IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 14, 2022, accepted May 6, 2022, date of publication May 17, 2022, date of current version May 26, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3175816

Adversary-Aware Multimodal Neural Networks for Cancer Susceptibility Prediction From Multiomics Data

MD. REZAUL KARIM[1,2], TANHIM ISLAM[1], CHRISTOPH LANGE[1,2], DIETRICH REBHOLZ-SCHUHMANN[3,4], AND STEFAN DECKER[1,2]

[1]Fraunhofer Institute for Applied Information Technology FIT, 53757 Sankt Augustin, Germany
[2]Computer Science 5–Information Systems and Databases, RWTH Aachen University, 52074 Aachen, Germany
[3]ZB MED - Information Centre for Life Sciences, 50931 Cologne, Germany
[4]Faculty of Medicine, University of Cologne, 50931 Cologne, Germany

Corresponding author: Md. Rezaul Karim (rezaul.karim@fit.fraunhofer.de)

ABSTRACT Artificial intelligence (AI) systems are increasingly used in health and personalized care. However, the adoption of data-driven approaches in many clinical settings has been hampered due to their inability to perform in a reliable and safe manner to leverage accurate and trustworthy diagnoses. A critical and challenging usage scenario for AI is aiding the treatment of cancerous conditions. Providing accurate diagnosis for cancer is a challenging problem in precision oncology. Although machine learning (ML)-based approaches are very effective at cancer susceptibility prediction and subsequent treatment recommendations, ML models can be vulnerable to adversarial attacks. Since adversarially weak models can lead to wrong clinical recommendations, such vulnerabilities is critical – especially when AI-guided systems are used to aid medical doctors. Therefore, it is indispensable that healthcare professionals employ trustworthy AI tools for predicting and assessing disease risks and progression. In this paper, we propose an *adversary-aware multimodal convolutional autoencoder* (MCAE) model for cancer susceptibility prediction from multi-omics data consisting of copy number variations (CNVs), miRNA expression, and gene expression (GE). Based on different representational learning techniques, the MCAE model learns multimodal feature representations from multi-omics data, followed by classifying the patient cohorts into different cancer types on multimodal embedding space that exhibit similar characteristics in end-to-end setting. To make the *MCAE* model robust to adversaries and to provide consistent diagnosis, we formulate *robustness* as a property, such that predictions remain stable with regard to small variations in the input. We study different adversarial attacks scenarios and take both proactive and reactive measures (e.g., adversarial retraining and identification of adversarial inputs). Experiment results show that the *MCAE* model based on *latent representation concatenation* (LRC) exhibits high confidence at predicting cancer types, giving an average precision and Matthews correlation coefficient (MCC) scores of 0.9625 and 0.8453, respectively and shows higher robustness when compared with state-of-the-art approaches against different attack scenarios w.r.t. *ERM* and *CLEVER* scores. Overall, our study suggests that a well-fitted and adversarially robust model can provide consistent and reliable diagnosis for cancer.

INDEX TERMS Cancer genomics, cancer type prediction, adversarial machine learning, out-of-distribution detection, deep learning, representation learning, multimodal information fusion.

I. INTRODUCTION

Cancer is caused when cells turn abnormal, divide rapidly, and spread to other tissues and organs and may be further driven by a series of genetic mutations of genes induced by selection pressures of carcinogenesis in the cells [1], [2]. The so-called marker genes including oncogenes and tumor suppressor genes are often responsible for cancer growth. When a gene is over- or under-expressed as a differentially expressed gene, the gene becomes uncontrollable proliferation or immortality of cancer cells [1], [2]. Although the difference in the average of expression values between two

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

sample classes is frequently employed in transcriptomics analyses, such difference is not the only way a gene can be expressed differentially [3]. With more than 200 different types identified to date, cancer has become the second leading cause of death worldwide [4]. According to the National Cancer Institute[1] an estimated 17.35 million new cancer cases were diagnosed in the United States in 2018, of which 609,640 people died; while there were 18 million new cases of which 9.5 million deaths were reported worldwide. The number of new cases in the United States is expected to rise to 23.6 million by 2030, which again is anticipated to increase by 70% by 2035.

Early detection of tumors is particularly important for better treatment of patients. Further, knowing the types or subtypes of cancer is a prerequisite for recommending the best possible treatment, a notable issue being the discrimination of tumor samples from normal ones [5], [29]. Once biologically relevant features have been identified, they can be used to improve the accuracy of diagnostic protocols [6]. A treatment paradigm that integrates molecular profiling approaches is accelerating clinical oncology, by which molecularly targeted agents can be identified that have improved the clinical outcomes of patients across multiple cancer types [6]. Therefore, providing accurate diagnosis for such a highly aggressive disease is a challenging problem [7].

Further, with the rapid development of new technologies in gene sequencing and with an increased importance of genetic knowledge in cancer treatment, several projects and diverse genomic data sets associated with cancer have emerged; e.g., the cancer genome atlas (TCGA) [8] is the most well-known source for omics data.[2] TCGA analyzed over 11,000 cancer cases from 33 prevalent forms of cancer as a part of the Pan-Cancer Atlas project [9] and makes available somatic mutation, gene expression (GE), DNA methylation (DNAm), copy number variations (CNVs), miRNA expressions data, as well as clinical and pathology information [10]. By acquiring insights from these data, treatment can be focused on preventive measures [11]. However, this requires the omics data, pathological reports, and imaging data to be integrated and analysed to understand the genetic and epigenetic causes of cancer before recommending appropriate treatment [12].

Artificial intelligence (AI)-based systems are increasingly getting better at optimizing treatment decision making for cancer patients [6], [12]–[15]. Machine learning (ML), in particular deep learning (DL) techniques can process large-scale data to learn biologically relevant patterns, by addressing issues such as the curse of dimensionality and heterogeneity [6]. Compared to ML-based approaches, DL approaches based on deep neural network (DNN) architectures found to be more effective to provide a more reliable and accurate diagnosis of cancer. Similar to our multimodal real-world experience [16], clinical decision-making in oncology also involves multimodal data [17]. However, since omics data are generated from multiplatform and heterogeneous sources, the high-dimensionality and heterogeneity impose great challenges to bioinformatics tools and algorithms [18].

Further, accurate diagnoses of cancer may be specific to patients with particular cancer subtypes and molecular traits. Since one type of omics data may not cover these biomarkers, providing diagnoses solely based on unimodal data may not be accurate and reliable. For example, breast cancer patients can be broadly categorized into different groups depending on the presence of ER, PGR, and HER2/neu proteins in normal cell growth, where ER, PGR, and HER2 represent a subset of breast cancer with different biological behavior and are mainly involved in determining breast cancer subtypes. This signifies the necessity of multimodal learning paradigms to provide reliable diagnosis of cancer by combining omics, bioimaging, and clinical outcomes. Therefore, numerous studies have been proposed to accurately diagnose cancer based on different types of data such as omics, radiology scans, molecular profiling, histopathology slides, and clinical factors by employing ML and DL techniques [17]. As of omics data, CNVs, DNAm, GE, and miRNA expression data are more widely used [19] in uni- and multimodal ML settings.

The multimodal information fusion (MIF) – a concept of integrating information from multiple modalities is widely used to predict an outcome measure [20]. Subsequently, computing methods for cancer diagnosis, survival analysis, and prognosis have been developed based on multiple data sources. The MIF provides several benefits over unimodal machine learning paradigm [20]:

- Multiple modalities of the same phenomenon enable making more reliable predictions.
- Complementary information from multiple modalities can be captured, e.g., the integration of genomics, proteomics, bioimaging, texts, or even clinical outcomes to support each other.
- A multimodal neural network can still operate when one of the modalities is missing (e.g., we can still rely on GE, miRNA, and CNVs modalities in case of missing bioimaging modalities).
- Since a multimodal neural network can be trained end-to-end to represent the data, both supervised and unsupervised learning tasks can be accomplished on learned representations.

Further, discriminative models may offer very limited performance guarantees when trained on a dataset, which has not been generated by the same process as the training distribution [21]. Suppose a DNN model is trained on omics data. Assuming the model that shows high confidence when evaluated w.r.t. the performance measures on the test set is then deployed for cancer diagnosis in a clinical setting. Similarly, it is not guaranteed that the model performs as expected or equally, as both training and test sets would be sampled from the same population. As a result, the model may tend to show

[1] https://www.cancer.gov/about-cancer/understanding/statistics

[2] A collection of biomolecules inside living organisms, e.g., genomics, metabolomics, and proteomics

average or poor performance on unseen data. If the test set that represents the population to a large extent and coming from the same distributions as the training set, statistical learning theory can answer this question through assumptions or manual inspection. Further, statistical learning theory would fail to relate the errors unless additional information, e.g., domain knowledge is available. Even though the DNN model may be an efficient predictive model, it is not guaranteed what would be the prediction in the case of input samples that slightly differ from the training set [22]. Using a brute-force approach, it possible to find small variations of specific samples that change the model's behavior (e.g., predictions and explanations [23], such instances become so-called *adversarial examples* (AEx) when unwanted noises are added in input samples with major content moderation. In adversarial ML theory, AEx are used to introduce different types of adversarial attacks on the models, either to sabotage their internal functionality.

Adversarial inputs (e.g., images) that look almost identical to original clean images[3] with small perturbations may be enough to misguide or fool even a robust model, unless the model does not have minimum level of adversarial robustness. Bendale et al. [25] show that it is easy to generate images that humans would never classify as a particular object class, whereas a model may still classify such images as that given class with high confidence. Consequently, a user may end up with an incorrect image label predicted, which becomes even more extreme if the test data comes from another distribution than the training set called out-of-distribution (OOD).[4] Suppose a convolutional neural network (CNN) that has been trained on millions images to solve the dog vs. cat classification problem shows high confidence at accurately recognizing images of both classes. In an adversarial scenario, the classification task can be described as follows: a user inputs an image of an elephant to get the prediction; there is a high probability that a user will get a response of an incorrect image label, i.e., either dog or cat, forcing the classifier to make such a wrong prediction. However, if we are given an independently sampled set generated from the same distribution, the model is less likely to make wrong predictions, at lest, on more than a certain number of samples with a high probability.

The ability of a model to recognize unknown or malicious inputs is important for many classification-based systems [26]. Therefore, high accuracy alone is not enough to provide trustworthy diagnosis in clinical settings, as it is also desirable that a deployed model is capable of detecting AEx or anomalous inputs [27]. Further, the use of more complex inputs in DNN magnifies the difficulty of distinguishing anomalous and in-distribution (ID) inputs [27]. Assuming a less complex CNN model trained on GE data shows 90% confidence at predicting different cancer types correctly when evaluated on a sufficiently large test set. Once the model is deployed and ready for inferencing, predicting[5] for a single instance is trivial. Suppose for a given representation of a GE example, the model predicts breast cancer with a probability of 90%, "how confident are we that: i) the input to the model is a GE example and not a CNV?, ii) the model will consistently make the correct prediction?, iii) the diagnosis will not end up with a wrong diagnosis decision?." Since the deployed model may be vulnerable to adversarial attacks, ML security researchers recommend not to trust the model blindly – especially in the presence of an adversary. Even if we assume there are no adversaries, we still would not have certainty that the input image is a GE sample.

The *robustness* of a model is difficult to interpret [23]. The only model that is fully robust for all inputs is the trivial model that returns the same prediction for all outputs, even in the presence of the strongest adversarial attacks [23]. For all other models, there is a decision boundary and some data points will be close to the decision boundary and are hence not robust, given that some parts of the neighborhood of inputs near the decision boundary will always be on each side of the decision boundary [23]. Therefore, no model is fully robust in practice. If a model can detect if an input belongs to the population distribution of the training data, it can be considered robust enough [23]. However, adversarial attacks are more critical in healthcare, especially when AI-guided systems are used to provide diagnosis aid to a doctor, e.g., in the case of an OOD data point, a learning algorithm may not only make wrong or erroneous prediction but also misguide medical doctors in clinical diagnosis [21]. Since diagnosis decisions provided by an AI system are critical and wrong decisions are not acceptable, it is essential to ensure that the model is robust to adversaries. This requires improving the adversarial robustness of the model by employing both proactive and reactive defense measures.

To date many studies have focused on multimodal diagnostic, prognostic, or survival analysis, but to our knowledge, no study has focused on improving the robustness of such multimodal neural network architectures for a similar purpose. In this paper, we propose an efficient approach to provide reliable diagnosis of cancer by means of cancer susceptibility prediction. We train a multimodal convolutional autoencoder (MCAE) architecture on multi-omics data by employing different representation learning (RL) techniques, followed by training classifiers on the embedding space to classify the cohorts into specific cancer groups. To improve the adversarial robustness of the MCAE, we take both proactive (e.g., adversarial retraining) and reactive (e.g., identification of AEx) measures. We impose content moderation and OOD adversarial attacks on the MCAE model with generated AEx, followed by performing adversarial retraining. Finally, we assess the model's robustness in a non-targeted attack scenario. Since it is impossible to build a fully robust

[3]For example, humans could easily distinguish the adversarial MNIST digits [24], often barely recognizable by humans if trivial amounts of noises are added into an image.

[4]Samples that are far away from the training distribution.

[5]In this paper, we use 'cancer type prediction' and 'cancer susceptibility prediction' interchangeably.

classifier, we model "robustness" as a property to make sure that the predictions remain stable to small variations in the input (minor perturbations), i.e., that such input variations do not flip the prediction to a completely different cancer type. We hypothesize that a well-fitted and adversarially robust model can provide consistent and reliable cancer diagnosis based on multi-omics data. The overall contributions of this paper can be summarized as follows:

- We tackle the curse of dimensionality of omics data by employing autoencoder-based representational learning techniques. We learn and construct a multimodal latent space from multimodal feature space (from multi-omics data), where both shared latent representation (SLR) and latent representation concatenation (LRC) techniques based on convolutional autoencoder (CAE) were employed.
- We not only study and observe different adversarial attack scenarios, but also take both proactive and reactive measures (e.g., adversarial retraining and identification of AEx) to ensure models are robust to adversaries and behave as intended. To the best of our knowledge, we are the first to improve adversarial robustness of DNN models towards providing trustworthy diagnosis of cancer.
- We provide comprehensive evaluations of our approach, both quantitative and qualitatively. Further, we provide comparative analyses with baseline models and state-of-the-art approaches.
- We prepare a labeled multimodal omics dataset for cancer type prediction task, which can be used for predictive modeling and to develop explainable AI systems for cancer diagnosis.

The rest of the paper is structured as follows: Section II reviews some related work, covering both unimodal and multimodal ML-based approaches for cancer diagnosis, and outlines a short overview of different adversarial attacks scenarios. Section III describes our proposed approach. Section IV discusses our experimental results both quantitative and qualitatively. Section V summarizes our research contribution, discussing potential limitations and pointing out possible outlooks, before concluding the paper.

II. RELATED WORK

In this section, unimodal and multimodal ML approaches for cancer diagnosis are discussed. Besides, "reactive" and "proactive" countermeasures against different adversarial attacks scenarios (e.g., content moderation and OOD attacks) are covered.

A. UNIMODAL APPROACHES

Numerous approaches have been proposed for analyzing genomic profiles of patient cohorts for treatment decision making [18], [28]. Genomics, bioimaging, and clinical data are used to identify rare and common transcripts, isoforms, and non-coding RNAs in cancer [19]. Besides, different types of somatic mutation data such as point mutation, single nucleotide variation (SNV), small insertion and deletion (a.k.a. INDELs), copy number aberration (CNA), translocation, and CNVs are also used. While RNA-Seq data is more widely used to identify rare and common transcripts, isoforms, and protein-coding RNAs in cancer, single nucleotide polymorphism data is used to identify segmental variations across multiple cancer genomes [13]. In a previous approach [29], we considered copy number segmentation as an important feature. We assumed the higher the segmentation means, the higher the copy number in that region. Based on this assumption, we calculated the length of a copy number and its value based on a difference between start and end positions of a CNV. Then, we represented copy number loss and gain w.r.t. negative segmentation and positive segmentation means, respectively. Copy numbers with segmentation values between a specific range were considered as noise and discarded from the rest of the calculation. Since a manual approach for CNV extraction like this often fails to extract recurrent CNV features in the case of simultaneous analysis of multiple samples [30], we used the MSeq-CNV tool for more efficient CNV extraction [31]. The extracted features we then used to train a Convolutional-LSTM (Conv-LSTM) network for cancer type prediction based on a snapshot neural ensemble method.

Sun et al. [32] proposed GeneCT, which constrains input genes into the oncogenes and tumor suppressors categories to determine the cancerous status and transcription factors to classify cohorts into the tissue of origin, achieving an overall accuracy of 97.8%. Lyu et al. [10] and Mostavi et al. [33] embedded the RNA-Seq data from the Pan-Cancer atlas project into 2D images and trained a CNN to classify 33 tumor types. Lyu et al. [10] trained a CNN model with a 2D mapping of the GE samples, achieving an accuracy of 95%. Besides, they provided a data interpretation approach based on guided-gradient class activation maps (Grad-CAM) [34]. Based on different designs of gene embeddings and convolution schemes, Mostavi et al. [33] implemented 1D-, 2D-Vanilla-, and 2D-Hybrid CNN models, the latter of which achieved a classification accuracy of 93.9%. Further, they extended the 1D-CNN model for the prediction of breast cancer subtypes and achieved an average accuracy of 88.42% among subtypes.

Inspired by these approaches, we proposed an explainable approach called *OncoNetExplainer* [18] for cancer type prediction based on GE data of 9,074 cancer patients covering 33 different cancer types from the Pan-Cancer Atlas. Similar to their approach, our approach first embeds high dimensional RNA-Seq data into 2D images, followed by training vanilla CNN and VGG16 networks on an embedding space with the Grad-CAM++ [35] technique. By averaging all the normalized heatmaps from the same class, we generated class-specific heatmaps, where a higher intensity pixel represents a higher significance to the final prediction, which indicates higher importance of corresponding genes and the GE values. Relevant driver- and ranked top-k genes are then identified

across cancer types w.r.t. intensity rankings and mean absolute importance threshold. To provide a comparison with baselines, we validated the findings based on annotations provided by TumorPortal[6] and the SHAP [36] interpretability technique.

The majority of these approaches did not systematically evaluate the effectiveness of different representation learning techniques from high dimensional data (e.g., qualitative evaluation of the learned embeddings) and assess their impact on the classification accuracy, even though they outperformed previous approaches. In a recent approach Itzhacky et al. [37], devised a novel deep learning method for predicting gene dependencies and drug sensitivities from GE measurements. By combining dimensionality reduction strategies, they are able to learn accurate models that outperform shallow DNN or ML models. Chen et al. [38] proposed gene superset autoencoder (GSAE), a multilayer autoencoder with a priori defined gene sets that retain the crucial biological features in the latent layer. They introduced the concept of the gene superset, an unbiased combination of gene sets with weights trained by the autoencoder, where each node in the latent layer is a superset in order to reduce the number of weights to be estimated. Using a GSAE model at its latent layer, they demonstrated that gene supersets retain sufficient biological information w.r.t. tumor subtypes and clinical prognostic significance.

B. MULTIMODAL APPROACHES

Although the development of unimodal representations has been widely studied and focused on unimodal RL, many current clinical methods fail to effectively utilize the large-scale multimodal data available today for cancer patients [39]. Therefore, there has been a shift to multimodal learning [16]. Further, the performance of unimodal information fusion architectures is greatly limited by their inability to detect and combine useful and complementary information from heterogeneous representations stemming from a set of distinctive modalities [40], [41]. The ability to represent multimodal data efficiently forms the backbone of many predictive models. The concept of multimodal ML has emerged, which aims to build a ML model capable of processing and relating information from multiple modalities [16]. Subsequently, recent approaches are more focused on multimodal representations from different types of data involving simple or shared concatenation of unimodal ones [16]. A multimodal representation is a representation of data using information from multiple entities [16].

Nevertheless, when it comes to multimodal genomics data for precision oncology, the diagnosis decision may rely on different types of data since one type of omics data may not cover the patient's biomarkers. Such datasets can be characterized as multimodal. From a biological perspective, providing cancer diagnosis based on multiple input types (e.g., genomics, proteomics, or imaging data) is analogous to creating knowledge together from multimodal perspectives [42]. Further, how to combine different types of data from heterogeneous sources, how to deal with different levels of noise and artifacts, and how to deal with missing data are a few challenges in the multimodal RL. The majority of multimodal methods have so far focused on representation fusions [41], either by combining representations before the classification, called feature level fusion, or by combining the results of classifications performed in single-mode representations in another analysis, called *decision level fusion* [43].

A substantial disadvantage of information fusion, however, is that the model is not able to handle missing data (a.k.a. mode collapse[7]). Another issue is that the reconstruction losses can go out of bounds during the pre-training phase of modality-specific RL. To overcome these limitations, it is common to pre-train such representations using an autoencoder on unsupervised data [16]. Further, a well-trained model is subject to enough data, as DNNs require a lot of labeled training data. A major advantage of DNN-based joint representations comes from their often superior performance and the ability to pre-train the representations in an unsupervised manner [44]. The multi-layer nature of a DNN and its successive layers are hypothesized to represent the data in a more abstract way [16]. Each modality starts with several individual neural layers followed by a hidden layer that projects the modalities into a joint space [45]. Then, it is common to capture the output of the deepest layer as a form of data representation from individual modalities [16], [45]. The joint representation is then passed through multiple hidden layers or used directly for prediction. Such multimodal RL generates distributed vectors by mapping multiple modalities of information to a single mathematical space based on a distance or similarity measure [41]. In such a network topology, multimodal representation and fusion are jointly learned [20] and the latent representation concatenation architecture learns a single latent representation from each modality.

A multimodal autoencoder architecture (MAE) proposed by Ngiam et al. [46] extended the idea of using an autoencoder in the MIF setting. They used stacked denoising autoencoders to represent each modality individually and then fused them into a multimodal representation using an encoding layer [16], [45]. Besides, it is common to fine-tune the representation on supervised learning tasks. This helps an MAE to learn from the prior distribution flexibly by capturing features from a target distribution [45], [47], [48]. Therefore, MAE-based approaches have been applied in a variety of settings such as natural language understanding (e.g., document and dialogue modeling) [45], emotion recognition [47], and near-infrared spectroscopy resting state prediction from multimodal electroencephalographic signals [49]. Besides, to make the most of multimodal data, a large amount of literature is devoted to the

[6] http://www.tumorportal.org/

[7] A phenomenon in multimodal ML, where a model may fail due to missing modalities (one or multiple) or corrupt modality.

construction of integration methods for predicting cancer survival [48]. In particular, the multimodal system proposed by Wang et al. [48] is extended by adding the capability of handling multimodalities, by discarding a small portion of patient data during the multimodal fusion approach, which does not have all modalities in the multimodal network. A fully connected layer is then added for supervised learning tasks such as cancer subtypes predictions. Furthermore, by concatenating multimodal data to one matrix, Zhang et al. [50] developed multiple kernel ML methods by combining min-redundancy max-relevance (mRMR) feature selection algorithm for glioblastoma multiforme prognosis prediction. Some integration strategies such as joint strategy and alignment strategy have been proposed to deal with multimodal data. While methods based on joint strategy fuse multiple data sources through concatenation (e.g., Sun et al. [51] present a triple modal DNN to learn effective representation from gene expression, CNV, and clinical data; Gao et al. [52] use GE, CNVs, and clinical data to construct a multimodal graph neural network), methods based on alignment strategy maximize the common information learned from different data sources through alignment (e.g., to boost prediction performance, Wang et al. [53] design a novel cluster-boosted multitask learning framework for survival analysis). Cheerla et al. [39], who developed an unsupervised encoder to combine four data sources into one single feature matrix, can force different modalities from same patient cohorts, while avoiding mode collapse. To achieve a modality-invariant representation from multimodal data, Tong et al. [54] proposed a cross-modality autoencoder to maximize the consensus among modalities. However, by utilizing only common information, these approaches could miss valuable complementary properties among multiple modalities.

Therefore, research simultaneously utilizes the common and complementary information between multimodal data. Wang et al. [55] create a sample similarity network for each data source, which is then fused to construct an integrated patient view. In another approach [56], the similarity network fusion algorithm is utilized to generate a sample similarity matrix, and mRMR is utilized to conduct a feature selection to obtain sample feature matrix. Based on these two matrices, they construct a graph convolutional network (GCN) to predict cancer survival. As the clinical decision-making in oncology involves multimodal data such as radiology scans, molecular profiling, histopathology slides, and clinical factors, a new approach called Deep Orthogonal Fusion (DOF) model has been proposed by Braman et al. [17]. To predict the overall survival of glioma patients from diverse multimodal data, the DOF model first learns to combine information from multimodal inputs into a comprehensive multimodal risk score, by combining embeddings from each modality via attention-gated tensor fusion. Then, to maximize the information gained from each modality, they introduce a new loss function called multimodal orthogonalization loss that increases model performance by incentivizing constituent embeddings to be more complementary. Recent approaches [17], [19], [39], [40] that benefited from MIF have shown that cancer diagnosis based on multimodal data is both clinically and biologically more accurate than approaches based on unimodal information fusion. However, the majority of MIF approaches fundamentally rely on autoencoder-based RL. In particular, convolutional, variational, and generative adversarial autoencoders are more widely used to learn representations from multimodal data [44] to be used for different downstream learning tasks such as cancer type prediction, survival predictions, prognostic biomarker discovery, etc.

C. ADVERSARIAL ATTACKS AND DEFENSES

Numerous approaches have been proposed to creating AEx, and some of them are already tackled by a countermeasure [24]. The concept of AEx was first formulated by Dalvi et al. [57] as a game between adversary and classifier, in which the attack and defence on AEx can be correlated as an iterative game. The *first gradient* approaches are the earliest approaches to generating AEx and attacking linear support vector machines (SVM) [58]. Szegedy et al. [59] introduced the concept of AEx in neural networks, where AEx was generated using an L-BFGS method[8] as follows [24]:

$$\begin{aligned} &\min_{x'} \; c\|\eta\| + J_\theta\left(x', l'\right) \\ &\text{s.t.} \quad x' \in [0, 1], \end{aligned} \tag{1}$$

where c is a constant used to approximate values of AEx by linear-searching with $c > 0$, making it computationally expensive to find the optimal value of c. Subsequently, Goodfellow et al. [60] proposed a fast method called Fast Gradient Sign Method (FGSM) to generate AEx. The simplicity and effectiveness of FGSM lie in the fact that it needs to perform a one-step gradient update along the direction of the sign of gradient at each pixel, where the perturbation is expressed as [60]:

$$\eta = \epsilon \, \text{sign}\left(\nabla_x J_\theta(x, l)\right), \tag{2}$$

where the magnitude of the perturbation ϵ is computed using back-propagation. The corresponding adversarial example x' for x is calculated as $x' = x + \eta$ [24]. Then the sign of gradient in FGSM is replaced with the raw gradient: $\eta = \nabla_x J(\theta, x, l)$. Since a one-step attack is not only easy to transfer to another domain but also easy to defend [24], Dong et al. [61] improved FGSM by employing momentum to generate AEx more iteratively, where the gradients were calculated as follows:

$$\mathbf{g}_{t+1} = \mu \mathbf{g}_t + \frac{\nabla_x J_\theta\left(x'_t, l\right)}{\left\|\nabla_x J_\theta\left(x'_t, l\right)\right\|}. \tag{3}$$

AEx are then subsequently generated by $x'_{t+1} = x'_t + \epsilon \; \text{sign} \; g_{t+1}$. DeepFool [62] is another method used to generate AEx by finding the closest distance from the original input to the decision boundary. To overcome the non-linearity

[8]Broyden-Fletcher-Goldfarb-Shanno is an iterative method for solving nonlinear optimization problems.

in high dimensional feature space, iterative attack is introduced with a linear approximation: starting from an affine classifier (AC), the minimal perturbation for an AC is the distance to the separating affine hyperplane $F = \{x : w^T x + b = 0\}$, where the perturbation of an AC f is $\eta^{(x)} = -\frac{f(x)}{\|w\|_{\text{js}}^2} w$ [24]. As a binary differentiable classifier, following iterative method is applied to approximate the perturbation by considering f linearized around x_i at each iteration [24]:

$$\begin{aligned} &\arg\min_{\eta_i} \|\eta_i\|_2 \\ &\text{s.t.} \quad F(x_i) + \nabla F(x_i)^T \eta_i = 0. \end{aligned} \tag{4}$$

For multi-class setting, the closest hyperplanes evolve for a more general ℓ_p norm, $p \in [0, \infty)$. Thus, DeepFool provides less perturbation compared to FGSM [24].

D. OOD ATTACKS AND DETECTORS

Several approaches have been proposed to identify OOD accurately. Sehwag et al. [63] categorized them into unsupervised and supervised approaches: unsupervised approaches include reconstruction-error based approaches using autoencoders, classification based approaches and probabilistic models. Further, some research works have treated OOD attacks and subsequent detection as anomaly detection problems. The majority of outlier detectors based on unlabeled data fail to scale up to complex data modalities [63]. In contrast, supervised detectors have been found most successful with complex input modalities like images and language [63]. Supervised approaches model features of ID data at output or in the feature space for detection. Lee et al. [64] explain that choosing test samples from far away from the training distribution is a fundamental requirement for deploying a model in many real-world applications. However, DNNs with the softmax classifier is known to produce highly overconfident posterior distributions even for AEx. Jie et al. [21] proposed a likelihood ratio method for deep generative models for OOD detection. Since OOD is heavily affected by population-level background statistics, they demonstrated that the likelihood ratio method could rectify this issue.

Vernekar et al. [65] proposed an efficient approach to generate OOD samples based on a manifold learning network. OOD samples are used to train a classifier with an extra class (i.e., $n + 1$ classes, where the $(n + 1)^{\text{th}}$ class represents OOD samples). DeVries et al. [22] show that, after training a DNN using IsoMax, ODD samples can be identified by simply calculating the entropy of the network's output probabilities. Their combined approach (IsoMax for training and ES for ODD during inference) was found to be fast, scalable, and unexposed. Based on the outlier exposure (OE), Rajati et al. [66] proposed a novel loss function Outlier Exposure with Confidence Control (OECC). They show that efficient optimization of OECC achieves SotA results in OOD detection with OE on both image and text classification tasks without needing many OOD samples. Shalev et al. [67] proposed to use multiple semantic dense representations instead of using sparse representations w.r.t the target labels. Choi et al. [68] proposed to scale OOD detection to high-dimensional data to learn a tractable likelihood approximation of training distribution and use it to reject unlikely inputs. They also exposed that the likelihood models on natural data are either susceptible to OOD errors or assign large likelihoods to samples from other datasets [68].

Liang et al. [64] propose OOD image detection in neural networks (ODIN), a simple and effective method that does not require any change to a pre-trained model. ODIN is based on the observation that using feature scaling and adding small perturbations to the input can separate the softmax score distributions between in- and OOD images, allowing for more effective detection. Qing et al. [69] proposed two-head deep CNN architectures by maximizing the discrepancy between two classifiers. The two-head CNN consists of one common feature extractor and two classifiers with different decision boundaries but can correctly classify ID samples. Further, since many discriminatively trained DNN classifiers only produce reliable predictions for ID samples, detecting OOD samples is challenging. Assuming an OOD example is outside of the closed boundary of an ID instance, typical DNN classifiers have no knowledge of this boundary. Since OOD samples are expected to be outside of the closed boundary of ID samples, a model is incapable of detecting OOD samples during the inferencing, recent approaches have been proposed to embed the logic into the learning algorithm and explicitly train the classifier with OOD samples close to the ID boundary. However, SotA approaches often fail to cover the entire ID boundary effectively, thus resulting in a sub-optimal OOD detector [67], [70].

Yu et al. [71] proposed an unsupervised OOD detection approach based on maximum classifier discrepancy (UMCD). In their approach, a two-head neural network is constructed consisting of n extractors and two gradient-based classifiers. Using a two-step process, supervised training for classifying ID samples correctly is followed by unsupervised training to maximize the discrepancy to detect OOD samples. In a very recent approach, an outlier detection approach called self-supervised outlier detection (SSD) is proposed by Sehwag et al. [63]. SSD works based on only unlabeled ID data in which self-supervised representation learning followed by a Mahalanobis distance-based detection is employed in the feature space. Further, they formulate a few-shot OOD detection technique in which the detector has access to only 1 to 5 samples from each class of the targeted OOD dataset. Yet, they show that only a few OOD samples are sufficient to guide the classifier's decision boundary to be bounded around the ID regions evidenced by their OOD detection results.

III. PROPOSED APPROACH

In this section, we cover our approach in detail.

A. PROBLEM STATEMENT

The problem of classifying the patients is grouping them into a specific cancer type based on their unimodal or

multimodal genomic profiles consisting of CNVs, miRNA, and GE. It involves: i) cancer diagnosis utilizing cancer type prediction from the individual patients' genomic profiles, ii) improving adversarial robustness to defend different types of adversarial attacks. We formulate the first problem as a multi-class classification problem and train a *classifier* in both unimodal and multimodal settings. In the following, we introduce essential terminology, notations, and definitions.

An *instance* x is an M-tuple: $x = (v_1, v_2, \ldots, v_M)$, where a_i is a feature name and v_i for $1 \leq i \leq M$ is the value of a_i for x from a real-valued domain $\mathbb{R}$. A pair $D = (\tilde{X}, \tilde{Y})$ is called a dataset, where $\tilde{X}$ is an N-tuple of M-instances and $\tilde{Y}$ is an N-tuple of labels $l \in L$. Each y_i is called the *label* of $x \in \tilde{X}$. The set of all labels Y of $\tilde{Y}$ is called the *decision space*, where $\tilde{Y} = (y_1, \ldots, y_N)$. Let $D = (\tilde{X}, \tilde{Y})$ be a dataset, let $\tilde{X}$ be an N-tuple of M-instances, X be the set of all instances in $\tilde{X}$, and $\tilde{Y}$ the N-tuple of labels $l \in L$.

Let Θ is a set of parameters, where a parameter is a pair (*param*, *value*). A *classifier* is a parameterized function $f : X \times \Theta \rightarrow \mathbb{R}$ that maps an input instance x from a feature space X to a decision $y \in L$ and returns an output called *prediction*. A prediction $\hat{y} = f(x_i, \theta)$ is accurate for model f and parameter θ if and only if $\hat{y} = \tilde{Y}[i]$ for $x = \tilde{X}[i]$, where $1 \leq i \leq M$. If classifier f requires k different input types (i.e., unimodal datasets) with shared labels Y to generate the decisions, we form a multimodal dataset by combining multiple unimodal datasets, where modalities refer to the way to integrate multiple input modalities [16]. We write $D = \{d_1, d_2, \ldots, d_k\}$ a multimodal dataset, where each input type d_k represents individual modalities in which $d_k \in \mathbb{R}^{N \times M}$ is equal to a unimodal dataset. Figure 1 shows how different input modality combinations are used to create unimodal and multimodal datasets.[9]

Literature has defined robustness as an invariant over the relation of two inputs without having to rely on specifications of the model and the ground truth [23]. Since we aim to improve both overall robustness of the model and individual diagnosis robustness, we formulate both local and global robustness. For an input sample x, the problem of local robustness is to ensure that the prediction $\hat{y}$ for model f remains consistently the same for all inputs x in the neighborhood of x, where the neighborhood is defined w.r.t. a distance function δ, where the maximum distance Δ is computed as [23]:

$$\forall x'.\delta\left(x, x'\right) \leq \Delta \Rightarrow f(x) = f\left(x'\right). \tag{5}$$

The above distance can be formulated as the amounts of noise to all input features or arbitrary changes to a few input features or more complicated transformations (e.g., OOD samples). Since this definition does not require knowledge about the ground truths for $f(x)$ or $f\left(x'\right)$, it is reasonable to hypothesize that the prediction will stay consistent within a neighborhood, regardless it is correct or not [23]. In addition to local robustness, the global robustness property for model f is formulated as an average robustness for all inputs in modality k, thereby measuring the global robustness as the average distance from each input to the nearest AEx [23].

[9]Considering the length of the manuscript, we cover the details of data selection and preprocessing in supplementary materials.

B. CONSTRUCTIONS AND TRAINING OF MCAE

Data for individual modalities is first fed into the encoder f_θ module of MCAE, such that each input modality X_k is transformed into a modality-specific latent representation Z_k, with a nonlinear mapping $f_\theta : X \rightarrow Z_k$, where θ are the learnable parameters and $Z_k \in \mathbb{R}^K$ is the learned embedding in which $K \ll X_k$. To parameterize f_θ, we employ neural network-based RL called convolutional autoencoders (CAE), owing to their function approximation properties and feature learning capabilities from genomic data.

The learned representation is then fed into the decoder module to reconstruct $X_k{}'$, similar to the original input X_k. Parameters for the modality-specific network are optimized to minimize the error between the decoder's output and the input signal. The latent representations of all modalities are then concatenated into a single representation Z. Then we construct the MCAE classifier by feeding the latent space to a fully connected softmax layer f. Eventually, f maps each data point x to an output $f(x) \rightarrow y$ in the embedded space z to classify the samples based on patient profiles.

1) REPRESENTATION LEARNING AND UNSUPERVISED PERTAINING

In a recent approach [19], we performed the RL based (it was a MAE architecture, technically) on the concept of SLR (see fig. 2b) in which the shared feature representation was generated from multimodal genomics data. Learned representations were then used for breast cancer subtype prediction. The MAE model performed moderately well for the sub-typing tasks. The key reason behind such a low performance was lack of enough data and high pretraining and reconstruction losses during the LR phase. Recently Patrick et al. [42] proposed another multimodal concept of learning called LRC (see fig. 2a). Based on several studies, covering text classification, sequence data, and imaging, they identified the following potential limitations of SLR:

- The reconstruction loss for LRC is significantly lower compared to SLR.
- When a classifier is trained on features learned by LRC, accuracy improves significantly, which is largely backed by lower reconstruction loss.

Considering the limitations of MAE and SLR, MCAE is constructed based on a CAE and LRC-based RL, as shown in fig. 2a Besides, to provide a comparative analysis, another variant of MCAE is trained based on SLR without covering the training details, as shown in fig. 2b. We treat them as two different models, namely f_1 for $MCAE_{slr}$ and f_2 for $MCAE_{lrc}$. A simple AE can be used to reconstruct an output similar to the original input. However, it cannot handle multimodal

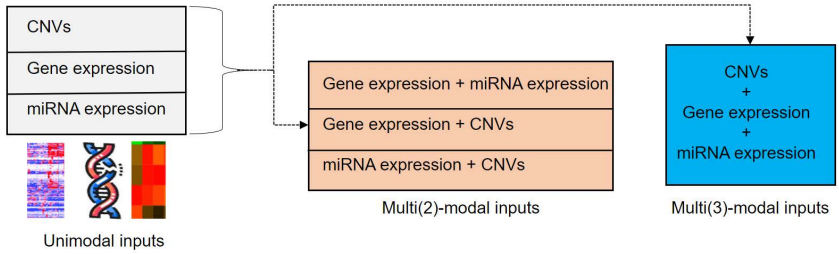


FIGURE 1. Unimodal and multimodal input combination: CNVs, GE, or miRNA expression.

inputs. Weights of the encoder module are learned from both non-corrupted and unlabeled data. Subsequently, noisy supervised data with missing modalities is not suitable for learning latent features. Nevertheless, the difference among input modalities is very large in terms of dimensionality. For example, sometimes GE and CNVs modalities come with about 20,000 features, but miRNA data has only 5,000 features. A non-trivial challenge in modality-specific RL and limitations of MAE architecture is enough motivation to employ multimodal RL and classification based on MCAE architecture.

Although CNNs are known as effective feature extractors, instead of using manually engineered convolutional filters in CNN, convolutional and pooling layers can be stacked together to construct a stacked convolutional autoencoder (SCAE), to leverage better feature extraction capability [72]. This makes CAE, compared to vanilla AE based multimodal learning, more effective for very high dimensional data [44]. In particular, CAE learns more optimal filters by minimizing the reconstruction loss, which results in more abstract features from the encoder (e.g., pixel-level features, when genomic samples are embedded into 2D pixel-space [33]). This helps stabilize the pre-training, and the network converges faster by avoiding corruption in the feature space [73]. Since the individual latent representation is required to have the same dimensionality [42], the MCAE architecture is used to generate a combined representation for all input modalities, instead of one latent representation for each input modality. From the network topological point of view, MCAE creates hierarchical hidden units by stacking multiple CAEs, which have a strong connection between nodes across the modalities. Training MCAE involves pre-training to leverage the RL, followed by supervised fine-tuning on the learned representation.

Pre-training the MCAE model is similar to training a two-stage CAE network: the first stage represents modality-specific learning. The second stage corresponds to cross-modality. Pre-training is performed greedily on each layer of the network, which corresponds to RL with individual CAEs. The individual modality of MCAE represents a specific modality for each type of data. Individual modality CAE is not only a one-layer CAE but also a multilayer and gradually shrinking CAE with a different number of layers per modality. Assuming input $X_k \in \mathbb{R}^D$ for each of $k \in \mathbb{R}^K$ modalities is consisting of n samples, a convolutional layer of CAE calculates the convolutional feature map. Max-pooling operation is then performed, which downsamples the output of the convolutional layer by taking the maximum value in each non-overlapping sub-region. Thus, X_k is mapped and transformed into a lower-dimensional embedding space Z_k. The latent-space representation $Z_k = g_\phi(X_k)$ is learned in the bottleneck layer [42]:

$$Z_k = h_k = g_\phi\left(X_k\right) = \sigma\left(W_k \oslash X_k + b_k\right), \tag{6}$$

where the encoder is a sigmoid function $g(.)$ parameterized by ϕ, while the decoder function $f(.)$ is parameterized by Θ. The final feature maps Z_k are latent variables, specific to modality k. In eq. 6, where ϕ are trainable parameters (including a weight matrix $W_k \in \mathbb{R}^{p\times q}$ and a bias vector $b_k \in \mathbb{R}^q$ specific to respective modality k, where p and q are the numbers of input and hidden units), $\oslash$ is the convolutional operation, and σ is the exponential linear unit (ELU) [74] activation function. The decoder module reconstructs the original input X_k from the latent representation Z_k using the decoder function $f(.)$. The hidden representation h_k is mapped back as a reconstructed version $\hat{X}_k$, similar to the original input X_k, as follows [42]:

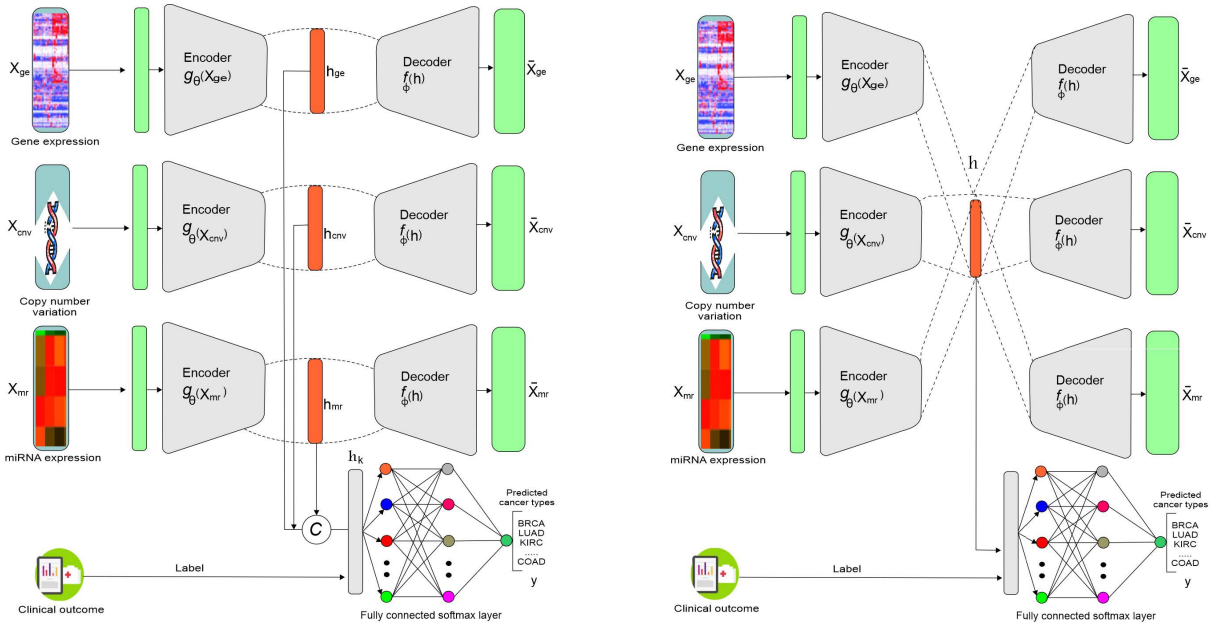
$$\hat{X_k} = f_\theta\left(Z_k\right) = f_\theta\left(g_\phi(X_k)\right), \tag{7}$$

where the parameters (θ, ϕ) are jointly learned to output a reconstructed version of the original input. As this is analogous to learning an identity function, such that $\hat{X_k} \approx f_\theta\left(g_\phi(X_k)\right)$, $f_\theta\left(g_\phi(X_k)\right)$ is equivalent to $\Psi\left(\hat{W}_k * h_k + \hat{b}_k\right)$, which changes eq. 7 into:

$$\hat{X_k} = \Psi\left(\hat{W}_k \odot h_k + \hat{b}_k\right), \tag{8}$$

where $\odot$ is the transposed convolution operation, θ are trainable parameters, i.e., a weight matrix $\hat{W}_k \in \mathbb{R}^{n\times p}$, a bias vector $\hat{b}_k$) specific to the modality k, and a sigmoid activation function Ψ. Let X_c, X_e, and X_r be the CNVs, GE, and miRNA expression input modalities, respectively. As shown in fig. 2, samples for each input modality are first embedded into 2D images, i.e., each genomic profile is reshaped into a 144×144 image by adding zero padding around the edges and normalizing the pixel values to [0,255]. Subsequently, each X_k is transformed into the following hidden representations [19].

$$\begin{aligned} h_c &= \sigma\left(W_c \oslash X_m + b_c\right) \\ h_e &= \sigma\left(W_e \oslash X_e + b_e\right) \\ h_r &= \sigma\left(W_r \oslash X_r + b_r\right), \end{aligned} \tag{9}$$



(a) Representation learning and classification based on LRC (b) Representation learning and classification based on SLR

FIGURE 2. **The fusion architectures for multimodal representation of multi-omics data and cancer type prediction.**

where $\{W_c, W_e, W_r\}$ are encoder's weight matrices, $\{b_c, b_e, b_r\}$ are bias vectors for CNV, GE, and miRNA modalities, respectively. Last element of the hidden dimension is the dimensionality of the modality-specific latent representation. As each of the X_c, X_e, and X_r input modalities are very high dimensional, with huge difference w.r.t. dimensionality, the mean squared error is used as the reconstruction loss:

$$L_k(\theta, \phi) = \frac{1}{n}\sum_{i=1}^{n}\left(X_k - \hat{X}_k\right)^2 + \lambda \left\|W_k\right\|_2^2, \tag{10}$$

where λ is the activity regularizer and W_k are network weights specific to input modality k. In stage-2, which is a cross-modality, a concatenation layer is placed to concatenate individual latent representations h_m, h_e and h_r into a single representation dimensionality ϕ:

$$h_{mcae} = \sigma\left(W_{mcae}\left[h_m \oplus h_e \oplus h_r\right] + b_{mcae}\right), \tag{11}$$

where $\oplus$ is the concatenation operation. The whole MCAE is pre-trained such that the outputs of the final de-convolution operation on the hidden representation h_{mcae} generate the original representation [48]:

$$\left[\hat{h}_m \oplus \hat{h}_e \oplus \hat{h}_r\right] = \Psi\left(\hat{W}_{mcae} \odot h_{mcae} + \hat{b}_{mcae}\right). \tag{12}$$

The above equation can be decomposed into the following individual modality-specific representations:

$$\begin{aligned}\hat{X}_c &= \Psi\left(\hat{W}_c \odot \hat{h}_m + \hat{b}_m\right)\\ \hat{X}_e &= \Psi\left(\hat{W}_e \odot \hat{h}_e + \hat{b}_e\right)\\ \hat{X}_r &= \Psi\left(\hat{W}_r \odot \hat{h}_r + \hat{b}_r\right),\end{aligned} \tag{13}$$

where $\left\{\hat{W}_c, \hat{W}_e, \hat{W}_r\right\}$ are the decoder's weight matrices, $\left\{\hat{b}_c, \hat{b}_e, \hat{b}_r\right\}$ are bias vectors for the CNVs, GE, and miRNA modalities, and Ψ is the sigmoid activation function. However, fusing such multimodalities involves a variable number of densely connected sigmoid layers.

2) SUPERVISED FINE-TUNING AND CLASSIFICATION

The latent vector h_{mcae} is feed into a fully-connected softmax layer for the classification, by optimizing the following categorical cross-entropy (CE) loss using an AdaGrad optimizer during back-propagation:

$$L_{ce} = -\sum_{m=1}^{C} y_k \log\left(\hat{y}_k\right), \tag{14}$$

where $m \in \mathbb{R}^C$ is the number of classes, y_k are ground-truths for modality k and $\hat{y}_k$ are the predicted outputs. The softmax activation function (used in the last dense layer to transforms the output into a vector of real numbers within the range of $(0, 1)$). This can be considered as a probabilistic interpretation, i.e., the probability distribution over the classes, before computing the CE loss. Further, effects of adding Gaussian noise layers is observed to improve model generalization for unseen test data. Gaussian noise is a natural choice as a corruption process for real-valued inputs, which makes it suitable for introducing noise to input values or between hidden layers. The reconstruction loss of individual modality L_k and CE loss L_{ce} of the entire MCAE architecture are then combined and optimized jointly [42]:

$$L_{mcae} = \sum_{i=0}^{n} \alpha_r L_k + \alpha_c L_{ce}, \tag{15}$$

where α_r and α_c are the regularization weights assigned to modality-specific (i.e., L_k) and CE specific (L_{ce}) loss functions, respectively.

C. FORMULATING THREAT MODELS

Yuan et al. [24] decomposed threat models from several different aspects: i) *adversarial falsification* – negative and positive samples are generated with false positive and false negative attacks, respectively, ii) *adversary's knowledge* – if the deployed model is a white-box or a black-box, iii) *adversarial specificity* - targeted or non-targeted attacks, and *attack frequency* – one-time or iterative attacks. These aspects can be categorized into black-box and white-box attacks. In a white-box attack scenario, it is assumed that the adversary has sufficient knowledge about the model, including the training data, model architectures, hyper-parameters, numbers of layers, activation functions, and model weights [24]. To introduce successful attacks on a white-box model, AEx is generated by calculating model gradients.

In a black-box attack, it is assumed that the adversary has no access to or knowledge about the model, but knows what the model is for (e.g., model's confidence score). Although the majority of adversarial attacks are white-box attacks, they can be applied to a black-box scenario too, due to the transferability of AEx [75]. Targeted attacks misguide a model to a specific class in a multiclass classification problem, e.g., an adversary can fool the cancer types classifier to predict all the AEx of type breast cancer. Targeted attacks maximize the probability of a targeted adversarial class, while a non-targeted attack does not assign a specific class to the model's output, i.e., the class output can be arbitrary [24]. Since non-targeted attacks are easier to implement compared to targeted attacks, we introduced only non-targeted attacks in a black-box scenario, thereby introducing only two types of adversarial attacks for each target model, as outlined in Figure 3. We generate AEx from existing train samples using FGSM and DeepFool. Then, we perform the adversarial attacks, before retraining and evaluating the models.

D. ADVERSARIAL ATTACKS TO MODELS

We generate AEx with the content moderation across samples by crafting original examples with FGSM and DeepFool. As for the first one, gradients are computed using backpropagation. That is, for a given trained model f and original input data sample x, generating an adversarial example x' can be formulated as a box-constrained optimization problem [24]:

$$\begin{aligned} &\min_{x'} \ \|x' - x\| \\ &\text{s.t.} \ \ f\left(x'\right) = y' \\ &\qquad f(x) = y \\ &\qquad l \neq y' \\ &\qquad x' \in [0, 1], \end{aligned} \tag{16}$$

where y and y' are the predicted labels of x and x', and $\|\cdot\|$ is the distance between two samples. Let $\eta = x' - x$ be the perturbation added on x to minimize the perturbation while misclassifying the prediction, θ is the model parameter, x is the input, y is the target associated with x, and $J(\theta, x, y)$ is the cost used to train f. The cost function around the current value of θ can be linearized by obtaining an optimal max-norm constrained perturbation as outlined in eq. 2.

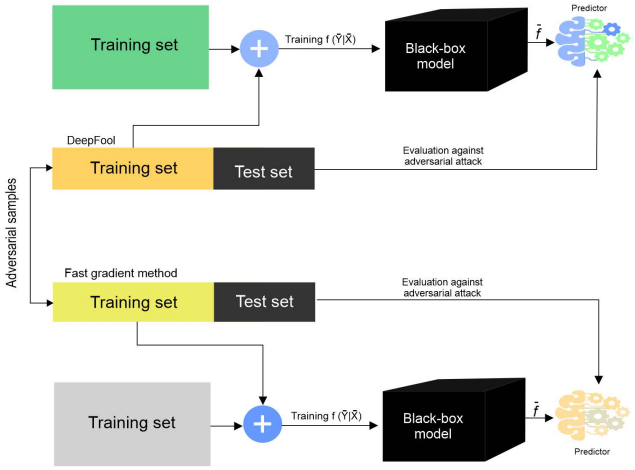


FIGURE 3. Different types of attack scenarios: up: content moderation with fast gradient sign method [60], below: crafting adversarial samples with DeepFool.

For the latter, DeepFool is employed to generate AEx. DeepFool is an iterative optimization-based approach, which provides less perturbation compared to FGSM [24], yet it has higher success rates under the same norm objective in a white-box setting [24]. Therefore, AEx generated with FGSM are used to introduce OOD attacks. We expect the AEx to be close to the original samples and be imperceptible to a human. Nevertheless, the same AEx are often misclassified by a variety of classifiers with different architectures or trained on different subsets of the training data [24], [60].

E. DEFENCES AGAINST ADVERSARIAL ATTACKS

Bendale et al. [69] introduced "reactive" and "proactive" countermeasures against adversarial attacks. The former deals with detecting AEx after an ML model is deployed. Examples include adversarial detecting, input reconstruction, and network verification. The latter is about making an ML more robust before an adversary generates and introduces an attack. Examples include network distillation, adversarial (re)training, and classifier robustifying. We employed only the adversarial retraining and input reconstruction, as shown in fig 4.

1) PROACTIVE MEASURE: ADVERSARIAL RETRAINING

We add noises in all input modalities with zero mean and standard deviations of 0–200% (k) of i^{th} per level average (μ_i), or $N(0, k\mu)$ per feature. This helps models learn robust features with little variation, which we hope to improve the generalization for multimodal learning scenario. Although introducing minor noises to the input helps improve model generalization, adversarial retraining is also necessary to

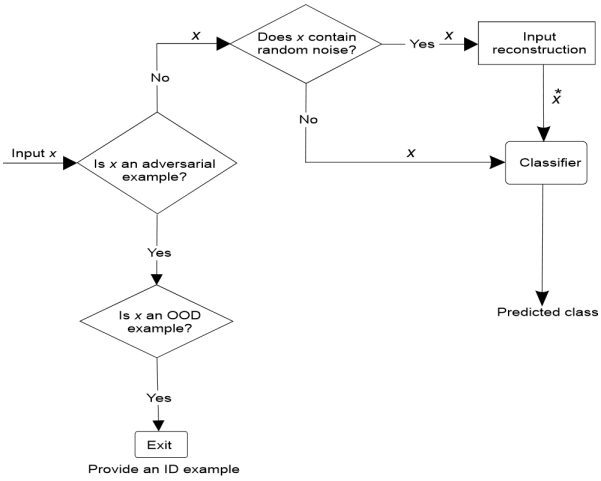


FIGURE 4. Workflow for the reactive and proactive measures against adversarial attacks at inference time.

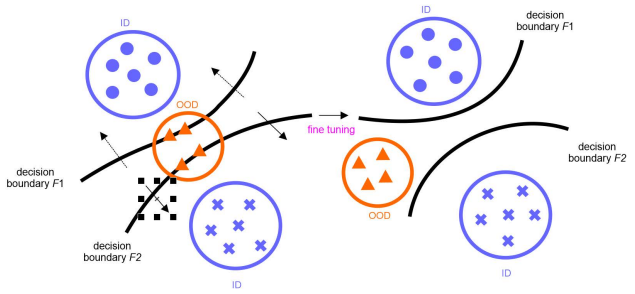


FIGURE 5. OOD sample detection w.r.t. discrepancy between two different classifiers based on [69].

improve the consistency of the model.[10] For model $f : \mathbb{R}^D \rightarrow \{1, \ldots, k\}$,[11] assuming $f(x) = \text{argmax}_i\,(z(x)_i)$, where $z(x) \in \mathbb{R}^k$ is the final layer output, and $z(x)_i$ is the prediction score w.r.t. the i^{th} class. Similar to the literature [76], the objective is formulated as the following optimization problem w.r.t. finding the minimum perturbations [76]:

$$\underset{x}{\text{argmin}}\,\{d\,(x, x_0) + cL(f(x), y)\}\,, \tag{17}$$

where $d(\cdot, \cdot)$ is a distance measure ℓ_2 (i.e., Euclidean distance), $L(\cdot)$ is the CE loss function and c is a balancing factor. However, searching for possible AEx is an expensive and non-trivial problem. The projected gradient descent is a commonly used method, which searches for the minimum perturbation from the test set of allowable perturbations $S \subseteq \mathbb{R}^D$ as follows [76]:

$$x^{t+1} = \Pi_{x+\mathcal{S}}\left(x^t + \alpha\,\text{sgn}\left(\nabla_x \mathcal{L}\left(f_{\theta*}(x), y\right)\right)\right). \tag{18}$$

The above relation holds until x^t is misclassified by the model. As the ℓ_2-norm perturbation distance is averaged over n test samples by taking greater effort for an attacker to evade detection, the larger the average perturbation, the more robust the model is [24], [76].

2) REACTIVE MEASURE: INPUT RECONSTRUCTION

Kästner et al. [23] state that evaluating the robustness at inference time is the most plausible usage scenario, i.e., checking whether a prediction made by a deployed ML model is robust. Depending upon the data generation platform (e.g., human methylation 450K vs. 27K), adversarial samples would be slightly different from original samples. Thus, we assume that an adversarial sample (i.e., after the transformation) will not be able to affect the prediction of a trained model very severely. Research [77] has shown that AEx can be transformed to corresponding clean input via reconstruction, e.g., Meng et al. [77] trained a denoising AE to reconstruct from the AEx the original samples by removing the added adversarial perturbations. The deep contractive autoencoder is proposed by introducing a penalty to both first-order and second-order derivatives of the mapping. This helps to improve the stability of the learned representations around training points w.r.t. classification errors [78].

Song et al. [79] showed that the reconstruction of a cleaner version of an adversarial sample is possible in case of minor perturbation added, where PixelCNN is proposed to reconstruct the AEx back to the training distribution. Inspired by these measures, we formulate the reconstruction robustness such that it can generate meaningful predictions by correcting or reconstructing a clean representation of the input AEx. For a perturbed input x', we reconstruct a cleaner version x^c using PixelCNN. The cleaner version is then used for classification using model $f : X^{(m)} \rightarrow Y$ (e.g., either f_1 or f_2 for multimodal inputs and CNN for unimodal input). Model f maps instance x from a feature space $X^{(m)}$ with m input features to label y in a target space Y, where $f(x) = y$ denotes the decision y predicted by f.

3) REACTIVE MEASURE: OUT-OF-DISTRIBUTIONS DETECTION

Similar to literature [21], we consider an input (x, y) to be OOD if $y \notin Y$ (i.e., class y does not belong to any of ID classes). Subsequently, we formulate the OOD detection task is to accurately predict if x is of OOD or not [21], [22], [65]. Assuming each sample consisting of 3 modalities (i.e., GE, CNV, and miRNA) and numeric input profiles $x_d \in \mathbb{R}$, the OOD detection capability of a model is performed w.r.t. UMCD and the contrastive self-supervised learning approach SSD [63].

On the other hand, we construct a two-head neural network architecture based on UMCD, which has a multimodal feature extractor, unlike the original network in [71]. The *feature extractor* module E takes inputs x_{in} or x_{ul} and produces the latent representation for each input modality, where X_{ul} signifies unlabeled samples. The classifier module consists of *two classifiers* f_1 and f_2, where each classifier takes features from E and maps them into one of K classes. Training the whole

[10] AEx are mostly added with higher perturbations.

[11] For a gradient-based classifier, e.g., f_1 or f_2, the robustness of each content moderation model is measured by the minimum perturbations required for an input sample to evade detection.

network consists of a pre-training step and two repeating fine-tuning steps. In the pre-training step, labeled ID samples $\{X_{\text{in}}, Y_{\text{in}}\}$ for individual modalities are used, out of which the network learns discriminative features before classifying the ID samples by optimizing the CE loss that maintains the manifold of ID samples [71]:

$$L_s = -\frac{1}{|X_{in}|} \sum_{\mathbf{x}_{\text{in}} \in X_{in}} \sum_{i=1}^{2} \log \left(p_i \left(y_{in} | \mathbf{x}_{\text{in}} \right) \right). \tag{19}$$

Once the network converges, fine-tuning is performed in order to detect OOD samples. During these steps, both $\{X_{in}, Y_{in}\}$ and unlabeled samples X_{ul} are used to train the network for separating ID and OOD samples while keeping the correct classification of ID samples. In order to enable the network to classify the labeled ID samples correctly by supervised learning, we maintain the manifold of ID samples by optimizing eq. (19). In the next alternating step, the network is trained to increase the discrepancy in an unsupervised manner in order to make the network detect the OOD samples without having support of the ID samples, where the unsupervised loss is computed as follows [71]:

$$L_u = \max \left(m - \frac{\sum_{\mathbf{x_{u}}1 \in X_{\text{ul}}} d \left(p_1 \left(\mathbf{y} | \mathbf{x_{ul}} \right), p_2 \left(\mathbf{y} | \mathbf{x_{ul}} \right) \right)}{|X_{\text{ul}}|}, 0 \right), \tag{20}$$

where m is the margin used to prevent overfitting.[12] Overall, this step involves combining and joint optimization of supervised and unsupervised losses [71]:

$$L = L_s + L_u \tag{21}$$

During inference time, to distinguish between ID and OOD samples, the discrepancy is computed w.r.t. the L_1 distance between f_1 and f_2 classifier's outputs:

$$\sum_{i=1}^{K} |p_1 \left(y_i | \mathbf{x} \right) - p_2 \left(y_i | \mathbf{x} \right)| > \delta, \tag{22}$$

where x is a unimodal or multimodal input. When the distance is above a detection threshold δ, we consider the sample to be OOD. We compute the discrepancy loss as the measure of the divergence between the two softmax class probabilities for an input [69]:

$$d \left(p_1(\mathbf{y}|\mathbf{x}), p_2(\mathbf{y}|\mathbf{x}) \right) = H \left(p_1(\mathbf{y}|\mathbf{x}) \right) - H \left(p_2(\mathbf{y}|\mathbf{x}) \right), \tag{23}$$

where $H(\cdot)$ is the entropy over the softmax distribution, and $p_1(\mathbf{y}|\mathbf{x})$ and $p_2(\mathbf{y}|\mathbf{x})$ are K dimensional softmax class probabilities. We train the model to maximize the discrepancy loss such that the model pushes OOD samples outside the manifold of ID samples. For an input x, classifiers f_1 and f_2 yield an output of a K dimensional vector of logits for each classifier. Since f_1 is optimized to maximize the discrepancy loss, its output entropy tends to maximize, whereas f_2's output entropy tends to be minimized. Consequently, f_2 is expected to predict a high probability of one class by pushing OOD samples outside the support of the ID samples. In the case of OOD samples, the discrepancy between model outputs, e.g., f_1 and f_2 would be much larger.

On the other hand, inspired from the fact that contrasting between instances using self-supervised learning is effective at outlier detection without labels [63] and since instance-based contrastive training by incorporating labels further improves the learned representations, we follow to incorporate labels in contrastive self-supervised setting of SSD to improve the quality of learned representations (in multimodal data setting) and OOD detection. In the contrastive self-supervised setting, instead of jointly optimizing supervised and unsupervised losses, we optimize the NT-Xent loss function proposed by Chen et al. [80]. The NT-Xent loss function is parameterized by a temperature variable τ [80].

IV. EXPERIMENT RESULTS

In this section, we discuss and analyse the results of quantitative as well as qualitative experiments.

A. DATASETS

Multi-omics data covering CNVs, miRNA, and GE profiles of 9,074 patients from the Pan-Cancer Atlas project covering 33 tumour types are considered. The preprocessed version of the dataset contributes to 15 GB of data.[13] The sample distribution across modalities is shown in table 1.

B. EXPERIMENT SETUP

Throughout several experiments, we assess the following:

- Which uni- and multimodal input combinations are more suitable for cancer susceptibility prediction?
- Can the MCAE model detect if a supplied input is normal samples or of adversarial examples?
- Can the robust MCAE model reconstruct a cleaner version of an AEx example from its noise input?
- Which neural network architecture is more robust against different adversarial attacks?
- What types of countermeasures are useful for a model trained on multimodal genomic data?

Keras with the TensorFlow backend was used to implement all neural network architectures.[14] The MCAE classifier has two modules: the autoencoder and the classifier. For both $MCAE_{lrc}$ and $MCAE_{slr}$ architecture, the CAE head is a 20-layer network. We use batch normalization before non-linearities (i.e., convolutional and dense layers), ReLU activation function in hidden layers, and softmax activation function in the fully-connected layer. A convolutional layer of the encoder calculates the feature map by taking the max-

[12] If the average discrepancy of unlabeled samples is greater than the margin m, the unsupervised loss tends to zero [71].

[13] Due to a data sharing agreement, we cannot make publicly available the data. However, the first author can be contacted to receive the data for review and research purposes.

[14] https://github.com/rezacsedu/Adversary_Aware_Multimodal_Neural_Networks.

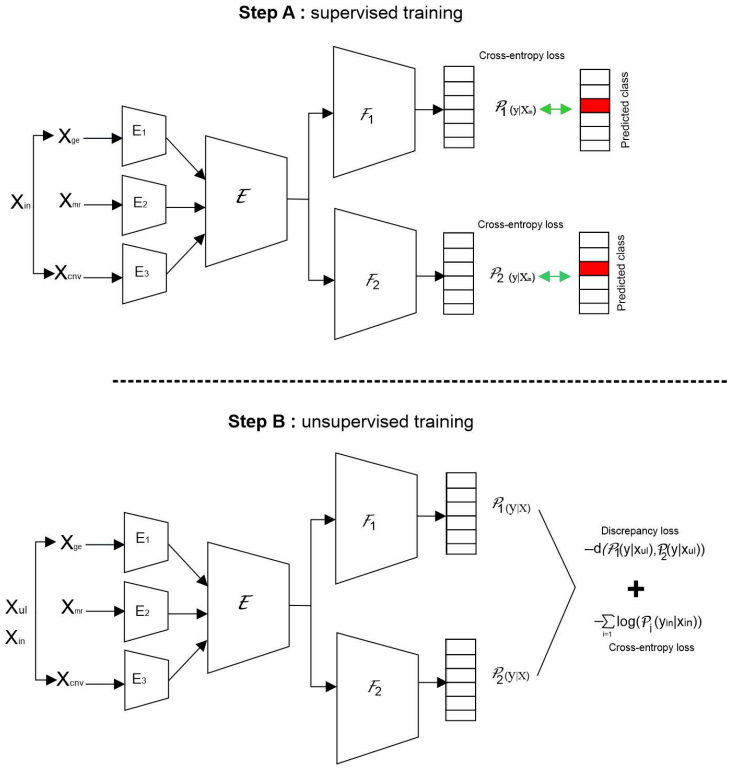


FIGURE 6. Supervised pre-training and unsupervised fine-tuning steps for the OOD detection method, which we extended for multimodal scenarios, based on [71]. The network consists of an extractor E and two classifiers f_1 and f_2. Step A: supervised training to classify ID samples correctly. Step B: f_1 and f_2 learn to maximize the discrepancy in an unsupervised manner to detect OOD samples.

imum value in each non-overlapping sub-region. The CAE part has the following layer-wise structure:

- Input layer: each input sample (e.g., CNVs) is reshaped from $1 \times 20,736$ to 144×144
- Convolutional layer: $32 \times 20,736$ (i.e., 144×144)
- Batch normalization layer: of size $32 \times 20,736$
- Convolutional layer: of size $32 \times 20,736$
- Batch normalization layer: of size $32 \times 10,368$
- Max-pooling layer: of size $32 \times 10,368$
- Convolutional layer: of size $64 \times 10,368$
- Batch normalization layer: of size $64 \times 10,368$
- Convolutional layer: of size $64 \times 10,368$
- Batch normalization layer: of size $64 \times 10,368$
- Convolutional layer: of size $128 \times 10,368$
- Batch normalization layer: of size $128 \times 10,368$
- Convolutional layer: of size $128 \times 10,368$
- Batch normalization layer: of size $128 \times 10,368$
- Convolutional layer: of size $128 \times 10,368$
- Batch normalization layer: of size $128 \times 10,368$
- Convolutional layer: of size $64 \times 10,368$
- Batch normalization layer: of size $64 \times 10,368$
- Upsampling layer: of size $64 \times 20,736$
- Convolutional layer: of size $1 \times 20,736$.

After pre-training the whole MCAE network, only the encoder part is used for the classification. On top of the encoder, a flattening layer, followed by a fully-connected layer of size 128, and a Softmax output unit of 33 (i.e., the number of classes) are added. The whole MCAE network is trained on an Nvidia Titan Xp GPU for 500 epochs in a similar cosine annealing cycling by setting the batch size to 128. The activity regularization term λ is set between $[0.001, 0.005]$, while the regularization weights of the loss function are set as $\alpha_r = 0.1$, and $\alpha_c = 0.25$. Further, the effect of minor noise (using Gaussian noise layers) and dropout regularization is observed. The Gaussian noise parameters are empirically set to a standard deviation of 0.1 and a mean of 0. In case of the pre-training phase of MCAE during the adversarial study, the learning rate is set to 0.1 and dropped by a factor of 10 at 50% and 75% of the training progress over 200 epochs. Each fit is iterated for 50 epochs during the fine-tuning, by setting margin $m = 1 : 2$ for the OOD detection.

To provide a fair comparison, we consider $MCAE_{slr}$ and DOF_o as baselines against the $MCAE_{lrc}$ model to provide comparative analysis. As shown in fig 7, we customized the DOF_o model for the classification setting. We perform pre-training using MMO loss for the multimodal fusion, followed by supervised fine-tuning of the network. Unlike the Cox Partial Likelihood Loss used in DFO, we optimize categorical CE loss in a multiclass classification setting. In the case of unimodal input, we provide a comparative analysis with GE-based diagnosis approaches by Mostavi et al. [33] and Lyu et al. [10].

TABLE 1. Sample distribution across tumour types: rows: tumour type, column: data types [81].

Cohort	CNVs	miRNA	Gene expression	Carcinoma type
ACC	345	332	227	Adrenocortical carcinoma
BLCA	164	143	71	Bladder urothelial carcinoma
BRCA	578	501	495	Breast invasive carcinoma
CESC	198	187	179	Cervical and endocervical cancers
CHOL	310	309	303	Cholangio carcinoma
COAD	126	121	96	Colon adenocarcinoma
DLBC	457	442	431	Lymphoid neoplasm diffuse large B-cell lymphom
ESCA	511	497	333	Esophageal carcinom
GBM	357	365	355	Glioblastoma multiforme
HNSC	577	454	581	Head and neck squamous cell carcinoma
KICH	887	870	817	Kidney chromophobe
KIRC	422	407	192	Kidney renal clear cell carcinoma
KIRP	198	187	179	Kidney renal papillary cell carcinoma
LAML	310	309	303	Acute myeloid leukemia
LGG	126	121	96	Brain lower grade glioma
LIHC	457	442	431	Liver hepatocellular carcinoma
LUAD	511	497	333	Lung adenocarcinoma
LUSC	357	365	355	Lung squamous cell carcinoma
MESO	577	454	581	Mesothelioma
OV	887	870	817	Ovarian serous cystadenocarcinoma
PAAD	422	407	192	Pancreatic adenocarcinoma
PCPG	577	454	581	Pheochromocytoma and paraganglioma
PRAD	887	870	817	Prostate adenocarcinoma
READ	422	407	192	Rectum adenocarcinoma
SARC	198	187	179	Sarcoma
SKCM	310	309	303	Skin cutaneous melanoma
STAD	126	121	96	Stomach adenocarcinoma
TGCT	457	442	431	Testicular germ cell tumors
THCA	511	497	333	Thyroid carcinoma
THYM	357	365	355	Thymoma
UCEC	577	454	581	Uterine corpus endometrial carcinoma
UCS	887	870	817	Uterine carcinosarcoma
UVM	422	407	192	Uveal melanoma

For consistent comparison of OOD detection performance, we compare UMCD [71] and self-supervised SSD [63] approaches. For consistency, we re-implement contrastive self-supervised learning in multimodal data setting. As omics data are significantly different from imaging data, we focus on a five-shot OOD detection such that the SSD model have access to 5 instances from each class of the targeted OOD datasets. Further, as NT-Xent loss requires a much larger batch size compared to the supervised cross-entropy loss function in order to contrast with a large number of negatives, the model was trained with a batch size of 512. Previous studies found that the performance of OOD detection degrades[15] with a lower temperature. Therefore, an optimal selection of τ has a significant effect on the OOD detection capability of the model, e.g., the OOD detection accuracy and AUROC scores was found to be highest when τ was set to 0.1 or 0.5, across a number of datasets. Therefore, we experiment by setting τ value in the range of [0.1, 0.5].

Results based on random search and 5-fold cross-validation are reported with macro-averaged precision and recall. We did not report F1 scores since they are significant only when precision and recall are very different, whereas it is important for cancer diagnosis to have both high precision and recall [82]. Further, since the dataset is moderately imbalanced, Matthews correlation coefficient (MCC) scores are reported. As Kästner et al. [23] recommended evaluating the average robustness of a model for arbitrary inputs as the global robustness measure, the robustness of the models is computed in terms of Empirical Robustness Metric (ERM) [62] and CLEVER [83] scores over the test set for each adversarial crafting attack. The ERM, which is equivalent to computing the minimal perturbation that the attacker must introduce for a successful attack, is formulated as [62]:

$$\hat{\rho}^{\infty}_{\text{adv}}(f) = \frac{1}{|D|} \sum_{x \in D} \frac{\|\hat{r}(x)\|_{\infty}}{\|x\|_{\infty}}, \tag{24}$$

where $\hat{r}(x)$ is computed using DeepFool (with $p = \infty$) and FGSM, respectively. The higher the ERM value, the higher is the classifier. The CLEVER score, proposed by Weng et al. [83], is attack-agnostic and computationally feasible for large neural networks, as it is aligned with the robustness indication measured by the ℓ_2 and ℓ_∞ norms of AEx from powerful attacks. We compute the CLEVER score in a non-targeted attack scenario, while false-positive rates (FPR) (at a 95% true positive rate), detection error (DE), Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall (AUPR), are computed to assess the robustness against OOD attacks.

[15] A smaller value of temperature quickly saturates the loss, discouraging it to further improve the feature representations.

TABLE 2. Class-specific classification results for $MCAE_{lrc}$, DOF_o, and $MCAE_{slr}$ classifiers.

	$MCAE_{slr}$ (89.75%)			DOF_o (92.32%)			$MCAE_{lrc}$ (96.25%)		
Type	**Precision**	**Recall**	**MCC**	**Precision**	**Recall**	**MCC**	**Precision**	**Recall**	**MCC**
BRCA	0.8785	0.8612	0.7564	0.8815	0.8775	0.7667	0.9437	0.9511	0.8465
LGG	0.9254	0.8926	0.8330	0.9292	0.8996	0.8413	0.9311	0.9402	0.8421
UCEC	0.8753	0.8819	0.7835	0.8823	0.8869	0.7887	0.9562	0.9429	0.8445
LUAD	0.8235	0.8354	0.7136	0.8316	0.8394	0.7195	0.9865	0.9823	0.8624
HNSC	0.8520	0.8743	0.7851	0.8617	0.8811	0.78892	0.9730	0.9822	0.8765
THCA	0.8528	0.8323	0.7275	0.8623	0.8421	0.7333	0.9138	0.9154	0.8125
PRAD	0.8827	0.8778	0.7847	0.8911	0.8817	0.7911	0.9233	0.9347	0.8207
LUSC	0.8726	0.8634	0.7625	0.8817	0.8722	0.7712	0.9434	0.9472	0.8524
BLCA	0.8956	0.9037	0.8075	0.9013	0.9079	0.8135	0.9656	0.9537	0.8475
STAD	0.8253	0.8156	0.6932	0.8279	0.8271	0.7015	0.9653	0.9556	0.8532
SKCM	0.8853	0.8711	0.8025	0.8916	0.8819	0.8055	0.9046	0.9136	0.8168
KIRC	0.8967	0.9123	0.8237	0.9047	0.9183	0.8273	0.9578	0.9689	0.8531
LIHC	0.8194	0.8085	0.6945	0.8243	0.8113	0.7011	0.9572	0.9664	0.8537
COAD	0.8368	0.8245	0.7679	0.8436	0.8287	0.7754	0.9776	0.9690	0.8514
CESC	0.8785	0.8743	0.7964	0.8832	0.8792	0.8017	0.9873	0.9885	0.8664
KIRP	0.8254	0.8032	0.7043	0.8312	0.8092	0.7133	0.9681	0.9782	0.8430
SARC	0.8753	0.8671	0.7835	0.8841	0.8743	0.7915	0.9365	0.9435	0.8421
OV	0.8825	0.8733	0.7936	0.8911	0.8821	0.8011	0.9725	0.9773	0.8262
ESCA	0.8913	0.8719	0.7951	0.8978	0.8851	0.8017	0.8956	0.8834	0.8076
PCPG	0.8537	0.8611	0.7875	0.8661	0.8695	0.7918	0.9875	0.9987	0.8735
PAAD	0.9629	0.9567	0.8407	0.9681	0.9612	0.8455	0.9452	0.9500	0.8325
TGCT	0.8736	0.8722	0.7825	0.8832	0.8775	0.7911	0.9890	0.9724	0.8434
GBM	0.8952	0.8845	0.8075	0.9046	0.8931	0.8153	0.9362	0.9453	0.8436
THYM	0.9255	0.9123	0.8232	0.9285	0.9135	0.8284	0.9775	0.9678	0.8622
READ	0.6795	0.6857	0.6225	0.6854	0.6919	0.6255	0.8874	0.8733	0.7525
LAML	0.8697	0.8567	0.8237	0.8842	0.8673	0.8311	0.9576	0.9632	0.8513
MESO	0.8991	0.9028	0.8076	0.9067	0.9078	0.8114	0.9534	0.9456	0.8457
UVM	0.8765	0.8623	0.7979	0.8855	0.8715	0.8033	0.9136	0.9089	0.8184
ACC	0.9217	0.9345	0.8225	0.9279	0.9395	0.8317	0.9623	0.9731	0.8611
KICH	0.9335	0.9475	0.8425	0.9392	0.9515	0.8483	0.9690	0.9625	0.8439
UCS	0.9157	0.9064	0.8125	0.9198	0.9111	0.8180	0.8726	0.8675	0.7869
DLBC	0.8678	0.8729	0.7005	0.8772	0.8782	0.7045	0.9347	0.9421	0.8389
CHOL	0.8838	0.8975	0.7979	0.8934	0.9033	0.8017	0.8455	0.8342	0.6821
Average	**0.8975**	**0.9065**	**0.8052**	**0.9232**	**0.9278**	**0.8217**	**0.9625**	**0.9542**	**0.8453**

C. ANALYSIS OF SUSCEPTIBILITY PREDICTION

We analyse the performance of individual models, covering both multimodality input combinations. The analyses will lead us to select the best model and input modality combination for providing a more reliable diagnosis. We observed a mean precision of 89.75%, 92.32%, and 96.25% for $MCAE_{slr}$, DOF_o, and $MCAE_{lrc}$ models, respectively. However, as classes are moderately imbalanced, we report class-specific classification reports along with corresponding MCC, precision, and recall scores in table 2 and fig 9. As shown, precision and recall for the majority of cancer types were high in which the $MCAE_{lrc}$ model performed consistently better than that of both $MCAE_{slr}$ and DOF_o models. Notably, the $MCAE_{lrc}$ model classifies BRCA, UCEC, LUAD, HNSC, LUSC, THCA, PRAD, BLCA, STAD, KIRC, LIHC, COAD, CESC, KIRP, SARC, OV, PCPG, TGCT, GBM, READ, LAML, MESO, and DLBC cancer cases with higher confidence, whereas, both $MCAE_{slr}$ and DOF_o model classify PAAD, CHOL, and UCS cancer cases more accurately than $MCAE_{lrc}$, except for some misclassifications.

According to the confusion matrix in fig 8, $MCAE_{lrc}$ classifies HNSC and LUSC tumour samples accurately in only 79% and 81% of the cases, while the $MCAE_{slr}$ model made more mistakes, particularly to classify STAD, HNSC, LUSC, and LGG tumour samples. Overall, both classifiers performed moderately well except for certain types of tumour cases such as STAD, HNSC, BLCA, THCA, UCEC, LUAD, LUSC, and LGG. As observed, the ROC curves for the $MCAE_{lrc}$ model show that AUC scores are consistent across the folds showing stable predictions, which shows about a 5% boost in AUC scores for the $MCAE_{lrc}$ model. This performance increase signifies that predictions made by the $MCAE_{lrc}$ model are much better than random guessing. Further, class-specific MCC scores of the $MCAE_{lrc}$ model are 4% higher than the $MCAE_{slr}$ model, which suggests that the predictions were strongly correlated with the ground truth, yielding a Pearson product-moment correlation coefficient higher than 0.70 for all the classes except for the CHOL tumour samples. The downside is that both classifiers made a number of mistakes, e.g., $MCAE_{lrc}$ can classify ESCA, READ, UCS, and CHOL tumour cases in only 89% of the cases accurately, while the $MCAE_{slr}$ model made more mistakes for the READ, LUAD, LIHC, KIRP, COAD, and STAD samples.

To show the effectiveness of our approach, we provide fair comparison between $MCAE_{lrc}$, $MCAE_{slr}$, and DOF_o models, covering both multimodal and unimodal settings. The $MCAE_{lrc}$ model clearly outperforms both $MCAE_{slr}$ and DOF_o models w.r.t. precision (scores of 0.8975, 0.9217, and 0.9625 for the $MCAE_{lrc}$, $MCAE_{slr}$, and DOF_o model, respectively) and MCC (scores of 0.8052, 0.8235, and 0.8453 for

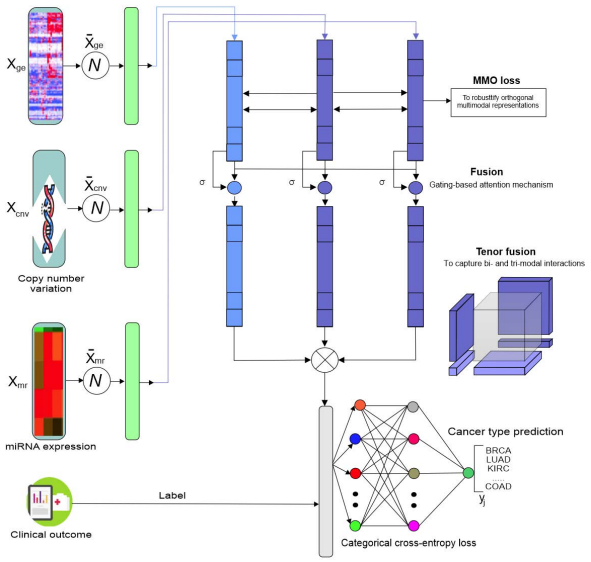


FIGURE 7. The DOF model based on Braman et al. [17], where the output of fusion based on MMO is used as the input to fully-connected classification layer.

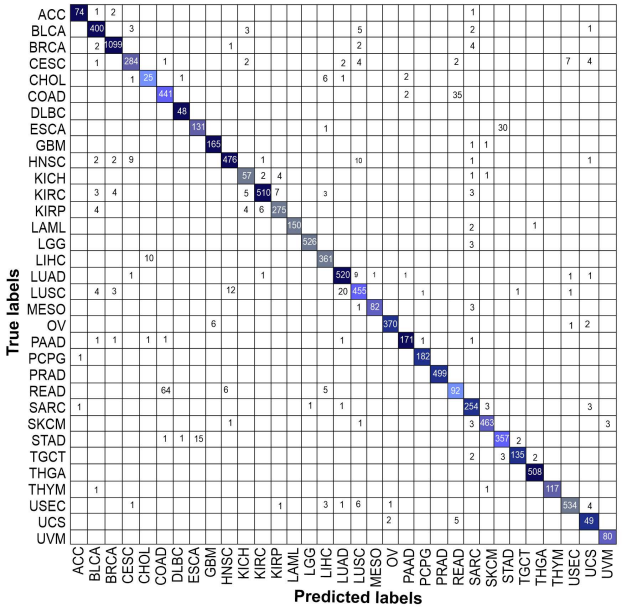


FIGURE 8. Confusion matrix for the CAE classifier when trained on GE samples.

the $MCAE_{lrc}$, $MCAE_{slr}$, and DOF_o model, respectively). The $MCAE_{slr}$ model performed worst for every combination of input modalities, as shown in fig 10 and fig 11, irrespective of reconstruction loss and f1-score. On the other hand, the GE+miRNA input combination gave the highest f1-scores than GE+CNV, GE+CNV+miRNA or CNV+miRNA input modalities. Our empirical study finds several potential reasons for such low performance:

- The number of samples were increased from 1000 to 10,000, across individual modalities.
- Reconstruction errors for both multimodal architectures based on CAE have decreased, while the MMO loss for DOF_o model also converged.
- Compared to $MCAE_{slr}$ model, both DOF_o and $MCAE_{lrc}$ models learned more complex concepts from the GE + miRNA multimodal features.

As deep architecture requires more training samples to converge well [31], adding more training samples helps increase the generalization of $MCAE_{lrc}$ model by mitigating bias, i.e., lowering the biases results higher training scores than the validation scores for the maximum number of samples. The second and third reasons helped all the models learn more abstract features towards improving the classification accuracy. Further, the Wilcoxon signed-rank test is performed to compare both $MCAE_{lrc}$ and $MCAE_{slr}$ classifiers w.r.t. reconstruction losses and accuracies at a significance level of 5%. Within each box plot in fig 10 and fig 11, mean and median of reconstruction losses and f1-scores are depicted with dots and horizontal lines. The GE+miRNA input combination yields the lowest reconstruction errors,[16] whereas GE+CNV+miRNA generates the highest errors.

[16]We did not include DOF_o and SDD models in this comparison as both use different types of loss called MMO loss and supervised contrastive training loss, respectively.

On the other hand, since the MMO loss forces unimodal embeddings to provide independent and complementary information to the fused prediction, the diagnosis prediction performance was expected to surpass $MCAE_{lrc}$ model. However, as our datasets consist of multi-omics data, with the subtle difference between feature levels, the DOF model was not befitted like the original study incorporating radiology, histology, and genomic data. Yet, it consistently outperformed the $MCAE_{slr}$ model across all the cancer types. Further, the ROC curves of the $MCAE_{lrc}$ model show consistent AUC scores across the folds showing stable predictions, giving a 4% boost in AUC scores. This signifies that the predictions by the $MCAE_{lrc}$ model are: i) clearly better than random guessing as well as both $MCAE_{slr}$ and DOF_o classifier, ii) strongly correlated with the ground truth, yielding a Pearson product-moment correlation coefficient that is higher than 0.70 for all the classes.

The precision plot (ref. fig. 12a), outlining the relation between the predicted probability (that an index belongs to the positive) and the percentage of the observed index in the positive class. As seen, the fraction of positive increases with the predicted probability. The observations get binned together in groups of roughly equal predicted probabilities, and the percentage of positives is calculated for each bin. While a perfectly calibrated model would show a straight line from the bottom left corner to the top right corner, a better-fitted model would classify most observations correctly with close to 0% or 100% probability. This indicates that the $MCAE_{lrc}$ model would be more suitable fora more reliable diagnosis, giving a precision score of 0.968. The lift curve for the $MCAE_{lrc}$ model (ref. fig. 12b) shows the percentage of positive classes when observations with a score above the cutoff are selected vs. random selection: the model is able

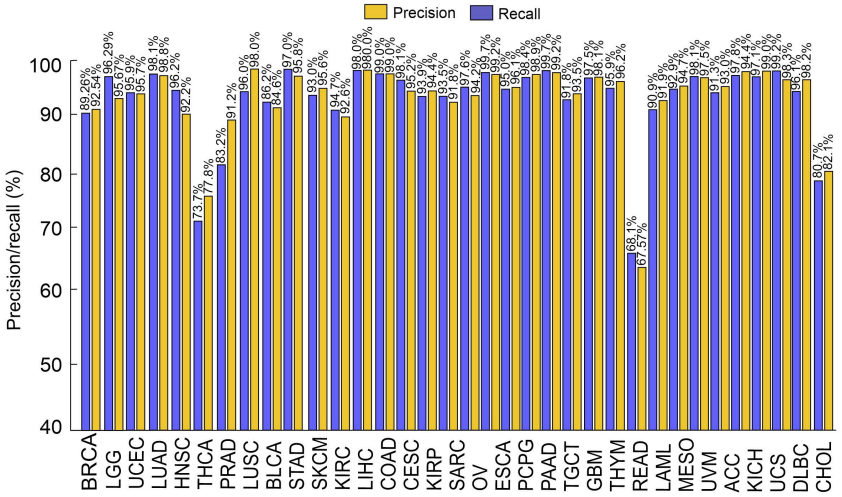
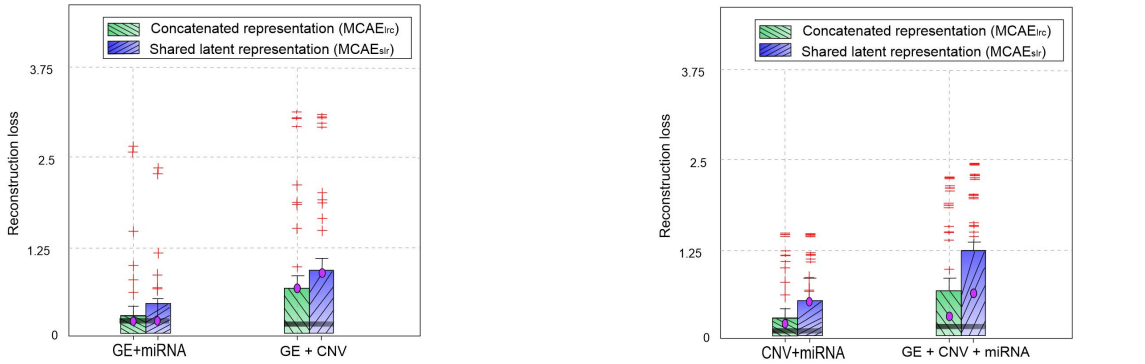


FIGURE 9. **Precision (blue) and recall (yellow) scores for the CAE classifier when trained on GE samples.**



(a) GE+miRNA vs. GE+CNV modality (b) CNV+miRNA vs. GE+CNV+miRNA modality

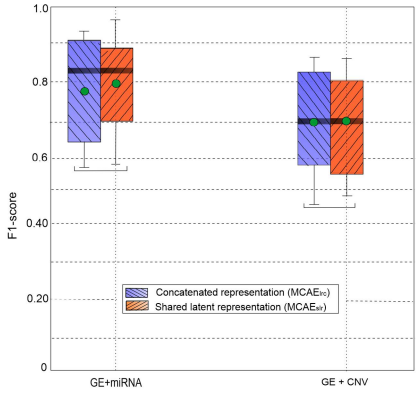
FIGURE 10. **Comparison of reconstruction losses for $MCAE_{lrc}$ and $MCAE_{slr}$ during pretraining.**

to identify 95 observations accurately out of 100, whereas only 17 observations are correctly predicted when done with random selection.[17] Overall, both $MCAE_{lrc}$ and $MCAE_{slr}$ models gave competitive results for the majority of the cancer types. However, it turns out that GE + miRNA expression input combination gives the best results.

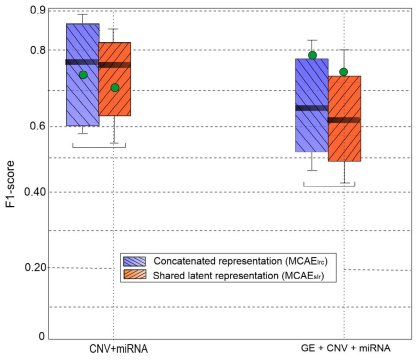
For the comparative analysis, we pick the GE modality for the CAE classifier. That is, when trained on GE data, the CAE model slightly outperforms the approach of Boyu et al. [10] and is about 6.5% better than the approach by Yuanyuan et al. [28], yielding that the model predicts 96.25% of the cases accurately. Besides, it reduces the misclassification rate for READ, UCS, ESCA, and CHOL tumour samples. Against 35%, 81%, 77%, and 56% of the correctly predicted cases by [10], the CAE classifier predicts 88.74%, 87.26%, 89.56%, and 84.55% of cases correctly, outlining a significant improvement. In contrast, the CAE classifier slightly underperforms than [10] at classifying BRCA, THCA, and PRAD. However, the CAE classifier shows consistent performance for the majority of cancer types, indicating potential high generalizability for unseen GE data.

We investigate the reasons why the CAE model outor underperforms existing approaches across certain cancer types. Mostavi et al. [33] have outlined that samples from the kidney (KICH, KIRC and KIRP), liver (CHOL and LIHC), lung (LUAD and LUSC) or digestive systems (ESCA and STAD) are clearly grouped together. As their model probably learns to recognize tissues of origin, the major classification errors (44 and 42 kidney and lung-related instances were misclassified) are also within kidney, lung, colon, and rectum adenocarcinomas. In contrast, only 28 and 30 kidney and lung-related instances were misclassified in our approach. Our approach made only 64 and 15 misclassification for the READ and STAD samples, their approach made 84 and 18 mistakes. This indicates significant improvement, at least, for some selected classes. The downside of our approach, however, is that against 1 misclassification, our approach made 6 mistakes in classifying USEC cancer types. For the READ tumor samples, our approach managed to reduce the confusion too, making only 75 mistakes (in contrast to 85 mistakes made by [33]).

[17] i.e., $MCAE_{lrc}$ model is much better than random guessing.

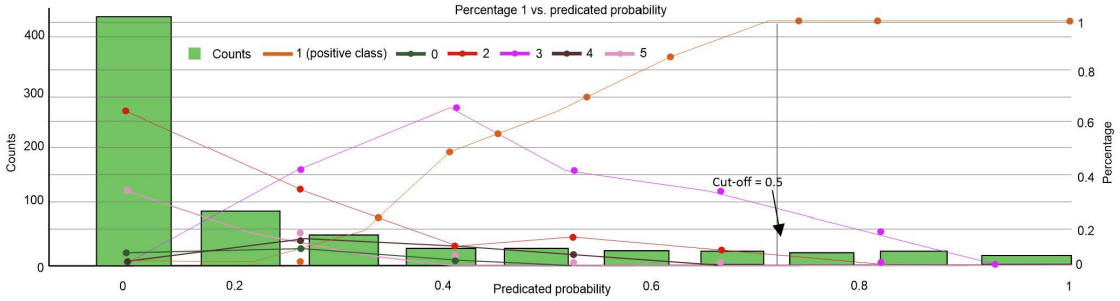


(a) GE+miRNA vs. GE+CNV modality

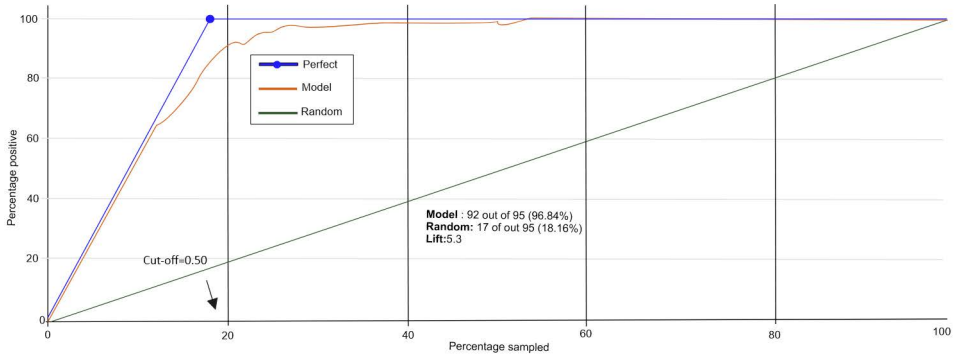


(b) CNV+miRNA vs. GE+CNV+miRNA modality

FIGURE 11. F1 score comparison for $MCAE_{lrc}$ and $MCAE_{slr}$ during supervised finetuning.



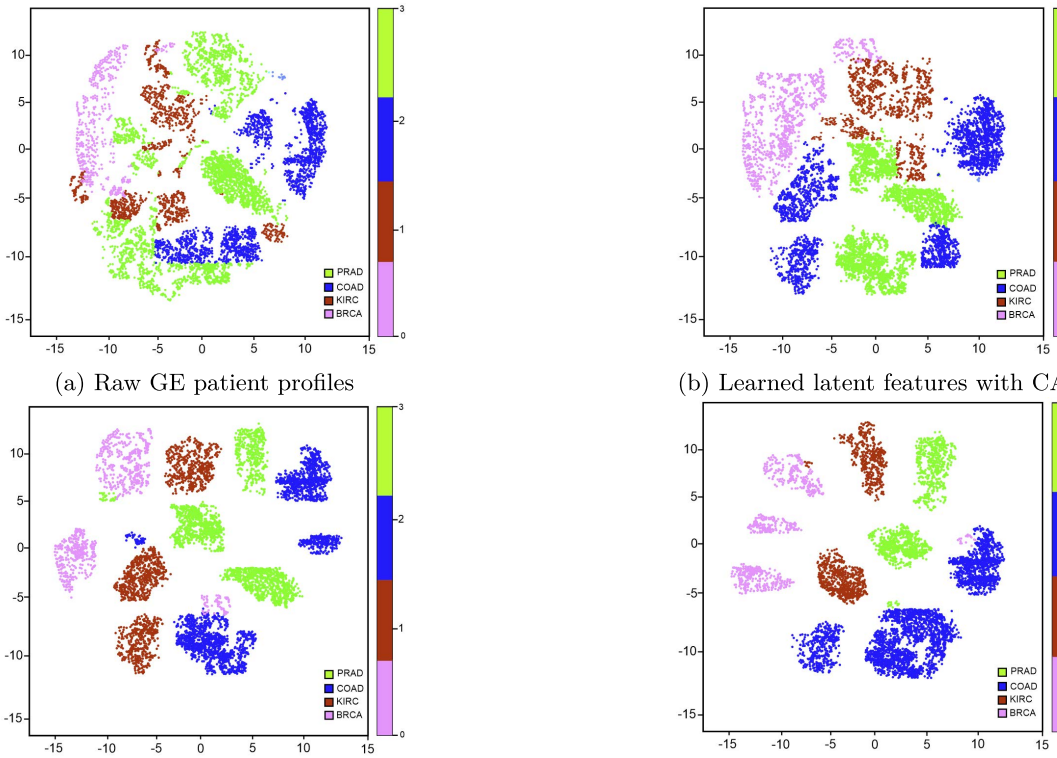
(a) Precision plot: if the fraction positive increases with the predicted probability



(b) Lift curve: the percentage of positive classes when observations with a score above cutoff is selected vs. random selection

FIGURE 12. Precision plot and the lift curve for the $MCAE_{lrc}$ model.

Like other omics data, GE is very high dimensional, where a significant number of genes have a small or no effect on cancer, making them very weak features [18]. Our study suggests that CAE-based RL can be more effective at learning hierarchical features from such high dimensional data than CNN-based approaches. To qualitatively analyse why CAE model outperforms these approaches for certain cancer types, we observe whether learned representations express biological characteristics of cohorts. We plot both embeddings and raw GE profiles in fig 13. Since plotting the same for all the cancer types would be cumbersome, we provide the t-SNE plots for breast carcinoma (BRCA), renal kidney carcinoma (KIRC), colon adenocarcinoma (COAD), and prostate adenocarcinoma (PRAD) cancer types only. Further, as each input modality has a high dimension, the association between each feature is considered. We can observe moderately high distinctive patterns among these cancer patients in fig. 13a, which, however, are not clearly visible for raw GE profiles. BRCA and COAD patients are clearly separated, even though PRAD and KIRC patients are moderately mixed and hence did not separate well. The latent space of CAE is slightly better than the abstract feature representation with CNN, hence the CAE classifier tends to achieve slightly better separability of the GE profiles, which is reflected in the classification results. This is an indication that both CAE and MCAE models have learned the latent molecular properties better in coded form than that of patient raw expression profiles. Overall, our study suggests that CAE-based RL can be more



(a) Raw GE patient profiles

(b) Learned latent features with CAE

(c) Classification using hybrid-2D CNN (on raw feature space)

(d) Classification using CAE classifier (on latent space)

FIGURE 13. t-SNE plots of different stages of classification using hybrid-2D CNN and CAE architectures.

effective at learning hierarchical features than CNN-based approaches.

D. ADVERSARIAL ROBUSTNESS

We analyse the performance of adversarially retrained models against adversarial and OOD attack scenarios.

1) ROBUSTNESS AGAINST ADVERSARIAL ATTACKS

Table 3 reports the ℓ_∞ robustness to adversarial perturbations for 4 different models based on DeepFool and FGSM with 90% of misclassification. The retrained $MCAE_{slr}$, DOF_o, and $MCAE_{lrc}$ models exhibit higher robustness compared to their originally trained versions. Although every model tend to achieve the smallest ℓ_∞ distortion for individual sample for $\epsilon \in \{0.01, 0.02, 0.03, 0.04, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, the retrained $MCAE_{lrc}$ model shows higher robustness than that of one with SLR or MMO loss. Classifier $MCAE_{lrc}$ was found robust even when ϵ is increased up to 0.4, giving up to 85% accuracy. In contrast, $MCAE_{slr}$ shows fragile performance, giving a drastic reduction of the accuracy by 25%. As a higher ERM value indicates a more robust classifier, $MCAE_{lrc}$ was found more robust to DeepFool and FGSM attacks. While DOF_o exhibits moderately high robustness, $MCAE_{slr}$ found to be fragile against these attacks.

Table 4 outlines the average untargeted CLEVER scores and distortion based on the FGSM and DeepFool untargeted attacks. As a lower CLEVER score indicates a more robust classifier, $MCAE_{lrc}$ was found more robust against both DeepFool and FGSM attacks. While DOF_o exhibits moderately high robustness, $MCAE_{slr}$ is found fragile against both DeepFool and FGSM attacks as the CLEVER scores are smaller than the distortions of adversarial samples in most cases. Since CLEVER is independent of attack algorithms, reported CLEVER scores roughly indicate the distortion of the best possible attack w.r.t. a specific ℓ_p distortion. CLEVER scores can be considered as a security checkpoint for unseen attacks. In addition, there are significant gaps in the distortion between the CLEVER score and the attack algorithms considered, suggesting that there is a more effective attack that can fill the gap [83].

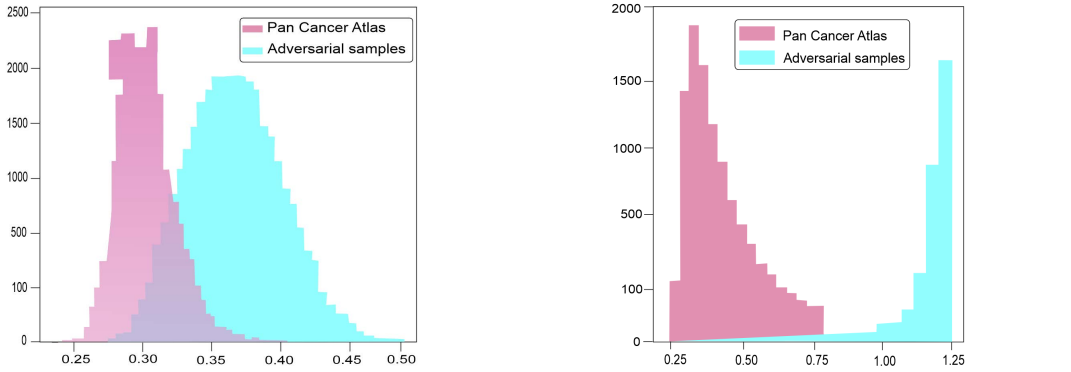
Further, adding Gaussian noises during retraining helped both classifiers gain minimal robustness against minor

TABLE 3. ERM scores ($\hat{\rho}^{\infty}_{\text{adv}}$) for different models based on DeepFool and FGSM.

Classifier	DeepFool	FGSM
$MCAE_{lrc}$	0.29	0.09
DOF_o	0.23	0.07
$MCAE_{slr}$	0.21	0.05

TABLE 4. Untargeted CLEVER scores vs. distortion for different models based on DeepFool and FGSM.

	DeepFool		FGSM	
Classifier	ℓ_2	ℓ_∞	ℓ_2	ℓ_∞
$MCAE_{lrc}$	83.21	82.32	86.16	85.37
DOF_o	85.17	84.45	85.22	84.43
$MCAE_{slr}$	89.25	87.54	89.25	87.54



(a) Histogram of maximum softmax scores (b) Histogram of ID and OOD detection scores

FIGURE 14. Histogram of the maximum softmax and ID/OOD detection scores of the classifiers.

TABLE 5. Results of distinguishing ID and OOD of two different classifiers w.r.t. DE, FPR, AUROC, AUPR In (AUPRI) and AUPR out (AUPRO); the ↑ indicates larger value is better, while the ↓ indicates lower value is better; FPR* at 95% TPR shows the FPR at 95% TPR.

	DOF_o					SSD					$MCAE_{lrc}$				
Modality	FPR*↓	DE↓	AUROC↑	AUPRI↑	AUPRO↑	FPR*↓	DE↓	AUROC↑	AUPRI↑	AUPRO↑	FPR*↓	DE↓	AUROC↑	AUPRI↑	AUPRO↑
GE	0.61	0.85	0.91	0.92	0.91	0.55	0.83	0.93	0.95	0.94	0.53	0.82	0.94	0.96	0.95
miRNA	0.69	0.81	0.82	0.88	0.87	0.65	0.82	0.89	0.90	0.89	0.62	0.84	0.91	0.92	0.90
CNV	0.85	0.88	0.85	0.83	0.84	0.77	0.90	0.86	0.85	0.85	0.73	0.92	0.87	0.87	0.86
GE+miRNA	0.81	0.69	0.79	0.78	0.77	0.77	0.71	0.87	0.86	0.85	0.72	0.73	0.89	0.88	0.87
GE+CNV	0.85	0.86	0.79	0.77	0.81	0.87	0.90	0.79	0.85	0.82	0.89	0.92	0.82	0.87	0.84
miRNA+CNV	0.58	0.87	0.76	0.81	0.79	0.92	0.96	0.78	0.82	0.81	0.95	0.97	0.79	0.83	0.82
GE+miRNA+CNV	1.37	1.55	0.75	0.76	0.75	1.11	1.17	0.76	0.77	0.76	0.70	1.25	0.74	0.73	0.74

perturbations. This makes the PixelCNN moderately effective at reconstructing cleaner versions of the AEx. The reason is that the $MCAE_{lrc}$ model learns more abstract features from the reconstructed inputs from PixelCNN than both $MCAE_{slr}$ and DOF_o models, which helps in the overall classification task.

2) ROBUSTNESS AGAINST OOD ATTACK

Owing to high confidence in cancer susceptibility prediction and higher adversarial robustness, the $MCAE_{slr}$ is excluded from the OOD detection test. The results of the OOD detection performance of $MCAE_{lrc}$, DOF_o, and SSD models are reported in table 5. As shown, our approach w.r.t. $MCAE_{lrc}$ model can distinguish ID samples from OOD with moderately high confidence. The histogram of the maximum softmax and ID/OOD detection scores of the classifiers are shown in fig. 14, which shows that the distribution of the score is moderately different w.r.t. whether a sample is an ID or OOD after the fine-tuning. We observed moderately high disagreement between two classifiers' outputs, which is based on unlabeled ID and OOD samples after training the model on labeled ID samples in a supervised way. Further, the discrepancy loss of ID samples is smaller than OOD samples in all settings.

The DE is lower when the difference between the discrepancy loss of ID and OOD samples is larger, which means that the divergence between two classifiers' output can separate ID and OOD samples. Since this method needs to fine-tune the classifiers to detect the OOD samples, the decision boundary changed slightly. A significant drop in accuracy (by 8%) compared to the original score signifies that the GE and miRNA are more sensitive to making changes compared to CNV, which reflects in the GE, miRNA, GE+miRNA modalities. As demonstrated, more test samples help decrease both DE, ID/OOD discrepancy losses, up to a certain point, e.g., when we evaluate the models on 1,800 samples, all the 3 metrics increased, which is probably because of higher discrepancy loss.

All the models show moderately high robustness at OOD detection task when the attacks are introduced with DeepFool, where $MCAE_{lrc}$ and SSD models outperformed both $MCAE_{slr}$ (result based on SLR is not shown in the table, though) and DOF_o models. On the other hand, all the models show quite limited robustness when the OOD attacks are introduced with FGSM. The reason is that DeepFool introduces weaker adversarial attacks as it provides fewer perturbations, while FGSM provides much stronger attacks as it provides much higher perturbations than DeepFool. In particular, the UCMD (based on $MCAE_{lrc}$) and SSD achieve per or comparable performance across individual pairs of in and out-distribution datasets. In particular, using labels in the instance-based contrastive training improves (i.e., combining SSD with a five-shot OOD detection method) further brings a gain of 2.0 in the average AUROC. However, owing to subtle differences among modalities in our multi-omics data, the DOF model was not befitted much during the representation learning stage.

Consequently, UCMD based on DOF turns out to be adversarially weakest than $MCAE_{lrc}$ and SSD. Overall, self-supervised representations found to be quite effective to boost the OOD detection performance compared to CAE-based representations. To investigate why contrastive self-supervised learning is effective in OOD detection,

we focus on the NT-Xent loss function, which is parameterized by the temperature variable τ. When experimented by setting the value of τ between 0.1 and 0.5, a temperature value of 0.4 turns out to be the optimal parameter, which helped pull positive instances such that different transformations of sample instances together while pushing away from other instances.

V. CONCLUSION AND OUTLOOK

We proposed an adversary-aware MCAE classifier for cancer susceptibility prediction from multi-omics data. Experiment results show, covering both single modality and multimodality, that omics data are useful at predicting different cancer types with high confidence w.r.t. precision score (of up to 96.25% for single modality and 92.45% for GE+miRNA input combination, respectively). The MCAE model is also found effective at tackling the curse of dimensionality of high dimensional omics data. Our approach also suggests that multimodal models (e.g., $MCAE_{lrc}$, $MCAE_{slr}$, and DFO_o models, in our case) may surpass the MAE model with the right combination of input modalities (e.g., GE + miRNA multimodality combination is found to be more suitable for cancer type prediction). This is supported by the fact that the learned representations from the GE + miRNA input modality can express biological characteristics of cohorts, which is reflected in the classification results.

Further, we made multimodal models adversarially robust by introducing different attacks on it, followed by taking proactive and reactive countermeasures. Besides, we trained two classifiers to detect OOD samples that are far from ID samples' support. Overall, our approach can identify if the supplied samples are of AEx, ID, or OOD with moderately high confidence. We outline some potential limitations of our study: i) our study is hindered due to the limited amount of labeled data used for training the multimodal models, while neural networks typically require many samples to converge well towards generalization, ii) although DL-based approaches are useful in cancer diagnosis and subsequent treatment recommendations, due to high non-linearity and higher-order interactions among a large number of features, complex DNN models are perceived as *black-box* methods [44], iii) a tricky drawbacks of multimodal embedding is that different types of data are conflated into a single representation in the semantic space [44], hence the learned representations from autoencoder architectures are not easily interpretable [19], iv) using a black-box model would not allow tracing how and why inputs are mapped to certain decisions [44].

This makes interpretability an essential requirement to provide insights into what features were captured during the RL and what sample attributes are the classifier based on [44]. On the other hand, a well-interpretable *white-box* model that can identify statistically significant features, can be used to explain the way they affect the model's outcome and whether they interact. Interpretable ML techniques are getting more adoption in many healthcare use cases. In particular, perturbing (e.g., sensitivity analysis and feature interactions), probing (e.g., Grad-CAM/++ [35], layer-wise relevance propagation (LRP) [84], and attention mechanism), and model surrogation strategies can be applied in order to generate insights on why and how a certain prediction has been made by the model (e.g., identifying important biomarkers that exhibit shared characteristics, which may help in recommending more reliable treatments and drug repositioning [85]. Nevertheless, explaining diagnosis decisions with plots and charts is helpful for exploration and discovery, but interpreting them may be difficult for non-domain experts and patients, unless they are not explained in natural languages or human-interpretable way (e.g., decision rules). Holzinger et al. [86] have shown that multimodal embeddings and interactive explainability can provide the foundations for effective human-AI interfaces. Since causal links between features can be defined using graph structures, they outlined, by constructing a multimodal feature space of images, text, and genomics data, that graph neural networks (GNNs) can be useful for enabling information fusion for multimodal causability.[18]

Further, neuro-symbolic reasoning techniques can be employed to explain the decision for the cancer diagnosis with domain knowledge. This can be achieved by combining a neural network (e.g., MCAE) with a domain-specific knowledge graph (by integrating domain knowledge, scientific literature, and omics data) [44]. Inspired by these techniques and methods, we intend to focus on: i) semi-supervised learning to reduce the need for a large number of labeled examples and instead utilize unlabeled ones, ii) employing the model ensemble method by training multiple model snapshots during a single training, followed by combining their predictions to make an ensemble prediction.[19] iii) improving the interpretability of diagnosis decisions by identifying biologically relevant biomarkers (i.e., genes) and providing both global and local explanations (e.g., identification of biomarkers based on relevance/importance and ranking of top genes across cancer types) in the future.

REFERENCES

[1] A. A. Ghazani, N. M. Oliver, J. P. S. Pierre, A. Garofalo, I. R. Rainville, E. Hiller, D. J. Treacy, V. Rojas-Rudilla, S. Wood, and E. Bair, "Assigning clinical meaning to somatic and germ-line whole-exome sequencing data in a prospective cancer precision medicine study," *Genet. Med.*, vol. 19, no. 7, pp. 787–795, Jul. 2017.

[2] S. G. Baker, "A cancer theory kerfuffle can lead to new lines of research," *JNCI J. Nat. Cancer Inst.*, vol. 107, no. 2, Dec. 2014, Art. no. dju405.

[3] X.-P. Xie, Y.-F. Xie, Y.-T. Liu, and H.-Q. Wang, "Adaptively capturing the heterogeneity of expression for cancer biomarker identification," *BMC Bioinf.*, vol. 19, no. 1, p. 401, Dec. 2018.

[4] R. A. Smith, K. S. Andrews, D. Brooks, S. A. Fedewa, D. Manassaram-Baptiste, D. Saslow, and R. C. Wender, "Cancer screening in the United States, 2019: A review of current American cancer society guidelines and current issues in cancer screening," *CA, Cancer J. Clinicians*, vol. 69, no. 3, pp. 184–210, May 2019.

[18] A measurable extent to which an explanation to humans achieves a specified level of causal understanding [86].

[19] This technique is known as snapshot neural ensemble method [87]. In a previous approach [31], model ensemble technique was found very effective compared to individual models as it helps to reduce error by increasing model generalization.

[5] J. Liu, X. Wang, Y. Cheng, and L. Zhang, "Tumor gene expression data classification via sample expansion-based deep learning," *Oncotarget*, vol. 8, no. 65, pp. 109646–109660, Dec. 2017.
[6] P. J. Ballester and J. Carmona, "Artificial intelligence for the next generation of precision oncology," *NPJ Precis. Oncol.*, vol. 5, no. 1, pp. 1–3, 2020.
[7] R. Delgado, J. D. Núñez-González, J. C. Yébenes, and Á. Lavado, "Survival in the intensive care unit: A prognosis model based on Bayesian classifiers," *Artif. Intell. Med.*, vol. 115, May 2021, Art. no. 102054.
[8] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1A, p. A68, 2015.
[9] J. N. Weinstein, E. A. Collisson, and G. B. E. A. Mills, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, p. 1113, 2013.
[10] B. Lyu and A. Haque, "Deep learning based tumor type classification using gene expression," in *Pro. ACM Intl. Conf. Bioinf., Comput. Biol., Health Informat.*, 2018, pp. 89–96.
[11] C. Piyawajanusorn, L. C. Nguyen, G. Ghislat, and P. J. Ballester, "A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling," *Briefings Bioinf.*, vol. 22, no. 6, Nov. 2021, Art. no. bbab312.
[12] S. Albaradei, M. Thafar, A. Alsaedi, C. Van Neste, T. Gojobori, M. Essack, and X. Gao, "Machine learning and deep learning methods that use omics data for metastasis prediction," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 5008–5018, Jan. 2021.
[13] K. Lee, J. H. Lockhart, M. Xie, R. Chaudhary, R. J. C. Slebos, E. R. Flores, C. H. Chung, and A. C. Tan, "Deep learning of histopathology images at the single cell level," *Frontiers Artif. Intell.*, vol. 4, p. 137, Sep. 2021.
[14] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," *Genome Med.*, vol. 13, no. 1, pp. 1–17, Dec. 2021.
[15] S. Klein and D. G. Duda, "Machine learning for future subtyping of the tumor microenvironment of gastro-esophageal adenocarcinomas," *Cancers*, vol. 13, no. 19, p. 4919, Sep. 2021.
[16] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
[17] N. Braman, J. W. Gordon, E. T. Goossens, C. Willis, M. C. Stumpe, and J. Venkataraman, "Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2021, pp. 667–677.
[18] M. R. Karim, M. Cochez, O. Beyan, S. Decker, and C. Lange, "OncoNetExplainer: Explainable predictions of cancer types based on gene expression data," in *Proc. IEEE 19th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2019, pp. 415–422.
[19] M. R. Karim, G. Wicaksono, I. G. Costa, S. Decker, and O. Beyan, "Prognostically relevant subtypes and survival prediction for breast cancer based on multimodal genomics data," *IEEE Access*, vol. 7, pp. 133850–133864, 2019.
[20] P. Thiam, H. Kestler, and F. Schwenker, "Multimodal deep denoising convolutional autoencoders for pain intensity classification based on physiological signals," in *Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods*, 2020, pp. 289–296.
[21] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, p. 14680–14691.
[22] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," 2018, *arXiv:1802.04865*.
[23] C. Kästner. (2021). *Why robustness is Not Enough for Safety and Security in Machine learning*. Accessed: May 18, 2021. https://towardsdatascience.com/why-robustness-is-not-enough-for-safety-and-security-in-machine-learning-1a35f6706601
[24] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019.
[25] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. ICLR*, 2017, pp. 1–13.
[26] M. Kliger and S. Fleishman, "Novelty detection with GAN," 2018, *arXiv:1802.10560*.
[27] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," 2018, *arXiv:1812.04606*.
[28] Y. Li, K. Kang, J. Krahn, and L. Li, "A comprehensive genomic pan-cancer classification using the cancer genome Atlas gene expression data," *BMC Genomics*, vol. 18, no. 1, p. 508, Jul. 2017.
[29] M. R. Karim, M. A. Rahman, and O. Beyan, "Cancer Risk and type prediction based on copy number variations with LSTM and deep belief networks," in *1st Int. Artif. Intell. Conf. (A2IC)*, vol. 1, 2018.
[30] S. A. Malekpour, H. Pezeshk, and M. Sadeghi, "MSeq-CNV: Accurate detection of copy number variation from sequencing of multiple samples," *Sci. Rep.*, vol. 8, no. 1, p. 4009, Dec. 2018.
[31] M. R. Karim, A. Rahman, S. Decker, and O. Beyan, "A snapshot neural ensemble method for cancer type prediction based on copy number variations," *Neural Comput. Appl.*, vol. 2, pp. 21–45, Oct. 2019.
[32] K. Sun, J. Wang, H. Wang, and H. Sun, "GeneCT: A generalizable cancerous status and tissue origin classifier for pan-cancer biopsies," *Bioinformatics*, vol. 34, no. 23, pp. 4129–4130, Dec. 2018.
[33] M. Mostavi, Y.-C. Chiu, Y. Huang, and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression," 2019, *arXiv:1906.07794*.
[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[35] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.
[36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.
[37] N. Itzhacky and R. Sharan, "Prediction of cancer dependencies from expression data using deep learning," *Mol. Omics*, vol. 17, no. 1, pp. 66–71, 2021.
[38] H.-I.-H. Chen, Y.-C. Chiu, T. Zhang, S. Zhang, Y. Huang, and Y. Chen, "GSAE: An autoencoder with embedded gene-set nodes for genomics functional characterization," *BMC Syst. Biol.*, vol. 12, no. S8, pp. 45–57, Dec. 2018.
[39] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, Jul. 2019.
[40] J. H. Phan, R. Hoffman, S. Kothari, P.-Y. Wu, and M. D. Wang, "Integration of multi-modal biomedical data to predict cancer grade and patient survival," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Feb. 2016, pp. 577–580.
[41] F. T. Ito, H. de Medeiros Caseli, and J. Moreira, "The effects of unimodal representation choices on multimodal learning," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1–8.
[42] P. Thiam, H. Kestler, and F. Schwenker, "Multimodal deep denoising convolutional autoencoders for pain intensity classification based on physiological signals," in *Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods*, 2020, pp. 289–296.
[43] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010.
[44] M. R. Karim, O. Beyan, A. Zappa, I. G. Costa, D. Rebholz-Schuhmann, M. Cochez, and S. Decker, "Deep learning-based clustering approaches for bioinformatics," *Briefings Bioinf.*, vol. 22, no. 1, pp. 393–415, Jan. 2021.
[45] Y. Liu, X. Feng, and Z. Zhou, "Multimodal video classification with stacked contractive autoencoders," *Signal Process.*, vol. 120, pp. 761–766, Mar. 2016.
[46] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*. Madison, WI, USA: Omnipress, 2011, pp. 689–696.
[47] Y. Liu, X. Feng, and Z. Zhou, "Multimodal video classification with stacked contractive autoencoders," *Signal Process.*, vol. 120, pp. 761–766, Mar. 2016.
[48] S. Wang, J. Zhang, and C. Zong, "Associative multichannel autoencoder for multimodal word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 115–124.
[49] P. Sirpal, R. Damseh, K. Peng, D. K. Nguyen, and F. Lesage, "Multimodal autoencoder predicts fNIRS resting state from EEG signals," *Neuroinformatics*, pp. 1–22, 2021.
[50] Y. Zhang, A. Li, C. Peng, and M. Wang, "Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 5, pp. 825–835, Sep. 2016.

[51] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 841–850, Jun. 2018.

[52] J. Gao, T. Lyu, F. Xiong, J. Wang, W. Ke, and Z. Li, "MGNN: A multimodal graph neural network for predicting the survival of cancer patients," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1697–1700.

[53] L. Wang, M. Chignell, H. Jiang, and N. Charoenkitkarn, "Cluster-boosted multi-task learning framework for survival analysis," in *Proc. IEEE 20th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2020, pp. 255–262.

[54] L. Tong, J. Mitchel, K. Chatlin, and M. D. Wang, "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–12, Dec. 2020.

[55] 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," *Nature Methods*, vol. 11, no. 3, pp. 333–337, Jan. 2014.

[56] C. Wang, J. Guo, N. Zhao, Y. Liu, X. Liu, G. Liu, and M. Guo, "A cancer survival prediction method based on graph convolutional network," *IEEE Trans. Nanobiosci.*, vol. 19, no. 1, pp. 117–126, Jan. 2020.

[57] N. Dalvi, P. Domingos, S. Sanghai, and D. Verma, "Adversarial classification," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 99–108.

[58] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2013, pp. 387–402.

[59] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[60] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[61] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[62] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.

[63] V. Sehwag, M. Chiang, and P. Mittal, "SSD: A unified framework for self-supervised outlier detection," 2021, *arXiv:2103.12051*.

[64] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14680–14691.

[65] S. Vernekar, A. Gaurav, V. Abdelzad, T. Denouden, R. Salay, and K. Czarnecki, "Out-of-distribution detection in classifiers via generation," 2019, *arXiv:1910.04241*.

[66] A. A. Papadopoulos, M. R. Rajati, N. Shaikh, and J. Wang, "Outlier exposure with confidence control for out-of-distribution detection," *Neurocomputing*, vol. 441, pp. 138–150, Jun. 2021.

[67] G. Shalev, Y. Adi, and J. Keshet, "Out-of-distribution detection using multiple semantic label representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7386–7396.

[68] H. Choi and E. Jang, "Generative ensembles for robust anomaly detection," *arXiv preprint arXiv1810.01392*, 2018.

[69] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 14680–14691.

[70] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," 2018, *arXiv:1802.04865*.

[71] Q. Yu and K. Aizawa, "Unsupervised out-of-distribution detection by maximum classifier discrepancy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9518–9526.

[72] M. Alirezaie, M. Längkvist, M. Sioutis, and A. Loutfi, "Semantic referee: A neural-symbolic framework for enhancing geospatial semantic segmentation," 2019, *arXiv:1904.13196*.

[73] X. Guo, X. Liu, and E. E. A. Zhu, "Deep clustering with convolutional autoencoders," in *Proc. Int. Conf. Neural Inf. Process.*, Guangzhou, China: Springer, 2017, pp. 373–382.

[74] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[75] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*.

[76] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 648–657.

[77] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 135–147.

[78] S. Rifai, G. Mesnil, P. Vincent, X. Müller, Y. Bengio, Y. Dauphin, and X. Glorot, "Higher order contractive auto-encoder," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2011, pp. 645–660.

[79] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "PixelDefend: Leveraging generative models to understand and defend against adversarial examples," 2017, *arXiv:1710.10766*.

[80] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[81] J. N. Weinstein, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, p. 1113, 2013.

[82] S. Naulaerts, C. C. Dang, and P. J. Ballester, "Precision and recall oncology: Combining multiple gene mutations for improved identification of drug-sensitive tumours," *Oncotarget*, vol. 8, no. 57, 2017, Art. no. 97025.

[83] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel, "Evaluating the robustness of neural networks: An extreme value theory approach," 2018, *arXiv:1801.10578*.

[84] J. Gu, Y. Yang, and V. Tresp, "Understanding individual decisions of CNNs via contrastive backpropagation," in *Proc. Asian Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2018, pp. 119–134.

[85] M. R. Karim, M. Cochez, J. B. Jares, M. Uddin, O. Beyan, and S. Decker, "Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network," in *Proc. 10th ACM Int. Conf. Bioinf., Comput. Biol. Health Informat.*, Sep. 2019, pp. 113–123.

[86] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI," *Inf. Fusion*, vol. 71, pp. 28–37, Jul. 2021.

[87] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get M for free," 2017, *arXiv:1704.00109*.