Bachelor Thesis Project (BTP) Report

Naresh Kumar Kaushal

Roll Number: 170030027
Month of Submission: May 2021

Supervisor: **Dr. Clint P. George**, Computer Science and Engineering

This report is submitted towards partial fulfillment of the requirements for the award of the Bachelor of Technology (B.Tech) degree in Computer Science and Engineering at the Indian Institute of Technology Goa.

Feature Selection Using Deep Neural Networks in Bioinformatics

Naresh Kumar Kaushal

May 19, 2021

DNN (Deep Neural Networks) has achieved a recent breakthrough in various applications ranging from NLP to RNN. But all these applications require sufficient number of samples for training which renders DNN approach useless in HDLSS (High Dimension Low Sample Size settings) such as cancer sub type classification from gene expression values in Bioinformatics. Simple DNN architecture suffers from over fitting and high variance of gradients which makes learning process difficult and our model fails to generalise. In this paper I implement DNN specifically tailored for HDLSS settings to classify Breast, lung and Prostate Cancer sub types using gene expression data. This method was first proposed by Bo Liu, Ying Wei, Yu Zhang, Qiang Yang (Hong Kong University of Science and Technology).[1] Using this approach we can tackle the problem of over fitting and after training we can make predictions using only k features where k « total dimensions. For analysis purpose we will compare DNP algorithm with simple FNN and PCA.

[1] Liran Shen, Meng Joo Er, and Qingbo Yin. The classification for high-dimension low-sample size data, 2021. URL https://arxiv.org/abs/2006.13018

Introduction and Overview

The growing concern for Cancer and its complications has urged the world scientists to come up with a machine learning algorithm which can predict the possibility of having cancer with high probability and to also reduce the False negatives. Breast cancer is the most pervasive cancer and statistics have shown that in India (94.1 per 100,000)males and (103.6 per 100,000)females are affected with breast cancer for the year 2020. One in 68 males (lung cancer), 1 in 29 females (breast cancer), and 1 in 9 Indians will develop cancer during their lifetime.[2] These number shows the scale of this disease. Early detection of cancer can save the patients life but the incubation periods of these cancers are 2 to 5 years so by the time you detect cancerous lump the situation is already out of control. In order to help medical personnel in cancer detection various machine learning algorithms are proposed but most of them suffer from HDLSS since we do not have enough medical data to train our models. DNP algorithm is one such approach which can prove to be useful even when sample size is very small compared to dimensions. In this paper we will see how DNP can help in predicting cancer sub types from gene expression data without letting our model to over fit.

[2] Prashant Mathur, Krishnan Sathishkumar, Meesha Chaturvedi, Priyanka Das, Kondalli Lakshminarayana Sudarshan, Stephen Santhappan, Vinodh Nallasamy, Anish John, Sandeep Narasimhan, and Francis Selvaraj and Roselind. Cancer statistics, 2020: Report from national cancer registry programme, india. *JCO Global Oncology*, (6):1063–1075, 2020. DOI: 10.1200/GO.20.00122. URL https://doi.org/10.1200/GO.20.00122. PMID: 32673076

Related Work

Various machine learning algorithms are proposed which can help in classifying data suffering from HDLSS but the most fundamental one is PCA (Principal Component Analysis) used by NguYen [3] which can help in dimensionality reduction. But PCA is strongly discouraged by Keith E Muller et al [4]. Another dominant method is to use sparsity-inducing regularizer like lasso but it only considers linear input output dependency and ignores non linearity which restricts its usage on a wider scale.

[3] Danh V. Nguyen and David M. Rocke. Tumor classification by partial least squares using microarray gene expression data . *Bioinformatics*, 18 (1):39–50, 01 2002. ISSN 1367-4803. DOI: 10.1093/bioinformatics/18.1.39. URL https://doi.org/10.1093/bioinformatics/18.1.39

[4] Keith Muller, Yueh-Yun Chi, Jeongyoun Ahn, and J Marron. Limitations of high dimension, low sample size principal components for gaussian data. 03 2008

Methodology

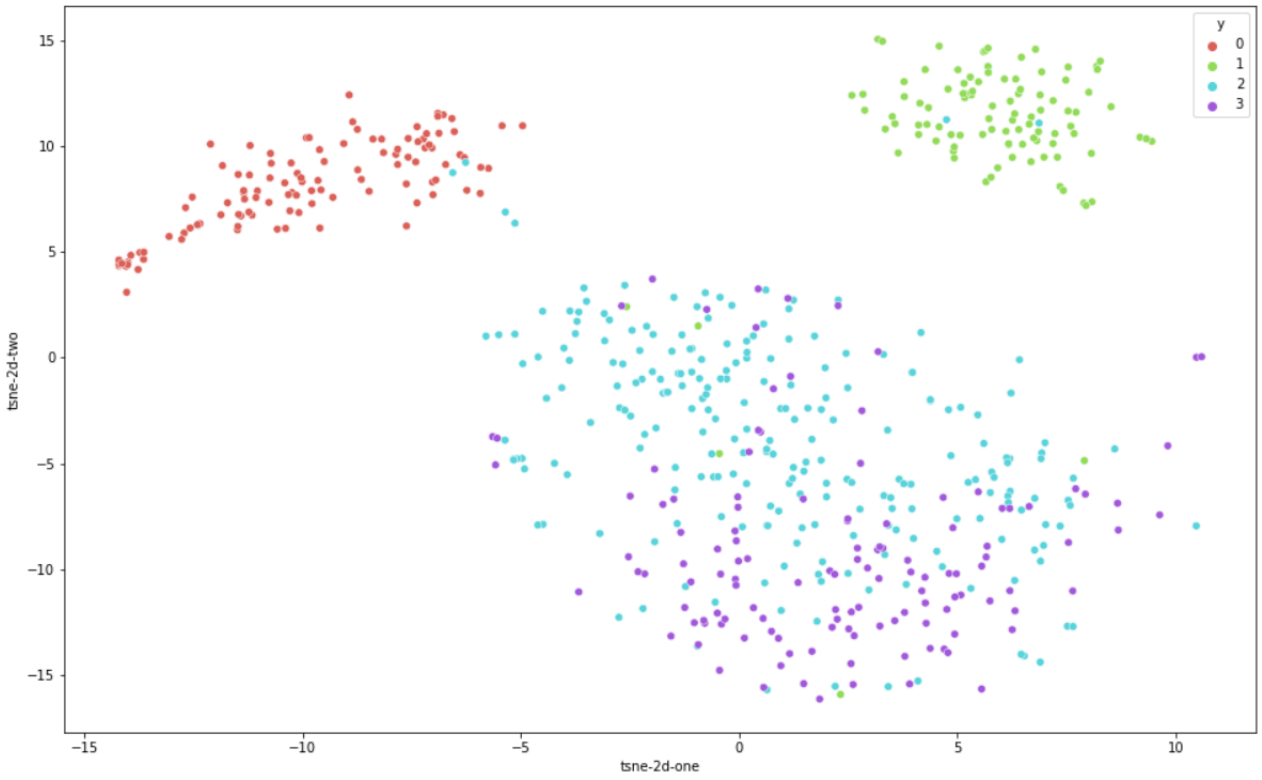


Figure 1: t-SNE plot for breast cancer dataset

1. **Dataset** - We chose 3 biological datasets (Prostate-GE, ALLAML and Lung) provided on this *link* and one breast cancer dataset from *link* All these datasets suffer from HDLSS. Few details of these data sets are - Breast(4 subtypes) 574 samples and 1519 features, Lung Cancer(5 subtypes) with 203 samples and 3312 features, Prostate Cancer(2 subtypes) with 102 samples and 5966 features, Leukemia

Cancer(2 subtypes) with 72 samples and 7129 features. These datasets are already preprocessed and for better prediction we normalised (mean = 0, std = 1) gene expression data. Every dataset has 2 tables one for gene expression values and other for subtype information. After downloading we split our datasets with 80% training 10% testing and 10% validation.

Subtype Cardinality					
Dataset	subtype_1	subtype_2	subtype_3	subtype_4	subtype_5
Breast Cancer	113	101	233	127	-
Prostate Cancer	50	52	-	-	-
Lung Cancer	139	17	21	20	6
Leukemia Cancer	47	25	-	-	-

Here special attention needs to be given to Leukemia and Lung Cancer datasets. Looking at their subtype cardinality we can infer that they both suffer from class imbalance problem. If we train our dataset using this distribution our model will train but accuracy will be a mere illusion as one class with higher subtype cardinality will dominate and predictions of test set will incline towards this particular class. So to remove this imbalance I used the technique called **SMOTE** [5](Synthetic Minority Oversampling Technique) which produces synthetic data using interpolation.

[5] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. Smote: Synthetic minority oversampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 06 2002. DOI: 10.1613/jair.953

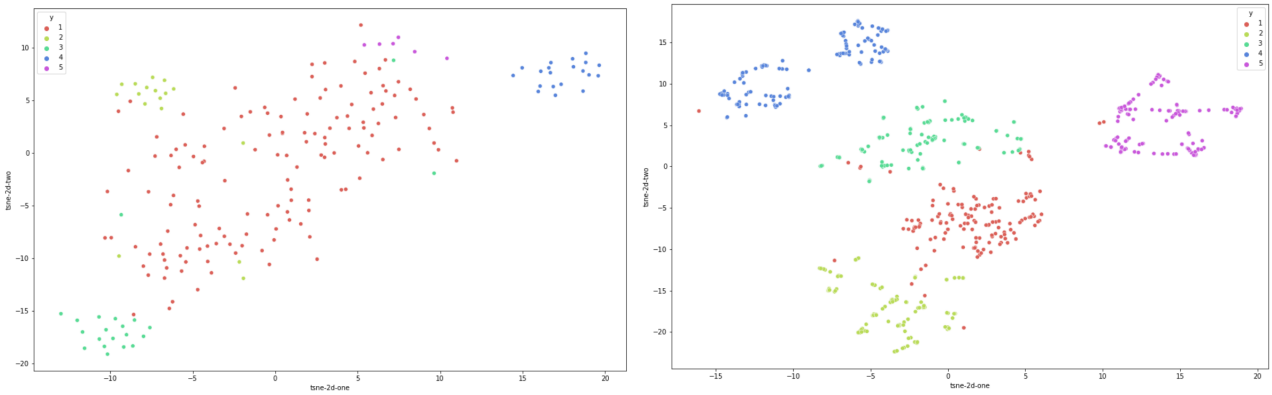


Figure 2: t-SNE plot of Lung Cancer (1) before applying SMOTE (2) After applying SMOTE.

Model Summary - The key challenges faced by HDLSS are :

- Risk of **Overfitting** (**High dimensions**) and risk of **High variance** gradients (**low sample size**).

- To counter Overfitting we will use feature selection and will optimize our loss function in greedy and incremental manner.
- To counter High Variance we will use multiple dropouts technique in our Neural Net.

DNP algorithm has 4 phases:

- **Train** phase - Train the subnetwork and learn the weights.
- **Select** phase - Dropout multiple times to select the new Feature based on norm of gradients.
- **Update** phase - Update the learning rates according to Adagrad.
- **Transfer** phase - Transfer the newly selected feature from the C set to S set.

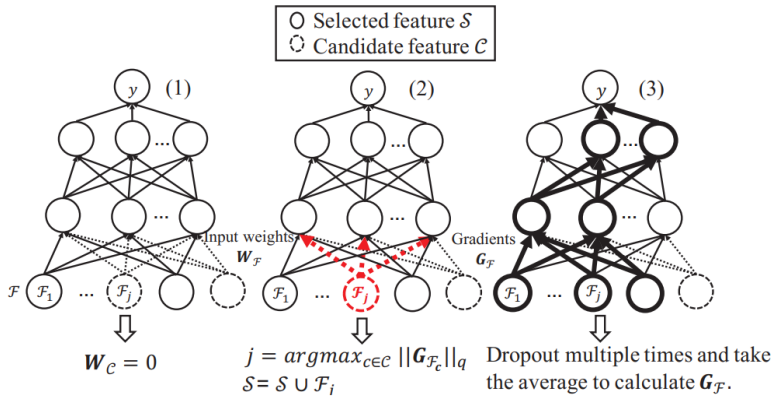


Figure 3: 1. The selected features and the corresponding subnetwork. 2. The selection of a single feature. 3. calculate gradients with lower variance via multiple dropouts.

(a) **Trains Phase** :

- This phase begins with 2 sets S and C.
- **S** set contains the features which we have already selected and that will be used in training the neural network.
- **C** set is the candidate set and from this set we will select the features in **SELECTION** Phase and transfer the newly selected feature from C set to S set in **Train** Phase.
- Now we will train the sub network until convergence and update the weights.

Selection Phase :

- Iterate over all the features in the C set and after multiple dropouts compute average of the gradients for this current feature.
- Finally select the feature with the maximum norm of its gradients.

Update Phase :

- Update the learning rates using Adagrad.
- For each weight in the neural network its learning rate decreases with the sum of gradients in all the past iterations.
- By this approach the newly selected feature will have more impact on the learning and on the same time other selected weights will also have non zero contribution.
- This can solve the problem of vanishing gradients during learning and training in Deep Neural Networks.

Transfer Phase :

- The newly selected feature in the **SELECT** Phase will be added to the S set and deleted from the C set.
- Features in the subnetwork will be increased by one and we will initialise the wights coming from this new feature using Xavier initialization (Normal distribution -1 to 1).
- Weights for the previous features and hidden nodes must be the weights we got after training in the previous **TRAIN** Phase.

Experimental Analysis

For experimental analysis I have compared my algorithm with Principal Component Analysis(PCA). I am using 3 important metrics accuracy, f1_score and auc_roc score. The formula for the standard F1-score is the harmonic mean of the precision and recall. A perfect model has an F-score of 1. [6]

$$2 * \frac{precision * recall}{precision + recall}$$

AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model.

[6] Meysam Vakili, Mohammad Ghamsari, and Masoumeh Rezaei. Performance analysis and comparison of machine and deep learning algorithms for iot data classification, 01 2020

Deep Neural Pursuit			
Dataset	accuracy	auc score	f1_score
Breast Cancer	77.6	90.8	77.0
Prostate Cancer	81.8	88.4	81.8
Lung Cancer	87.14	98.49	80.75
Leukemia Cancer	85.45	83.33	87.91
Principal Component Analysis with Classification Trees			
Dataset	accuracy	auc score	f1_score
Breast Cancer	79.95	85.32	79.92
Prostate Cancer	80.36	80.25	79.14
Lung Cancer	88.67	87.50	87.36
Leukemia Cancer	81.96	83.02	84.47

Here accuracy, auc_score and f1_score values of PCA are averaged over K-folds see last page for more info.

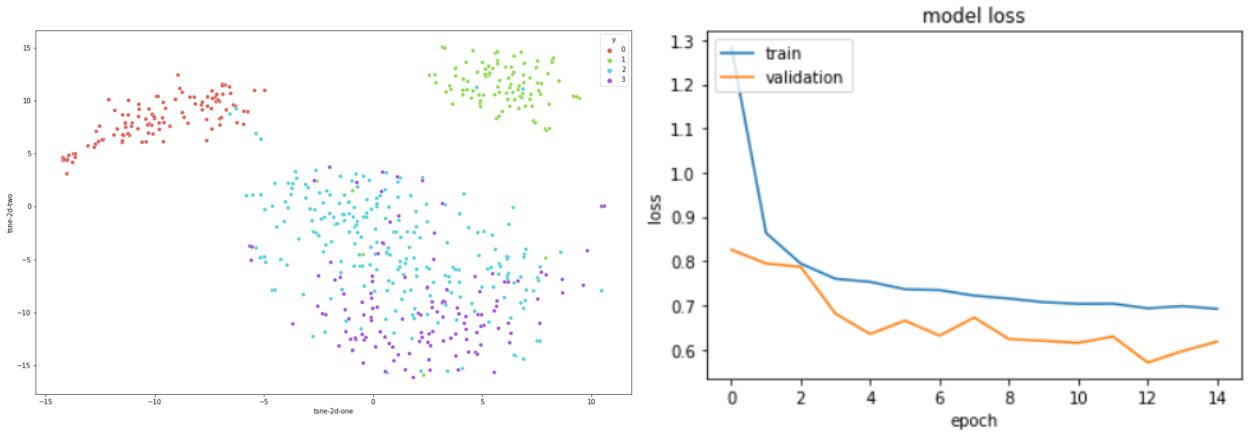


Figure 4: t-SNE plot and loss function for breast cancer.

From the t-SNE plot we can see that the class 2 (Luminal A) and the class 3 (Luminal B) are hard to classify compared to class 1 (Basal like) and class 0 (No Cancer). Validation loss is lower than training loss which signifies that I might be training my model harder than what is required or my model is too complex for my dataset. So I can reduce the complexity by reducing number of hidden layers or hidden neurons to make 2 loss functions overlap.

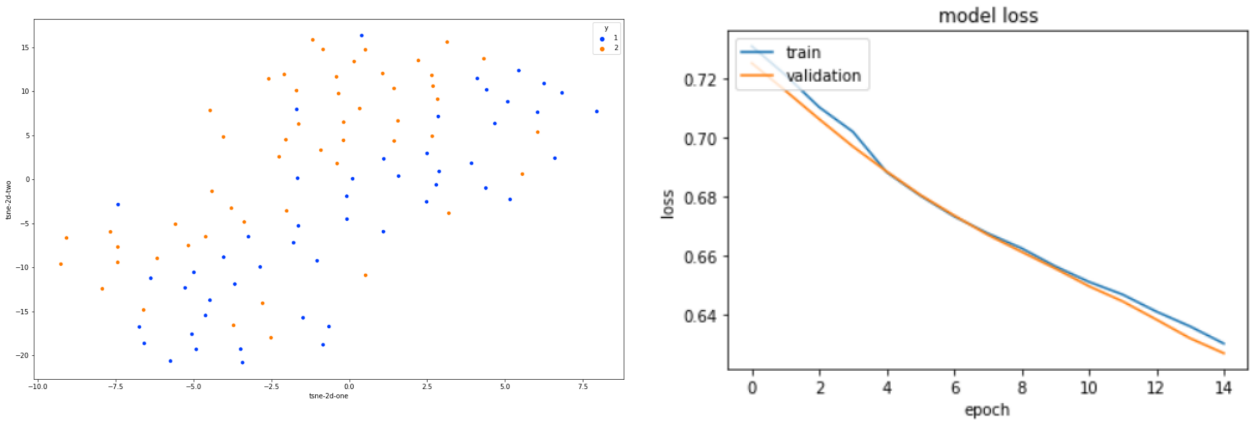


Figure 5: t-SNE plot and loss function for prostate cancer.

From the t-SNE plot we can see that 2 classes are almost distinguishable hence it is easier to get a linear seperable line. We can also see that loss function has not plateaued yet and it can decrease more which

can be solved by increasing the epoch numbers. This represent under fitting but it can be easily removed by increasing number of epochs. I haven't increased the number of epochs here because the dataset is huge and it will take significantly longer to converge.

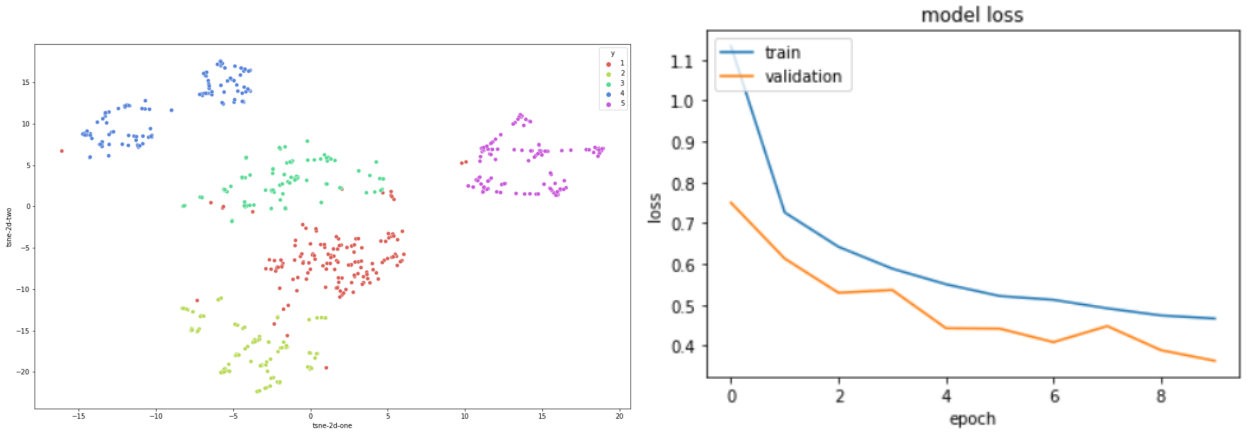


Figure 6: t-SNE plot and loss function for lung cancer.

As we can see from the t-SNE plot that all the classes are properly distinguishable with enough data points but this is possible only after applying SMOTE to our training set. SMOTE removes class imbalance problem which helps my algorithm to generalise better. A similar gap between 2 loss functions can be observed here which can be eliminated by relaxing my model a bit.

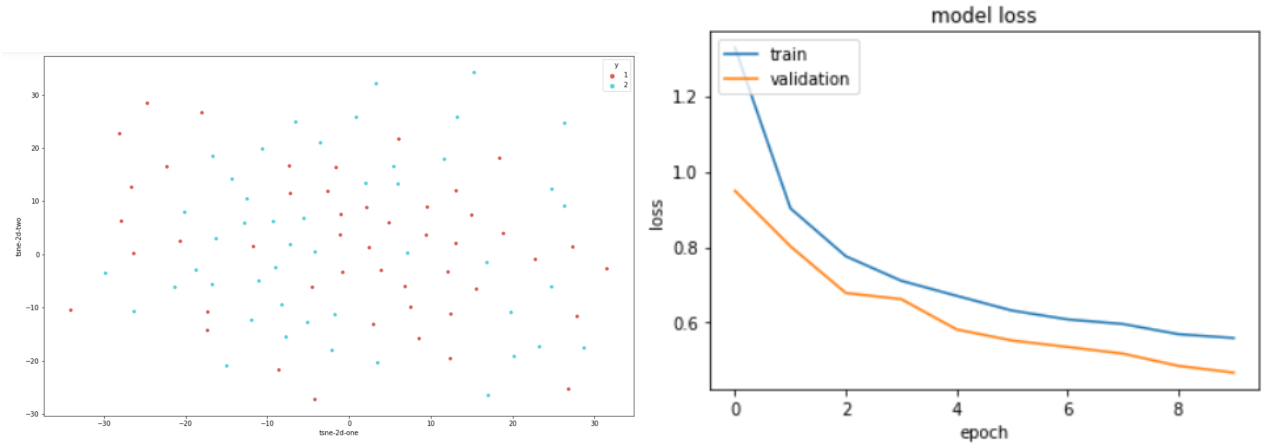


Figure 7: t-SNE plot and loss function for leukemia cancer.

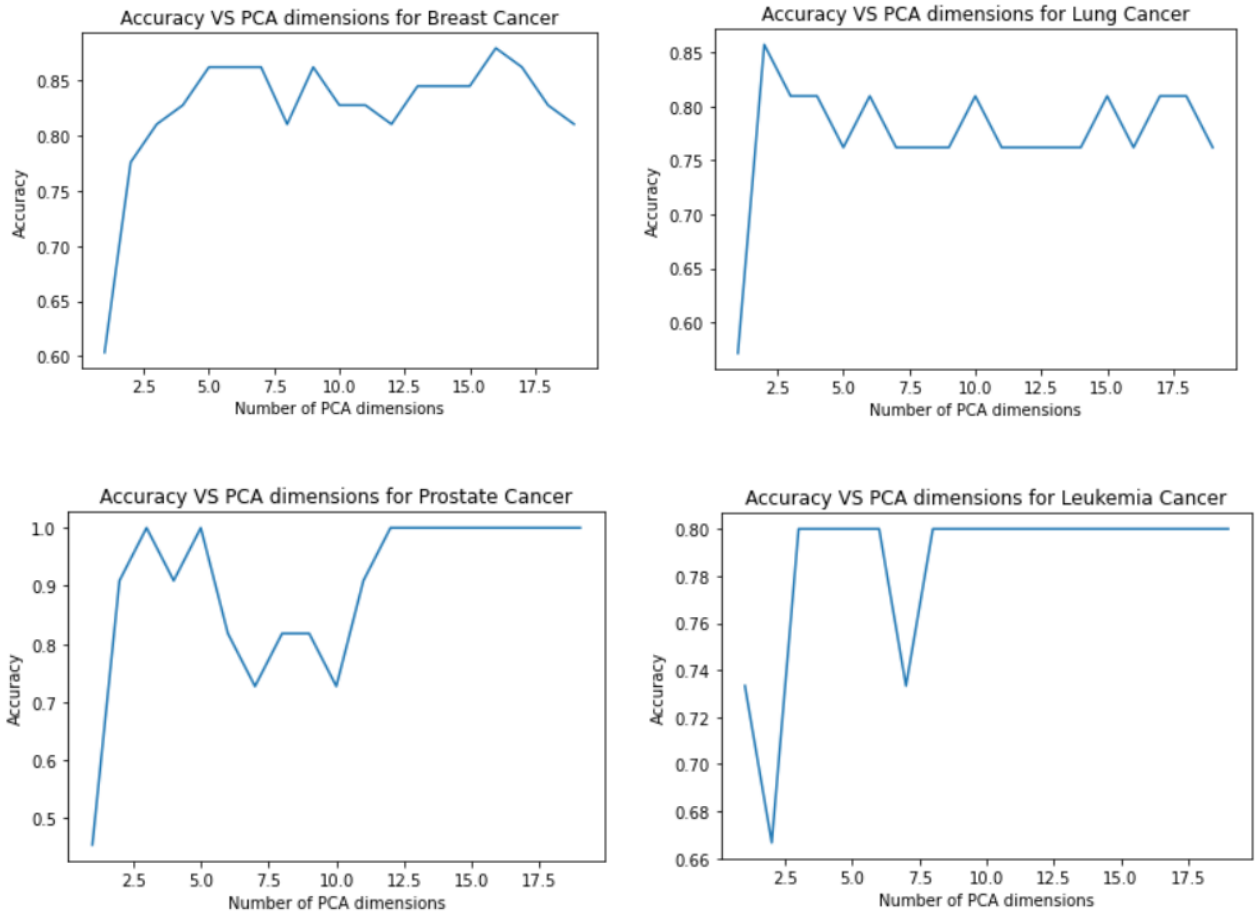
The 2 classes are difficult to distinguish by just looking at the graph but DNP algorithm learns to generalise. The loss functions are smooth

and there is no overfitting observed. This dataset also suffers from class imbalance problem hence SMOTE comes as a rescue. It makes cardinality of 2 classes equal in train set.

Great Thing about DNP - All these datasets reached to this higher level of accuracy by using only 10 features slected by DNP while some can be stopped even earlier with similar higher accuracies. Secondly DNP proves to be quite stable algorithm with accuracies being nearly similar for every train/test split of the dataset.

PCA (Principal Component Analysis)

I also observed that PCA attains high level of accuracy with very few principal components i.e. Lung 2, Prostate 3, Leukemia 3 and Breast 17. Below graphs shows Accuracy vs number of principal components of Breast, Lung, Prostate and Leukemia Cancer. Using these graphs we calculated minimum number of principal components required to get the maximum accuracy.



But I also observed that PCA is not stable in HDLSS setting since the accuracy fluctuates a lot with different splits of train and test sets. But DNP prooved to be stable because of multiple dropouts and average gradients technique while selecting new features.

PCA with Classification Trees										
Dataset	fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
Breast Cancer	87.93	82.75	81.03	72.41	77.19	78.94	77.19	78.94	82.45	80.7
Prostate Cancer	90.9	63.63	60.0	90.0	90.0	70.0	100.0	80.0	70.0	60.0
Lung Cancer	85.29	91.17	89.55	-	-	-	-	-	-	-
Leukemia Cancer	75.0	87.5	100.0	85.71	85.71	71.42	57.14	85.71	85.71	85.71

This table shows unstability of PCA on HDLSS data.

Simple Deep Neural Network

Another comparison is made with simple deep Neural Network but here 3 out of 4 datasets suffered from severe over fitting and predictions were not quite stable for different test/train split which shows high variance of this model.

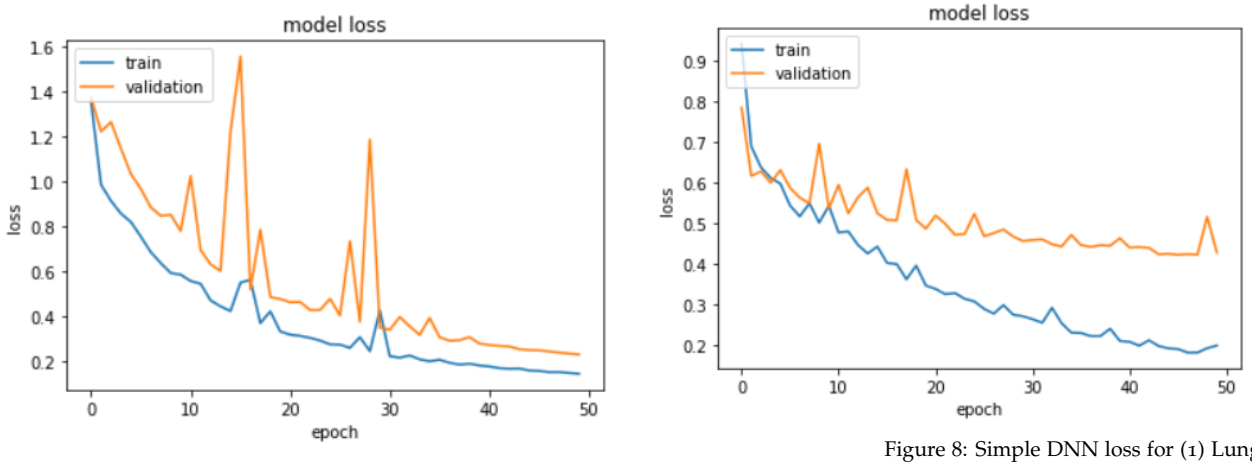


Figure 8: Simple DNN loss for (1) Lung and (2) Prostate Cancer .

Discussion

The DNP (Deep Neural Pursuit) algorithm is promising but requires tweaking based on the dataset because number of hidden layers, number of hidden nodes, learning rate, split ratio and other hyper parameters are all dependent on the type of dataset. DNP takes huge amount

of time as compared to PCA or simple feed forward network and the bottleneck is **Selection** phase. We can improve this phase by first reducing dimensions using PCA or some other dimensionality reduction technique and then selecting features based on DNP but this method needs to be verified for its stability. I have explored 4 types of cancers with significant variability and now next step will be to analyse the hyper parameters and model architecture properly so as to get a generalised DNP model and to reduce its time complexity.

Supporting information

The code for this problem statement is available on github link

References

Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 06 2002. DOI: 10.1613/jair.953.

Prashant Mathur, Krishnan Sathishkumar, Meesha Chaturvedi, Priyanka Das, Kondalli Lakshminarayana Sudarshan, Stephen Santhappan, Vinodh Nallasamy, Anish John, Sandeep Narasimhan, and Francis Selvaraj and Roselind. Cancer statistics, 2020: Report from national cancer registry programme, india. *JCO Global Oncology*, (6):1063–1075, 2020. DOI: 10.1200/GO.20.00122. URL https://doi.org/10.1200/GO.20.00122. PMID: 32673076.

Keith Muller, Yueh-Yun Chi, Jeongyoun Ahn, and J Marron. Limitations of high dimension, low sample size principal components for gaussian data. 03 2008.

Danh V. Nguyen and David M. Rocke. Tumor classification by partial least squares using microarray gene expression data . *Bioinformatics*, 18(1):39–50, 01 2002. ISSN 1367-4803. DOI: 10.1093/bioinformatics/18.1.39. URL https://doi.org/10.1093/bioinformatics/18.1.39.

Liran Shen, Meng Joo Er, and Qingbo Yin. The classification for high-dimension low-sample size data, 2021. URL https://arxiv.org/abs/2006.13018.

Meysam Vakili, Mohammad Ghamsari, and Masoumeh Rezaei. Performance analysis and comparison of machine and deep learning algorithms for iot data classification, 01 2020.