Deep learning based methods for cancer subtype discovery.

Khamkar Runad Rajkumar
M200416CS
khamkar_m200416cs@nitc.ac.in

Nagateja Banothu
M200288CS
banothu_m200288cs@nitc.ac.in

Pritam Kumar
M200286CS
pritam_m200286cs@nitc.ac.in

***Abstract*—The Cancer is group of diseases which are caused by abnormal growth and spread of affected/mutated cells or group of different cells.With technological advancement in bioinformatics and abundant genomics data,we will try to find sub-types of breast cancer.In our project we will use RNA sequences as with advancement of cancer activities mutation occurs in RNA sequences increases.Affected RNA sequences are responsible for growth of cancer.So it is vicious cycle of mutations .Feature extraction is very tedious task in any machine learning algorithm and often very crucial step as entire prediction is depend on features/knowledge in dataset.Modern deep learning techniques provides technique to extract most significant features in huge dataset such as cancer.Thus to extract the features we will use deep learning,specifically neural networks.Also to improve the performance of deep learning phase we will use 2 stages (unsupervised learning and supervised fine tuning) .In this way we able to reduce the dimensions of data as well as extract features in given data to feed in different classification algorithms or series of classifiers to predict the sub types of cancer type.With the help of weighted features we will able to predict sub types more efficiently.**

Keywords—neural network, deep learning, genetic data, feature extraction.

1. Introduction

The Cancer is abnormal growth of infected/cancer causing cells which does damage and spread to other parts of body.There are over 100 types of cancer affecting humans.According to International Agency for Research on Cancer(IARC),1 in 5 people develop cancer on their lifetime.Furthermore 1 in 8 men and 1 in 11 female die from cancer worldwide.According to new estimates suggest that more than 50 million people are living within five years of a past cancer diagnosis which lead to high death rate of cancer.The ageing population and socio-economics are the main reason to rising death rates of cancer.
The severity of cancer is divided in 5 stages.Treatment of cancer is depend on stage,location,grade,DNA structure.If we are able to diagnose cancer in early stage the cost and treatment will be minimal.As severity of cancer increases the cost and treatment as well as chances of fatality increases. Breast cancer is the type of cancer which forms in breast cells.Every 1 in 4 cancer in women is breast cancer (24.5%).Breast cancer can occur in both men and women,but far more likely to affect to women.
Due to support and research funding has helped in advancement in breast cancer diagnosis and treatment.The survival rate of breast cancer has increased substantially recently due to early detection,self awareness and personalize approach and better understanding of disease.
Common types of breast cancers:

- Angiosarcoma
- Ductal carcinoma in situ (DCIS)
- Inflammatory breast cancer
- Invasive lobular carcinoma
- Lobular carcinoma in situ (LCIS)
- Male breast cancer
- Paget's disease of the breast
- Recurrent breast cancer

About 5% to 10% of breast cancers are caused by abnormal genes passed from parent to child.The levels of risk in genes are likely to cause breast cancer:

- High risk mutation:PTEN, PALB2, TP53
- moderate to high:ATM, CDH1
- moderate:CHEK2, NBN, NF1, STK11
- Inherited abnormal genes:BRCA1, BRCA2

Positive result in above mutation are likely due to breast cancer.

1.1. RNA sequence

RNA-SEQ is a sequencing technique used to quantity RNA in a biological sample and NGS (Next Generation Sequencing) to reveal the presence. It has also analyzed the continuously changing cellular transcriptome. RNA sequencing is used to the capability of high throughput sequencing methods to provide insight into the transcriptome of a cell. RNA sequencing is used for cancer diagnostic based on molecular profiling If RNA sequencing improves the precision of cancer diagnosis by providing comprehensive tumor characterization. RNA sequencing has been utilized in many features of cancer research and therapy. Like, It used in breast cancer. In RNA sequencing-based on molecular profiling of primary breast cancer tumors can directly replace.

Microarray technology also used in gene expression profiling analysis but more prefer in RNA sequencing because:-

- Ability to detect novel transcripts:- If RNA sequence doesn't require species or transcript-specific probes. It can detect novel transcripts, gene fusions single nucleotide variants.
- Wider dynamic range:- In RNA sequencing technology produce discrete, digital sequencing read counts and quantify the expression in a larger dynamic range. Its range is $> 10^5$ for RNA sequencing.
- Higher specific and sensitivity:- RNA sequencing technology can detect a higher percentage of differentially expressed genes. It is especially detected in low expressed genes.
- Simple detection of rare and low - abundance transcripts:-Sequencing coverage depth can easily be increased to detect rare transcripts, single transcripts per cell, or weakly expressed genes.

1.2. Feature extraction with deep learning

Deep Learning is an Artificial Intelligence function the working as a human brain in processing data and creating a model to help in decision making, detecting object recognizing speech, and translating language. Deep Learning is a subfield of Machine Learning that is concerned with the structure of the human brain and it has capable of learning unsupervised from data that unlabelled. A deep learning algorithm also detects the beast cancer. It accurately detects beast cancer in mammography images and generalized well to the population not represented in the training dataset. In breast cancer detection the feature extraction is an important step because it helps discriminate between benign and malignant tumors Deep Learning works like a human being. It is a neural network layer of a node like a human brain. Node is an individual layer connected to an adjacent layer. The Deep Learning network is said to be a layer. A single neuron in the human brain is received thousand of signals from other neurons. In a neural network, a signal travels between nodes and assigns the weight. If the weighted node is heavier then, it will exert more effect on the next layer of the node. The final layer compiles the weighted input to produce an output

1.3. Dataset:

In our work, we wanted to use public RNA-seq dataset named Breast invasive Carcinoma(BRCA) collected from cBioportal and consists of 20,439 genes with 605 samples. BRCA dataset has been labeled into five molecular sub-types such as LumA, LumB, Basal, Triple negative and HER2-positive.
We also choose UCEC (TCGA, Nature 2013) 2 dataset which has four molecular sub-types, namely

- copy number high (CNH)
- copy number low (CNL)
- hypermutated (HYM)
- ultra-mutated (ULM)

are used to validate biological insights. A total of 230 samples and 20482 genes per sample are there in UCEC.

1.4. Related work:

There have been several promising studies done to classify and cluster gene expression using various machine learning algorithms. Algorithms like staked auto-encoder(SAE), denoising auto-encoder(DAE) and variational auto-encoder(VAE) are unsupervised deep feature extraction algorithms and are applied to extract features. Moreover, the above algorithms are compared with traditional machine learning based algorithms such as principal component analysis(PCA) to measure the performance.

These techniques can be used in applications like to predict protein-protein interaction, discover new drugs, and fix class imbalance problem. Various unsupervised algorithms have been employed to classify the diseases which includes auto-encoder based clustering, partition based clustering, K-means clustering,hierarchical clustering, expectation maximization (EM) and nonnegative matrix factorization (NMF) clustering algorithms. Many more algorithms such as LR, DT, SVM, RF etc. equally exploited in cancer research in order to classify diseases from microarray, RNA-Seq, and integrative multi omics data.For breast cancer survival prediction and biomarker identification a variety of feature selection methods are applied, such as Chi2, mRMR and Info-Gain, are used along with classification methods, such as RF, SVM and NB..

In machine learning, in order to get high number of features (genes) with limited number of samples (patients) we may get significant amount of noise which is known as "curse of dimensionality". In order to overcome these challenges, feature extraction method such as AAE can be used to lower the dimension. This technique also helps in the visualization of high dimensional data and prevents manifold fracturing problem. With the introduction of next generation sequencing(NGS) technologies, modern deep learning algorithms try to find out some basic bilogical questions such as prediction of single nucleotide polymorphisms(SNPs), identification of biologically relevant gene, design of protein and predicting its structures. With addition to this, personalised treatment such as targated therapies, and predicting drug response could disrupt traditional treatment in future.

Feature extraction, an essential part of data analysis is used to reduce high-dimensional data to amenable number that helps to interpret results. Analysing the hidden nodes within neural networks potentially high light important genetic behaviour.Underlying patterns in genomics and biomedical domains can be revealed by latent space features learned through unsupervised pre-training and supervised fine-tuning. To identify genes that are critical for the diagnosis of breast cancer, DAE is applied on RNA-Sed data from the cancer genome atlas(TCGA) and latent space is used. Also in similar fashion VAE is also applied on pan-cancer gene expression data and compared with other feature extraction method, and VAE proved to be more informative.

2. Problem Definition:

In human health, cancer is a very dangerous disease. It has multiple subtypes of cancer in terms of pathological or clinical features of the tumor. It needs to identify the subtype of cancer and take better precision in cancer diagnosis and therapy. Different sequencing techniques are used to high throughput and provide a large number of publicly accessible omics data related to cancer. The Cancer Genome Atlas(TCGA) is one such public database that hosts various types of cancer-related sequencing data. The cancer subtype data present in the public, so it has very helpful for researchers to improve the prediction of cancer subtypes using computational techniques. It has integrating multi-omics data for cancer subtype prediction will improve the accuracy of the prediction. I used a different machine learning technique to improve the accuracy of prediction in the cancer subtype.

3. Proposed Methodology

We are implementing given problem in 3 stages:

1) PreProcessing
2) Feature extraction
3) Classification

3.1. PreProcessing

3.1.1. SMOTE. (Synthetic minority oversimplifying technique)
Class imbalance lead to over-fitting and low predictive score.The class imbalance occurs due to less data in minority class therefore our machine learning model not able to make clear class boundaries.In this method,we synthesis new feature vector to minority class using nearest neighbour and synthesising new features in between original points and nearest neighbour point randomly.

$$s = x + \mu.(x^R - x)$$

Where,$0 \leq \mu \leq 1$,x^R is randomly chosen 5 closest minority neighbours of x.

3.1.2. Quantile Transformer. It is used to remove data redundancy and noise.It spread data items Gaussian normal form or uniform normal form. This technique spreads out the most frequent values.In this way we normalize the data to their mean values without changing their respective order in the genes.

$$F_X(x) := \Pr(X \leq x) = p.$$
$$Q(p) = \inf\{x \in R : p \leq F(x)\}$$
which insures,
$$Q(F(X)) = X \text{ i.e.,}$$
$$Q = F^{-1}$$

Where,$p(x)$ is probability of sample x,
$F(x)$ is applied function on x
$Q(x)$ is Quantile Transform of x

3.2. Feature Extraction

3.2.1. Unsupervised Pretraining. With supervised learning the neural networks tents to get more biased in the direction to get correct labels.Thus they may ignore important gene/feature which plays important role in mutation.By using this method we are able to extract features with required dimensions without loosing any important gene.Drawback of this method is we need huge data to capitalise abilities of unsupervised pretraining.

3.2.2. Fine tuning using supervised learning. For this process we add 4 node output layer to encoded layer (4 nodes to represent 4 sub types in data).After that ,we freeze input and hidden layer so that features learned by pre trained phase do not change.In fine tuning phase only weights and bias of encoded layer is changed in supervised learning method.This helps encoded layer to relearn low level feature to help in accurate prediction.After this process output layer is removed and we are ready to classify features extracted by encoder.

3.3. Classification

After extracting features from deep learning stage we are going to feed the given features to the different classifiers and choose best among them.
Different classifiers like:

- KNeighborsClassifier
- DecisionTreeClassifier
- RandomForestClassifier
- XGBClassifier
- GradientBoostingClassifier
- GaussianNB
- LinearDiscriminantAnalysis
- QuadraticDiscriminantAnalysis
- SVC
- LogisticRegression
- MLPClassifier
- VotingClassifier

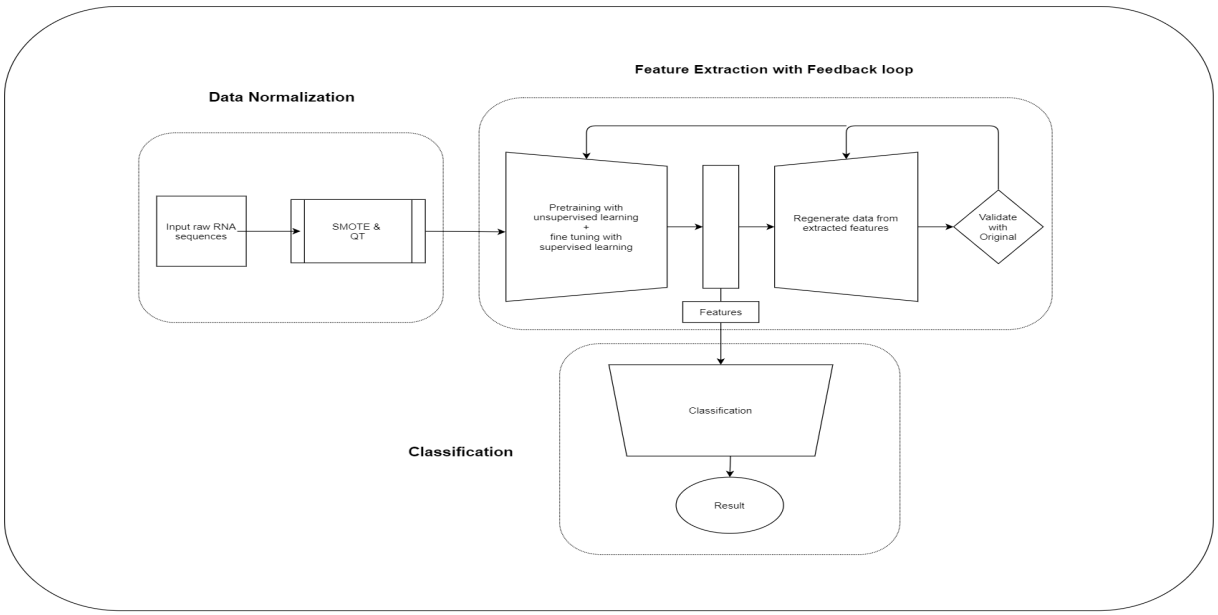

Figure 1: Block diagram of workflow

4. Work Plan

4.1. System Requirement

Below are the libraries required for the proposed project.

- Keras 2.0.6
- Keras-adversial(0.0.3)
- Tensorflow(1.13.1)
- Scikit-Learn(0.20.3)
- Numpy(1.16.3)
- Imbalanced-Learn(0.4.3)

4.2. Datasets

For our analysis we used two datasets

- Breast Invasive Carcinoma(BARC) - Clinical information used to label various molecular subtypes.
- cBioPortal - Cancer Geonomics dataset

Breast Invasive Carcinoma shortly named as BRCA is a public RNA-seq dataset collected from cBioportal. This data consists of 20,439 genes with 605 samples. According to the clinical information, BRCA dataset has been labeled into five molecular sub-types such as LumA, LumB, Basal, Triple negative and HER2-positive. As Basal and Tri-Neg share very similar biological patterns we have merged both into one subtype.

We also choose UCEC (TCGA, Nature 2013) 2 dataset which comprised four molecular sub-types, namely copy number high (CNH), copy number low (CNL), hypermutated (HYM) and ultra-mutated (ULM) to validate results regarding biological insights. UCEC has total number of 230 samples and each sample contains 20482 genes.

Molecular Subtypes	Number of Patients	Label
Luminal A	304	0
Luminal B	121	1
Basal and Triple Negative	137	2
Her 2 Positive	43	3

TABLE 1:
Labels for various molecular subtypes of Barc Dataset

Molecular Subtypes	Number of Patients	Label
copy number high (CNH)	90	0
copy number low (CNL)	64	1
hypermutated (HYM)	60	2
ultra-mutated (ULM)	13	3

TABLE 2:
Labels for various molecular subtypes of UCEC dataset

Dataset	Total Number of Samples (Patients)	Total Number of Features (Genes)
BRAC	605	20439
UCEC	230	20482

4.3. Proposed Model flow

We propose a feature extraction method to extract most relevant features from high dimensional gene expression data. The system takes gene expression data and then prosessing techniques such as sampling and normalization are applied before passing it for feature extraction. After the feature extraction the extracted feature is fed to 12 classifiers to evaluate the performance of feature extraction.

4.4. Common preprocessing:

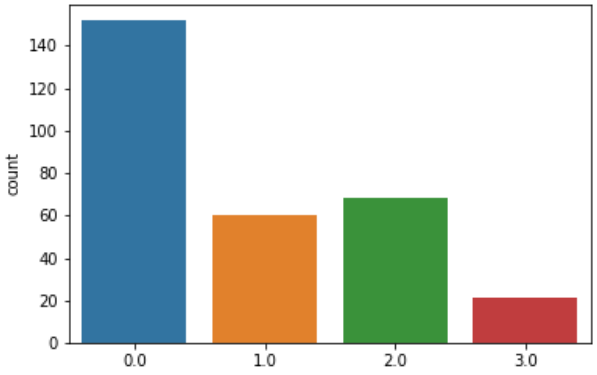

Figure 2: Dataset

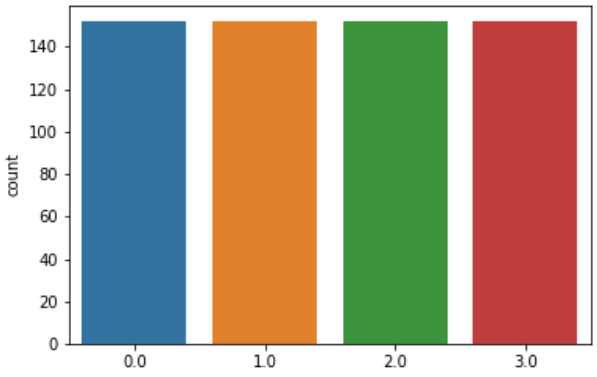

Figure 3: Oversampled Dataset

In preprocessing for normalization and sampling we are using SMOTE(Synthetic minority over simplifying technique) and QT(Quantile Transformer).
SMOTE is used to balance the classes in the dataset.We are using QT for Normalization,normalization is used to eliminate data redundancy and noise. Quantile transformer (QT) normalization technique which transforms the features to follow a uniform or normal distribution. This technique not only reduces the impact of outliers, but also spreads out the most frequent values.

4.5. Feature extractor

4.5.1. PCA. Principal Component Analysis (PCA) is one of the most commonly used unsupervised machine learning algorithms across a variety of applications such as exploratory data analysis, dimensionality reduction, information compression, data de-noising, etc.In the pre processing phase PCA is used for Reduce the number of dimensions in the training dataset and to De-noise the data.

Steps to compute PCA:

- Feature standardization: We standardize each feature to have a mean of 0 and a variance of 1.
- Obtain the covariance matrix computation: The covariance matrix is a square matrix, of d x d dimensions, where d stands for "dimension" (or feature or column). It shows the pairwise feature correlation between each feature.
- Calculate the eigen decomposition of the covariance matrix: We calculate the eigen vectors (unit vectors) and their associated eigenvalues (scalars by which we multiply the eigen vector) of the covariance matrix.
- Sort the eigen vectors from the highest eigenvalue to the lowest:The eigen vector with the highest eigenvalue is the first principal component.
- Select the number of principal components: Select the top N eigen vectors (based on their eigenvalues) to become the N principal components. The optimal number of principal components is both subjective and problem-dependent. Usually, we look at the cumulative amount of shared variance explained by the combination of principal components and pick that number of components, which still significantly explains the shared variance.

4.5.2. Select Best with Linear Support Vector Classification.. Similar to SVC with parameter kernel='linear', but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples.

This class supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme.

4.5.3. Pre-Training in Feature extraction. In this work we are using auto encoder and decoder.
Autoencoders are a specific type of feedforward neural networks where the input is the same as the output. They compress the input into a lower-dimensional code and then reconstruct the output from this representation. This compression is also called as latent-space representation. Autoencoders are considered an unsupervised learning technique since they don't need explicit labels to train on. But to be more precise they are self-supervised because they generate their own labels from the training data.

Steps for feature extraction using Unsupervised method:

- Create input layer of the size of gene
- Create hidden encoder layer with dense neurons of converging sizes. In our model it is 1000 to 150 to 50.
- Create bottleneck(feature rich layer) of size of 50, of dense neurons.

- Create hidden dense layer of dense layer of dense neurons by expanding the sizes. In our case it is 50 to 150 to 1000
- Add output layer of size of input gene.
- Train the model, in which it first reduces it dimension till ”50”(Features) and generates ”50” features to its original input.
- Train till we get minimum loss.
- Split the original model till bottle neck layer to use it further.

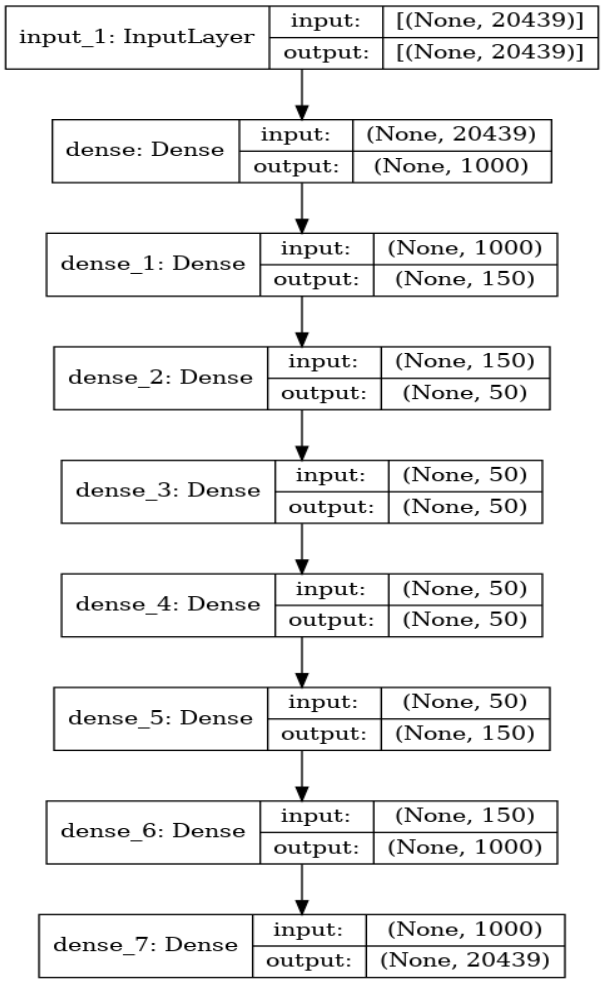

Figure 4: Neural Network structure of unsupervised model

4.5.4. Fine tuning. We fine-tune the trained encoder by adding a temporary output layer of four neurons, representing the four breast cancer sub-types.. Input layer and hidden layer are kept frozen when reusing the pre-trained model so that features learned from pre-training don't change. During fine tuning, only weights and biases of encoded layer are updated to help the encoded layer re-learn low-level features enabling more accurate predictions. Finally, output layer is removed and encoded layer holds the extracted features which is then evaluated through twelve different classifiers.
Here we need to note that feature extraction is performed in unsupervised manner and fine tuning is done in supervised manner.

Steps for fine tunning model

- Load the unsupervised encoder model
- freeze hidden layers
- Add output layer of size of the subtype in our case the number of subtypes are four subtypes.
- Train the given model with x-train and y-train till max epchos
- Remove the output layer
- Unfreeze the hidden layer

4.5.5. Variational multilayer autoencoder. A Variational Autoencoder(VAE) provides a probabilistic manner for describing an observation in latent space. Thus, rather than building an encoder which outputs a single value to describe each latent state attribute, we'll formulate our encoder to describe a probability distribution for each latent attribute.
For variational autoencoders, the encoder model is sometimes referred to as the recognition model whereas the decoder model is sometimes referred to as the generative model.
Suppose that there exists some hidden variable z which generates an observation x.
We can only see x, but we would like to infer the characteristics of z. In other words, we'd like to compute $p(z|x)$.

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

To calculate x from $p(z|x)$

$$p(x) = \int \mathrm{p(x}\|z)p(z)dz$$

4.6. Classification

For classfication we have used 12 different classifiers

4.6.1. K Neighbors Classifier. K Nearest Neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure(e.g., distance functions). The classes formed here are depending on some similarity.
Pseudocode for KNN-classifier:

- Load the data
- Initialise the value of k
- For getting the predicted class, iterate from 1 to total number of training data points
  - Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method.
  - Sort the calculated distances in ascending order based on distance values.
  - Get top k rows from the sorted array.
  - Get the most frequent class of these rows.
  - Return the predicted class.

4.6.2. Decision Tree Classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

4.6.3. Random Forest Classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

4.6.4. XGB Classifier. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. The boosting ensemble technique consists of three simple steps:

- An initial model F0 is defined to predict the target variable y. This model will be associated with a residual (y – F0)
- A new model h1 is fit to the residuals from the previous step
- Now, F0 and h1 are combined to give F1, the boosted version of F0. The mean squared error from F1 will be lower than that from F0:
  –
$$F_1(x) < -F_0(x) + h_1(x)$$
- To improve the performance of F1, we could model after the residuals of F1 and create a new model F2:
  –
$$F_2(x) < -F_1(x) + h_2(x)$$
- This can be done for 'm' iterations, until residuals have been minimized as much as possible:
  –
$$F_m(x) < -F_{m-1}(x) + h_m(x)$$
- Here, the additive learners do not disturb the functions created in the previous steps. Instead, they impart information of their own to bring down the errors.

4.6.5. Gradient Boosting Classifier. Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.Decision trees are usually used when doing gradient boosting.

4.6.6. Gaussian Naive Bayes. :A Gaussian Naive Bayes algorithm is a special type of Naive Bayes algorithm. It's specifically used when the features have continuous values.

4.6.7. Linear Discriminant Analysis. A classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule.

4.6.8. Quadratic Discriminant Analysis. A classifier with a quadratic decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule.

4.6.9. Support Vector Classification(SVC). The implementation is based on libsvm. The fit time scales at least quadratically with the number of samples and may be impractical beyond tens of thousands of samples.

4.6.10. Logistic Regression. In the multiclass case, the training algorithm uses the one-vs-rest (OvR) scheme if the 'multi-class' option is set to 'ovr', and uses the cross-entropy loss if the 'multi-class' option is set to 'multinomial'.

4.6.11. MLP Classifier. This Multi-layer Perceptron classifier model optimizes the log-loss function using LBFGS or stochastic gradient descent.

4.6.12. Voting Classifier. A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

5. Results

We use 2 different scale of cancer subtypes dataset for this experiment with 5 feature extractors and 12 classifiers.We used following performance parameters in our experiment.

Accuracy: This function computes subset accuracy . The set of labels predicted for a sample must exactly match the corresponding set of labels in ytrue. Parameters ytrue array or label indicator array.
F1-SCORE: F1-SCORE is the harmonic mean between precisin and recall.It is used as a statistical measure to rate performance.
RECALL: It is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.
Precision: It is the fraction of relevant instances among the retrieved instances
AUC: Compute Area Under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores. This implementation can be used with binary, multiclass, and multilabel classification.
MCC: The Matthews correlation coefficient is used in machine learning as a measure of the quality of binary and multiclass classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.
Kappa: This function computes Cohen's kappa 1, a score that expresses the level of agreement between two annotators on a classification problem.
Log-Loss: This is the loss function used in (multinomial) logistic regression and extensions of it such as neural networks, defined as the negative log-likelihood of a logistic model that returns ypred probabilities for its training data ytrue.

5.1. BRCA dataset

BRCA dataset has been labeled into four molecular subtypes as LumA, LumB, Basal, Triple negative, and HER2-positive. In this dataset,data consists of 20,439 genes with 605 samples an efficient feature extraction technique is necessary which can handle high dimensions of data and can perform efficiently with any classifier.
According to Table 2, in BARC Dataset, PCA feature extraction method is performing best compared to the other feature extraction methods in VotingClassifier. The accuracy of VotingClassifier using feature extraction method is 87.4 % . But, KNeighborsClassifier, LogisticRegression, QuadraticDiscriminantAnalysis, SVC, XGBClassifier and GaussianNB classifiers have low accuracy compared to the other feature extractors in PCA Feature extraction.
LinearSVC feature extractor have overall lowest accuracy among all other feature extractor.
Compared to PCA, WithFineTunning model performed well in all other classifications with highest accuracy 86.77% in MLPClassifier classifier.Compared to WithoutFineTunning we see small accuracy buff in WithFineTunning model of 2%-3%. Variational autoencoder perform consistent on BRAC dataset with highest accuracy score 85.61% in LinearRegression classifier
For BARC dataset,we show that WithFineTunning give highest overall performance in all classifiers.

5.2. UCEC dataset

It is comprised of four molecular subtypes, namely copy number high (CNH), copy number low (CNL), hypermutated (HYM), and ultra-mutated (ULM) to validate results regarding biological insights. UCEC has total number of 230 samples and each sample contains 20482 genes
According to Table 3, in UCEC Dataset, PCA and LinearSVC feature extractor are performed better compared to the other feature extraction methods votingclassifier and randomForest respectively.
The performance of encoder-decoder architectures like WithFineTunning,WithoutFineTunning and Variational autoencoder has very less accuracy.

5.3. Compare between UCEC and BRCA dataset

According to table2, table 3, In BRCA dataset is a much higher performance for all feature extraction in all classifiers compares to the UCEC dataset. The number of samples BRCA dataset is 605 and the samples in UCEC dataset is 230.
In deep learning performance is directly depends on size of data.

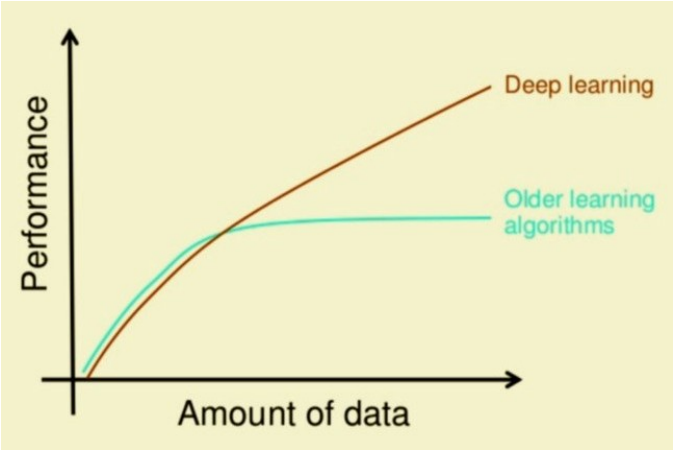

Figure 5: Size to performance ratio in deep learning.

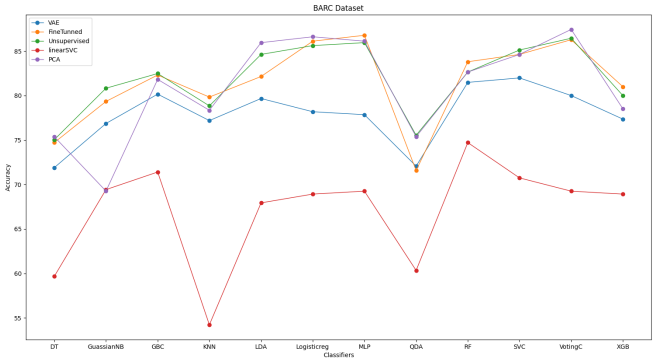

Figure 6: Accurarcy plot of BARC Dataset

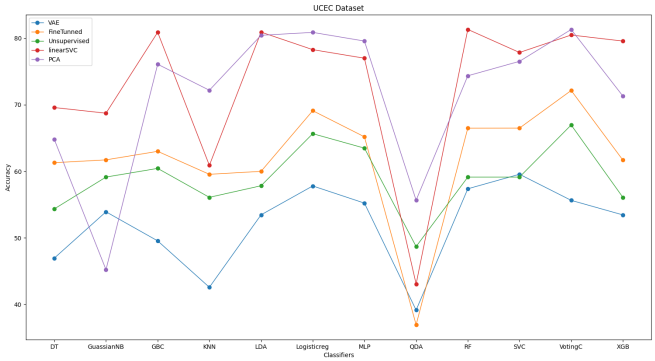

Figure 7: Accurarcy plot of UCEC Dataset

TABLE 3: Accuracies of the classifiers using different feature extraction methods on BARC Dataset

Classifier	WithFineTunning	WithoutFineTunning	PCA	LinearSVC	VEA
DecisionTreeClassifier	74.7122	75.0413	75.3739	59.6719	71.8979
GaussianNB	79.3409	76.8615	69.4259	69.2601	80.8264
GradientBoostingClassifie	82.3136	82.4793	81.8178	71.4028	80.1618
KNeighborsClassifier	79.8302	77.1866	54.2214	78.3458	78.8429
LinearDiscriminantAnalysis	82.1445	84.6280	85.9506	67.9334	79.6717
LogisticRegression	86.1156	78.1811	68.9218	**86.6123**	**85.6198**
MLPClassifier	**86.7773**	85.9504	86.1164	69.2502	77.8524
QuadraticDiscriminantAnalysis	71.5720	72.0629	60.3361	75.3739	75.5371
RandomForestClassifier	83.802	82.6446	82.6478	74.7122	81.4820
SVC	84.6280	81.9852	70.7428	84.6263	85.1239
VotingClassifier	86.2814	**86.4462**	**87.4398**	69.2511	80.0026
XGBClassifier	80.9968	77.3525	76.9202	78.5133	80.0

TABLE 4: Accuracies of the classifiers using different feature extraction methods on UCEC Dataset

Classifier	FineTunned	Unsupervised	linearSVC	VAE	PCA
DecisionTreeClassifier	61.2895	54.3478	69.5773	46.9526	64.7826
GaussianNB	61.7167	59.1304	68.7286	53.9131	45.2173
GradientBoostingClassifie	62.9984	60.4347	80.8840	49.5443	76.0869
KNeighborsClassifier	59.5465	56.0869	60.8908	42.5723	72.1739
LinearDiscriminantAnalysis	59.9908	57.8260	80.9125	53.4575	80.4347
LogisticRegression	69.10457	65.6521	78.2752	57.7979	80.8695
MLPClassifier	65.1799	63.47826	76.9879	55.2175	79.5652
QuadraticDiscriminantAnalysis	36.9332	48.6956	43.0622	39.1433	55.6521
RandomForestClassifier	66.4843	59.1304	**81.3169**	57.3706	74.3478
SVC	66.4900	59.1304	77.8480	59.5522	76.5217
VotingClassifier	**72.1405**	**66.9565**	80.4739	55.6334	**81.3043**
XGBClassifier	61.6826	56.0869	79.5625	53.4461	71.3043

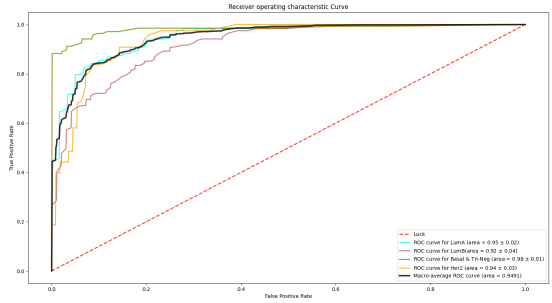

Figure 8: BARC dataset ROC curve

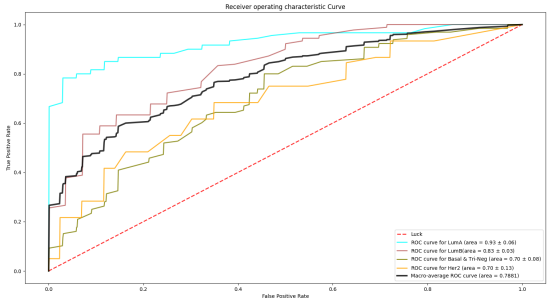

Figure 9: UCEC Dataset ROC Curve plot

6. Conclusion

In this paper we used encoder-decoder architecture to extract features as well as dimensionality reduction for classification of breast cancer sub-types. Proposed method exhibited high performance on multiclass classification using five-fold cross-validator. Then performance comparison of our proposed architecture with other methods has been carried out using BRCA datasets as well as UCEC datasets, and it is shown that Unsupervised deep learning architecture and PCA outperforms state-of-the-art feature extractors for BRCA dataset and linearSVC and VAE outperforms other feature extractor for UCEC dataset. We successfully show that encoder-decoder architecture,variational autoencoder is able to use to extract features.As dataset is growingexponentially, in the future it is possible to develop higher capacity model to identify cross-cancer bio-markers in more heterogeneous dataset. Most importantly, our proposed feature

extraction method reveals important bio-markers that could be used in forecasting diseases progression and determining treatment strategy.

References

[1] R. K. Mondol, N. D. Truong, M. Reza, S. Ippolito, E. Ebrahimie and O. Kavehei, "AFExNet: An Adversarial Autoencoder for Differentiating Breast Cancer Sub-types and Extracting Biologically Relevant Genes," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, doi: 10.1109/TCBB.2021.3066086.

[2] Jiang Limin, Xiao Yongkang, Ding Yijie, Tang Jijun, Guo Fei,Discovering Cancer Subtypes via an Accurate Fusion Strategy on Multiple Profile Data in Frontiers in Genetics,DOI=10.3389/fgene.2019.00020.

[3] Md. Mohaiminul Islam, Shujun Huang, Rasif Ajwad, Chen Chi, Yang Wang, Pingzhao Hu,An integrative deep learning framework for classifying molecular subtypes of breast cancer, Computational and Structural Biotechnology Journal, Volume 18, 2020, Pages 2185-2199, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2020.08.005. (https://www.sciencedirect.com/science/article/pii/S2001037020303585)

[4] Zhao, Shanrong, Zhang, Baohong, Zhang, Ying, Gordon, William, Du, Sarah, Paradis, Theresa, Vincent, Michael, von Schack, David,Bioinformatics for RNA-Seq Data Analysis,2016 https://doi.org/10.5772, doi:10.5772/63267.

[5] Roman Rosipal,Mark Girolami,Leonard J.Trejo, Andrzej Cichocki,Kernel PCA for Feature Extraction and De-Noising in Nonlinear Regression,231–243 (2001)

[6] A.M.Jade, B.Srikanth, V.K.Jayaraman, B.D.Kulkarni, J.P.Jog, L.Priya, Feature extraction and denoising using kernel PCA, https://doi.org/10.1016/S0009-2509(03)00340-3

[7] J.C.B. Melo, G.D.C. Cavalcanti, K.S. Guimaraes, PCA feature extraction for protein structure prediction, DOI: 10.1109/IJCNN.2003.1224040

[8] Muchenxuan Tong,Kun-Hong Liu, Chungui Xu, Wenbin Ju, An ensemble of SVM classifiers based on gene pairs, https://doi.org/10.1016/j.compbiomed.2013.03.010

[9] Yuanfang Guan, Chad L Myers, David C Hess, Zafer Barutcuoglu, Amy A Caudy, Olga G Troyanskaya, Predicting gene function in a hierarchical context with an ensemble of classifiers, DOI: https://doi.org/10.1186/gb-2008-9-s1-s3

[10] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, Fabian J. Theis, Deep learning: new computational modelling techniques for genomics, DOI https://doi.org/10.1038/s41576-019-0122-6