IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 December 2022, accepted 20 January 2023, date of publication 30 January 2023, date of current version 6 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3240515

METHODS

Moanna: Multi-Omics Autoencoder-Based Neural Network Algorithm for Predicting Breast Cancer Subtypes

RICHARD LUPAT[1,2], **RASHINDRIE PERERA**[1,3], **(Member, IEEE), SHERENE LOI**[1,2], **AND JASON LI**[1,2], **(Member, IEEE)**

[1]Division of Cancer Research, Peter MacCallum Cancer Centre, Melbourne, VIC 3000, Australia
[2]The Sir Peter MacCallum Department of Oncology, The University of Melbourne, Parkville, VIC 3010, Australia
[3]Optimization and Pattern Recognition Group, Faculty of Engineering and IT, The University of Melbourne, Parkville, VIC 3010, Australia

Corresponding author: Jason Li (jason.li@petermac.org)

This work was supported by the Peter MacCallum Cancer Foundation. The work of Sherene Loi was supported by the National Breast Cancer Foundation of Australia Endowed Chair and the Breast Cancer Research Foundation, New York, NY, USA.

ABSTRACT Cancer subtyping delivers valuable insights into the study of cancer heterogeneity and fulfills an essential step toward personalized medicine. For example, studies in breast cancer have shown that cancer subtypes based on molecular differences are associated with different patient survival and treatment responses. However, recent studies have suggested inconsistent breast cancer subtype classifications using alternative approaches, suggesting that current methods are yet to be optimized. Existing computation-based methods have also been limited by their dependency on incomplete prior knowledge and ineffectiveness in handling high-dimensional data beyond gene expression. Here, we propose a novel deep-learning-based algorithm, Moanna, that is trained to integrate multi-omics data for predicting breast cancer subtypes. Moanna's architecture consists of a semi-supervised Autoencoder attached to a multi-task learning network for generalizing the combination of gene expression, copy number and somatic mutation data. We trained Moanna on a subset of the METABRIC breast cancer dataset and evaluated the performance on the remaining hold-out METABRIC samples and a fully independent cohort of TCGA samples. We evaluated our use of Autoencoder against other dimensionality reduction techniques and demonstrated its superiority in learning patterns associated with breast cancer subtypes. The overall Moanna model also achieved high accuracy in predicting samples' ER status (96%), differentiating basal-like samples (98%), and classifying samples into PAM50 subtypes (85%). Moreover, Moanna's predicted subtypes show a stronger correlation with patient survival when compared to the original PAM50 subtypes.

INDEX TERMS Feature extraction, cancer subtyping, artificial neural networks, machine learning, classification algorithms, cancer genomics, bioinformatics, genetic expression, deep learning, artificial intelligence.

I. INTRODUCTION

Cancer is characterised by abnormal cells that are invasive and growing out of control [1]. Each cancer type, such as breast cancer, can be further categorised into multiple subtypes through histopathological and clinical characteristics, and more recently, through molecular profiling of the primary tumour [2], [3], [4], [5], [6].

Cancer subtyping provides valuable molecular insights that help achieve personalised treatments. In breast cancer, multiple studies have demonstrated that tumours with different pathological and molecular features display different biological characteristics despite originating from the same site [2], [3], [4], [5], [6]. These studies have identified four main primary breast cancer intrinsic subtypes, namely luminal A,

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

luminal B, HER2-enriched and basal-like subtypes, through unbiased hierarchical clustering of gene expression patterns among the samples [2], [3], [4], [5], [6]. The primary characteristics of the subtypes are based on the expression levels of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and proliferation indicator Ki67 [2], [3], [4], [5], [6].

Breast cancer subtypes have been associated with distinctive clinical presentations, risk factors, responses to treatments and prognosis profile [7], [8]. The ER-positive group has higher 5-year overall survival and relapse-free survival than the ER-negative tumours, and better response to hormonal therapy such as tamoxifen [7], [9]. Luminal A is the most common subtype of breast cancer and has a better prognosis compared to luminal B, which occurs in $10\%-20\%$ of breast cancer cases [9]. The HER2-enriched group, which happens in $5\% - 15\%$ of breast cancer, proliferates faster with worse prognosis but is more likely to respond to HER2-targeted therapy, such as trastuzumab or lapatinib [9]. Triple-negative breast cancer (TNBC), which includes most basal-like tumours, tends to be more aggressive and has the worst prognosis among all other subtypes with few targeted therapy available [9].

In this study, we introduce a neural network algorithm for predicting breast cancer subtypes using the combination of gene expression, copy number variation and somatic mutation data. Apart from gene expression profiles, studies have shown that breast cancer subtypes show different patterns of mutations and copy number aberrations [19], [20], [21], [22]. Basal-like breast cancer is characterised by a high prevalence of *TP53* mutations, and deletion of RB1 and BRCA1, while ERBB2 amplification is often associated with HER2-enriched subtypes [19]. On the other hand, the two luminal subtypes are frequently observed with *PIK3CA* mutations, with luminal B also showing a higher frequency of mutated *TP53* gene than luminal A [19]. Recent advancements in deep learning technologies for gene expression, copy number variation and somatic mutation data analysis have shown success in using deep learning for omics data analysis [23]. Therefore, we hypothesise that integrating these different sources of omics data through a deep learning model will improve prediction for subtype classification. However, as discussed in detail in our related work (section II), existing work on breast cancer subtype classification have not utilized the advancements in deep learning for the integration of multiple omics data in subtype classification.

Our proposed solution in this paper is to develop a multi-omics neural network-based algorithm (Moanna) to classify molecular breast cancer subtypes using a semi-supervised Autoencoder layer that is jointly trained with supervised feed-forward neural network multi-task classification layers. It is important to note that the main aim of this study is not to identify new clusters, but rather to further refine subtype classification provided by current methodology with the help of state-of-the-art neural network models in integrating copy number and somatic mutation data on top of the well-evaluated gene expression data. The employed dimensionality reduction technique is designed to computationally generalise the high-dimensional multi-omics data, away from the limitation of the prior knowledge method. Thus, the implementation will then serve as a proof of concept for future Moanna's application in predicting other breast cancer biomarkers, such as the percentage of Tumour Infiltrating Lymphocytes (TILs) and for building a deep-learning-based prognosis model.

II. RELATED WORK

There are many published methodologies to identify the intrinsic subtypes of breast cancer. Two of the most frequently used methods in the clinical settings are either immunohistochemistry (IHC)-based markers or gene expression-based assays. PAM50 (50-gene signature), MammaPrint (70-gene signature) and BluePrint (80-gene signature) are examples of assays based on gene expression [10], [11], [12]. Subtypes identified by these methods are able to predict prognosis and potential targeted therapies that benefit patients [13]. However, multiple studies have shown that breast cancer subtypes identified by these methodologies do not always align, with as high as 25% discordance rate between the IHC-based method and MammaPrint/ BluePrint [11] and 38.4% between IHC-based subtype and PAM50 [14]. The inconsistencies could also be attributed to intra-tumour heterogeneity, where samples are composed of multiple subtypes [15], [16], [17]. In addition, the PAM50-classifier has been demonstrated to have limitations if ER status is not balanced within the dataset [18]. Therefore, there is a scope to further improve the precision of the methodologies used to identify subtypes.

Recent advances in the field of machine learning have enabled deep learning algorithms to be applied more widely on cancer data. Specifically, innovations in computer vision and artificial intelligence have assisted developments in radiographic imaging and digital pathology [24], [25], [26], [27]. For instance, deep learning techniques have been applied to diagnose metastasis in lymph nodes of breast cancer patients from whole-slide pathology images [25] and to automatically classify lung cancer tissue into its specific lung cancer subtypes [26]. Algorithms such as DeepSurv [28] and Cox-nnet [29] built prognosis predictors using artificial neural network extension of the Cox regression model. Other deep learning-based methods such as Tybalt uses Autoencoders, an unsupervised neural network approach, to extract biologically relevant features from gene expression data [30].

One of the difficulties of deep learning applications in genomics is its high-dimensional data. The number of genes available is significantly larger than the availability of training data, leading the model to often overfit. Deep learning implementations, such as DeepCC [31], use function pathways to transform input gene expression data, while DeepTRIAGE [32] converts its input features through Gene

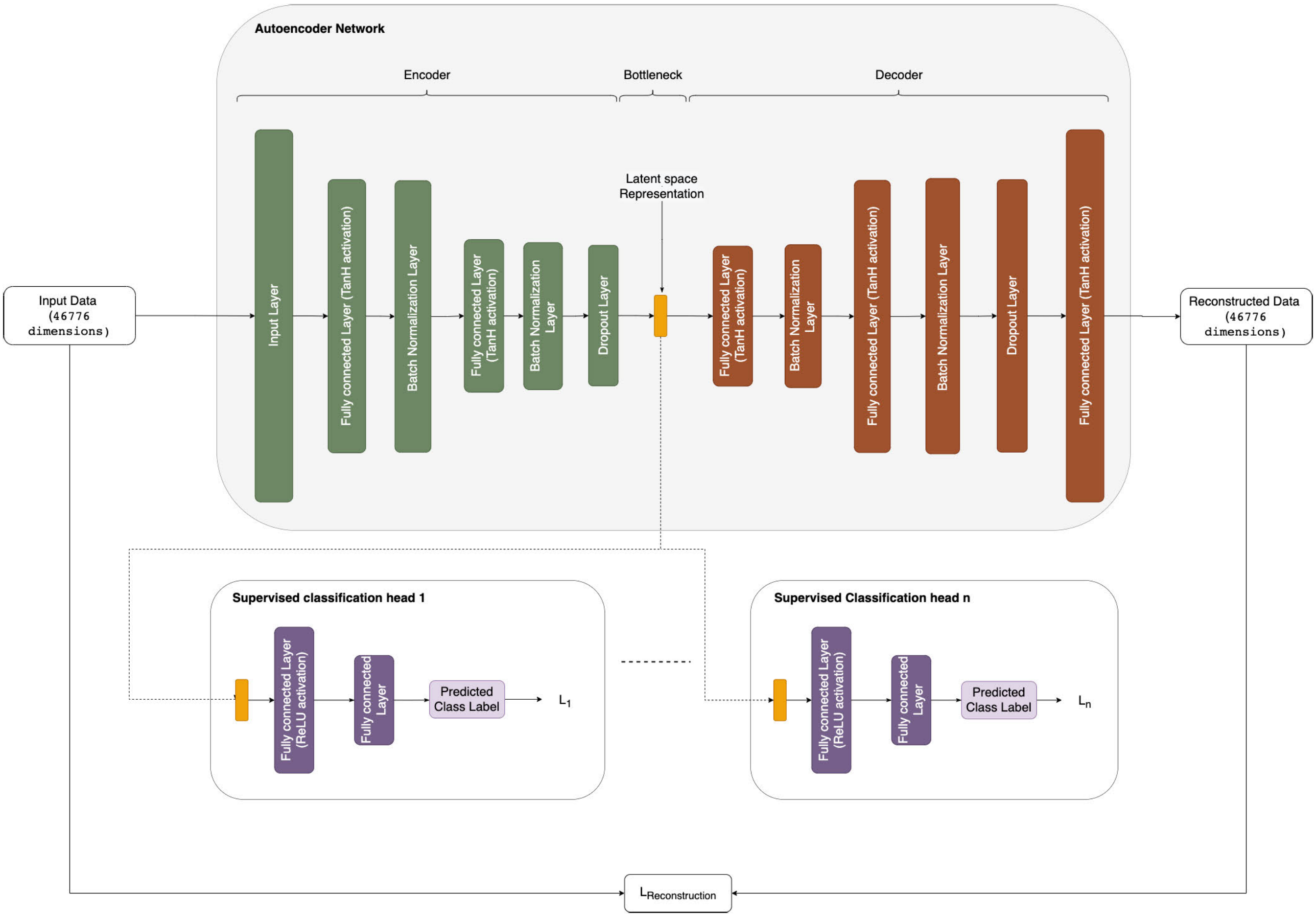

Autoencoder Network
Encoder
Bottleneck
Decoder
Latent space Representation
Input Data (46776 dimensions)
Input Layer
Fully connected Layer (TanH activation)
Batch Normalization Layer
Fully connected Layer (TanH activation)
Batch Normalization Layer
Dropout Layer
Fully connected Layer (TanH activation)
Batch Normalization Layer
Fully connected Layer (TanH activation)
Batch Normalization Layer
Dropout Layer
Fully connected Layer (TanH activation)
Reconstructed Data (46776 dimensions)
Supervised classification head 1
Fully connected Layer (ReLU activation)
Fully connected Layer
Predicted Class Label
L1
Supervised Classification head n
Fully connected Layer (ReLU activation)
Fully connected Layer
Predicted Class Label
Ln
LReconstruction

FIGURE 1. Overview of Moanna's neural network architecture for predicting breast cancer subtypes using multi-omics data. The input to Moanna's Autoencoder network is processed through several fully connected, batch-normalization and activation layers (encoder) to produce a latent space vector representation of 64 dimensions. The decoder will then take this bottleneck layer representation and up-sample its dimensions to reconstruct the input data using a reverse replica of the encoder network. Next, the bottleneck layer representation of the input data is extracted and fed as input to several feed-forward neural networks for supervised classification. Each supervised classification head handles the classification of a specific breast cancer biomarker.

Ontology (GO). These prior-knowledge-based dimensionality reduction techniques have an excellent advantage in their interpretability [33], [34], [35], [52]. However, they have also been described to have some limitations, particularly around bias on the knowledge that is still incomplete, as well as the inability to include all genes in the datasets [33], [34]. Moreover, they often only work for a single point of data, in this case, only gene expression data [33], [34], and thus not applicable to multiple omics data integration. In contrast, Moanna attempts to overcome such limitations in current work by employing recent advancements in deep learning for integration of multi-omics data to refine the subtype classifications provided by existing work.

III. MATERIALS AND METHODS

In this section, we outline the detailed description of our proposed deep neural network architecture, Moanna, as well as the datasets used to train, validate and test the breast cancer subtyping model.

A. MOANNA

Moanna is a deep learning framework that combines multiple supervised and unsupervised neural network architectures. This setup is adapted from the idea of semi-supervised Autoencoders, or also known as ladder network, where a supervised learning method is attached to a deep Autoencoder to assist in filtering irrelevant features [36]. This allows both networks to be jointly trained, instead of only utilising the Autoencoder as a separate pre-training model for dimensionality reduction [36], [37], [38]. For supervised biomarker classification, we employ multi-task learning which has been described to be useful in improving independent multi-class classifications by reducing overfitting in general [39]. In addition, breast cancer samples' hormone receptor status and subtypes have been studied to be correlated and it is therefore intuitive that the classification neural networks should share common variables. This led to the design of Moanna, where the classification tasks share some mutual hidden layers and parameters. The two major components in Moanna are shown in Fig. 1 and described below:

1) Semi-supervised Autoencoder layer: Each of the samples in the datasets consists of approximately 47, 000 features, containing the details of gene expression, copy number and somatic mutation profiles from over 15, 000 genes. Small datasets with a large number of features (large p; small n problems) is a common obstacle of deep learning application, where the feature engineering step is required to prevent overfitting [33]. As a solution, Moanna employs an Autoencoder in the network architecture. An Autoencoder is an unsupervised machine learning technique consisting of an encoder function and a decoder function. The encoder function maps the high dimensional input features to a compressed, latent internal representation while the decoder function attempts to recreate the original data using only the latent representation [40]. During training, the network optimises itself to better compress the input data in a meaningful manner such that the decoder can reconstruct the original data using only the compressed representation. In our implementation, we selected the number of layers for the encoder and the number of neurons in each layer using hyperparameter tuning (III-A2) while the decoder was constructed as a reverse replica of the encoder network. This setup converts the original data of 47000 dimensions into a latent vector of 64 dimensions which is fed into the multi-task classification heads.
2) Multi-task classifications layers: The 64 dimensional latent feature vector from the bottleneck layer of the autoencoder is carried into several feed-forward neural networks for supervised classification. For this study, we are using multiple breast cancer biomarkers, including ER status, HER2 status and PAM50 subtypes, as our training labels. A separate classification head was added in parallel to handle the classification of each biomarker in the dataset.

1) NEURAL NETWORK TRAINING

A joint supervised and unsupervised neural network training allows better generalisation in data learning [37]. The sum of loss functions from the two components becomes the objective function that is used to train this model (eq. 1). For the semi-supervised autoencoder, Moanna measures the mean-squared error between the input and reconstructed layer ($L_{\text{reconstruction}}$). On the other hand, cross-entropy loss between training and predicted classification labels were calculated for the classification tasks (L_{i}). This objective function was jointly optimised with a single backpropagation using a stochastic gradient descent algorithm, eliminating the necessity to set up multiple independent sets of training. Therefore, apart from better generalisation, this neural network architecture is also more efficient computationally [37].

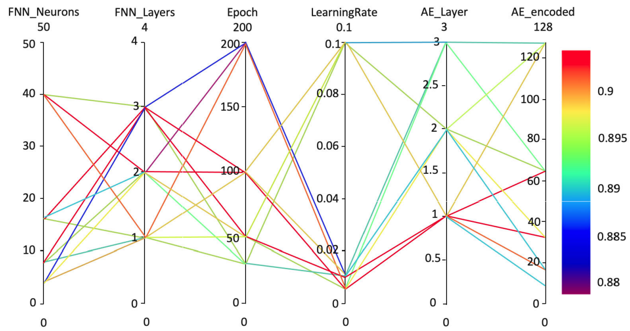


FIGURE 2. Parallel coordinates plot of different Moanna's parameters combination. Red arrows represent the combinations that Moanna employs in its final model. These are the parameters with the highest PAM50 subtype prediction accuracy from doing a grid search on our validation data.

$$L_{total} = \underbrace{L_1 + \ldots + L_n}_{\text{loss from n classification tasks}} + \underbrace{L_{reconstruction}}_{\text{loss from autoencoder network}} \tag{1}$$

2) HYPERPARAMETER TUNING

We performed a grid search to select the optimum Moanna parameters with the highest classification accuracy on our hold-out validation data (Fig. 2). The final designed model consists of 1 Autoencoder and 5 supervised classifiers. The encoder part of this Autoencoder was designed with 2 hidden layers of 256 and 128 neurons, a representation layer of 64 encoded neurons and a Tanh activation function. The decoder part of the model mirrored the encoder setup on the other side. The classifiers took these 64 encoded features through a hidden layer of 40 neurons. Moanna used Stochastic Gradient Descent (SGD) as its optimiser for backpropagation with a learning rate of 0.005 and momentum of 0.9, over 100 epochs.

B. DATASETS

Moanna was trained on Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [21], [22] datasets downloaded from cbioportal [41], [42]. METABRIC is a comprehensive breast cancer study from over 2000 primary tumours, including gene expression and copy number profiles of 25, 160 genes alongside somatic mutations of 173 frequently mutated breast cancer genes. This dataset also comprises clinical data and long-term follow-up information, including the PAM50 subtypes, estrogen receptor (ER) and HER2 status that Moanna uses as its training label. We excluded samples that are not one of the four intrinsic subtypes (Basal-like, HER2-enriched, Luminal A and Luminal B) and samples that do not have all three genomics profiles (gene expression, copy number and somatic mutation). This left us with a total of 1689 samples which are then randomly split into 70% training and 30% hold-out validation data. The distribution of subtypes from the METABRIC dataset is shown in Table 1. While we use a single hold-out

TABLE 1. PAM50 subtype samples distribution from our training, validation and testing datasets.

PAM50 Subtype	Training (70% METABRIC) n=1182	Validation (30% METABRIC) n=507	Independent Test (TCGA) n=631
Basal-like	18% (n=213)	17.6% (n=89)	17.7% (n=112)
HER2-enriched	12.8% (n=151)	16.2% (n=82)	9.4% (n=59)
Luminal A	41% (n=485)	39.8% (n=202)	52.8% (n=333)
Luminal B	28.2% (n=333)	26.4% (n=134)	20.1% (n=127)

split to report results in the main text, we also ran a stratified k-fold cross-validation experiment to test the robustness of Moanna across different dataset splits. The results of these runs are provided in Supplementary Tables 1, and 2.

To evaluate the robustness of Moanna, we use the METABRIC-trained Moanna model for predicting subtypes of independent breast cancer datasets from The Cancer Genome Atlas (TCGA) [19], [20]. This TCGA dataset was also retrieved from cbioportal [41], [42], where a total of 954 samples were selected using the same criteria that we applied for METABRIC. The majority of these samples come with PAM50 subtype, ER, HER2 status and long-term follow-up information. The distribution of subtypes from these TCGA datasets is shown in Table 1.

1) DATA PRE-PROCESSING

Some of the major issues when dealing with gene expression profiles are the different platforms used to generate these data and possible batch effects associated with the experiments. Gene expression data from METABRIC were obtained through microarray data on the Illumina HT-12 v3 platform while TCGA transcriptomic profiles were from RNA-sequencing performed on Illumina HiSeq. Hence, we used the relative expression (z-score transformed) calculated by cbioportal where expression values have been further normalised based on the distribution of the diploid samples in the datasets.

For copy number variation (CNV) and somatic single nucleotide polymorphism (SNP) data, information is summarised into a matrix form of gene and sample combination. CNV data has a range of $[-2, 2]$, where 0 is copy number neutral; -1 represents heterozygous deletion; -2 indicates homozygous loss; 1 and 2 are low-level gain and high-level amplification respectively. SNP data is constructed in a binary format where 0 indicates no detected somatic mutation in that gene, and 1 represents the mutated gene. For METABRIC, any genes that are not sequenced by the targeted panel will be assigned 0 for its somatic mutation status.

The combinations of these pre-processed data were used as the input features to Moanna. An equal number of features from each 'omics type (gene expression, CNV, SNP) were included in the overall neural network design. For the results presented in this paper, we only include genes that have expression values in both METABRIC and TCGA datasets. After filtering, our input features consisted of approximately 47, 000 input features from over 15, 000 genes.

IV. EXPERIMENTS AND RESULTS

A. AUTOENCODERS AS THE BEST DIMENSIONALITY REDUCTION METHOD THROUGH BIOMARKER CLUSTER ANALYSIS

To address large p small n problems [33] on our datasets, we evaluated multiple dimensionality reduction techniques to prevent overfitting or poor generalisation to new data. The strategy of using Autoencoders for feature extraction is comparable to applying principal component analysis (PCA), which is another widely used dimensionality reduction technique. In PCA, high dimensional data is transformed to a series of eigenvectors and eigenvalues such that the top N principal components represent the majority of the variance of the original data [18], [33]. The data used in this work is non-linear as it is hypothesized that the expression of a gene can be driven by the expression of many other genes, as well as copy number changes [43]. Therefore, we believe non-linear transformations as such found in neural networks like Moanna may be better suited to handle omics data than linear transformations such as PCA. Additionally, alternative strategies through feature selection based on prior knowledge or level of activities have also been widely applied [32]. To cover such alternatives, we have compared Moanna's extracted features against randomly selected genes, PAM50 genes, top differentially expressed genes (DEG), and features extracted from the top 64 PCA principal components.

We first projected the input data into two-dimensional space with t-SNE [44] and compared the sample distribution with the t-SNE plot of the extracted features from Moanna's Autoencoder. Fig. 3 reports multiple clusters from Moanna's extracted features annotated by PAM50 subtypes and ER status. This indicates that the 64 neurons from the neural network model's representation layer have extracted important biological characteristics of the 47, 000 input features for the purpose of subtyping, even before going through the final classification layers. We observed the same result when we repeated the exercise on TCGA breast cancer datasets, showing a vast improvement when compared to the clusters from the original input features.

To further evaluate the performance of Moanna's Autoencoder, we performed clustering analysis on different selected and extracted features. The comparison includes: 1) gene expression of 50 genes from PAM50, 2) top 200 differentially expressed genes (DEG), 3) first 50 principal components (from PCA) of all input features, 4) first 50 principal components (from PCA) of all gene expression input features, and 6) randomly selected 64 genes. Following the clustering evaluation strategy from Geddes et al. [45], we calculate three metrics for assessing the performance of these dimensionality reduction strategies in retaining relevant features required for clustering breast cancer samples to their subtypes. These metrics are Fowlkes-Mallows index (FM), Adjusted Rand Index (ARI) and normalised mutual information (NMI) score, which was calculated for each method after running k-means clustering on its selected/extracted features. In addition, we apply these features to Moanna's

TABLE 2. Results of dimensionality reduction evaluation based on clustering metrics: ARI, NMI and FM index.

Method	Dataset	ARI	NMI	FM
PAM50	V T	0.470 0.536	0.487 0.476	0.628 0.718
TopDEG (200)	V T	0.262 0.311	0.303 0.293	0.467 0.718
PCA (All)	V T	0.323 0.306	0.347 0.376	0.536 0.718
PCA (EXPR)	V T	0.242 0.264	0.300 0.389	0.474 0.718
Random	V T	0.206 0.250	0.259 0.310	0.422 0.559
Moanna	V T	**0.628** **0.621**	**0.629** **0.630**	**0.733** **0.752**

* **Bold denotes the best in its category.**

TABLE 3. Results of classification accuracies on ER status, HER2 status and PAM50 subtype classification tasks in comparison to other feature extraction and feature selection strategies.

Method	Dataset	ER status	HER2 status	PAM50 subtype
PAM50	V T	0.966 0.935	**0.974** 0.853	0.838 **0.851**
TopDEG (200)	V T	0.921 0.905	0.880 0.773	0.755 0.791
PCA (All)	V T	0.961 0.941	0.937 0.854	0.805 0.810
PCA (EXPR)	V T	**0.968** 0.937	0.945 0.859	**0.854** 0.843
Random	V T	0.935 0.926	0.878 0.773	0.694 0.754
Moanna	V T	0.964 **0.946**	0.959 **0.864**	0.850 0.848

* **Bold denotes the best in its category.**

TABLE 4. Classification metrics on multiple tasks predicted by Moanna on validation (V) and testing (T) datasets.

Classification task	Dataset	Accuracy	Precision*	Recall*	F1-Score*
ER Status	V T	0.964 0.946	0.965 0.947	0.964 0.946	0.965 0.947
HER2 Status	V T	0.959 0.864	0.960 0.872	0.959 0.863	0.959 0.844
PAM50 Subtype	V T	0.850 0.848	0.857 0.864	0.850 0.848	0.852 0.852
Basal vs other subtypes	V T	0.984 0.989	0.984 0.989	0.984 0.989	0.984 0.989

* weighted-average, calculated by scikit-learn package [46].

feed-forward neural network, by replacing the Autoencoder layer, to measure their usefulness when employed to solve classification problems. The result of this evaluation on both validation (V) and testing (T) datasets (see Table 2) indicates that Moanna's Autoencoder performed the best in clustering samples to their subtypes.

We also compare Moanna with other strategies such as feature extraction with PCA and feature selection of 1) PAM50, 2) top 200 DEG and 3) randomly selected 64 genes. We observe that Moanna achieves an overall better accuracy when deployed alongside a neural network classifier, in comparison to the other dimensionality reduction techniques tested (see Table 3). Moanna's extracted features are better at clustering samples to subgroups, and significantly improved clusters that are only based on 50 genes from PAM50. We used feature selection on PAM50 genes as our benchmark for this evaluation, given that our subtype training labels originated from this 50-gene signature, and that they were expected to perform the closest to the label. On the other hand, although it struggled to separate the clusters of luminal samples, unsupervised feature extraction using PCA achieved reasonable high classification accuracy when paired with Moanna's multi-task learning (see Fig. 3). The results shown in Fig. 3 indicate that Moanna's autoencoder performed best in clustering samples to their subgroups even before entering the multi-task learning layer.

B. MOANNA ACHIEVES HIGH ACCURACY IN PREDICTING ER-STATUS, HER2-STATUS AND PAM50 SUBTYPES

We applied the proposed method on our training datasets (70% METABRIC, $n = 1182$) and evaluated the classification accuracy, precision and recall on our validation samples (30% METABRIC, $n = 507$). Table 4 summarises Moanna classification performance on the METABRIC dataset splits where it accurately differentiates well-characterised markers, for instance, differentiating ER-positive (ER+) and ER-negative (ER-) samples (96.5% accuracy), as well as the difference between basal and non-basal-like samples (98.4% accuracy). In addition, the majority of the subtypes predicted by Moanna (85.6%) agree with the original subtypes identified by PAM50. From a total of 507 validation samples, Moanna classifies 16.6% ($n = 84$) basal-like, 13.6% ($n = 69$) HER2-like, 32.7% ($n = 166$) LumA-like and 22.1% ($n = 112$) LumB-like subtype.

We then further evaluated the 76 samples that were classified differently by Moanna in comparison to PAM50 (see Fig. 4a). We found that 28.9% ($n = 22$) of the dissimilarities are on ER+/HER2- High Proliferation samples that were classified as Luminal B-like by Moanna, but predicted as Luminal A in PAM50. There were also 15.8% ($n = 12$) samples that are ERBB2 amplified and classified as HER2-enriched by Moanna but called differently in PAM50. This discordance suggests that this Moanna's subtype prediction model did not only fit the training subtypes label but also integrated information learned from ER and HER2-status predictions.

C. APPLICATION OF MOANNA ON INDEPENDENT DATASETS SHOW THE MODEL DOES NOT OVERFIT

We next applied METABRIC-trained Moanna on the TCGA breast cancer dataset to evaluate the robustness of the architecture when dealing with new data from different experiments. Table 4 shows the precision and recall from this classification are consistent with the previous result Moanna

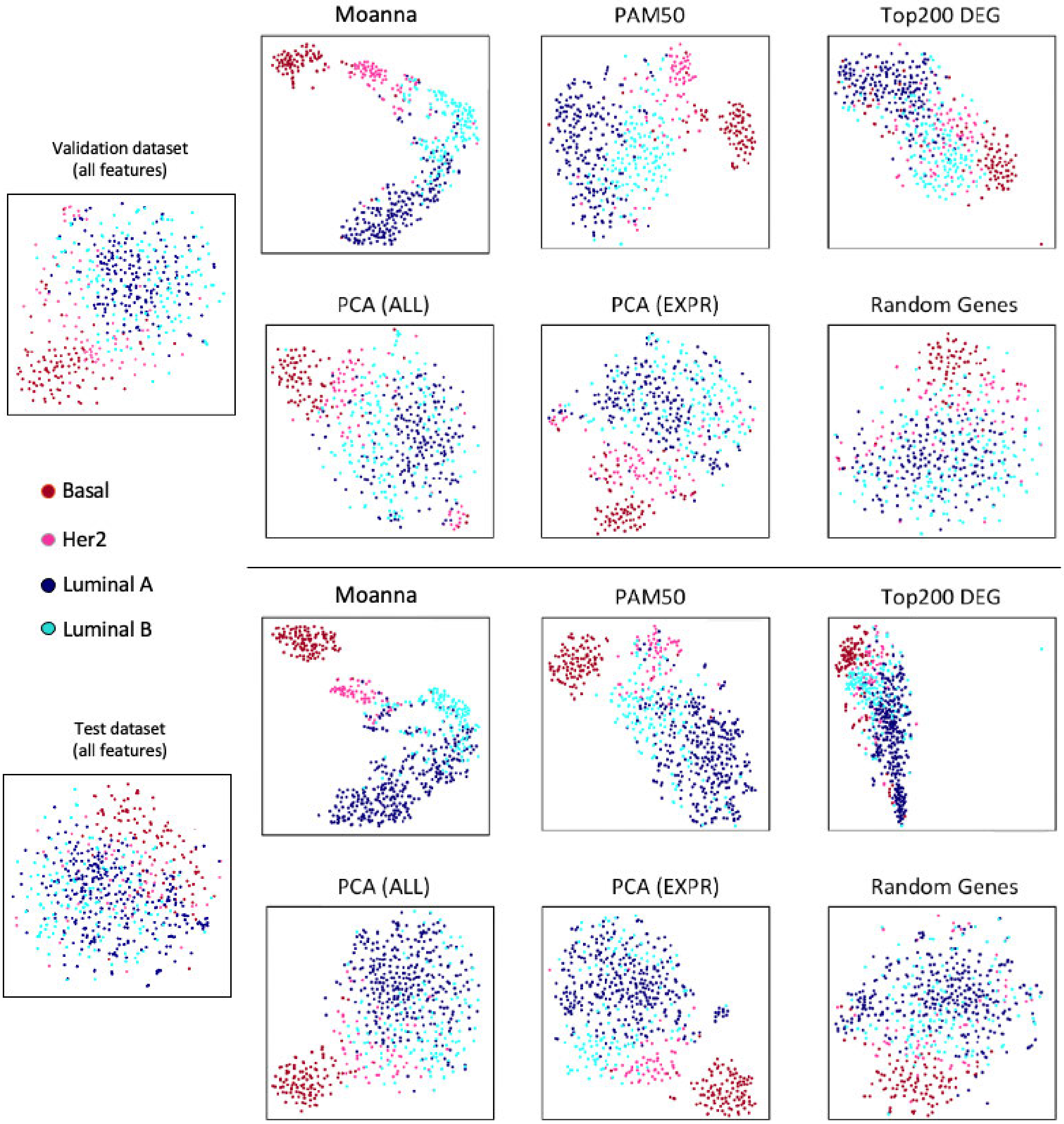


FIGURE 3. T-SNE plots of all the extracted/selected input-features through various dimensionality techniques described in Table 2 and Table 3. Top plots are from validation dataset, while bottom half plots are from testing dataset.

achieved on the METABRIC validation dataset. The model predicted the ER status at 94.7% accuracy when compared to the label acquired from cbioportal. It also managed to differentiate basal-like samples from the other subtypes at 98.9% accuracy while 86.4% of the subtypes predicted are concordant with the PAM50 subtype from TCGA. From a total of 631 test samples, Moanna classifies 17.6% ($n = 111$) as basal-like, 7.6% ($n = 48$) as HER2-like, 52.8% ($n = 333$) LumA-like and 20.1% ($n = 127$) LumB-like.

Fig. 4b shows the confusion matrix of Moanna's classification from both METABRIC and TCGA datasets, where it is obvious that the proportion of samples' subtypes are not balanced. HER2-enriched subtype has the least number of samples while luminal A samples represent almost half of the cases on both datasets. Imbalance class training has been studied to affect classifiers' performance [47], and we hypothesised that this would be one of the reasons for the lower concordance between the predicted HER2-like subtype and the training label. The other major dissimilarities are concentrated between the classification of the two luminal subtypes. A few studies on the same datasets have identified potential admixed cases in luminal A and luminal B samples, as well as further subclasses due to heterogeneity of luminal breast cancer [15], [16], [17].

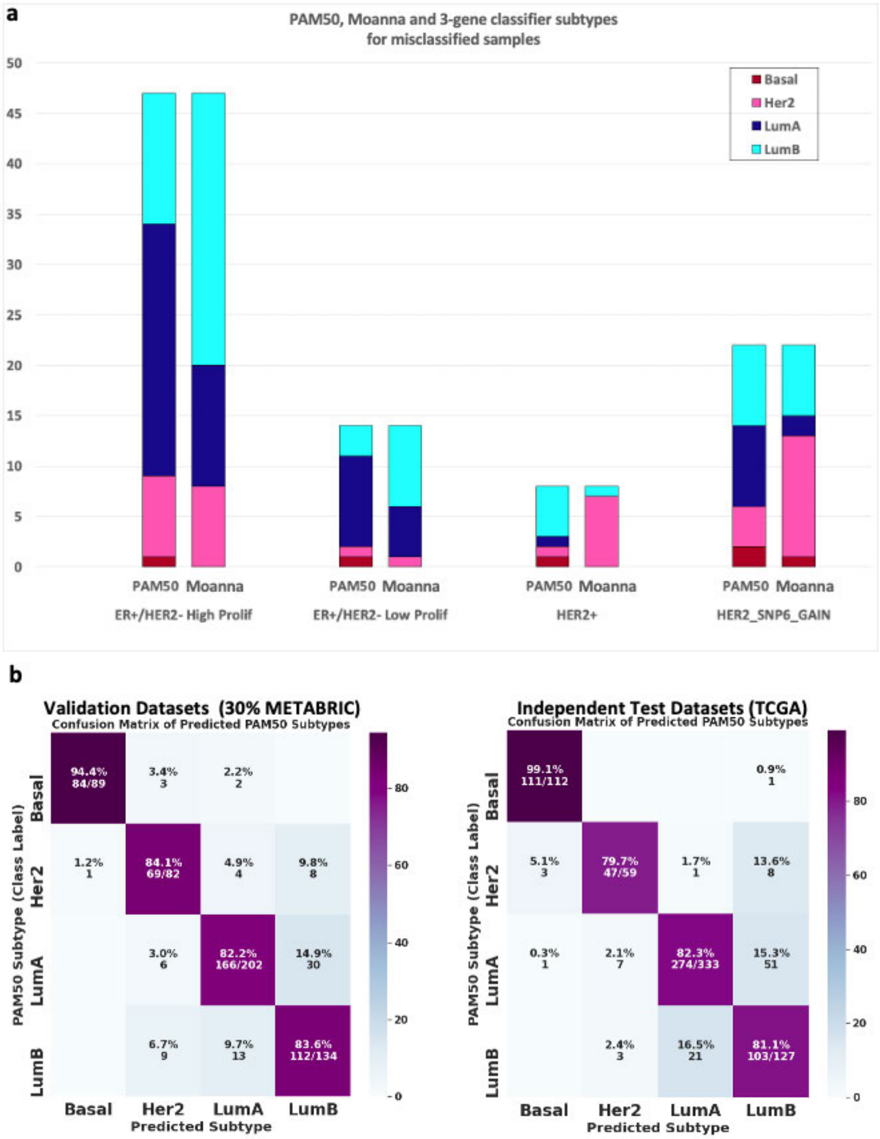


FIGURE 4. a) The stacked bar plots show the differences between PAM50 and Moanna's predicted subtypes from the 76 misclassified samples, grouped by the 3-gene classifier and the HER2 copy number gain status from SNP6 data. b) Heatmap visualisation of Moanna's confusion matrix (left: validation data; right: testing data).

D. MOANNA'S PREDICTED SUBTYPES SHOW BETTER CORRELATION TO PATIENTS' SURVIVAL

To validate the clinical significance of Moanna's classification, we perform disease-free-survival analysis using these predicted subtypes using Kaplan-Meier, a metric commonly used for survival analysis [51]. Kaplan-Meier plots (Fig. 5) show that Moanna's predicted subtypes display a more distinct separation of survival patterns compared to the original subtypes. To assess this further, we compare the prognosis between the two luminal subtypes (LumA-like vs LumB-like), which is one of the main dissimilarities between Moanna's and the original PAM50 classes. Cox proportional hazard ratio from our analysis shows a stronger correlation to patient survival between luminal A and luminal B samples ($HR = 2.95$, $CI = 1.45 - 6.00$, $p < 0.005$) when compared to the original subtypes ($HR = 1.98$, $CI = 1.03 - 3.82$, $p < 0.005$). This is consistent with literature where luminal A has a better prognosis than luminal B patients [9]. This result also implies subtypes that were predicted differently by Moanna were not necessarily misclassified, but rather a potential improvement to the original subtyping.

E. MOANNA PERFORMS MORE CONSISTENTLY THAN OTHER MACHINE LEARNING CLASSIFIERS

To further benchmark Moanna's performance, we constructed four others widely used machine learning algorithms

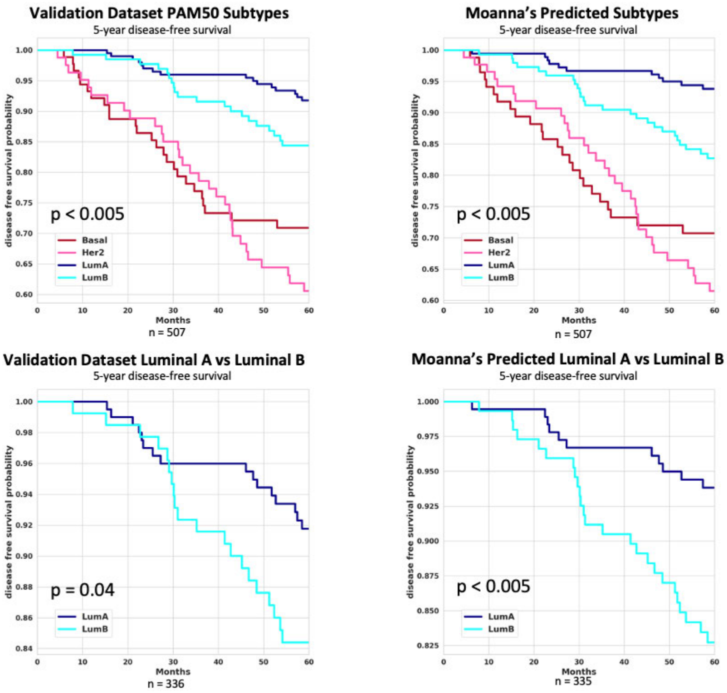


FIGURE 5. Kaplan-Meier survival-plot of 5-year disease-free-survival (DFS) of the predicted subtypes from all patients in our validation datasets. The top plots show the differences between all four PAM50 subtypes predicted by Moanna (right) and the original label (left). The bottom plots compare the survival analysis between Luminal A and Luminal B subtypes. Cox proportional hazard ratio analysis was performed using the Python package *lifelines* (https://doi.org/10.5281/zenodo.3267531).

for classification tasks based on random forest (RF), support vector machine (SVM), multinomial logistics regression, and stochastic gradient descent (SGD) based classifier. These algorithms were trained with an identical setup, including datasets split, number of samples and input features. Fig. 6 summarises the performance of all these machine algorithms when compared to the original hormone status and PAM50 subtypes. The precision and recall values indicate similar performance across all of these machine-learning implementations with Moanna and SVM being the top performers. The average F1-score (harmonic mean of precision of recall), calculated as the average of F1-score across all three classifications on independent testing datasets, shows that Moanna outperforms SVM and other methods (see Table 5).

TABLE 5. Comparisons of F1-score across all five machine learning algorithms on independent datasets classifications.

Task	Moanna	SGD Classifier	RF	Logistic Regression	SVM
ER Status	**0.947**	0.926	0.917	0.936	0.941
HER2 Status	**0.844**	0.816	0.827	0.838	0.841
PAM50 Subtype	0.852	0.851	**0.858**	0.850	**0.858**
Average	**0.881**	0.864	0.868	0.875	0.880

*** Bold denotes the best in its category.**

F. MOANNA'S MAIN DRIVER IS CORRELATED WITH THE GENOMIC DATA TYPE THAT DRIVES PAM50 SUBTYPE CLASSIFICATION

To assess the benefit of using multi-omics data over a single type of genomics data, we re-evaluated the classification accuracy of Moanna when trained with the individual omics data type. We set up multiple models trained on input features consisting of gene expression profiles (EXPR), copy number variation (CNV), and somatic mutation (SNP) data, and multiple combinations between them. The final evaluated Moanna referred to throughout this manuscript was trained and evaluated using a combination of all three data types.

The contribution of each data type and their combinations towards the classifying breast cancer subtypes on our datasets is summarised in Table 6. We completed this evaluation on both validation (V) and testing (T) datasets. Looking at individual data, it is clear that the gene expression profile is a better classifier in comparison to CNV and SNP data. This is not surprising given that many studies have demonstrated the utility of gene expression assays in capturing different breast cancer subtypes, including the PAM50 label that is being used for this study [3], [12], [13]. In addition, while CNV data alone do not have the same predictive power, the combined

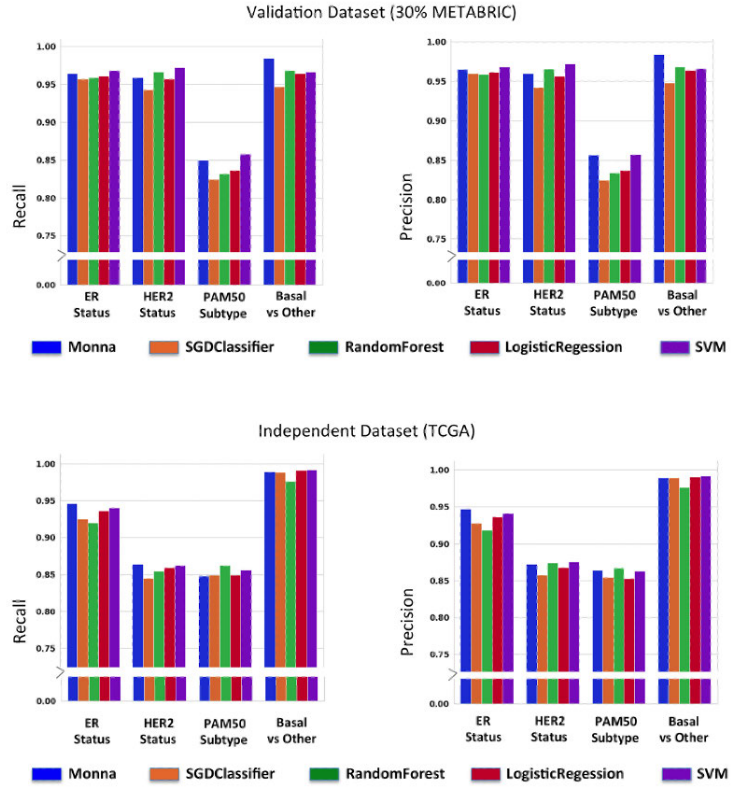


FIGURE 6. **Precision and recall summary of Moanna's evaluation against other machine learning algorithms (top: validation dataset; bottom: testing dataset). Moanna's classification accuracy is comparable to other widely used machine learning algorithms including Stochastic Gradient Descent (SGD) classifier, random forest, logistic regression classifier, and support vector machine.**

TABLE 6. **Classification accuracy of Moanna trained with various combinations of genomics data (EXPR = gene expression profile; CNV = copy number variation; SNP = somatic mutation data).**

Task	Dataset	EXPR	CNV	SNV	EXPR-CNV	EXPR-SNV	CNV-SNV	EXPR-CNV-SNV
ER status	V	**0.951**	0.901	0.807	**0.970**	0.968	0.919	**0.964**
	T	**0.946**	0.908	0.810	**0.948**	0.941	0.903	0.946
HER status	V	**0.957**	0.955	0.866	**0.968**	0.955	0.947	0.959
	T	**0.872**	0.853	0.773	**0.862**	0.859	0.848	**0.864**
PAM50 subtype	V	**0.815**	0.649	0.513	0.826	**0.842**	0.645	**0.850**
	T	**0.851**	0.686	0.517	**0.857**	0.853	0.669	0.848
Basal vs other subtypes	V	**0.961**	0.931	0.834	**0.976**	0.972	0.929	**0.984**
	T	**0.987**	0.959	0.838	**0.987**	**0.987**	0.954	**0.989**

*** Bold denotes the best in its category.**

data classification result suggests that CNVs are complementing the gene expression data in improving the classification accuracy. This is consistent with literature that studies how CNVs on certain genes cause them to be up-or-down regulated [48], [49]. On the other hand, we observe that the presence of SNP data as part of our input features contributes towards differentiating basal-like subtypes from the other subtypes. This is aligned with SNP analysis of these datasets where different breast cancer subtypes were described with different frequently mutated genes. For example, basal-like datasets have a higher frequency of *TP53* mutations, while luminal subtypes samples tend to see more *PIK3CA* mutations [19]. This analysis indicates that Moanna's neural network architecture setup provides a mechanism for combining the knowledge from different resolutions of omics data to achieve good classification accuracy.

V. DISCUSSIONS & CONCLUSION

Breast cancer is a heterogeneous disease with various subtypes that exhibits different characteristics. The four main molecular subtypes are Basal-like, HER2-enriched, Luminal A and Luminal B. These subtypes have been studied extensively to show differences in prognosis, incidence rate, and response to treatments and therapies [3], [4], [9]. Gene expression-based assays, such as the 50-gene panel called PAM50, are one of the well-established methods to infer molecular breast cancer subtypes [10]. However, there have been many studies analysing the discordance between gene expression and IHC-based subtypes. Various explanations have been proposed, such as the limitations of these assays and the presence of intra-tumour heterogeneity [11], [14], [15], [17], [18]. To evaluate this further, we developed a novel deep-learning-based framework, Moanna, to predict breast cancer subtypes by integrating gene expression, SNP and CNV data.

In this manuscript, we demonstrated that a trained Moanna model is capable of extracting biological patterns from its training datasets and predicting the biomarkers of breast cancer samples with high accuracy. Although not all of the predicted breast cancer subtypes agree with the provided labels on the validation and testing datasets, Moanna's predicted subtypes show a more significant correlation with patient survival when compared to the original subtype labels. This suggests that the mispredictions might not be necessarily incorrect, but rather a potential further investigation into the accuracy of the original labels.

The neural network architecture of Moanna is designed to handle the high-dimensionality of integrated 'omics data. It is a joint semi-supervised learning algorithm, based on the concept of a ladder network, combining the training of unsupervised Autoencoders and multi-task learning feed-forward neural networks. The ladder network design allows the Autoencoder to find relevant latent variables faster by discarding irrelevant features to the classification while maintaining a decoder that can reconstruct a representation of the input features. In addition, multitask learning setup improves the model generalisation, essentially equivalent to adding regularisation to the overall training by learning independent patterns using shared hidden layers. In combination, this implementation enables Moanna to be extended for other classifications beyond cancer subtyping.

There are, however, some limitations to this approach. First, the implementation of Moanna for breast cancer subtypes prediction currently does not work with a single sample as Moanna expected the gene expression data to be normalised against a control. Although this limitation can be addressed in future implementation by adding a baseline reference, it will still be largely restricted in the absence of normal samples in the cohort. This is an area that we are currently working on for the next iteration of Moanna. Second, Moanna currently integrates multi-omics data directly in its very first layer, despite dealing with discrete and continuous variables. While the chosen activation function could potentially deal with this limitation, various studies have proposed better approaches to dealing with different data types. One possible solution is to implement three different input channels before integrating post-neural-network features into the current architecture. In future work, we would explore options to extend Moanna for addressing these limitations.

In summary, we presented Moanna, a multi-omics neural network algorithm for predicting breast cancer subtypes. Through training and evaluation on public breast cancer datasets, we have demonstrated Moanna's performance in generalising knowledge extracted from gene expression, CNV and SNP data. Despite the heavy focus on breast cancer subtypes in this manuscript, Moanna's proof-of-concept implementation can be extended for predicting other biomarkers, such as the TILs or even for building a prognosis model. The generalised neural network architecture can also be deployed on other cancer types, extracting valuable information from vast amounts of public cancer datasets.

VI. AUTHOR CONTRIBUTIONS

Richard Lupat designed and implemented the neural network model, Rashindrie Perera contributed to model interpretation and manuscript writing, Sherene Loi provided clinical interpretation, and Jason Li supervised the study.

VII. CODE AVAILABILITY

Moanna was implemented as a collection of python scripts packaged in a docker image with all its python libraries dependencies. Moanna is developed with PyTorch [50] deep learning framework.

The source code and the trained model are available at https://github.com/rlupat/moanna. Preprocessed data used for the study are also available for download at https://doi.org/10.5281/zenodo.4326602.

ACKNOWLEDGMENT

The authors would like to thank all the patients enrolled in The Cancer Genome Atlas (TCGA) and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) studies.

CONFLICT OF INTEREST

Prof. Sherene Loi receives research funding to her institution from Novartis, Bristol Meyers Squibb, Merck, Puma Biotechnology, Eli Lilly, Nektar Therapeutics, AstraZeneca, and Seattle Genetics. She has acted as a Consultant (not compensated) to Seattle Genetics, Novartis, Bristol Meyers Squibb, Merck, AstraZeneca, Eli Lilly, Pfizer, Gilead Therapeutics, and Roche-Genentech. She has acted as a Consultant (paid to her institution) to Aduro Biotech, Novartis, GlaxoSmithKline, Roche-Genentech, AstraZeneca, Silverback Therapeutics, G1 Therapeutics, PUMA Biotechnologies, Pfizer, Gilead Therapeutics, Seattle Genetics, Daiichi Sankyo, Merck, Amunix, Tallac Therapeutics, Eli Lilly, and Bristol Meyers Squibb.

REFERENCES

[1] B. Alberts, *Molecular Biology of the Cell*, 6th ed. New York, NY, USA: Garland Science, 2015.

[2] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, and Ø. Fluge, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[3] T. Sørlie, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 19, pp. 10869–10874, 2001.

[4] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lønning, P. O. Brown, A.-L. Børresen-Dale, and D. Botstein, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 14, pp. 8418–8423, 2003.

[5] J. I. Herschkowitz, "Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors," *Genome Biol.*, vol. 8, no. 5, pp. 1–17, 2007.

[6] A. Prat, J. S. Parker, O. Karginova, C. Fan, C. Livasy, J. I. Herschkowitz, X. He, and C. M. Perou, "Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer," *Breast Cancer Res.*, vol. 12, no. 5, pp. 1–18, Oct. 2010.

[7] G. K. Malhotra, X. Zhao, H. Band, and V. Band, "Histological, molecular and functional subtypes of breast cancers," *Cancer Biol. Therapy*, vol. 10, no. 10, pp. 955–960, Nov. 2010.

[8] B. Weigelt, F. C. Geyer, and J. S. Reis-Filho, "Histological types of breast cancer: How special are they?" *Mol. Oncol.*, vol. 4, no. 3, pp. 192–208, Jun. 2010.

[9] Y. Feng, M. Spezia, S. Huang, C. Yuan, Z. Zeng, L. Zhang, X. Ji, W. Liu, B. Huang, W. Luo, B. Liu, Y. Lei, S. Du, A. Vuppalapati, H. H. Luu, R. C. Haydon, T.-C. He, and G. Ren, "Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis," *Genes Diseases*, vol. 5, no. 2, pp. 77–106, Jun. 2018.

[10] R. R. Bastien, "PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers," *BMC Med. Genomics*, vol. 5, no. 1, pp. 1–12, Dec. 2012.

[11] J. J. Gao and S. M. Swain, "Luminal a breast cancer and molecular assays: A review," *Oncologist*, vol. 23, no. 5, pp. 556–565, May 2018.

[12] P. Whitworth, L. Stork-Sloots, F. A. de Snoo, P. Richards, M. Rotkis, J. Beatty, A. Mislowsky, J. V. Pellicane, B. Nguyen, L. Lee, C. Nash, M. Gittleman, S. Akbari, and P. D. Beitsch, "Chemosensitivity predicted by BluePrint 80-gene functional subtype and MammaPrint in the prospective neoadjuvant breast registry symphony trial (NBRST)," *Ann. Surgical Oncol.*, vol. 21, no. 10, pp. 3261–3267, Oct. 2014.

[13] J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, and Z. Hu, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J. Clin. Oncol.*, vol. 27, no. 8, p. 1160, 2009.

[14] H. K. Kim, K. H. Park, Y. Kim, S. E. Park, H. S. Lee, S. W. Lim, J. H. Cho, J.-Y. Kim, J. E. Lee, J. S. Ahn, Y.-H. Im, J. H. Yu, and Y. H. Park, "Discordance of the PAM50 intrinsic subtypes compared with immunohistochemistry-based surrogate in breast cancer patients: Potential implication of genomic alterations of discordance," *Cancer Res. Treatment*, vol. 51, no. 2, pp. 737–747, Apr. 2019.

[15] N. Kumar, D. Zhao, D. Bhaumik, A. Sethi, and P. H. Gann, "Quantification of intrinsic subtype ambiguity in luminal a breast cancer and its relationship to clinical outcomes," *BMC Cancer*, vol. 19, no. 1, pp. 1–14, Dec. 2019.

[16] E. H. Allott, J. Geradts, X. Sun, S. M. Cohen, G. R. Zirpoli, T. Khoury, W. Bshara, M. Chen, M. E. Sherman, J. R. Palmer, C. B. Ambrosone, A. F. Olshan, and M. A. Troester, "Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification," *Breast Cancer Res.*, vol. 18, no. 1, pp. 1–11, Dec. 2016.

[17] L. G. Martelotto, C. K. Ng, S. Piscuoglio, B. Weigelt, and J. S. Reis-Filho, "Breast cancer intra-tumor heterogeneity," *Breast Cancer Res.*, vol. 16, no. 3, pp. 1–11, Jun. 2014.

[18] P.-K. Raj-Kumar, J. Liu, J. A. Hooke, A. J. Kovatich, L. Kvecher, C. D. Shriver, and H. Hu, "PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal a tumors as luminal b," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, May 2019.

[19] D. C. Koboldt, R. Fulton, and M. McLellan, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[20] G. Ciriello, "Comprehensive molecular portraits of invasive lobular breast cancer," *Cell*, vol. 163, no. 2, pp. 506–519, 2015.

[21] C. Curtis, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, pp. 346–352, Jun. 2012.

[22] B. Pereira, "The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes," *Nature Commun.*, vol. 7, no. 1, pp. 1–16, May 2016.

[23] J. Martorell-Marugán, *Deep Learning in Omics Data Analysis and Precision Medicine*. Brisbane, QLD, Australia: Exon Publications, Oct. 2019, pp. 37–53.

[24] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff, and M. Claassen, "Automated Gleason grading of prostate cancer tissue microarrays via deep learning," *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, Aug. 2018.

[25] B. E. Bejnordi, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *J. Amer. Med. Assoc.*, vol. 318, no. 22, pp. 2199–2210, Dec. 2017.

[26] N. Coudray, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nat. Med.*, vol. 24, no. 10, pp. 1559–1567, Sep. 2018.

[27] J. Saltz, "Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images," *Cell Rep.*, vol. 23, no. 1, pp. 181–193, 2018.

[28] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Med. Res. Methodol.*, vol. 18, no. 1, pp. 1–12, Dec. 2018.

[29] T. Ching, X. Zhu, and L. X. Garmire, "Cox-nNet: An artificial neural network method for prognosis prediction of high-throughput omics data," *PLOS Comput. Biol.*, vol. 14, no. 4, Apr. 2018, Art. no. e1006076.

[30] G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," in *Proc. Biocomput.*, Jan. 2018, pp. 80–91.

[31] F. Gao, W. Wang, M. Tan, L. Zhu, Y. Zhang, E. Fessler, L. Vermeulen, and X. Wang, "DeepCC: A novel deep learning-based framework for cancer molecular subtype classification," *Oncogenesis*, vol. 8, no. 9, pp. 1–12, Aug. 2019.

[32] A. Beykikhoshk, T. P. Quinn, S. C. Lee, T. Tran, and S. Venkatesh, "DeepTRIAGE: Interpretable and individualised biomarker scores using attention mechanism for the classification of breast cancer sub-types," *BMC Med. Genomics*, vol. 13, no. S3, pp. 1–10, Feb. 2020.

[33] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinf.*, vol. 2015, pp. 1–13, Jun. 2015.

[34] E. Glaab, "Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification," *Briefings Bioinf.*, vol. 17, no. 3, pp. 440–452, May 2016.

[35] J. Li, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, 2017.

[36] H. Valpola, "From neural PCA to deep unsupervised learning," in *Advances in Independent Component Analysis and Learning Machines*. New York, NY, USA: Academic, 2015, pp. 143–171.

[37] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko, "Semi-supervised learning with ladder networks," 2015, *arXiv:1507.02672*.

[38] M. Pezeshki, "Deconstructing the ladder network architecture," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2368–2376.

[39] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*.

[40] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning (Adaptive Computation and Machine Learning Series)*. Cambridge, MA, USA: MIT Press, 2017, pp. 321–359.

[41] E. Cerami et al., "The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data," *Cancer Discovery*, vol. 2, no. 5, pp. 401–404, 2012, doi: 10.1158/2159-8290.CD-12-0095.

[42] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci. Signaling*, vol. 6, no. 269, p. 11, Apr. 2013.

[43] H. K. Solvang, O. C. Lingærde, A. Frigessi, A.-L. Børresen-Dale, and V. N. Kristensen, "Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer," *BMC Bioinf.*, vol. 12, no. 1, pp. 1–12, Dec. 2011.

[44] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008=.

[45] T. A. Geddes, T. Kim, L. Nan, J. G. Burchfield, J. Y. H. Yang, D. Tao, and P. Yang, "Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis," *BMC Bioinf.*, vol. 20, no. S19, pp. 1–11, Dec. 2019.

[46] F. Pedregosa, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Jul. 2017.

[47] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2017.

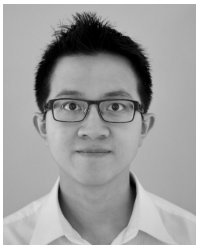
[48] X. Shao, N. Lv, J. Liao, J. Long, R. Xue, N. Ai, D. Xu, and X. Fan, ''Copy number variation is highly correlated with differential gene expression: A pan-cancer study,'' *BMC Med. Genet.*, vol. 20, no. 1, pp. 1–14, Dec. 2019.

[49] M. Zhao and Z. Zhao, ''Concordance of copy number loss and down-regulation of tumor suppressor genes: A pan-cancer study,'' *BMC Genomics*, vol. 17, no. S7, pp. 207–216, Aug. 2016.

[50] A. Paszke, ''Pytorch: An imperative style, high-performance deep learning library,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32. Vancouver, BS, Canada, Dec. 2019, pp. 8026–8037.

[51] J. Kishore, M. Goel, and P. Khanna, ''Understanding survival analysis: Kaplan-Meier estimate,'' *Int. J. Ayurveda Res.*, vol. 1, no. 4, p. 274, 2010.

[52] N. Almugren and H. Alshamlan, ''A survey on hybrid feature selection methods in microarray gene expression data for cancer classification,'' *IEEE Access*, vol. 7, pp. 78533–78548, 2019.



RICHARD LUPAT received the M.Phil. degree in machine learning from The Sir Peter MacCallum Department of Oncology, The University of Melbourne, Australia, in 2021. He is currently working as a Senior Bioinformatics Software Engineer with the Bioinformatics Core Facility, Peter MacCallum Cancer Centre, Australia. His research interests include machine learning, bioinformatics, cancer genomics, and data workflow optimization.

RASHINDRIE PERERA (Member, IEEE) is currently pursuing the Ph.D. degree with the Optimization and Pattern Recognition Group, Faculty of Engineering and IT, The University of Melbourne, Australia. She is also a Data Analyst with the Bioinformatics Core Facility, The Sir Peter MacCallum Cancer Centre, Australia. Her research interests include machine learning, deep learning, bioinformatics, and computer vision.

SHERENE LOI received the Ph.D. degree from the Institut Jules Bordet, Brussels, Belgium.

She is currently working as a Medical Oncologist specialized in breast cancer treatment and a Clinician Scientist with expertise in genomics, immunology, and drug development. She is also the holder of the Inaugural National Breast Cancer Foundation of Australia (NBCF) Endowed Chair and a Research Fellow of the Breast Cancer Research Foundation (BCRF), New York. She is recognized internationally as a leading Clinician Scientist whose work has led to new insights into the breast cancer immunology field. After completing Medical Oncology Specialist Clinical Training in Melbourne, she held a postdoctoral position at the Institut Jules Bordet. In 2013, she returned to a Group Leader position at the Peter MacCallum Cancer Centre in Melbourne and a Consultant Medical Oncologist in breast service and the Head of the Breast Cancer Clinical Trials Unit. She has published over 250 peer-reviewed research articles with a lifetime H-index of 90. Her recent work has been highly influential with 39,750 total citations and 30,730 (77%) within the past five years.

Dr. Loi is a Board Director and a member of the Scientific Advisory Committee of the Australia New Zealand Breast Cancer Trials Group (BCT Australia/New Zealand), which is the largest breast cancer clinical trials cooperative group in Australia. She also the Co-Chairs and the Scientific Executive Committee of the International Breast Cancer Study Group (IBCSG) based in Bern, Switzerland, which conducts academic global breast cancer clinical trials in over 16 countries. She is ranked in the top 1% of highly cited researchers globally by the Web of Science, since 2018.

JASON LI (Member, IEEE) received the Ph.D. degree from the Faculty of Engineering, The University of Melbourne, in 2007. He joined the Cancer Research Division of Peter Mac after his Ph.D. He was appointed as a Senior Core Facility Manager of bioinformatics, in 2017. He is currently a Senior Bioinformatician with the Peter MacCallum Cancer Centre, Australia, where he is the Head of Bioinformatics Core Facility. He has published highly-cited research papers in the area of DNA copy number analysis. His expertise lies in the analysis of large-scale genomics data derived from high-throughput sequencing/microarray experiments. His current research interests include the application of deep learning in radiology images and cancer genomics datasets.