort used in our study), discovering two subgroups with survival differences. In our work, we identified subgroups using all three CNN models, with the subgroups from both Inception (C-index = 0.628; P value = 7.39E-07) and ResNet (C-index = 0.582; P value = 0.001) models showing significant differences in OS. We note that this significance of the Inception model is lower than that achieved using subgroups identified using multiple omic data integration (C-index = 0.68 and P value = 7.13E-6) [169], although the C-index is also slightly lower. We also detected significant survival differences in DFS using all three models, which to our knowledge has not been previously investigated. Interestingly, the subgroups from Inception model were most significantly different in OS.

In the analysis of histopathological images, the large image size and different levels of resolution from whole slide images (WSIs) pose challenges to accurate and efficient feature selection [148]. To avoid information loss, WSIs are

often divided into small patches (e.g., 256 x 256 pixels) and each patch is analyzed individually as a Region of interest (ROI). These ROIs are first labeled using active learning [199] or by professionally trained pathologists [200]. Subsequently, averaged regions of patches representing WSIs are studied for specific tasks [148]. In our work, we randomly selected 20 patches of 256 x 256 and 512 x 512 pixels from each WSI and extracted features from the last layers of CNN models to represent each image for visualization and classification. To robustly deal with color variation and image artifact issues, we conducted color normalization and augmentation before applying CNN models. Color normalization adjusts pixel-level image values [201], and color augmentation generates more data by altering hue and contrast in the raw images [202]. We achieved very good classification performance, with AUCs between 0.99 and 1 for distinguishing between normal and tumor samples. To illustrate the power of a transfer learning approach using pre-trained CNNs, we also applied a simple (not pre-trained) CNN model (Figure S5) for classifying tumor and normal samples. This approach achieved a best validation accuracy of 87.8% (Figure S6), which was substantially worse than the transfer learning performance.

Comparing our performance to previous work, we note that in one study of histopathological images [203], classification performance reached 81.14% accuracy using the extracted features from a pre-trained VGG 19 (similar to VGG 16) network. In a similar study of histopathological images of breast cancer [204], classification performance on 400 H&E-stained images of 2048 1536 pixels each reached an AUC of 0.963 for distinguishing between non-carcinomas vs carcinomas samples. We note that our study uses higher resolution histopathological images (median 5601 x 2249.5 pixels), which may explain the better performance.

Recent related work in histopathological image analysis include a deep-learning-based reverse image search tool for images called SMILY (Similar Medical Images Like Yours) [158]. By building an embedding database using a

specialized CNN architecture called a deep ranking network, SMILY enables search for similar histopathological images based on a query image. SMILY's deep ranking network utilizes an embedding-computing module that compresses input image patches into a fixed-length vector. This module contains layers of convolutional, pooling, and concatenation operations. SMILY retrieves image search results with similar histological features, organ sites, and cancer grades, based on both large-scale quantitative analysis of annotated tissue regions and prospective studies with pathologists blinded to the source of the search results. SMILY's creators comprehensively assessed its ability to retrieve search results in two ways: using pathologist-provided annotations, and via prospective studies where pathologists evaluated the quality of SMILY search results.

Additional related work has made use of deep learning generative models to help delineate fundamental characteristics of histopathological images. Generative Adversarial Networks (GANs) have enjoyed wide success in image generation. GANs involve training a generator to fool a discriminator, while a discriminator is trained to distinguish the generated samples from real samples. This approach eventually produces high-quality images [205]. The creators of Pathology GAN recently demonstrated its abilities to create artificial histological images and learn biological representations of cancer tissues [206]. A second type of generative model known as a variational autoencoder (VAE) learns the distribution of latent variables and reconstructs images. VAEs have been successfully applied in image generation [90], and a specialized version known as a conditional VAE can be suitable for pathology detection in medical images [207].

We note that our study has several limitations, including limited interpretability of the most discriminative HCC image features and a lack of external validation datasets. We also did not address multiclass grading on the HCC samples, instead focusing on a binary classification. Despite using pre-trained CNN models for feature selection, our results may still be limited by

the somewhat small and unbalanced sample sizes of our dataset. Additional studies on other independent data sets should be evaluated to further explore the correlation between deep learning-based extracted images, clinical survival and biological pathways. Future work will involve experimenting with other CNN models, as well as improving the biological interpretation of features from pre-trained models.

B.6 Conclusions

The image features extracted from HCC histopathological images using pre-trained CNN models VGG 16, Inception V3 and ResNet 50 can accurately distinguish normal and cancer samples. Furthermore, these image features are significantly correlated with clinical survival and biological pathways.

Appendix C

Abbreviations of Clinical Features

ABM1: 1. Body Weight

ABM2: 2. BMI

ABM3A: 3. Standing Height

ABM3B: 3. Standing Height

ABM4: 4. Waist Circumference

ABM5: 5. Waist Circumference

ABM6: 6. Hip Circumference

ABM7: 7. Hip Circumference

ABM8: 8. Waist to Hip Ratio

ABM9-Syst: 9. Systolic Blood Pressure

ABM9-Dias: 9. Diastolic Blood Pressure

ABM10: 10. Pulse

DemoRel: Relationship Status

DemoRel-Married: Relationship Status

DemoRel-Living: Relationship Status

DemoRel-SteadyRel: Relationship Status

DemoRel-Divorced: Relationship Status

DemoRel-Widow: Relationship Status

DemoRel-Single: Relationship Status

DemoChi: Children Status

baso: Basophils

eosino: Eosinophils

hct: hematocrit

lymph: Lymphocytes

monos: Monocytes

neut: Neutrophils

rdw: Red cell distribution width

utox: ny urine tox

mcv: mean corpuscular volume

mpv: Mean platelet volume

ALB: albumin

hgb: hemoglobin

mchc: mean corpuscular hemoglobin concentration

ProTot: total serum protein

baso .: Basophils .

eosino.: Eosinophils .

lymph.: Lymphocytes .

monos.: Monocytes .

Neut.: Neutrophils .

plt: Platelet count

wbc: white blood cells

efgraa: Estimated Glomerular Filtration Rate African American

efgrnaa: Estimated Glomerular Filtration Rate Non-African American

rbc: red blood cells

Cl: chloride

CO2: bicarbonate

K: potassium

Na: sodium

BILITOT: bilirubin

bun: blood urea nitrotgen

Ca: calcium

cholest: total cholesterol

creatinine: creatinine

glucose: fasting glucose

hbA1c: glycosylated hemoglobin

hdl: HDL cholesterol

ldl: LDL cholesterol

triglyc: triglycerides

hscrp: C reactive protein

dotclinlabs: date of clinical labs

mch: mean corpuscular hemoglobin

alkphos: Alkaline phosphatase

alt: Alanine Aminotransferase

ast: Aspartate transaminase

ggt: Gamma-glutamyltransferase

bmkrid: subject ID

acth1: ACTH DST 1

acth2: ACTH DST 2

acthdif: acth1-acth2

acthsup: (acth1-acth2)/acth1

athf: 5a-tetrahydrocortisol

athftobthf: aTHF/bTHF

athftof: 5-reductase (aTHF/F)

bthf: 5b-tetrahydrocortisol

bthftof: 5-reductase (bTHF/F)

cor1: plasma cortisol DST 1

cor2: plasma cortisol DST 2

cordif: cor1-cor2

corsup: (cor1-cor2)/cor1

dex: Dexamethasone

dhea: DHEA

dheas: DHEA-S

dheatodheas: dhea/dheas

dotdsta: date of blood collection DST 1

dotdstb: Date of blood collection DST 2

doturn: urine collection date

e: cortisone

estrogen: estrogen

f: free cortisol

ic50: Lysosyme IC50

npy: Neuropeptide Y

the: tetrahydrocortisone

thetoe: the/e

totgluc: f+e+athf+bthf+the

urncatot: ne,da,epi total

urncor: Urine Cortisol

urncr: Urine Creatinine

urnda: Urine Dopamine

urnepi: Urine Epinephrine

urnne: Urine Norepinephrine

urnnetocor: urnne/urncor

urnvol: urine volume

psqi1a: 1. During the past month,when have you usually gone to bed at night?

psqi1b: 1. During the past month,when have you usually gone to bed at night?

psqi1c: 1. During the past month,when have you usually gone to bed at night?

psqi2: "2. During the past month, how long (in minutes) has it usually taken you to fall asleep each night?

" **psqi3a:** 3. During the past month, when have you usually gotten up in the morning?

psqi3b: 3. During the past month, when have you usually gotten up in the morning?

psqi3c: 3. During the past month, when have you usually gotten up in the morning?

psqi4a: 4. During the past month, how many hours of actual sleep did you get at night? (This may be different from the number of hours you spent in bed.)

psqi4b: 4. During the past month, how many hours of actual sleep did you get at night? (This may be different from the number of hours you spent in bed.)

psqi5a: "5. During the past month, how often have you had trouble sleeping because you . . .A. Cannot get to sleep within 30 minutes

" **psqi5b:** B. Wake up in the middle of the night or early morning

psqi5c: C. Have to get up to use the bathroom

psqi5d: D. Cannot breathe comfortably

psqi5e: E. Cough or snore loudly

psqi5f: F. Feel too cold

psqi5g: G. Feel too hot

psqi5h: H. Had bad dreams

psqi5i: I. Have pain

psqi5j: J. How often during the past month have you had trouble sleeping because of one or more problems NOT listed above?

psqi6: 6. During the past month, how would you rate your sleep quality overall?

psqi7: 7. During the past month, how often have you taken medicine (prescribed or "over counter") to help you sleep?

psqi8: 8. During the past month, how often have you had trouble staying awake while driving, eating meals, or engaging in social activities?

psqi9: 9. During the past week, how much of a problem has it been for you to keep up enough enthusiasm to get things done?

psqi10: 10. Do you have a bed partner or roommate?

:

WVocRS: Vocabulary:Raw score

WVocAS: Vocabulary:age scaked score

WVocPR: Vocabulary:percentile range

WEstimateIntelligence: Standard Score (Estimate of Intelligence)

WMemDigRS: Digit Span:Raw Score

WMemDigAgeSS: Digit Span:Age scaled score

WMemDigPR: Digit Span:Percentile Range

WMemLetterRS: Letter number:Raw Score

WMemLetterAgeSS: Letter number:Age Scaled Score

WMemLetterPR: Letter number:percentile Range

WProcSpeedCodRS: Coding:Raw Score

WProcSpeedCodAgeSS: Coding:Age Scaled Score

WProcSpeedCodPR: Coding: Percentile Range

WDigSpanForwardRS: Digit Span Forward: Raw Score

WDigSpanBackRS: Digit Span Backwards: Raw Score

WLongestForwRS: Longest Digit Span Forward: Raw Score

WLongestBackRS: Longest Digit Span Backwards:Raw Score

WForwardPercent: Longest Digit Span Forward:cumulative percentage

WBackPercent: Longest Digit Span Backwards:Cumulative percentage

WAuditMemSum: Auditory Memory Sum of Scaled Scores

WAuditMemIndex: Auditory Memory Index Score

WAuditMemPR: Auditory Memory PR

WAuditMemConf1: Auditory Memory Index 95% Confidence Interval

WAuditMemConf2: Auditory Memory Index 95% Confidence Interval

WAuditMemoryDesc: Auditory Memory Qualitative Description

WVisWorkMemSum: Visual Working Memory Sum of Scaled Scores

WVisWorkIndex: Visual Working Memory Index

WVisWorkMemPR: Visual Working Memory PR

WVisWorkMemConf1: Visual Working Memory Index 95% Confidence Interval

WVisWorkMemConf2: Visual Working Memory Index 95% Confidence Interval

WVisWorkMemQual: "Visual Working Memory Qualitative Description

" **WVisReprod-1-RS:** Visual Reproduction I:raw score

WVisReprod-1-SS: Visual Reproduction I:scaled score

WVisReprod-1-PR: Visual Reproduction I:percentile range

WVisReprod-2-RS: Visual Reproduction II: raw score

WVisReprod-2-SS: Visual Reproduction II:scaled score

WVisReprod-2-PR: Visual Reproduction II:percentile range

WLogicalMem-1-RS: Logical Memory I:Raw Score

WLogicalMem-1-SS: Logical Memory I:Scaled Score

WLogicalMem-1-PR: Logical Memory I:Percentile Range

WLogicalMem-2-RS: Logical Memory 2:Raw Score

WLogicalMem-2-SS: Logical Memory 2:Scaled Score

WLogicalMem-2-PR: Logical Memory 2:Percentile Range

WVerbalPairedAss-1-RS: Verbal Paired Associates I: Raw Score

WVerbalPairedAss-1-SS: Verbal Paired Associates I:scaled score

WVerbalPairedAss-1-PR: Verbal Paired Associates I:percentile range

WVerbalPairedAss-2-RS: Verbal Paired Associates 2: Raw Score

WVerbalPairedAss-2-SS: Verbal Paired Associates 2:scaled score

WVerbalPairedAss-2-PR: Verbal Paired Associates 2:percentile range

WSpatialAdd-RS: Spatial Addition:Raw Score

WSpatialAdd-SS: Spatial Addition:Scaled score

WSpatialAdd-PR: Spatial Addition:Perecntile Range

WSymbolSpan-RS: Symbol Span:Raw Score

WSymbolSpan-SS: Symbol Span:Scaled Score

WSymbolSpan-PR: Symbol Span:Percentile Range

WLongDSSRS: Longest Digit Span Sequence:Raw Score

WLongDSSPercent: Longest Digit Span Sequence:percentile

Age-A: A. Did you ever suffer a serious personal injury or illness? (occured once)

Age-B: B. Were you involved in a serious accident? (occured twice)

Age-C: C. Did your parents or primary caretaker have a problem with alcohol? (ongoing)

Army: Military Service

Navy: Military Service

AirForce: Military Service

Marine: Military Service

NatGuard: Military Service

Reserve: Military Service

DemoMilServiceTours: Number of Military Tours

DemoMilService-Iraq: Military Service

DemoMilServiceIraqFrom-1: Military Service

DemoMilServiceIraqTo-1: Military Service

DemoMilServiceIraqMOS-1: Military Service

DemoMilServiceIraqFrom-2: Military Service

DemoMilServiceIraqTo-2: Military Service

DemoMilServiceIraqMOS-2: Military Service

DemoMilServiceIraqFrom-3: Military Service

DemoMilServiceIraqTo-3: Military Service

DemoMilServiceIraqMOS-3: Military Service

DemoMilService-Afgh: Military Service

DemoMilServiceAfghFrom-1: Military Service

DemoMilServiceAfghTo-1: Military Service

DemoMilServiceAfghMOS-1: Military Service

DemoMilServiceAfghFrom-2: Military Service

DemoMilServiceAfghTo-2: Military Service

DemoMilServiceAfghMOS-2: Military Service

DemoMilServiceAfghFrom-3: Military Service

DemoMilServiceAfghTo-3: Military Service

DemoMilServiceAfghMOS-3: Military Service

DemoMilService-Other: Military Service

DemoLocationOtherSpecify: Military Service

DemoMilServiceOtherFrom-1: Military Service

DemoMilServiceOtherTo-1: Military Service

DemoMilServiceOtherMOS-1: Military Service

DemoMilServiceOtherFrom-2: Military Service

DemoMilServiceOtherTo-2: Military Service

DemoMilServiceOtherMOS-2: Military Service

DemoMilServiceOtherFrom-3: Military Service

DemoMilServiceOtherTo-3: Military Service

DemoMilServiceOtherMOS-3: Military Service

PCS-1-Dizzy: Feeling dizzy:

PCS-2-Balance: Loss of balance:

PCS-3-Coordination: Poor coordination, clumsy:

PCS-4-Vision: Vision problems, blurring, trouble seeing:

PCS-5-Light: Sensitivity to light:

PCS-6-Taste: Change in taste and/or smell:

PCS-Score1-Dizzy: Score for "Feeling dizzy"

PCS-Score2-Balance: Score for "Loss of balance"

PCS-Score3-Coordination: Score for "Poor coordination, clumsy"

PCS-Score4-Vision: Score for "Vision problems, blurring, trouble seeing"

PCS-Score5-Light: Score for "Sensitivity to light"

PCS-Score6-Taste: Score for "Change in taste and/or smell"

PCS-ScoreTotal: Total Score