CLUSTERING AND CLASSIFICATION OF PROTEIN TO PROTEIN GENE EXPRESSION BREAST CANCER

K K.L.V.G.K. Murthy[1*], Dr. R. J. Rama Sree[2*]

[1.]Research Scholar, Rayalaseema University, KURNOOL, A.P.

[2.]Research Supervisor, Professor, Rayalaseema University, KURNOOL, A.P.

Abstract

Breast cancer is one of the most common cancers all over the world, which bring about more than 450,000 deaths each year. Although this malignancy has been extensively studied by a large number of researchers, its prognosis is still poor. Since therapeutic advance can be obtained based on gene signatures, there is an urgent need to discover genes related to breast cancer that may help uncover the mechanisms in cancer progression. In the recent past, the Classifiers are based on genetic signatures in which many microarray studies are analyzed to predict medical results for cancer patients. However, the Signatures from different studies have been benefitted with low-intensity ratio during the classification of individual datasets has been considered as a significant point of research in the present scenario. Hence to overcome the above discussed issue, this paper provides a Deep Learning Framework that combines an algorithm of necessary processing of Linear Discriminant Analysis (LDA) and Auto Encoder (AE) Neural Network with Long Short Term Memory is used to classify different features within the profile of gene expression for breast cancer detection. Hence, an advanced ensemble classification has been developed based on the Deep Learning (DL) algorithm to assess the clinical outcome of breast cancer. Furthermore, numerous independent breast cancer datasets and representations of the signature gene, including the primary method, have been evaluated for the optimization parameters. Finally, the experiment results show that the suggested deep learning frameworks achieve 98.27% accuracy than many other techniques such as genomic data and pathological images with multiple kernel learning (GPMKL), Multi Layer Perception (MLP), Deep Learning Diagnosis (DLD), and Spatiotemporal Wavelet Kinetics (SWK).

Keywords: Genetic Algorithm, Gene Signature, Breast Cancer, Classifier, Predictor, Linear Discriminant Analysis, LSTM.

1. Introduction

Breast cancer seems to be the most common type of cancer that women around the world, and it is found in developed countries after lung cancer . In all over the world, 50% to 60% of cases of breast cancer occur in late stages, and patients have one of the lowest survival levels in the region. Hence there is a need to determine multiple factors that affect the survival rate of breast cancer patients. Clinicians are using basic Software programs for analyzing factors influencing breast cancer survival rates. Such traditional statistical methods cannot be modified to identify new variables or create innovative and inclusive visualizations. Different approaches to machine learning (ML) are used in this area as decision tree (DT) random forest (RF) approaches neural network, extreme boost, logistic reversal, and support vector machine (SVM).

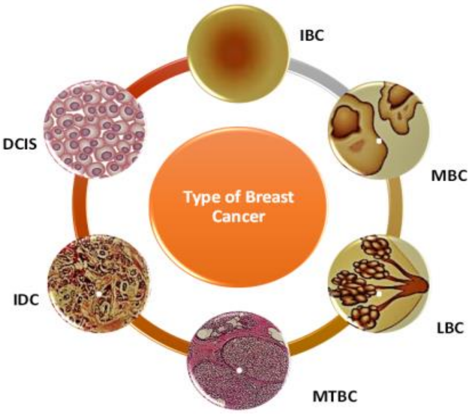

Fig 1: Demonstration of major types of Breast Cancer

The Decision Tree is a supervised learning algorithm representing the outcomes in a tree structure easily interpretable, where visualization is a significant factor in the analysis of large numbers of data. In both supervised and unsupervised mode, the Random forest (Breiman's algorithm) is a derivation of DT, which manages simultaneous and numerical data in classification or regression. Neural networks are complex and are frequently represented as black boxes that model data through training with known results and maximizing weights for a better estimation in situations with undetermined results.

Gene Expression is an ensemble of classification and regression trees that can be paralleled and generated with successful prediction accuracy, which is simple to use in various machine learning functions. Logistical regression assumes the distribution of Gaussian variables and describes all types of variables, such as continuous, discrete, and dichotomous, without assuming normality. The support vector machine will be used for supervised classification and performs by identification of the best decision-making limit, which separates data points from different groups and then the prediction of new observations on the basis of that classification limits. The machine learning ability is increasingly essential for the accuracy of clinical diagnostics. In addition to this, advanced state-of-the-art methods with the technology of machine learning also provide the ability to extract information from statistically significant complex medical imaging data sets.

Breast cancer is one of the most lethal and heterogeneous disease in this present era that causes the death of enormous number of women all over the world. It is the second largest disease that is responsible of women death. There are various machine learning and data mining algorithms that are being used for the prediction of breast cancer. Finding the most suitable and appropriate algorithm for the prediction of breast cancer is one of the important task. Breast cancer is originated through malignant tumours', when the growth of the cell got out of control. A lot of fatty and fibrous tissues of the breast start abnormal growth that becomes the cause of breast cancer. The cancer cells spread throughout the tumours' that cause different stages of cancer. There are different types of breast cancer which occurs when affected cells and tissues spread throughout the body. Ductal Carcinoma in Situ (DCIS) is type of the breast cancer that occurs when abnormal cells spread outside the breast it is also known as the non-invasive cancer.

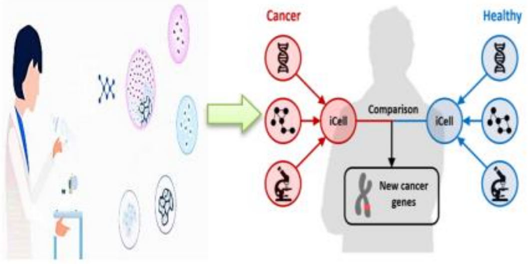

Fig 2: Deep Learning for Breast Cancer Identification

Biological systems' physiological status and gene activities can be revealed at the transcriptome level by gene expression i.e. the actively expressed genes at any given time are reflected by transcriptome. The transcriptome of an organism can be measured using RNA-Seq or DNA microarrays. It is used to denote all RNAs or just mRNA, which is a messenger RNA molecule. These molecules transfer genetic information from DNA, which encodes all the information needed to specify the features and functions of every single cell to the ribosome. Ribosomes connect amino acids, which are the building blocks of the protein together in the order specified by mRNA that encode proteins through the genetic code. RNA-Seq measures the transcription ofa specific gene by converting long RNAs into a library of complementary DNA (cDNA) fragments that generate an expression profile.

The genes that play a main role in the specification of the phenotype can be identified by comparing gene expression profiles from different tissues. i.e. comparing the diseased with the healthy tissues can reveal new insights over the genetic variables involved in pathology. Therefore, gene expression data can provide researchers with features that can be analyzed using computational methods to discover gene regularity targets, diagnosis disease, and develop drug. Studies show that these data can provide very important information regarding tumour characteristics, which provide options for the treatment, care, and management of the patient. Tumour characteristics offered deep insight into cancer detection problems. Identifying genes that are highly expressed in tumour cells but not in normal ones using gene expression data is considered as a challenge that needs to be addressed using computational methods. Gene expression data itself revealed other challenges for the use of the computational methods due to the high dimensionality associated with these data and the relatively very small number of samples as well as the high amount of noise. Many unsupervised and supervised learning methods have been developed for cancer classification using gene expression data.

2. Literature Review

Technological advancements in high-throughput genomics enable the generation of complex and large data sets that can be used for classification, clustering, and bio-marker identification. Modern deep learning algorithms provide us with the opportunity of finding most significant features in such huge dataset to characterize diseases (e.g., cancer) and their sub-types. Thus, developing such deep learning method, which can successfully extract meaningful features from various breast cancer sub-types, is of current research interest. In this paper, Mondol et al. [1] developed dual stage (unsupervised pre-training and supervised fine-tuning) neural network architecture termed AFExNet based on adversarial auto-encoder(AAE) to extract features from high dimensional genetic data. We evaluated the performance of our model through twelve different supervised classifiers to verify the usefulness of the new features using public RNA-Seq dataset of breast cancer.

The remarkable growth of multi-platform genomic profiles has led to the challenge of multiomics data integration. In this study, Pouryahya et al. [2] presented a novel network-based multiomics clustering founded on the Wasserstein distance from optimal mass transport. This distance has many important geometric properties making it a suitable choice for application in machine learning and clustering. The proposed method of aggregating multiomics and Wasserstein distance clustering (aWCluster) is applied to breast carcinoma as well as bladder carcinoma, colorectal adenocarcinoma, renal carcinoma, lung non-small cell adenocarcinoma, and endometrial carcinoma from The Cancer Genome Atlas project. Subtypes were characterized by the concordant effect of mRNA expression, DNA copy number alteration, and DNA methylation of genes and their neighbors in the interaction network. aWCluster successfully clusters all cancer types into classes with significantly different survival rates. Also, a gene ontology enrichment analysis of significant genes in the low survival subgroup of breast cancer leads to the well-known phenomenon of tumor hypoxia and the transcription factor ETS1 whose expression is induced by hypoxia. We believe aWCluster has the potential to discover novel subtypes and biomarkers by accentuating the genes that have concordant multiomics measurements in their interaction network, which are challenging to find without the network inference or with single omics analysis.

Nowadays, the heterogeneous characteristics of cancer patients throw a big challenge to precision medicine and targeted therapy. Identifying cancer subtypes shed new light on effective personalized cancer medicine, future therapeutic strategies and minimizing treatment-related costs. Recently, there are many clustering methods have been proposed in categorizing cancer patients. Although these methods obtained a certain achievement in cancer subtype identification, they still fail to fully use the prior known biological information in the model designing process to improve precision and efficiency. It is acknowledged that the driver gene always regulates its downstream genes in the network to perform a certain function. By analyzing the known clinic cancer subtype data, we found some special co-pathways between the driver genes and the downstream genes in the cancer patients of the same subgroup. Hence, Song et al. [3] proposed a novel model named DDCMNMF (Driver and Downstream gene Co-Module Assisted Multiple Non-negative Matrix Factorization model) that first for cancer subtypes by identifying co-modules of driver genes and downstream genes. We applied our model on lung and breast cancer datasets and compared it with the other four state-of-the-art models.

Breast cancer is a heterogeneous disease with many clinically distinguishable molecular subtypes each corresponding to a cluster of patients. Identification of prognostic and heterogeneous biomarkers for breast cancer is to detect cluster-specific gene biomarkers which can be used for accurate survival prediction of breast cancer outcomes. In this study, Li et al. [4] proposed a FUsion Network-based method (FUNMarker) to identify prognostic and heterogeneous breast cancer biomarkers by considering the heterogeneity of patient samples and biological information from multiple sources. To reduce the affect of heterogeneity of patients, samples were first clustered using the K-means algorithm based on the principal components of gene expression. For each cluster, to comprehensively evaluate the influence of genes on breast cancer, genes were weighted from three aspects: biological function, prognostic ability and correlation with known disease genes. Then they were ranked via a label propagation model on a fusion network that combined physical protein interactions from seven types of networks and thus could reduce the impact of incompleteness of interactome. We compared FUNMarker with three state-of-the-art methods and the results showed that biomarkers identified by FUNMarker were biological interpretable and had stronger discriminative power than the existing methods in differentiating patients with different prognostic outcomes.

The use of machine learning has increased over the years, especially in the world of molecular data. Generally, the inference of relationships between features is determined by statistical models. The phenotype (observable clinical characteristics) can result from the expression of the genotype (genetic code) or environmental factors. Molecular datasets have limited information, while supporting clinical data is ambiguous. There are no well-established approaches for combining clinical information with genomic repositories. The genomic tests that are available only use molecular data and give physicians a result which can be integrated clinically. In this article, Płaczek et al. [5] presented the strategy where clinical data, regardless of its limitations, is combined in one predictive model with molecular features. We predict the risk of malignancy in the thyroid nodules based on the results of fine-needle aspiration biopsy and expression of selected genes. We utilize a Bayesian network (BN) framework to discover relationships between molecular features and assess the impact of added clinical data quality on the performance of the chosen gene set. Bayesian network offering both prognostic and diagnostic perspectives is a perfect non-parametric technique for feature selection, feature extraction, and prediction purposes. We show that certain clinical factors could work as a synthetic feature and provide predictive abilities beyond what genes alone can offer.

Joint graphical lasso (JGL) approach is a Gaussian graphical model to estimate multiple graphical models corresponding to distinct but related groups. Molecular Apocrine(MA) breast cancer tumour has similar characteristics to luminal and basal subtypes. Due to the relationship between MA tumour and two other subtypes, Shahdoust et al. [6] investigated the similarities and differences between the MA genes association network and the ones corresponding to other tumours' by taking advantageous of JGL properties. Two distinct JGL graphical models are applied to two sub-datasets including the gene expression information of the MA and the luminal tumours and also the MA and the basal tumours. Then, topological comparisons between the networks such as finding the shared edges are applied. In addition, several support vector machine (SVM) classification models are performed to assess the discriminating power of some critical nodes in the networks, like hub nodes, to discriminate the tumours sample. Applying the JGL approach prepares an appropriate tool to observe the networks of the MA tumour and other subtypes in one map.

Breast Cancer is a highly aggressive type of cancer generally formed in the cells of the breast. Despite significant advances in the treatment of primary breast cancer in the last decade, there is a dire need to attempt of an accurate predictive model for breast cancer prognosis prediction. Researchers from various disciplines are working together to develop methods to save people from this fatal disease. A good predictive model can help in correct prognosis prediction of breast cancer. This accurate prediction can have several benefits like detection of cancer in the early stage, spare patients from getting unnecessary treatment and medical expenses related to it.

Previous works rely mostly on uni-modal data (selected gene expression) for predictive model design. In recent years, however, multi-modal cancer data sets have become available (gene expression, copy number alteration and clinical). Motivated by the enhancement of deep-learning based models, in the current study, Arya et al. [7] proposed to use some deep-learning based predictive models in a stacked ensemble framework to improve the prognosis prediction of breast cancer from available multi-modal data sets. One of the unique advantages of the proposed approach lies in the architecture of the model. It is a two-stage model. Stage one uses a convolution neural network for feature extraction, while stage two uses the extracted features as input to the stack-based ensemble model. The predictive performance evaluated using different performance measures shows that this model produces better result than already existing approaches.

Cancer sub typing delivers valuable insights into the study of cancer heterogeneity and fulfils 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, Lupat et al.

Records	Bad result	Better result	Final result	Samples were taken
GSE2990	84	154	238	12
GSE3494	32	101	133	125
GSE9195	25	87	112	186
GSE17705	34	168	202	15
GSE17907	27	189	216	61

[8] proposed 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.

3. Dataset Description

In this paper to execute the proposed model the 5 Gene Expression Datasets are considered. The description about them is GSE2990, GSE3494, GSE9195, GSE17705 and GSE17907. Each dataset contains the human breast cancer tumour genes collected from NCBI GEO Database.

4. Proposed Model

The proposed LDA and AE based deep learning framework have been discussed. Next, it defines five data sets for gene expression that uses a specific pre-processing method. Datasets and pre-processing of gene expression Information on gene expression has been downloaded from the GEO database. Every study includes 129,158 Genomic versions of the Affymetrix microarray platform, and each profile has 22,268 mutations, equivalent to 978 landmarks and 21,290 aim genes. Especially from the LINCS cloud, the alternative approach has tested in five different data sets for breast cancer. The information is given in-depth in Table 1.

Table 1: Samples and Results.

Patients use various immune and physiological factors in the five samples to affect the prognosis of outcomes. For example, all ER-positive patients in GSE3494 include ER-positive and ER-negative patients in addition to other data sets. In view of the classification mission, the pre-processing with the 5 data sets has carried out in two steps:

The first argument is to follow the dataset partitioning method with a weak prognostic (set to 1) and reliable predictive (set to 0) division of all people with cancer. Data on aid has eliminated from the review of patients receiving adjuvant or screened for 5 years. In addition, it has been quantified the five datasets with a MAS 5.0 algorithm and converted all of the samples into an Expression gene ID, because the microarray models are used to calculate gene expression value. The aim is to integrate both feature selection and feature elimination with profound learning basics that it learns from genomic profiles quite concisely and establish a more accurate classification of cancer prognoses. Two stages are comprised of this deep learning approach are discussed as follows

- LDA: Considering that data on gene expression is extensive, containing repetitive, noisy data, the LDA program (as defined in section I-A) is utilized as the tool for selecting features to reduce genomic model complexity. LDA conducts a linear estimation of the existing data and, in the meantime, maintains essential information.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 11, Nov 2022

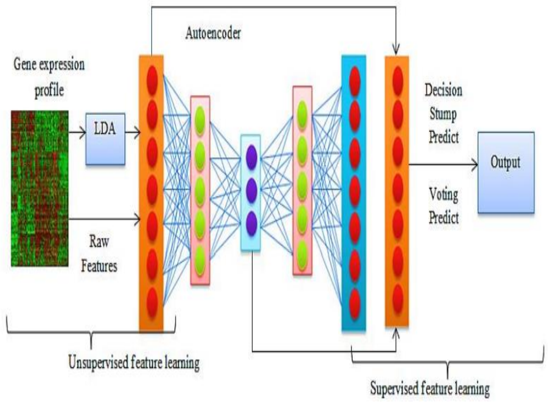

Fig 3: Flowchart of the proposed method

Autoencoder: The result is simply a continuous representation of the source data since LDA has been applied. An optimized version of LDA characteristics is subsequently incorporated in a feature extraction system in addition to raw attributes, to capture nonlinear interactions between the expressions of different genes. For feature extraction, it uses an Autoencoder neural network.

The features extracted from the proposed two-phase unregulated attribute learning methodology are ultimately applied to a set of labels for classifier learning to forecast clinical outcomes for cancer patients. The method of classification is subject to test labels that indicate directed classification preparation. Therefore, an LDA-Ada compartment is designed to use LDA compressed results as input properties to deal with a two-stage classifying feature learning system.

Auto encoders are comprised of an input layer, an output layer that has the same size as the input layer, and (in its simplest form) a single hidden layer [16]. In our case, we actually tried to pre-process the original dataset without generating new features and the same number of neurons in the input and output assured this requirement. An autoencoder has 2 steps: encoding and decoding. In our study, encoding was the step whereby the initial values of the predictors were transformed as a result of applying a sigmoid function to the input vector. Decoding is the procedure used to convert the values in the hidden layer into the ones shown as output vector in the output layer. This last transformation was also carried out by a sigmoid function.

An autoencoder can be seen as a double transformation which tries to provide an approximation of the identity function. By placing constraints on the network, such as by limiting or increasing the number of hidden units, interesting structure about the data can be discovered.

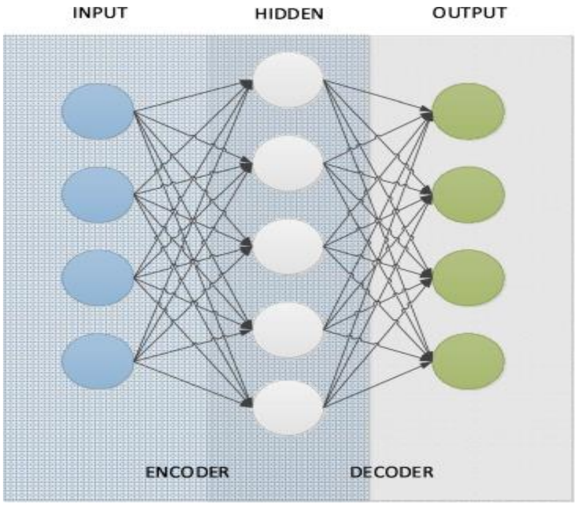

Fig 4: Auto Encoding and Decoding Process

Although the studies on genes that cause breast cancer have been fruitful, the molecular mechanisms that play an important role in the regulation of the disease are still not fully clarified [5]. Microarray technology, which has emerged as a result of developments in DNA analysis in recent years, allows measurements of interactions of tens of thousands of genes with each other. With this technology, early diagnosis can be performed by determining the underlying genetic factors of diseases [6]. However, the large number of features in gene expression data increases the data dimension, making the solutions of problems such as analysis and classification difficult. This necessitates operations such as dimension reduction and gene selection in order to work with microarrays. For this reason, the process of feature extraction in gene analysis is of great importance [7], [8].

In this study, Stacked Auto Encoder (SAE), which is a deep learning model used on the microarray data set for the diagnosis and classification of breast cancer, and the models created with different classifiers applied to the last layer to improve SAE performance was used. Each model was implemented on the dataset used in the study and performance evaluations were carried out. In order to improve the classification accuracy, the feature extraction method called autoencoder is implemented. The image dataset is fed into the autoencoder and the output of the autoencoder is fed into the input of the Convolution Neural Network (CNN). The autoencoder takes the relevant features from the given image as an input and the classification accuracy is monitored. LDA focuses primarily on projecting the features in higher dimension space to lower dimensions. You can achieve this in three steps:

Firstly, you need to calculate the reparability between classes which is the distance between the mean of different classes. This is called the *between-class variance*.

$$S_b = \sum_{i=1}^{g} N_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

Secondly, calculate the distance between the mean and sample of each class. It is also called the within-class variance.

$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T$$

Finally, construct the lower-dimensional space which maximizes the between-class variance and minimizes the within-class variance. P is considered as the lower-dimensional space projection, also called Fisher's criterion.

$$P_{lda} = \arg\max_{P} \frac{|P^T S_b P|}{|P^T S_w P|}$$

The intuition behind the LSTM architecture is to create an additional module in a neural network that learns when to remember and when to forget pertinent information.[15] In other words, the network effectively learns which information might be needed later on in a sequence and when that information is no longer needed.

The compact forms of the equations for the forward pass of an LSTM cell with a forget gate are:

$$\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f)\\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i)\\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o)\\
\tilde{c}_t &= \sigma_c(W_c x_t + U_c h_{t-1} + b_c)\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t\\
h_t &= o_t \odot \sigma_h(c_t)
\end{aligned}$$

Variables

- $x_t \in \mathbb{R}^d$: input vector to the LSTM unit
- $f_t \in (0,1)^h$: forget gate's activation vector
- $i_t \in (0,1)^h$: input/update gate's activation vector
- $o_t \in (0,1)^h$: output gate's activation vector
- $h_t \in (-1,1)^h$: hidden state vector also known as output vector of the LSTM unit
- $\tilde{c}_t \in (-1,1)^h$: cell input activation vector
- $c_t \in \mathbb{R}^h$: cell state vector
- $W \in \mathbb{R}^{h\times d}$, $U \in \mathbb{R}^{h\times h}$ and $b \in \mathbb{R}^h$: weight matrices and bias vector parameters which need to be learned during training

where the superscripts d and h refer to the number of input features and number of hidden units, respectively.

LSTM networks are the most commonly used variation of Recurrent Neural Networks (RNNs). The critical component of the LSTM is the memory cell and the gates (including the forget gate but also the input gate), inner contents of the memory cell are modulated by the input gates and forget gates. Assuming that both of the segue he are closed, the contents of the memory cell will remain unmodified between one time-step and the next gradients gating structure allows information to be retained across many time-steps, and consequently also allows group that to flow across many time-steps. This allows the LSTM model to overcome the vanishing gradient properly occurs with most Recurrent Neural Network models.

5. Results

Breast cancer is one of the most common forms of cancer that is diagnosed in most women and some rare cases even men. In recent years, breast cancer survival rates have increased significantly, due to factors such as earlier detection. Treatment for breast cancer largely depends on identifying the type of mass of tissue formed, which is known as a tumor. If normal cells grow in an uncontrollable manner the tumour is called benign (non-cancerous). But, if the cells' growth is out of control and their behavior is abnormal, then the tumour is called malignant (cancerous). During the invasive, (i.e. curable) stage of cancer, only 10–15% part of the breast contains cancerous cells. Therefore, it is difficult to diagnose it using mammography. However, the development of machine learning techniques has allowed early detection of breast cancer in clinical trials. Hence, Deep learning has been effectively eliminating the problem of uneven distribution of training and improves classifier's capacity to generalize breast cancer.

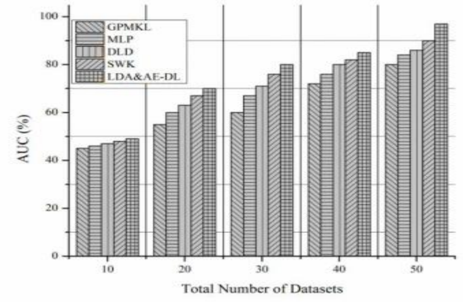

Fig 5: AUC Evaluation

Figure 5 shows that this advanced ensemble classification provides the best AUC (Area Under Curve) performance over several datasets, although the gene set methods are not based on the datasets. Hence, the gene-set ways that achieves higher AUC efficiency than that of the two gene classifiers. The MCC scores indicate a similar phenomenon. Table 2 shows the evaluation of the Matthews correlation coefficient of proposed LDA&AE-DL in comparison with GPMKL, MLP, DLD, SWK.

Total Number of Datasets	GPMKL	MLP	DLD	SWK	LDA&AE-DL
10	50.6	53.3	56.8	58.4	60.3
20	66.5	69.1	72.8	78.5	82.2
30	76.8	78.2	82.1	84.3	89.9
40	70.5	74.3	80.2	86.1	92.7
50	78.4	83.6	85.1	89.7	96.8

Table 2: AUC Comparison Levels

From Figure 6, the proposed classification system (especially those based on physical signature), performed well on a dataset. This is because these two datasets have both a negative lymph node and a positive lymph node, whereas the other datasets have only lymph node-negative patients. Remarkably, the high performance of specific advanced ensemble classifiers [24,25] achieves 96.7% MCC, which outperforms the others substantially.

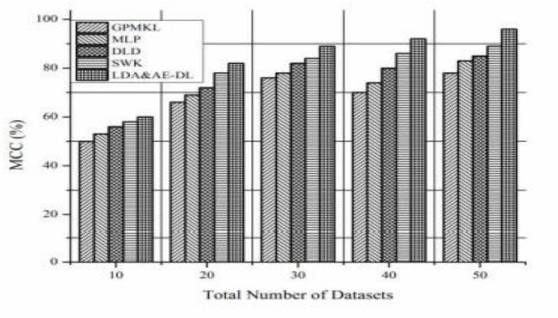

Fig 6: MCC Evaluation

The above observation shows that the method being proposed is less susceptible to unbalanced data sets and is indeed more stable, compared with the four other categories which show dramatic changes in the various datasets. However, the AUCs and MCCs are compared with further four combined methods in public datasets. Figure 7 shows the outcomes to demonstrate the efficiency of a particular way more in detail. Complete analysis indicates that this AUC classification reaches more than 75.3%, while the two genetic classification devices have a significantly worse outcome of

75.8%. However, both genetic classification devices can only achieve AUC 97.4%. The MCC indexes will demonstrate a similar phenomenon.

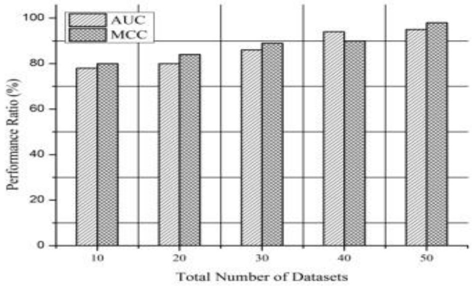

Fig 7: Performance Analysis

Deep learning in various machine learning tasks, including object detection, and classification, has proved breakthrough technology of recent years. Detailed learning methods from the input information concerning the target output following classical Machine Learning methods that involve a handcrafted feature extraction level. The highest performance networks achieve marginally higher efficiency and exceed all other tasks in terms of accuracy when measured using the same data set. The architecture, however, offers detailed lesion segmentation as an input to extract from features. The performance of the system is, therefore, highly dependent on the quality of the segmentation provided, which can be a task that takes by the user with a great deal of time and effort. In the meantime, the lack of any CNN results in an enhanced end-to-end clinical classification system. Figure 8 shows the accuracy of the proposed LDA&AE-DL when compared to GPMKL, MLP, DLD, and SWK.

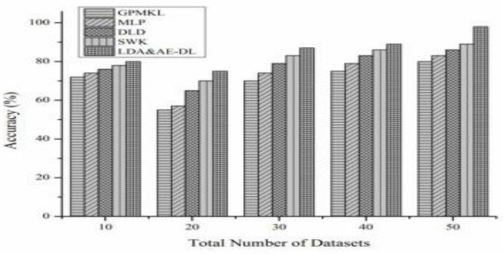

Fig 8: Accuracy Evaluation

Although several improvements in the training process and network architecture have been proposed in response to the difficulties caused by the increasing potential and complexity of models, there is a still significant amount of data required to provide adequate training which is not available for most medically focused applications, such as the present problem, i.e., the diagnosis of breast cancer.

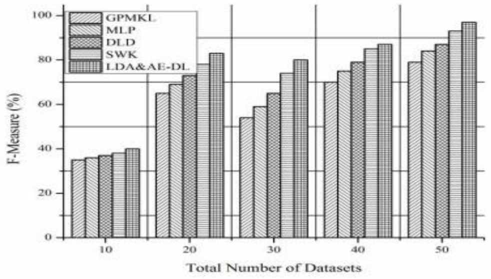

Fig 9: F-Measure Evaluation

The F measure is known as the harmonic mean of accuracy and recall of classification. In comparison to other examples, both FP and FN are included in the calculation of F-measure. The meaningful indicator of the consistency of binary classifications is the Matthews correlation coefficient. MCC calculates a correlation of the classification prediction with a more balanced index. Figure 9 shows the F-Measure of proposed LDA&AEDL when compared to GPMKL, MLP, DLD and SWK. According to the results and discussion section, the proposed method has high AUC, MCC, and accuracy for classifying breast cancer. At the end after applying the Advanced KNN Clustering and Auto Encoder proposed model on Class 1, the following 36 tumour genes are identified. Table 3 represents the identified hub genes from the given datasets.

TP53	CCNB2	FBX05	MCM10
KIFLA	TPX2	BRCA1	BRCA2
NUSAP1	MELK	CENRF	NEK2
TUP2A	PRARG	MAPK1	LMNB1
CCNA2	EEF1E1	IDUA	AACS
RFC2	KIFI4	EBP	CPZ
PALB2	CHEK2	CDH1	PTEN
STK11	CTLA4	CYP19A1	MAP3K1
CASP8	TERT	FGFR2	BARD1

Table 3. Final HUB Genes identified

6. Conclusion

One of the key advantages of deep learning over prior neural nets and machine-learning algorithms is its ability to infer new characteristics from a small collection of features in a training set. That is, it will look for and find more characteristics that are similar to the ones currently recognised. Deep learning's capacity to build features without being explicitly taught means that depending on these networks can save data scientists months of labour. It also means that data scientists can work with more complicated feature sets than machine learning techniques would allow. The main aim of the work was to design a deep neural network with convolution autoencoders that performs the task of detecting breast cancer and achieves results with accuracy that proves to be challenging to the existing work. This paper incorporates a linear discriminant analysis (LDA) with an Autoencoder neural network with LSTM deep learning techniques to learn from the gene expression information with most characteristic features. This uses the deep learning algorithm at the stage of classification to create an advanced ensemble classification for the prediction. Hence, the suggested system has more prediction capacity with deep learning classification compared to other techniques, which are shown in evaluation results. This analysis showed excellent ability to generalize quickly and explicitly improve the performance of the prediction of the results with 98.27% of accuracy, which has been automatically obtained from the network. This approach has great potential for generalization, and it must be further enhanced with more public data sets.

References

[1] Md. Mohaiminul Islam a,b, Shujun Huang c, Rasif Ajwad a,b, Chen Chi a, Yang Wang b, Pingzhao Hu, "An integrative deep learning framework for classifying molecular subtypes of breast cancer" in Elsevier, Computational and Structural Biotechnology Journal, Vol. No.18, 2020, pp. 2185–2199, August 2020.

[2] M. Pouryahya et al., "aWCluster: A Novel Integrative Network-Based Clustering of Multiomics for Subtype Analysis of Cancer Data," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 3, pp. 1472-1483, 1 May-June 2022, doi: 10.1109/TCBB.2020.3039511.

[3] X. Li, J. Xiang, J. Wang, J. Li, F. -X. Wu and M. Li, "FUNMarker: Fusion Network-Based Method to Identify Prognostic and Heterogeneous Breast Cancer Biomarkers," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 6, pp. 2483-2491, 1 Nov.-Dec. 2021, doi: 10.1109/TCBB.2020.2973148.

[4] A. Płaczek, A. Płuciennik, A. Kotecka-Blicharz, M. Jarzab and D. Mrozek, "Bayesian Assessment of Diagnostic Strategy for a Thyroid Nodule Involving a Combination of Clinical Synthetic Features and Molecular Data," in IEEE Access, vol. 8, pp. 175125-175139, 2020, doi: 10.1109/ACCESS.2020.3026315.

[5] M. Shahdoust, H. Mahjub, H. Pezeshk and M. Sadeghi, "A Network-Based Comparison Between Molecular Apocrine Breast Cancer Tumor and Basal and Luminal Tumors by Joint Graphical Lasso," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 5, pp. 1555-1562, 1 Sept.-Oct. 2020, doi: 10.1109/TCBB.2019.2911074.

[6] N. Arya and S. Saha, "Multi-Modal Classification for Human Breast Cancer Prognosis Prediction: Proposal of Deep-Learning Based Stacked Ensemble Model," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 2, pp. 1032-

1041, 1 March-April 2022, doi: 10.1109/TCBB.2020.3018467.

[7] Uzma, Feras Al-Obeidat, Abdallah Tubaishat, Babar Shah, Zahid Halim, "Gene encoder: a feature selection technique through unsupervised deep learning-based clustering for large gene expression data", Springer, 4 June 2020, Neural Computing and Applications, 00521-020-05101.

[8] Fariha Muazzam, "Multi-class Cancer Classification and Biomarker Identification using Deep Learning", do i: February 11, 2021, https://doi.org/10.1101/2020.12.24.424317;

[9] David Pratella 1, Samira Ait-El-Mkadem Saadi 2, Sylvie Bannwarth 2, Véronique Paquis-Fluckinger 2,† and Silvia Bottini, "A Survey of Autoencoder Algorithms to Pave the Diagnosis of Rare Diseases", 2021, Vol.No. 22, 10891, 7 October 2021.

[10] T. Nguyen et al., "PAN: Personalized Annotation-Based Networks for the Prediction of Breast Cancer Relapse," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 6, pp. 2841-2847, 1 Nov.-Dec. 2021, doi: 10.1109/TCBB.2021.3076422.

[11] S. Rajpal, M. Agarwal, V. Kumar, A. Gupta and N. Kumar, "Triphasic DeepBRCA-A Deep Learning-Based Framework for Identification of Biomarkers for Breast Cancer Stratification," in IEEE Access, vol. 9, pp. 103347-103364, 2021, doi: 10.1109/ACCESS.2021.3093616.

[12] X. Zhang et al., "Deep Learning Based Analysis of Breast Cancer Using Advanced Ensemble Classifier and Linear Discriminant Analysis," in IEEE Access, vol. 8, pp. 120208-120217, 2020, doi: 10.1109/ACCESS.2020.3005228.

[13] C. Peng, Y. Zheng and D. -S. Huang, "Capsule Network Based Modeling of Multi-omics Data for Discovery of Breast Cancer-Related Genes," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 5, pp. 1605-1612, 1 Sept.-Oct. 2020, doi: 10.1109/TCBB.2019.2909905.

[14] X. Liu and J. Tang, "Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method," in IEEE Systems Journal, vol. 8, no. 3, pp. 910-920, Sept. 2014, doi: 10.1109/JSYST.2013.2286539.

[15] F. Ismailoglu, R. Cavill, E. Smirnov, S. Zhou, P. Collins and R. Peeters, "Heterogeneous Domain Adaptation for IHC Classification of Breast Cancer Subtypes," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 1, pp. 347-353, 1 Jan.-Feb. 2020, doi: 10.1109/TCBB.2018.2877755.

[16] 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, vol. 19, no. 4, pp. 2060-2070, 1 July-Aug. 2022, doi: 10.1109/TCBB.2021.3066086.

[17] A. B. O. V. Silva and E. J. Spinosa, "Graph Convolutional Auto-Encoders for Predicting Novel lncRNA-Disease Associations," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 4, pp. 2264-2271, 1 July-Aug. 2022, doi: 10.1109/TCBB.2021.3070910.

[18] C. Pan, J. Luo and J. Zhang, "Computational Identification of RNA-Seq Based miRNA-

Mediated Prognostic Modules in Cancer," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 2, pp. 626-633, Feb. 2020, doi: 10.1109/JBHI.2019.2911528.

Authors' profile

K.L.V.G.K.MURTHY, pursuing Ph.D from Rayalaseema University in Computer Science, under the guidance of Dr.R.J.Rama Sree, HoD & Dean, CS Department, NSU, Thirupati. He received his M.Tech in CSE from JNTUH. He is having 16 years of teaching experience.He was published various research papers in national and International Journals. His research areas of interest are Machine Learning, Data Mining, Data warehousing and Data Mining.

Dr. R. J. Rama Sree, HOD & Dean, National Sanskrit University, Tirupati. She Received her Ph.D from S.V.Women's University, Thirupati, received M.S in Software System from BITS,Pilani. She has 20 years of teaching experience and published more than 25 National and International Journal Papers. She conducted so many workshops and conferences as a part of Curriculum. Her interested research areas are Data Mining, Natural Language Processing and Internet Of Things. She guided so many research students in their research areas.