Extracting High-Quality Features From Biomedical Datasets Using Multimodal Autoencoders *

Dmitry Kazhdan

January 15, 2019

Abstract

Biomedical datasets are typically ultra high-dimensional, often consisting of sub-datasets of different modalities, making them challenging to work with. This emphasises the importance of Multimodal Deep Learning techniques that are capable of learning more compact and informative representations of these datasets that are easier to work with in practice. This report explores one such approach by applying a Multimodal Autoencoder to the METABRIC dataset in order to extract higher-quality features from raw input data and consequently use these features for cancer subtype classification.

1 Introduction

Many healthcare-related domains have benefitted from Machine Learning (ML) techniques and strategies [3, 12]. This increasing synergy has largely been driven by a recent resurgence of progress in the field of Deep Learning (DL). DL techniques are capable of efficiently processing large, high-dimensional datasets, and are capable of automatic feature extraction, thus relieving the developers from the burden of feature engineering.

A subfield of ML that is of particular relevance to healthcare is Multimodal Machine Learning (MML). MML aims to build DL models that can process and relate information from multiple modalities (e.g. sound, image or text) by fusing them together, and has a wide range of possible applications [4, 14, 19] (a comprehensive overview is given in [1]). Many medical tasks rely on datasets that are multimodal (e.g. tomography scans and doctor's notes), and would thus benefit from MML techniques that could leverage shared modality representations. Recent work has thus focused on exploring MML techniques in the context of healthcare [24, 13].

Unfortunately, many techniques that apply ML to healthcare are either heavily reliant on expert knowledge (as is the case with many supervised approaches), or are limited to discerning only the most significant signals in the data (as is the case with many unsupervised approaches, such as clustering). In response to this issue, a range of existing work has explored applying Autoencoders (AEs) to healthcare tasks in order to learn compact, meaningful data representations that are more informative and easier to work with than the raw data [20, 14].

*Report word count: 2490

This report explores the advantages of Multimodal Autoencoders (MMAE) in the context of healthcare by applying a MMAE to the dataset provided by the Molecular Taxonomy of Breast Cancer International Consortium (referred to as *the METABRIC dataset* in the rest of this report). The contributions of this report are the following:

- Implementing a MMAE that combines Copy Number Alteration (CNA) and RNA gene expression data
- Using this MMAE as a feature extractor to extract higher-quality features from raw CNA and RNA data
- Evaluating the quality of these extracted features by performing cancer subtype classification

An overview of AE models is given in Section 2. Section 3 describes the METABRIC dataset. Section 4 and Section 5 include descriptions of the implementation and evaluation (respectively) of the MMAE model mentioned above. Possible directions for future work are discussed in Section 6. Finally, Section 7 gives some concluding remarks.

2 Autoencoders

AEs are Neural Network (NN) models that use unsupervised learning techniques for representation learning [2]. An AE typically refers to a NN architecture that contains 'bottleneck layers', forcing the NN to learn a compressed representation of the original input, learning any correlations and sub-structures of input features. For a given unlabelled dataset $\mathbf{X}$, AE learning is typically formulated as a supervised learning problem where the task is to learn an approximation of the identity function, minimising a suitably-defined *reconstruction error* loss function $\mathcal{L}(\mathbf{x}, \mathbf{x}')$, that measures the difference between the original input $\mathbf{x}$ and the AE output $\mathbf{x}'$ (which is a reconstruction of $\mathbf{x}$). Strictly speaking, AEs are *self-supervised learning techniques*, though in practice they are frequently referred to as *unsupervised* as well.

A simple AE is shown in Figure 1. The AE consists of an input layer that receives an input $\mathbf{x}$, a hidden layer that attempts to learn a compressed representation of the input, and an output layer that outputs the reconstructed value.

A wide range of different AE structures exist, offering various advantages. For instance, Denoising Autoencoders (DAEs) are capable of reconstructing initial input from corrupted data by incorporating noise during training, generating more robust features [22]. Variational Autoencoders (VAEs) learn a latent variable model of the input data by learning parameters of a probability distribution modelling the data [6]. New input data samples may thus be generated by sampling from this distribution (hence, a VAE is referred to as a *generative model*).

This work makes use of a MMAE that is capable of learning shared representations of multiple data modalities through data fusion [11, 14]. A detailed description of this model is given in Section 4.

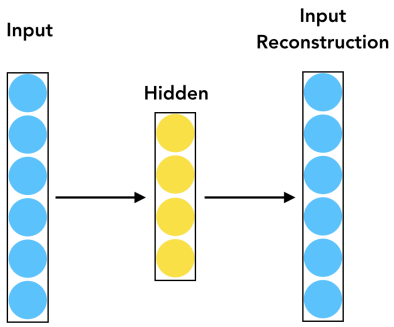

Figure 1: Simple Autoencoder Structure

3 METABRIC Dataset

The METABRIC dataset contains clinical traits, expressions, CNV profiles, and SNP genotypes derived from breast tumors collected from participants of the METABRIC trial [5]. This report relies on only a subset of this dataset, namely the RNA expression data, CNA data, and patient cancer subtypes (referred to as *IntClusts*).

The RNA gene expression data consists of expression log intensity levels of a set of genes measured for the selected patients. For the purposes of this report, it is sufficient to view this dataset as set of continuous variables (gene expressions) recorded for the patients. CNA refers to a phenomenon in which sections of a genome are repeated and the number of repeats in a genome varies between individuals in the human population. Elevated copy numbers of particular genes are frequently associated with certain types of cancer. CNA data consists of alteration levels of a set of genes measured for the selected patients. For the purposes of this report, this dataset may be viewed as a set of discrete variables (CNA levels) recorded for the patients. There are 5 different CNA categories in total, resembling the type of CNA abnormality present in the corresponding gene: homozygous deletion, hemizygous deletion, neutral/no change, gain, and high level amplification.

Further details regarding biological significance, collection and processing of this data is outside the scope of this report (for more details, see Supplementary Materials of [5]).

4 Implementation

This report explores the use of a MMAE to learn a compact, fused representation of the CNA and RNA data, that can then be used to identify patient cancer subtypes more accurately. Identifying cancer subtypes can reveal useful information regarding cancer pathogenesis, which can in turn be used for prediction of a patient's survival time, and for making a more informed decision when deciding on suitable patient therapy procedures. The following subsections describe the implemented MMAE structure, and the preprocessing steps applied to the METABRIC dataset.

4.1 Data Preprocessing

Firstly, only data for genes present in both the CNA and RNA datasets and having no missing values was selected. Secondly, only gene and IntClust data for patients present in both the CNA and RNA datasets was selected. The remaining pre-processing steps are described in the following sections.

4.1.1 RNA Preprocessing

In order to discard uninformative RNA data, the Median Absolute Deviation (MAD) score across all patients was computed for every gene, and RNA data of only the top 1200 genes was retained. The 50 genes used for PAM50 classification [16] (proven to be related to subtype diagnosis), were also added to this list of selected genes. Once these genes were extracted, their values were normalized to lie in the range $[-1, 1]$ (the $[0, 1]$ range was tried as well, but gave worse performance in practice).

4.1.2 CNA Preprocessing

In order to discard uninformative CNA variables, the entropy score across all patients was computed for every gene from the pre-selected RNA genes described in the previous section, and CNA data of only the top 300 genes was retained. Categories 'homozygous deletion' and 'hemizygous deletion' were merged into a single category 'low', and 'gain' and 'high level amplification' were merged into a single category 'high'. This made the CNA data more balanced whilst retaining key information. Finally, this dataset was converted to a one-hot encoding (thus having three features per original CNA variable).

4.1.3 IntClust Preprocessing

In order to make the IntClust patient groups more balanced, patients belonging to IntClust 2 and 6 were removed altogether (as there were considerably fewer patients with these IntClusts), and only 200 patients were randomly selected from IntClust groups that had over 200 patients. This produced much more balanced IntClust groups, with the largest group containing 200 patients, and the smallest group containing 132 patients.

4.2 Multimodal Autoencoder Architecture

The MMAE architecture used in this report was based on the MMAE architectures described in [11, 18], and is shown in Figure 2.

The first set of hidden layers allows the model to learn useful features for the separate modalities before they are combined. The intermediate layer combines these modalities, learning a joint representation, leveraging higher-order correlations across modalities. The output layers are then used to map the representation back to the original data.

Similarly to other existing work on MMAEs, all layers were dense, and applied sigmoid activation functions to their input. The sizes of the hidden layers were 800 for both the RNA and CNA data. The size of the combined layer was 1600. Initially, activity regularisers were employed in the hidden layers in an attempt to regularise hidden units. However, they did not

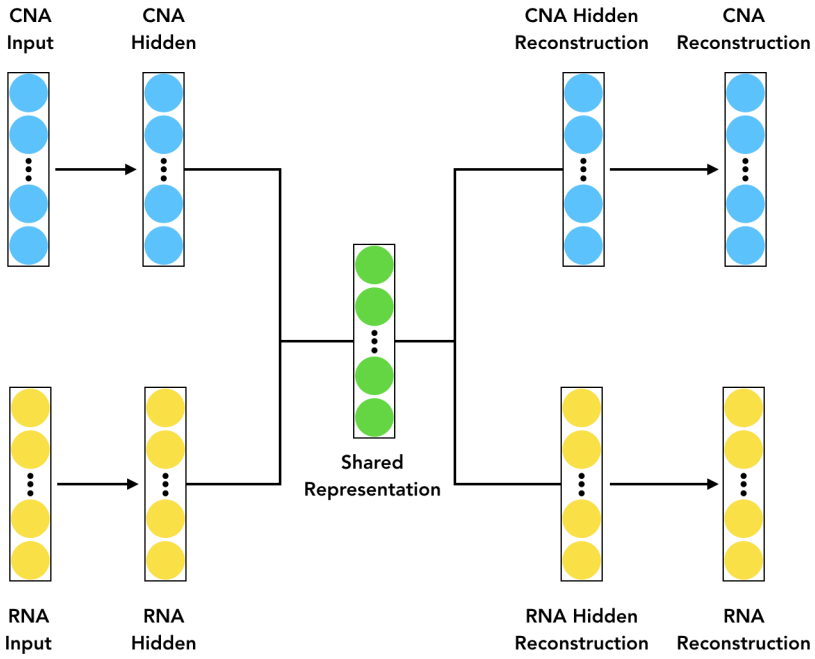

Figure 2: Multimodal Autoencoder structure

have an effect on overall performance and were thus consequently removed.

The MMAE was trained in a layer-wise fashion by training the layers in turn, as described in [18]. At every layer learning stage, parameter optimisation was performed by evaluating the trained model on a subset of the data using different optimizers (Adam and Stochastic Gradient Descent (SGD)), different loss functions (binary cross-entropy, categorical cross-entropy and mean squared error (MSE)), different numbers of epochs (50, 100, 200, 500) and different batch sizes (16, 32, 64, 128). This ensured that maximum performance was achieved by each individual layer.

5 Evaluation

The MMAE was evaluated by investigating the quality of the learned features. The preprocessed sample set (consisting of patient CNA and RNA data, labelled with the patient's IntClust) was randomly split into a training set (80% of samples) and test set (20% of samples). The MMAE was trained on the training set (as described in the previous section), and then used as a feature extractor on the test set (taking the features computed by the combined layer). The quality of extracted features was judged by comparing IntClust classification accuracies of a classifier trained and tested using either the extracted features, or the original input data.

The above procedure was run using the multiclass AdaBoost and the Gradient Tree Boosting (GTB) classifiers from Python's *sklearn* API [17] (both classifiers had *n_estimators* set to 100, and used default values for the other parameters). The multiclass classification AdaBoost al-

gorithm is a direct extension of the binary AdaBoost algorithm to the multi-class case without reduction to multiple two-class problems (for more details, see [25]). Like AdaBoost, GTB combines weak learners into a single strong learner in an iterative fashion, but also allows optimization of an arbitrary differentiable loss function (further details can be found in [8]). GTB is generally more robust and was shown to give better performance on biomedical datasets compared to AdaBoost [15].

All experimental setups described below report results (mean ± standard deviation) averaged over 15 independent re-runs.

5.1 Complete Data

Table 1 below shows average classification accuracy of the two classifiers on the sample sets obtained using the procedure described in Section 4.1.

Input Data	**AdaBoost**	**GTB**
MMAE Features	**46.5 ± 1.9 %**	67.4 ± 1.6 %
Preprocessed RNA data	40.4 ± 5.5 %	70.0 ± 2.6 %
Preprocessed CNA data	42.1 ± 7.7 %	64.1 ± 1.6 %
Preprocessed RNA and CNA data concatenated together	41.8 ± 7.6 %	**74.3 ± 1.5 %**

Table 1: Comparison using complete data

AdaBoost benefitted from the learned MMAE representation, with MMAE features achieving the highest average accuracy and lowest variance from the different types of input data. With GTB, the MMAE feature accuracy score ranked between the two individual modality scores, showing that the MMAE was likely able to slightly improve data from one modality using the other one. However, GTB accuracy was highest when using the concatenation of modalities.

5.2 Missing RNA Data

Table 2 below shows average classification accuracy of the two classifiers on the sample sets obtained using the procedure described in Section 4.1, in which a random 10% of RNA expression values were set to 0 for every patient in the test set (resembling missing values).

Input Data	**AdaBoost**	**GTB**
MMAE Features	**43.7 ± 3.2 %**	65.5 ± 2.8 %
Preprocessed RNA Data	33.1 ± 5.4 %	66.6 ± 3.2 %
Preprocessed CNA Data	39.4 ± 7.3 %	63.9 ± 2.7 %
Preprocessed RNA and CNA data concatenated together	34.4 ± 6.0 %	**72.6 ± 1.8 %**

Table 2: Comparison using missing RNA data

As before, AdaBoost achieved the highest accuracy when using the MMAE features. Crucially, MMAE accuracy was less affected by the missing data, compared to the concatenated representation, or the RNA-only representation. This demonstrates that the MMAE was able to partially induce the missing information by exploiting modality correlations. As before, the GTB classifier did not benefit from the MMAE feature representation. The MMAE feature accuracy score ranked between the two modality scores, with the concatenated modalities achieving highest accuracy.

5.3 Random Genes

Table 3 shows the classification accuracies obtained when random selection was used instead of MAD and entropy sorting during gene data selection.

Input Data	**AdaBoost**	**GTB**
MMAE Features	38.4 ± 3.7	56.8 ± 2.0 %
Preprocessed RNA Data	39.4 ± 5.7 %	71.8 ± 1.8 %
Preprocessed CNA Data	47.4 ± 5.2 %	60.2 ± 2.3 %
Preprocessed RNA and CNA data concatenated together	**46.0 ± 8.3 %**	**73.3 ± 2.0 %**

Table 3: Comparison using random genes

For both classifiers, performance of MMAE features was worse than using either or both of the modalities. This is likely because the samples had lower variance, compared to samples used in Section 5.1, meaning that any indicative information signals were much less pronounced. The MMAE was thus unable to learn these slight variations from the training data, and likely lost some of this crucial information during feature extraction (this was further exacerbated by the small size of the training set). Consequently, the classifiers were able to select informative features from the raw data, but were unable to recover information from the fused features. Furthermore, AdaBoost achieved a higher accuracy in this case than in Section 5.1 when using concatenated features, because it was easier for AdaBoost to partition the data more accurately when the features had lower variance (AdaBoost was less thrown off by variations in test sample features). Thus relying on raw input data was more beneficial in this case.

6 Future Work

One factor likely limiting overall model performance was the simplistic handling of CNA data (apart from initial pre-processing steps, it was handled in the same way as continuous data). Recent work has explored variants of multimodal VAEs for handling discrete AE input [23, 10, 9], however, these techniques are considerably more complex. Thus, incorporating these techniques into the MMAE model was outside the scope of this project, but may be explored in future work.

Existing work [20] showed that post-hoc interpretation of learned features may lead to novel insights about the original dataset and the types of interconnections learned by the AE. Thus, future work may focus on investigating the behaviour of individual nodes in the combined layer of the MMAE in an attempt to interpret the learnt features, potentially revealing useful information. More importantly, applying approaches described in [5] to the fused representation could potentially yield novel, more detailed cancer subtypes.

Finally, future work may explore applications of the above approach to richer datasets, having more modalities (e.g. cancer datasets found in [7]). Apart from further demonstrating the wide applicability of the approach described in this report, larger datasets with more modalities allow the MMAE representation to be utilised in a wider variety of ways, such as cross-modal learning [14, 21] (e.g. predicting one modality whilst training on another). Alternatively, future work may explore applying the learnt MMAE feature representation to other types of classification/regression tasks, such as predicting the patient survival rate.

7 Conclusions

To conclude, this report explored the usage of MMAEs for fusing biomedical data of different modalities. Extracted features produced by the MMAE successfully improved overall cancer subtype classification accuracy of the AdaBoost classifier, provided the data was varied enough, thus demonstrating the utility of shared representation learning. In case of the GTB classifier, however, the learnt MMAE features did not yield any benefits. Nevertheless, improving performance of simpler classifiers, such as AdaBoost, increases their usability, which is extremely important in safety-critical domains, such as healthcare, which often require ML models to be simple in order to be reliable and interpretable. Furthermore, numerous extensions to the MMAE architecture have been proposed, which could potentially make the MMAE representation useful in an even wider range of use-case scenarios (e.g. with more powerful classifiers, or with noisier data).

References

[1] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal Machine Learning: A Survey and Taxonomy". In: *CoRR* abs/1705.09406 (2017). arXiv: 1705.09406. URL: http://arxiv.org/abs/1705.09406.

[2] Y. Bengio, A. Courville, and P. Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (Aug. 2013), pp. 1798–1828. ISSN: 2160-9292. DOI: 10.1109/tpami.2013.50. URL: http://dx.doi.org/10.1109/TPAMI.2013.50.

[3] R. Bhardwaj, A. R. Nambiar, and D. Dutta. "A Study of Machine Learning in Healthcare". In: *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 2. July 2017, pp. 236–241. DOI: 10.1109/COMPSAC.2017.164.

[4] Catalina Cangea, Petar Velickovic, and Pietro Lio. "XFlow: 1D-2D Cross-modal Deep Neural Networks for Audiovisual Classification". In: *arXiv preprint arXiv:1709.00572* (2017).

[5] Christina Curtis et al. "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups". In: *Nature* 486 (Apr. 2012).

[6] Carl Doersch. "Tutorial on variational autoencoders". In: *arXiv preprint arXiv:1606.05908* (2016).

[7] *Genomic Data Commons Data Portal*. URL: https://portal.gdc.cancer.gov/.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2008.

[9] Eric Jang. *Tutorial: Categorical Variational Autoencoders using Gumbel-Softmax*. 2016. URL: https://blog.evjang.com/2016/11/tutorial-categorical-variational.html.

[10] Eric Jang, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax". In: *arXiv preprint arXiv:1611.01144* (2016).

[11] N. Jaques et al. "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction". In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. Oct. 2017, pp. 202–208. DOI: 10.1109/ACII.2017.8273601.

[12] Igor Kononenko. "Machine learning for medical diagnosis: history, state of the art and perspective". In: *Artificial Intelligence in Medicine* 23.1 (2001), pp. 89–109. ISSN: 0933-3657. DOI: https://doi.org/10.1016/S0933-3657(01)00077-X. URL: http://www.sciencedirect.com/science/article/pii/S093336570100077X.

[13] Donghuan Lu et al. "Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images". In: *Scientific reports* 8.1 (Apr. 2018), pp. 5697, 5697–5697.

[14] Jiquan Ngiam et al. "Multimodal Deep Learning". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, pp. 689–696. ISBN: 978-1-4503-0619-5. URL: http://dl.acm.org/citation.cfm?id=3104482.3104569.

[15] Randal S Olson et al. "Data-driven advice for applying machine learning to bioinformatics problems". In: *arXiv preprint arXiv:1708.05070* (2017).

[16] Joel S Parker et al. "Supervised risk predictor of breast cancer based on intrinsic subtypes". In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 27.8 (Mar. 2009), pp. 1160–1167.

[17] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[18] Patrick Poirson. *Multimodal Stacked Denoising Autoencoders*. 2013.

[19] Nitish Srivastava and Ruslan R Salakhutdinov. "Multimodal Learning with Deep Boltzmann Machines". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 2222–2230. URL: http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines.pdf.

[20] Jie Tan et al. “Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders”. In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 20 (2015), pp. 132–143.

[21] Petar Veličković et al. “X-CNN: Cross-modal convolutional neural networks for sparse datasets”. In: *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*. IEEE. 2016, pp. 1–8.

[22] Pascal Vincent et al. “Extracting and Composing Robust Features with Denoising Autoencoders”. In: *Proceedings of the 25th International Conference on Machine Learning*. ICML ’08. Helsinki, Finland: ACM, 2008, pp. 1096–1103. ISBN: 978-1-60558-205-4. DOI: 10.1145/1390156.1390294. URL: http://doi.acm.org/10.1145/1390156.1390294.

[23] Mike Wu and Noah Goodman. “Multimodal Generative Models for Scalable Weakly-Supervised Learning”. In: *CoRR* abs/1802.05335 (2018). arXiv: 1802.05335. URL: http://arxiv.org/abs/1802.05335.

[24] Yong Xia et al. “Machine Learning in Multimodal Medical Imaging”. In: *BioMed research international* 2017 (2017), pp. 1278329, 1278329–1278329. DOI: 10.1155/2017/1278329. URL: https://www.ncbi.nlm.nih.gov/pubmed/28357398.

[25] Ji Zhu et al. “Multi-class AdaBoost”. In: *Statistics and its interface* 2 (Feb. 2006). DOI: 10.4310/SII.2009.v2.n3.a8.